# Hypothesis Testing Procedures for Two-sample Data

Ryan Miller

# Outline

1. The two-sample $Z$-test
2. The two-sample $T$-test
3. Sample size conditions for $Z$ and $T$ tests

# Two-sample data

- So far, we've used the $Z$-test to evaluate hypotheses involving a *single proportion*, and the $T$-test to evaluate hypotheses involving a *single mean*
  - These are *one-sample tests*, as they treat all of the data as a single sample (group)

# Two-sample data

- So far, we've used the $Z$-test to evaluate hypotheses involving a *single proportion*, and the $T$-test to evaluate hypotheses involving a *single mean*
  - These are *one-sample tests*, as they treat all of the data as a single sample (group)
- The $Z$-test can also test hypotheses involving a *difference in proportions* (ie: $H_0 : p_1 - p_2 = 0$)
  - Similarly, the $T$-test can also test hypotheses involving a *difference in means* (ie: $H_0 : \mu_1 - \mu_2 = 0$)
- These applications are called *two-sample tests*, as they involve splitting the data into two groups

# Null hypotheses for one-sample and two-sample data

- For one-sample data, the null hypothesis *must* provide a specific value for the population parameter of interest
  - For example, $H_0 : p = 0.5$ or $H_0 : \mu = 0.4$

# Null hypotheses for one-sample and two-sample data

- For one-sample data, the null hypothesis *must* provide a specific value for the population parameter of interest
  - For example, $H_0 : p = 0.5$ or $H_0 : \mu = 0.4$
- For two-sample data, the null hypothesis could be satisfied by many different values
  - For example, $H_0 : p_1 - p_2 = 0$ is true when $p_1$ and $p_2$ are both 0.3, or when $p_1$ and $p_2$ are both 0.6

# Standard errors for two-sample data

For an observed difference in proportions, $\hat{p}_1 - \hat{p}_2$, CLT suggests:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

For a difference in means, $\bar{x}_1 - \bar{x}_2$, CLT suggests:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

These different SE formulas are the primary change from the earlier hypothesis tests we've worked with, though there are a few additional smaller details to consider.

# Example - the two-sample $Z$-test

Researchers randomly assigned 1000 fruit flies to one of two environments where they could eat only organically grown bananas, or only conventionally grown bananas. After 15 days:

- 345 of 501 fruit flies eating organic bananas were still alive
- 320 of 499 fruit flies eating non-organic bananas were still alive.

Let $p_1$ represent the proportion of fruit flies eating organic bananas that survive, and $p_2$ represent the same proportion for non-organic bananas.

Does this experiment provide convincing evidence that $p_1 \neq p_2$ (a difference in survival)?

# Example - the two-sample $Z$-test

- First, $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$

# Example - the two-sample $Z$-test

- First, $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
- Next, CLT states: $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - Under the null hypothesis, $p_1 = p_2$, so we should plug-in the same value for each into the standard error formula

# Example - the two-sample $Z$-test

- First, $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
- Next, CLT states: $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - Under the null hypothesis, $p_1 = p_2$, so we should plug-in the same value for each into the standard error formula
  - The best choice is the **pooled proportion**: $\hat{p}_0 = \frac{345+320}{501+499} = 0.665$

# Example - the two-sample $Z$-test

- First, $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
- Next, CLT states: $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - Under the null hypothesis, $p_1 = p_2$, so we should plug-in the same value for each into the standard error formula
  - The best choice is the **pooled proportion**: $\hat{p}_0 = \frac{345+320}{501+499} = 0.665$
  - Thus, $SE = \sqrt{\frac{0.065(1-0.665)}{501} + \frac{0.665(1-0.665)}{499}} = 0.03$

# Example - the two-sample $Z$-test

- First, $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
- Next, CLT states: $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - Under the null hypothesis, $p_1 = p_2$, so we should plug-in the same value for each into the standard error formula
  - The best choice is the **pooled proportion**: $\hat{p}_0 = \frac{345+320}{501+499} = 0.665$
  - Thus, $SE = \sqrt{\frac{0.065(1-0.665)}{501} + \frac{0.665(1-0.665)}{499}} = 0.03$
- So, $Z = \frac{\text{observed}-\text{null}}{SE} = \frac{(345/501-320/499)-0}{0.03} = 1.58$
  - Comparing this $Z$-value against a Standard Normal curve we get a $p$-value of 0.114 (two-sided)

# Example - the two-sample $Z$-test

- First, $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$
- Next, CLT states: $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
    - Under the null hypothesis, $p_1 = p_2$, so we should plug-in the same value for each into the standard error formula
    - The best choice is the **pooled proportion**: $\hat{p}_0 = \frac{345+320}{501+499} = 0.665$
    - Thus, $SE = \sqrt{\frac{0.065(1-0.665)}{501} + \frac{0.665(1-0.665)}{499}} = 0.03$
- So, $Z = \frac{\text{observed} - \text{null}}{SE} = \frac{(345/501 - 320/499) - 0}{0.03} = 1.58$
    - Comparing this $Z$-value against a Standard Normal curve we get a $p$-value of 0.114 (two-sided)
    - We cannot reject $H_0$, but there seems to be borderline evidence of high survival for organic bananas

# The two-sample $Z$-test (procedure)

1) State the null and alternative hypotheses (usually $H_0 : p_1 - p_2 = 0$)
2) Calculate the *pooled proportion*, $\hat{p}_0$,and use it to find the standard error, $SE = \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}$
3) Calculate the $Z$-value: $Z = \frac{\text{observed}-\text{null}}{SE} = \frac{(\hat{p}_1-\hat{p}_2)-0}{SE}$
4) Compare the $Z$-value against a Standard Normal distribution to find the $p$-value, then use the $p$-value to reach a conclusion.

# Practice #1

Until 2002, hormone replacement therapy (HRT) was commonly prescribed to postmenopausal women. This changed in 2002, when a large clinical trial was stopped early for safety concerns.

In the trial, 8506 women were randomized to take HRT and 8102 were randomized to take a placebo. Researchers observed 164 cases of cardiovascular disease (CVD) in the HRT group, but only 122 cases in the placebo group.

1) State the null and alternative hypotheses used to test whether the risk of CVD is higher in women taking HRT
2) Find the *pooled proportion*, and the *SE* for this application
3) Perform a two-sample $Z$-test

# Practice #1

1) $H_0 : p_1 - p_2 = 0$, where $p_1$ is the proportion of cases of cardiovascular disease in the HRT group, and $p_2$ is the equivalent proportion for the placebo group.
2) $\hat{p}_0 = \frac{164+122}{8506+8102} = 0.017$, so $SE = \sqrt{\frac{0.017(1-0.017)}{8506} + \frac{0.017(1-0.017)}{8102}} = 0.002$
3) $Z = \frac{(164/8506-122/8102)-0}{0.002} = 2.11$, the corresponding $p$-value (two-sided) is 0.034, which is strong evidence of a higher rate of cardiovascular disease in the HRT group

# Practice #2

Doctors are widely stereotyped as having messy handwriting. A 2010 study randomly assigned doctors to use either electronic prescription forms, or continue using written prescriptions. After 1 year, the error rate of each group was recorded:

| | Error | Non-errors | Total |
|---|---|---|---|
| Electronic | 254 | 3594 | 3848 |
| Hand-written | 1176 | 2370 | 3746 |

1) Propose appropriate the null and alternative hypotheses.
2) What is the *pooled proportion* in this application? What is the *SE*?
3) Using the *SE* and the observed difference in proportions, perform a two-sample $Z$-test.

# Practice #2 (solution)

1) $H_0 : p_1 - p_2 = 0$, where $p_1$ is the proportion of handwritten prescriptions resulting in errors and $p_2$ is the equivalent proportion for electronic prescriptions. $H_a : p_1 - p_2 \neq 0$
2) $\hat{p}_0 = \frac{254+1176}{3848+3746} = 0.188$, so
   $SE = \sqrt{\frac{0.188(1-0.188)}{3746} + \frac{0.188(1-0.188)}{3848}} = 0.009$
3) $Z = \frac{(1176/3746-254/3848)-0}{0.009} = 27.55$, the corresponding $p$-value is approximately zero, indicating overwhelming evidence of a lower rate rate for electronic prescription forms.

# The two-sample $T$-test

When testing a difference in means, we must make two major changes:

1) $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, which is based Central Limit theorem
2) Because the $SE$ relies upon $s_1$ and $s_2$ as estimates of $\sigma_1$ and $\sigma_2$ (population parameters), we now need to calculate a $T$-value and compare it to a $t$-distribution.

Because we've now got two groups (ie: two samples), the degrees of freedom are complicated. We'll use *the smaller group size minus 1* as a conservative approach.

# Practice #1

We've previously analyzed data from an experiment where 12 swimmers participated in a 1500m time trial with an without a scientifically designed wetsuit. In this example, we'll see what happens when we *ignore the paired study design*.

- When swimming with the wetsuit, the average velocity was $\bar{x}_1 = 1.507$ m/s, with a standard deviation of $s = 0.136$ m/s
- When swimming without the wetsuit, the average velocity was $\bar{x}_2 = 1.429$ m/s, with a standard deviation of $s = 0.141$ m/s

1) For $H_0 : \mu_1 - \mu_2 = 0$ (wetsuit - no wetsuit), report the observed sample statistic and its standard error
2) Perform a two-sample $T$-test

# Practice #1 (solution)

1) The observed difference in means is $\bar{x}_1 - \bar{x}_2 = 1.507 - 1.429 = 0.078$, the standard error is $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.136^2}{12} + \frac{0.141^2}{12}} = 0.057$
2) The $T$-value is $T = \frac{0.078-0}{0.057} = 1.37$, we need to use $df = 12 - 1 = 11$, so the two-sided $p$-value is 0.198. This seems to suggest insufficient evidence of a difference in velocity, but we need to remember that it's ignoring the paired design of the study!

# Practice #2

CDC researchers collected data on children aged 3-15 in El Paso, TX who lived near (within 1 mile) and far (more than 1 mile away) from a local lead smelter. One dependent variable they considered was the age-adjusted IQ score of these children.

These data are available on our course website as "Lead IQ", they're also available by clicking here

1) Using proper notation, state the null hypothesis for test comparing the mean age-adjusted IQ of the "near" and "far" groups.
2) Using StatKey, find the sample means, sample standard deviations, and sample sizes for each group.
3) Perform a two-sample $T$-test.

# Practice #2 (solution)

1) $H_0 : \mu_1 - \mu_2 = 0$, where $\mu_1$ is the mean age-adjusted IQ of children who live within 1 mile of a lead smelter, and $\mu_2$ is the equivalent mean for children who live 1 or more miles away.
2) $\bar{x}_1 - \bar{x}_2 = 89.193 - 92.687 =$ -3.494, $s_1 = 12.175$ and $s_2 = 15.975$, $n_1 = 115$ and $n_2 = 141$
3) $T = \frac{-3.494 - 0}{\sqrt{\frac{12.175^2}{115} + \frac{15.975^2}{141}}} = \frac{-3.494}{1.76} = -1.99$; using $df = 115 - 1 = 114$, the two-sided $p$-value is 0.048. We conclude that age-adjusted IQs are lower for children who live near a lead smelter.

# Comments - sample size assumptions

Both of these two-sample hypothesis testing approaches are built upon Central Limit theorem results:

1) The two-sample $Z$-test requires 10 “successes” and 10 “failures” in each of the two samples (ie: $n_1 p_1 \geq 10 \ldots$)
2) The two-sample $T$-test requires either Normally distributed data (if $n_1$ and $n_2$ are small), or sufficiently large samples of $n_1 \geq 30$ and $n_2 \geq 30$ (regardless of how the data are distributed)

If these conditions are not met, randomization tests are a reasonable alternative.

# Summary

In this presentation we focused on two specific hypothesis testing scenarios:

- Testing $H_0 : p_1 - p_2 = 0$ using $Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}}$
  - Notice the *pooled proportion*, $\hat{p}_0$
- Testing $H_0 : \mu_1 - \mu_2 = 0$ using $T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
  - We must compare this $T$-value against a distribution with either $n_1 - 1$ or $n_2 - 1$ degrees of freedom (whichever is smaller)