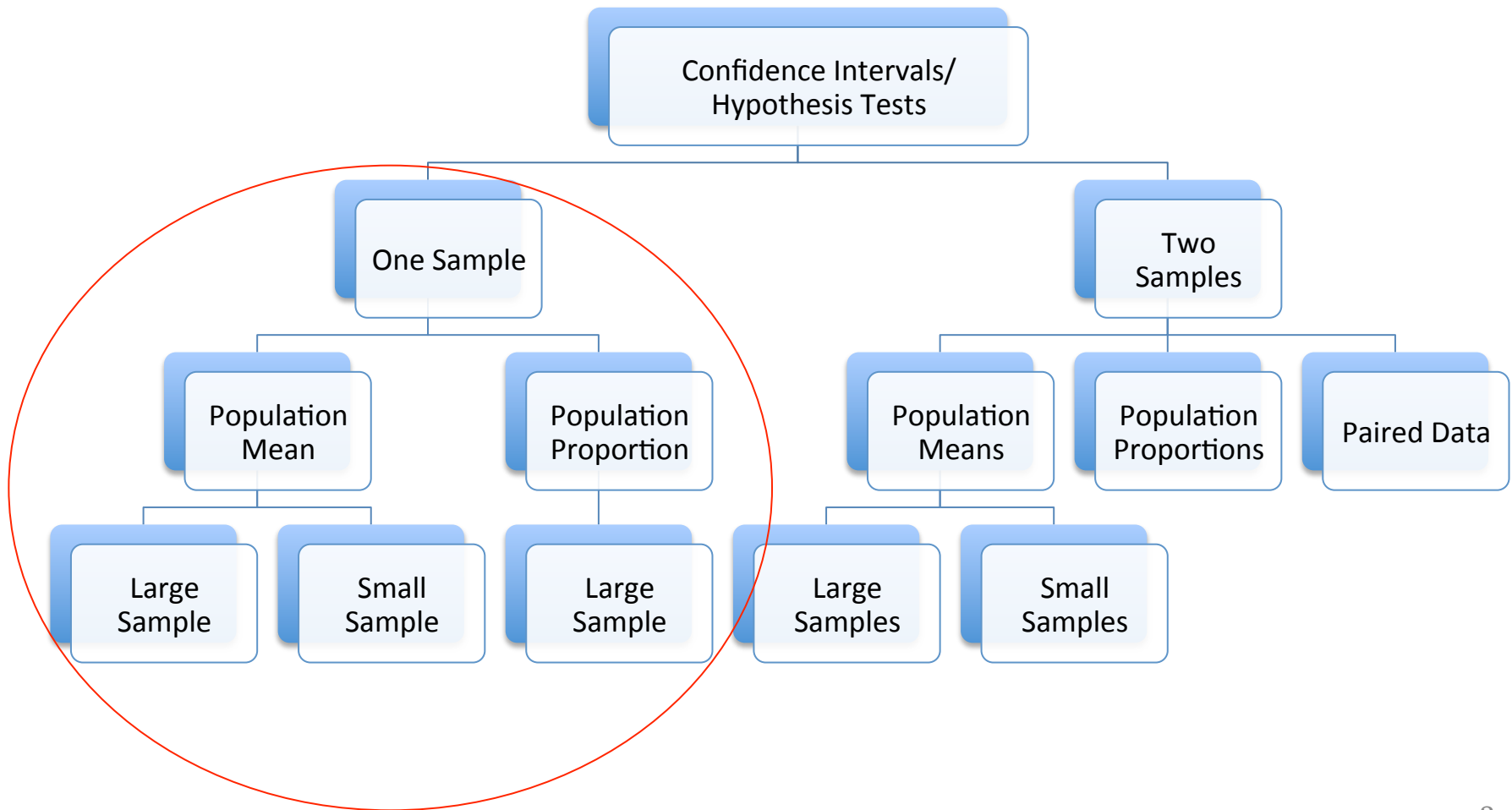


More Hypothesis Testing: Difference in Population Means

Keegan Korthauer
Department of Statistics
UW Madison

Outline



Recap: 5 Steps to Perform a HT

1. Define H_0 and H_1
2. State the level of significance α
3. Construct the test statistic
4. Assume H_0 is true and evaluate the test statistic by finding the p-value
5. Make a conclusion based on the p-value

Use these steps in your HW and on the exam

General Form of Test Statistic

- Recall the general form for the CI:

$$\text{point estimate} \pm \text{critical value} \times \text{standard deviation}$$

- We can write the general form of a test statistic as:

$$\text{test statistic} = \frac{\text{point estimate} - \text{hypothesized value}}{\text{standard deviation}}$$

TWO-SAMPLE TESTS FOR INDEPENDENT SAMPLES



Population means – Large samples and Small samples

Population proportions

Motivation

- Recall the example of reaction time for treatment versus control (six subjects each) – we were interested in a CI for the difference in means of the treatment and control groups $\mu_T - \mu_C$
- Can we formulate a HT for testing whether the mean difference in reaction time $\mu_T - \mu_C$ is different from zero?
- How would this procedure change if we had large samples (more than 30 subjects in each group)?

The Idea

- One-sample test statistic (population mean):
 - observed \bar{X} too far from μ_0 -> reject H_0
 - observed \bar{X} close to μ_0 -> do not reject H_0
 - Two-sample test (difference in population mean):
 - observed difference $\bar{X} - \bar{Y}$ too far from 0 -> reject H_0
(means are not equal)
 - observed difference $\bar{X} - \bar{Y}$ close to 0 -> do not reject H_0
(means are equal)
-  We are not restricted to the null hypothesis that the difference in means is 0
-  Construct a test statistic using $\bar{X} - \bar{Y}$

Large Sample Case: Deriving the Test Statistic

- Recall that the difference of 2 Normal RVs is Normal

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

- And by the CLT, the means (of a large sample) are also normal

$$\bar{X} \sim N(\mu_X, \sigma_X^2 / n_X) \text{ and } \bar{Y} \sim N(\mu_Y, \sigma_Y^2 / n_Y)$$

- Combining these, we get the following result, which will come in handy when we compute the p-value

$$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma_X^2 / n_X + \sigma_Y^2 / n_Y), \text{ so}$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}} \sim N(0, 1)$$

HT for Large-Sample Difference in Means

Let X_1, \dots, X_{n_X} and Y_1, \dots, Y_{n_Y} be large ($n_X > 30, n_Y > 30$) **independent** samples from any population with means μ_X and μ_Y and standard deviations σ_X and σ_Y . Approximate σ_X and σ_Y with s_X and s_Y when unknown.

1. Set up the null H_0 and alternative H_1 hypotheses (see table below)
2. State the level of significance α you will use

3. Calculate the **z-score** (test statistic):

$$z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}}$$

Where Δ_0 is the hypothesized difference

4. Assume H_0 is true and calculate the P-value:

H_0	H_1	P-value
$\mu_X - \mu_Y \leq \Delta_0$	$\mu_X - \mu_Y > \Delta_0$	Area to the right of z
$\mu_X - \mu_Y \geq \Delta_0$	$\mu_X - \mu_Y < \Delta_0$	Area to the left of z
$\mu_X - \mu_Y = \Delta_0$	$\mu_X - \mu_Y \neq \Delta_0$	Area to the left of $-z$ plus area to the right of z

5. Make a conclusion based on the P-value

Example – Exercise 6.5.5

- In a test to compare the effectiveness of two drugs designed to lower cholesterol levels,
 - 75 randomly selected patients were given drug A
 - Average cholesterol reduction of 40 mg/dl
 - Standard deviation of 12 mg/dl
 - 100 randomly selected patients were given drug B
 - Average cholesterol reduction of 42 mg/dl
 - Standard deviation of 15 mg/dl
- Can we conclude that the mean reduction using drug B is greater than that of drug A?

What About Small Samples?

- We just learned how to conduct a HT involving the difference in sample means based on large samples
 - the test statistic is normally distributed by the CLT so the p-value is found by finding areas under the normal curve
- What if we want to test a hypothesis about the difference in means **based on a small sample?**
 - CLT no longer applies
 - if both populations are approximately normal then the means will be approximately normal
 - sample sds s_x and s_y no longer approximate σ_x and σ_y well



Make use of the t distribution!


Small-sample HT for Difference in Population Means

- If the populations of X and Y are approximately normal, then

$$\frac{\bar{X} - \mu_X}{s_X / \sqrt{n_X}} \sim t_{n-1} \quad \text{and} \quad \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n_Y}} \sim t_{n-1}$$

- Then we can use the following test statistic to measure the evidence against $H_0: \mu_X - \mu_Y = \Delta_0$ using the difference in means and standard deviations of our small samples of size n_X and n_Y :

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{s_X^2 / n_X + s_Y^2 / n_Y}}$$

- Assuming H_0 is true, t will have a t distribution with **v** degrees of freedom  find p-value using areas under t curve

Degrees of Freedom of Test Statistic t

Recall that the degrees of freedom for a t statistic calculated from two small independent samples (with unequal variance) is not so straightforward:

$$v = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \right)^2}{\frac{(s_X^2 / n_X)^2}{n_X - 1} + \frac{(s_Y^2 / n_Y)^2}{n_Y - 1}}$$

(Rounded **down** to the nearest integer)

Small-Sample HT for the Difference in Population Mean

Let $X_{1,\dots,X_{n_X}}$ and $X_{1,\dots,X_{n_Y}}$ be small ($n_X < 30$ and $n_Y < 30$) **independent** samples from **normal** populations with means μ_X and μ_Y (unknown, **unequal** standard deviations σ_X and σ_Y).

1. Set up the null H_0 and alternative H_1 hypotheses (see table below)
2. State the level of significance α you will use

3. Calculate the test statistic and df:

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{s_X^2 / n_X + s_Y^2 / n_Y}} \quad \nu = \frac{(s_X^2 / n_X + s_Y^2 / n_Y)^2}{\frac{(s_X^2 / n_X)^2}{n_X - 1} + \frac{(s_Y^2 / n_Y)^2}{n_Y - 1}} \text{ (rounded down)}$$

4. Assume H_0 is true and calculate the P-value using areas under the t curve with ν degrees of freedom:

H_0	H_1	P-value
$\mu_X - \mu_Y \leq \Delta_0$	$\mu_X - \mu_Y > \Delta_0$	Area to the right of t
$\mu_X - \mu_Y \geq \Delta_0$	$\mu_X - \mu_Y < \Delta_0$	Area to the left of t
$\mu_X - \mu_Y = \Delta_0$	$\mu_X - \mu_Y \neq \Delta_0$	Area to the left of $-t$ plus area to the right of t

5. Make a conclusion based on the P-value

Notes

- Recall that using the t-table, we can typically only say that the p-value is between two values
 - On the exam, this would be the final answer for the p-value
 - On the homework, use R to evaluate the p-value
- In the (rare) case that the population standard deviations are known – use Z test (large-sample) instead of t test
- The previous HT assumes that the population variances are not equal
 - If they are equal ($\sigma_x = \sigma_y$) then use the method on the next slide
 - Caution - only use when explicitly stated; ***cannot conclude that the population variances are equal just because the sample standard variances are close***

Small-Sample HT for the Difference in Population Mean (Special Case: Equal Population Variance)

Let X_1, \dots, X_{n_X} and X_1, \dots, X_{n_Y} be small ($n_X < 30$ and $n_Y < 30$) **independent** samples from **normal** populations with means μ_X and μ_Y (unknown, **equal** standard deviations $\sigma_X = \sigma_Y$).

1. Set up the null H_0 and alternative H_1 hypotheses (see table below)
2. State the level of significance α you will use
3. Calculate the test statistic:

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{s_p \sqrt{1/n_X + 1/n_Y}} \quad \text{where } s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}$$

4. Assume H_0 is true and calculate the P-value using areas under the t curve with $n_X + n_Y - 2$ degrees of freedom:

H_0	H_1	P-value
$\mu_X - \mu_Y \leq \Delta_0$	$\mu_X - \mu_Y > \Delta_0$	Area to the right of t
$\mu_X - \mu_Y \geq \Delta_0$	$\mu_X - \mu_Y < \Delta_0$	Area to the left of t
$\mu_X - \mu_Y = \Delta_0$	$\mu_X - \mu_Y \neq \Delta_0$	Area to the left of $-t$ plus area to the right of t

5. Make a conclusion based on the P-value

Example – Arsenic in Drinking Water

Arsenic concentration in public drinking water supplies is a potential health risk. An article in the *Arizona Republic* (May 2001) reported drinking water arsenic concentrations in parts per billion for 10 metro Phoenix communities and 10 rural communities in Arizona:

Metro Phoenix ($\bar{x}_1 = 12.5, s_1 = 7.63$)	Rural Arizona ($\bar{x}_2 = 27.5, s_2 = 15.3$)
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apache Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

Assume the boxplots for the two samples show no evidence of skew so that it is plausible that the populations are normally distributed. Can we conclude that there is a difference in mean arsenic concentrations for metro Phoenix and rural Arizona communities?

Next

- Hypothesis testing for the difference in two population proportions
- Hypothesis testing for paired data
- HW7 Due on Friday at the Beginning of Lecture