Chapter 10: Comparing Two Groups

Yu Yang

School of Statistics
University of Minnesota

October 31, 2022

Motivation

Instead of focusing on one population mean or proportion, we might be interested in comparing parameters from 2 distinct populations.

Examples:

- compare average GRE scores before and after taking a prep course
- compare the proportion of Minnesotans who own guns to the proportion of Wisconsinites who own guns

Comparing Two Means

Comparisons of two means are based on either two *independent* samples or two *dependent* samples.

Independent Samples:

- ex1. compare mean heights of men and women
- ex2. compare average exam scores for two different stat classes

Comparing Two Means

Comparisons of two means are based on either two *independent* samples or two *dependent* samples.

Independent Samples:

- ex1. compare mean heights of men and women
- ex2. compare average exam scores for two different stat classes

Dependent Samples: The dependent samples we will consider result from *matched pairs* experiments.

- ex1. compare the mean weight loss resulting from 2 different diet pills.

  Sample 2n people who are "matched" by weight. Randomly assign one person in each pair to pill 1 and the other pill 2.

  This natural pairing causes the samples to be dependent.

- ex2. compare mean weights before and after taking a diet pill.

Comparing Two Means - Independent Samples

Goal: Compare two (unknown) means corresponding to two *independent* samples.

We independently collect a random sample from each population to make inferences about the differences between two means ($\mu_1 - \mu_2$) through point estimation, interval estimation, and hypothesis testing.

Population	Pop. Mean	Pop. St. Dev.	Sample Size	Sample Mean	Sample St. Dev.
1	μ_1	σ_1	n_1	$\overline{x}_1$	s_1
2	μ_2	σ_2	n_2	$\overline{x}_2$	s_2

NOTE: we will assume that σ_1 and σ_2 are unequal. The book also looks at the case when $\sigma_1 = \sigma_2$. We will not cover that case in class.

Point Estimation for $\mu_1 - \mu_2$

The obvious point estimate of $\mu_1 - \mu_2$ is

Point Estimation for $\mu_1 - \mu_2$

The obvious point estimate of $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$

The statistic varies from sample to sample, so has its own sampling distribution.

$$\bar{x}_1 - \bar{x}_2 \overset{.}{\sim} N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

σ_1 and σ_2 are unknown, we estimate the standard deviation using the standard error:

$$se(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Then we have

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \overset{.}{\sim} t_{df},$$

where df = smaller of $(n_1 - 1)$ and $(n_2 - 1)$.

Confidence Interval for $\mu_1 - \mu_2$

Assumptions:

1. Two independent random samples
2. Approximately normal distribution for **each** group, though the two-sample t-procedure is robust to non-normality for larger sample sizes $n_1, n_2 \geq 30$

The $1 - \alpha$ confidence level confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $df =$ smaller of $(n_1 - 1)$ and $(n_2 - 1)$.

The Two-Sample t-Test for Comparing μ_1 and μ_2

Assumptions

1. Two independent random samples
2. Both populations are approximately normal, though the two-sample t-procedure is robust to non-normality for larger sample sizes $n_1, n_2 \geq 30$

Hypotheses

The null hypothesis is always $H_0 : \mu_1 = \mu_2$.

1. $H_0 : \mu_1 = \mu_2$ 2. $H_0 : \mu_1 = \mu_2$ 3. $H_0 : \mu_1 = \mu_2$
   $H_a : \mu_1 < \mu_2$ $H_a : \mu_1 > \mu_2$ $H_a : \mu_1 \neq \mu_2$

Test-Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \overset{.}{\sim} t_{df}$$

df = smaller of $n_1 - 1$ and $n_2 - 1$.

The Two-Sample t-Test for Comparing μ_1 and μ_2

p-value

Hypothesis 1: p-val $= P(t_{df} < t)$
Hypothesis 2: p-val $= P(t_{df} > t)$
Hypothesis 3: p-val $= 2P(t_{df} > |t|)$

Conclusion

If p-value $< \alpha$, reject H_0
If p-value $\geq \alpha$, fail to reject H_0

Example 10.5

Verbal SAT scores were recorded for independent samples of students who intend to major in engineering and students who intend to major in literature. Suppose histograms of both samples show no strong skewness and no outliers. From the data, we calculate:

Intended Major	Sample Size	Sample Mean	Sample St. Dev.
(1) Engineering	44	446.9	42.0
(2) Literature	44	534.2	45.5

Let:
μ_E = true mean verbal score for intended engineering majors
μ_L = true mean verbal score for intended literature majors

Example 10.5

(a) Calculate a 90% confidence interval for $\mu_E - \mu_L$, the true difference in mean verbal scores for students intending to major in engineering and students intending to major in literature.

Example 10.5

(a) Calculate a 90% confidence interval for $\mu_E - \mu_L$, the true difference in mean verbal scores for students intending to major in engineering and students intending to major in literature.

$n_1 - 1 = 43$ and $n_2 - 1 = 43$ so df $= 43$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.05,43}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (446.9 - 534.2) \pm 1.68 \cdot \sqrt{\frac{42^2}{44} + \frac{45.5^2}{44}}$$

$$= -87.3 \pm 15.72$$

$$= (-103.02, -71.58)$$

We are 90% confident that the true mean verbal score for intended engineering majors is between 71.58 and 103.02 *less* than the mean verbal score for intended lit majors.

Example 10.5

(b) Determine if there is strong evidence at the .01 level that the true mean verbal score of intended engineering majors is less than that of intended lit majors.

Example 10.5

(b) Determine if there is strong evidence at the .01 level that the true mean verbal score of intended engineering majors is less than that of intended lit majors.
Assumptions: independent random samples, large sample sizes
Hypotheses: $H_0 : \mu_E = \mu_L$ vs $H_a : \mu_E < \mu_L$
Test statistic:

Example 10.5

(b) Determine if there is strong evidence at the .01 level that the true mean verbal score of intended engineering majors is less than that of intended lit majors.

Assumptions: independent random samples, large sample sizes

Hypotheses: $H_0 : \mu_E = \mu_L$ vs $H_a : \mu_E < \mu_L$

Test statistic:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{se(\overline{x}_1 - \overline{x}_2)} = \frac{446.9 - 534.2}{9.33} = -9.35$$

p-**value:**

Example 10.5

(b) Determine if there is strong evidence at the .01 level that the true mean verbal score of intended engineering majors is less than that of intended lit majors.
Assumptions: independent random samples, large sample sizes
Hypotheses: $H_0 : \mu_E = \mu_L$ vs $H_a : \mu_E < \mu_L$
Test statistic:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{se(\overline{x}_1 - \overline{x}_2)} = \frac{446.9 - 534.2}{9.33} = -9.35$$

p-**value:** *p*-val $= P(t_{43} < -9.35) = \text{pt(-9.35, df = 43)} = 3.2 \times 10^{-12}$
Conclusion:

Example 10.5

(b) Determine if there is strong evidence at the .01 level that the true mean verbal score of intended engineering majors is less than that of intended lit majors.

Assumptions: independent random samples, large sample sizes

Hypotheses: $H_0 : \mu_E = \mu_L$ vs $H_a : \mu_E < \mu_L$

Test statistic:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{se(\overline{x}_1 - \overline{x}_2)} = \frac{446.9 - 534.2}{9.33} = -9.35$$

p-**value:** *p*-val $= P(t_{43} < -9.35) = $ `pt(-9.35, df = 43)` $= 3.2 \times 10^{-12}$

Conclusion: *p*-val $< 0.01 \Rightarrow$ Reject H_0. We have strong enough evidence to conclude that mean verbal scores of intended eng majors is significantly less than that for intended lit majors at the 0.01 level.

Example 10.6: Two-Sample t-test in R

Consider a data set collected from a recent survey of U of M students. The students were asked about the money spent on their latest haircut.

Example 10.6: Two-Sample *t*-test in R

(a) Construct and interpret a 95% confidence interval for the true difference in the mean amount of money men and women spend on a haircut.

```
> # Sample of Male Haircuts
> MaleHair <- dat$Haircut[dat$Gender == "M"]
> # Sample of Female Haircuts
> FemaleHair <- dat$Haircut[dat$Gender == "F"]
> # Confidence Interval for the Mean Difference
> t.test(x = MaleHair, y = FemaleHair, alternative = "two.sided",
+         conf.level = 0.95)

    Welch Two Sample t-test

data:  MaleHair and FemaleHair
t = -6.5631, df = 92.461, p-value = 3.011e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.333509  -9.280328
sample estimates:
mean of x mean of y
 12.46667  25.77358
```

95% CI: $(-17.333509, -9.280328)$

Example 10.6: Two-Sample t-test in R

(b) Test at the .05 level whether male students spend less money on average than female students on their haircuts

$H_0 : \mu_M = \mu_F$
$H_a : \mu_M < \mu_F$

```
> # Two-Sample t-test
> t.test(x = MaleHair, y = FemaleHair, alternative = "less",
+        conf.level = 0.95)


Welch Two Sample t-test

data:  MaleHair and FemaleHair
t = -6.5631, df = 92.461, p-value = 1.506e-09
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -9.938168
sample estimates:
mean of x mean of y
 12.46667  25.77358
```

p-value $= 1.5\text{e-}09 < .05$, thus we **reject** H_0. There is strong evidence that on average women spend more than men on haircuts.

Comparing Two Means - Matched Pairs

Goal: Determine if a response differs between "matched pairs"

Definition (matched pairs): Data are matched pairs if each subject in one sample is matched with a subject in another sample in some meaningful way.

Why use matched pairs?

Comparing Two Means - Matched Pairs

Goal: Determine if a response differs between "matched pairs"

Definition (matched pairs): Data are matched pairs if each subject in one sample is matched with a subject in another sample in some meaningful way.

Why use matched pairs? It controls for extra sources of variability in the response variable.

Method for studying matched pairs:

Comparing Two Means - Matched Pairs

Goal: Determine if a response differs between "matched pairs"

Definition (matched pairs): Data are matched pairs if each subject in one sample is matched with a subject in another sample in some meaningful way.

Why use matched pairs? It controls for extra sources of variability in the response variable.

Method for studying matched pairs: **Use the single sample of differences among the pairs.** Therefore, we can use our one-sample t-procedures.

Notation:
μ_D = true mean of paired differences
$\bar{x}_D$ = sample mean of paired differences
s_D = sample standard deviation of paired differences
n_D = number of pairs in the sample.

Example 10.3

Six adults are chosen for a blood alcohol content (BAC) study. Each adult is given 3 beers at sea level and 3 beers at high altitude. The order of two drinking sessions is chosen at random for each subject. We compute the differences (Sea Level - High Altitude)

Sea Level (SL)	High Altitude (HA)	Difference
0.007	0.013	-0.006
0.010	0.017	-0.007
0.009	0.015	-0.006
0.011	0.014	-0.003
0.008	0.010	-0.002
0.006	0.009	-0.003

Example 10.3

Based on this data, we can show that

$$\bar{x}_D = -0.0045 \quad \text{and} \quad s_D = 0.0021$$

Suppose the distribution of the differences is normal. At the .01 level, test the hypothesis that the average BAC at sea level is less than the average BAC at high altitude.

We can now use the 5 steps from the one sample test to do this hypothesis test!

Example 10.3

Assumptions: random sample and distribution of differences is normal

Hypotheses:

$$H_0 : \mu_D = 0$$
$$H_a : \mu_D < 0$$

Test statistic:

$$t = \frac{\overline{x}_D - \mu_0}{s_D/\sqrt{n_D}} = \frac{-0.0045}{0.0021/\sqrt{6}} = -5.25$$

p-**value:** p-val $= $ `pt(-5.25, df = 5)` $= 0.0017$

Conclusion: p-val $< 0.01 \Rightarrow$ Reject H_0. We have enough evidence (at the 0.01 level) to conclude that alcohol reaction is greater at HA than at SL.

Example 10.3 in R

```
> BAC.sl <- c(0.007, 0.010, 0.009, 0.011, 0.008, 0.006)
> BAC.ha <- c(0.013, 0.017, 0.015, 0.014, 0.010, 0.009)
> t.test(x = BAC.sl, y = BAC.ha, paired = TRUE, alternative = "less",
 conf.level = 0.95)

Paired t-test

data:  BAC.sl and BAC.ha
t = -5.3156, df = 5, p-value = 0.001576
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -0.002794137
sample estimates:
mean of the differences
-0.0045

> # or equivalently:
> t.test(x = BAC.sl - BAC.ha, alternative = "less", conf.level = 0.95)
```

Example 10.4

Suppose we are interested in the protein concentration (in gms/kg of weight) of a winter wheat and a spring wheat. The growth and protein level of the wheats may be influenced by the location in which they are grown. Therefore, we select 20 different locations and plant both a winter wheat and a spring wheat in each location. That is, winter and spring wheat observations are matched by location. The following is the partial data

Location	Spring Wheat Protein	Winter Wheat Protein	Difference (spring - winter)
1	122	87	35
2	171	145	26
3	144	116	28
⋮	⋮	⋮	⋮

Example 10.4

Data summary: $\bar{x}_D = 29.2$ and $s_D = 9.35117$.

Calculate a 95% confidence interval for μ_D, the true mean difference (spring - winter) in protein concentration.

Example 10.4

Data summary: $\bar{x}_D = 29.2$ and $s_D = 9.35117$.

Calculate a 95% confidence interval for μ_D, the true mean difference (spring - winter) in protein concentration.

$$
\begin{aligned}
\bar{x}_D \pm t_{\alpha/2, n_D - 1} \frac{s_D}{\sqrt{n_D}} &= 29.2 \pm 2.09 \cdot \frac{9.35117}{\sqrt{20}} \\
&= 29.2 \pm 4.38 \\
&= (24.82, 33.58)
\end{aligned}
$$

We are 95% confident that the average protein content for spring wheat is between 24.82 and 33.58 higher than for winter wheat.