LECTURE 3 – INFERENTIAL STATISTICS

Part D - Hypothesis Testing with Two Samples

Book Chapter 10

Comparing Means/Proportions of Two samples

- Many times, we are interested in comparing the means or proportions of two samples:
  - Comparing Mean number of heart attacks between those who take aspirin and those who did not
  - Comparing SAT scores of students who take a preparatory course against those who did not
  - Comparing Mean Income for Males and Females
  - Comparing Mean Blood Pressure before and after taking medication
  - Comparing proportion of Males vs females who vote for a particular political party
- The groups are classified either as **independent** or **matched pairs**.
- **Independent groups** consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population.
  - Comparing SAT scores of students who take a preparatory course against those who did not
- **Matched pairs** consist of two samples that are dependent.
  - Comparing Mean Blood Pressure before and after taking medication

Comparing Means of Two Independent Samples – Known Standard Deviations

- We are interested in testing whether the means of two populations ($\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$), with the known standard deviations ($\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$), are significantly different. We take samples from each population with sizes n_1 and n_2 and obtain sample means $\overline{X}_1$ and $\overline{X}_2$.

- If the population distributions are Normal, then the sample sizes don’t matter. Otherwise, if the population distribution is unknown, since the population standard deviations are known, we can use the *Central Limit Theorem* if both the sample sizes are >= 30.

- The sampling distribution of the difference in sample means $\overline{X}_d = (\overline{X}_1 - \overline{X}_2) \sim N(\boldsymbol{\mu_1 - \mu_2}, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$ where the standard error is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- Null Hypothesis: $\mathbf{H_o}$: $\boldsymbol{\mu_1 - \mu_2 = 0}$;

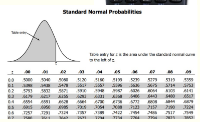
- Alternate Hypothesis: $\mathbf{H_1}: \boldsymbol{\mu_1 - \mu_2} <, \neq, > \ \mathbf{0}$

- Test statistic is:

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Practice with calculator



Practice with Tables

Comparing Means of Two Independent Samples – Known Standard Deviations – Book Problem 10.96 – Page 597 - (**Two-sample.R**)

- Some manufacturers claim that non-hybrid sedan cars have a lower mean miles-per-gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of seven mpg. Thirtyone non-hybrid sedans get a mean of 22 mpg with a standard deviation of four mpg. Suppose that **the population standard deviations are known to be six and three**, respectively. Conduct a hypothesis test to evaluate the manufacturers claim.

- Null Hypothesis: $\mathbf{H_o}$: $\boldsymbol{\mu_1 - \mu_2 = 0}$; Alternate Hypothesis: $\mathbf{H_a}$: $\boldsymbol{\mu_1 - \mu_2 > 0}$

- $n_1 = 21$, $\bar{X}_1 = 31$, $\sigma_1 = 6$; $n_2 = 31$, $\bar{X}_2 = 22$, $\sigma_2 = 3$

```
# 10.96 - Page 597
# Independent two-sample test - known population standard deviations
#
n1 = 21
n2 = 31
x1_bar = 31
x2_bar = 22
sigma1 = 6
sigma2 = 3
se1 <- (sigma1^2/n1)
se2 <- (sigma2^2/n2)
std_err <- sqrt(se1+ se2)
print(std_err)
z_statistic <- (x1_bar - x2_bar)/std_err
print(z_statistic)
# Lower-tailed test
#p_value <- (pnorm(z_statistic,deg_free))
#Upper-tailed test
p_value <- (1 - pnorm(z_statistic,0,1))
# Two-tailed test
#p_value <- 2*(1-*pnorm(abs(z_statistic,0,1)))
print(p_value)
```

```
> print(std_err)
[1] 1.415842
```

```
> print(z_statistic)
[1] 6.356642
```

```
> print(p_value)
[1] 1.03106e-10
```

10.2 Two Population Means with Known Standard Deviations

Normal Distribution:

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

Generally $\boldsymbol{\mu_1 - \mu_2 = 0}$.

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Because p-value = 0, at a significance level of 0.05, the samples provide evidence that the mean mpg of hybrid sedans is greater than non-hybrid sedans.

Effect of Variance on Difference in Means

- Whether the difference in sample means is statistically significant, depends on the variance reflected in the samples by the population variance.
- Consider the following data from 2 Normal populations.
- The means will be fixed and different.
- We will have the same variances for both, but we will show what happens to the test for mean differences, when the variances increase.
- Case 1:
  - $X_1 \sim N(\mathbf{5}, 2)$; $X_2 \sim N(\mathbf{3}, 2)$
- Case 2:
  - $X_1 \sim N(\mathbf{5}, 8)$; $X_2 \sim N(\mathbf{3}, 8)$
- We will simulate sample sizes of 50 in each case, and test for significant differences in mean at $\alpha = 0.05$

10.2 Two Population Means with Known Standard Deviations

Normal Distribution:

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

Generally $\boldsymbol{\mu_1 - \mu_2 = 0}$.

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Effect of Variance on Difference in Means

```
> #Case 1: X1 ~ N(5, 2); X2 ~ N(4,2) - sample size 50
> #
> x1 <- rnorm(50, 5, 2)
> x2 <- rnorm(50, 3, 2)
> #
> # Independent two-sample test - Known population standard deviations
> #
> n1 = 50
> n2 = 50
> x1_bar = mean(x1)
> print(x1_bar)
[1] 4.954474
> x2_bar = mean(x2)
> print(x2_bar)
[1] 2.910275
> sigma1 = 2
> sigma2 = 2
> se1 <- (sigma1^2/n1)
> se2 <- (sigma2^2/n2)
> std_err <- sqrt(se1+ se2)
> print(std_err)
[1] 0.4
> z_statistic <- (x1_bar - x2_bar)/std_err
> print(z_statistic)
[1] 5.110497
> # Lower-tailed test
> #p_value <- (pnorm(z_statistic,deg_free))
> #Upper-tailed test
> #p_value <- (1 - pnorm(z_statistic,0,1))
> # Two-tailed test
> p_value <- (2*(1-pnorm(abs(z_statistic),0,1)))
> print(p_value)
[1] 3.213132e-07
```

significant

```
> #Case 2: X1 ~ N(5, 2); X2 ~ N(4,2) - sample size 50
> #
> x1 <- rnorm(50, 5, 2)
> x2 <- rnorm(50, 3, 2)
> #
> # Independent two-sample test - Known population standard deviations
> #
> n1 = 50
> n2 = 50
> x1_bar = mean(x1)
> print(x1_bar)
[1] 4.861552
> x2_bar = mean(x2)
> print(x2_bar)
[1] 3.001746
> sigma1 = 8
> sigma2 = 8
> se1 <- (sigma1^2/n1)
> se2 <- (sigma2^2/n2)
> std_err <- sqrt(se1+ se2)
> print(std_err)
[1] 1.6
> z_statistic <- (x1_bar - x2_bar)/std_err
> print(z_statistic)
[1] 1.162378
> # Lower-tailed test
> #p_value <- (pnorm(z_statistic,deg_free))
> #Upper-tailed test
> #p_value <- (1 - pnorm(z_statistic,0,1))
> # Two-tailed test
> p_value <- (2*(1-pnorm(abs(z_statistic),0,1)))
> print(p_value)
[1] 0.2450818
```

Not significant

10.2 Two Population Means with Known Standard Deviations

Normal Distribution:

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

Generally $\mu_1 - \mu_2 = 0$.

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

- You can see that when the population variance increases, the p-value shows that the same difference in population means is no longer significant at α = 0.05!

Effect of Variance on Difference in Means

- Intuitively, when we compare the average scores of two basketball players who are very consistent (low variance) we can detect who is better based on their average scores.
- However, if both are very inconsistent (high variance), we cannot tell who is better based on their average scores.
- That is why, we want to look at or analyze the variances. Later we will see that this leads to ANOVA or “Analysis of Variance”
- You can also see this effect by looking at the test statistic (z-score).

Test Statistic (z-score):

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

- The denominator of the test-statistic contains the variance, and as the variance within each sample increases, increases the z-score decreases and you are less likely to find significant differences in the means. The variance within each group sample is called “within sample variance”
- For two groups, the numerator is the difference in mean between the two groups and represents the variation between the two group means.
- In other words, larger the ratio is “between samples variance”/”within sample variance”, the more likely that the difference in means is statistically significant.

Comparing Means of Two Independent Samples – Unknown Standard Deviations

- The two independent samples are simple random samples from two distinct populations.
- For the two distinct populations:
  - if the sample sizes are small, the distributions are important (should be normal)
  - if the sample sizes are large, the distributions are not important (need not be normal) (i.e., CLT applies even though we don't know population standard deviations)
- Null Hypothesis: $\mathbf{H_o}$: $\boldsymbol{\mu_1 - \mu_2 = d}$
- Alternate Hypothesis: $\mathbf{H_a}$: $\boldsymbol{\mu_1 - \mu_2} >, \neq, <\ \mathbf{d}$
- The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the *Aspin-Welch t-test*. The degrees of freedom formula was developed by Aspin-Welch.

10.1 Two Population Means with Unknown Standard Deviations

Standard error: $SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

Test statistic (t-score): $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

where:

s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

$\bar{x}_1$ and $\bar{x}_2$ are the sample means.

Comparing Means of Two Independent Samples – Unknown Standard Deviations – Book Problem 10.2 – Page 565 (**Two-sample.R**)

Practice with calculator

Practice with t-Tables

- A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes**, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

- Null Hypothesis: $\mathbf{H_o}: \boldsymbol{\mu_1 - \mu_2 = 0}$; Alternate Hypothesis: $\mathbf{H_a}: \boldsymbol{\mu_A - \mu_B > 0}$

-
```
# Book Problem 10.2 - Page 565
# Independent two-sample test - unknown population standard deviations
#
n1 = 11
n2 = 9
x1_bar = 4
x2_bar = 3.5
s1 = 1.5
s2 = 1
se1 <- (s1^2/n1)
se2 <- (s2^2/n2)
std_err <- sqrt(se1+ se2)
print(std_err)
t_statistic <- (x1_bar - x2_bar)/std_err
print(t_statistic)
deg_free_n <- (se1 + se2)^2
deg_free_d <- (se1^2/(n1 -1)) + (se2^2/(n2 -1))
deg_free <- deg_free_n/deg_free_d
print(deg_free)
# Lower-tailed test
#p_value <- (pt(t_statistic,deg_free))
# Upper-tailed test
p_value <- (1 - pt(t_statistic,deg_free))
# Two-tailed test
#p_value <- 2*(1-pt(abs(t_statistic,deg_free)))
print(p_value)
```

```
> print(std_err)
[1] 0.5618332
```

```
> print(t_statistic)
[1] 0.8899438
```

```
> print(deg_free)
[1] 17.39784
```

```
> print(p_value)
[1] 0.1928185
```

10.1 Two Population Means with Unknown Standard Deviations

Standard error: $SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

Test statistic (t-score): $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

where:

s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

$\bar{x}_1$ and $\bar{x}_2$ are the sample means.

Since p = 0.193, at a significance level of 0.01, the samples do not provide evidence to reject the null hypothesis that the mean number of students graduating is the same in the two neighboring colleges.

Practice with calculator

Practice with t-Tables

Comparing Means of Two Independent Samples – Unknown Standard Deviations – Book Problem 10.98 – Page 597 (**Two-sample.R**)

- R provides a *t.test() function* that can be used when raw data is available.

98. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from one (strongly agree) to five (strongly disagree). **Table 10.26** contains ten of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's Score	2	2	3	3	4	2	1	1	2	4
Husband's Score	2	2	1	3	2	1	1	1	2	4

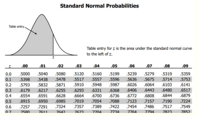
Table 10.26

```
> # Book Problem 10.98 - Page 597
> # Independent two-sample test - unknown population standard deviations
> wife <- c(2,2,3,3,4,2,1,1,2,4)
> husband <- c(2,2,1,3,2,1,1,1,2,4)
> t.test(wife,husband,paired=FALSE, alternative = "less")

        Welch Two Sample t-test

data:  wife and husband
t = 1.0797, df = 17.892, p-value = 0.8527
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.303276
sample estimates:
mean of x mean of y
      2.4       1.9
```

```
> n1 = length(wife)
> n2 = length(husband)
> x1_bar = mean(wife)
> x2_bar = mean(husband)
> s1 = sd(wife)
> s2 = sd(husband)
> se1 <- (s1^2/n1)
> se2 <- (s2^2/n2)
> std_err <- sqrt(se1+ se2)
> print(std_err)
[1] 0.4630815
> t_statistic <- (x1_bar - x2_bar)/std_err
> print(t_statistic)
[1] 1.079724
> deg_free_n <- (se1 + se2)^2
> deg_free_d <- (se1^2/(n1 -1)) + (se2^2/(n2 -1))
> deg_free <- deg_free_n/deg_free_d
> print(deg_free)
[1] 17.89193
> # Lower-tailed test
> p_value <- (pt(t_statistic,deg_free))
> print(p_value)
[1] 0.852697
```

Comparing Two Independent Population *Proportions* Book Problem 10.107 – Page 599 (**Two-sample.R**)

Practice with calculator 

Practice with Tables 

- Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in Table 10.27. Test at the 1% level of significance.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

Table 10.27

- Null Hypothesis: $\mathbf{H_o}$: $\mathbf{p_A - p_B = 0}$; Alternate Hypothesis: $\mathbf{H_a}$: $\mathbf{p_A - p_B < 0}$
- x_A= 67169, x_B = 42769; n_A= 248775, n_B = 155525;
- p'_A = 67169/248775;= 0.270, p'_B = 42769/155525= 0.275;

```
# Book Problem 10.107 - Page 599
# Independent two-sample test - Population Proportions
#
xA = 67169
nA = 248775
xB = 42769
nB = 155525
pprime_A = xA/nA
pprime_B = xB/nB
pc = (xA + xB)/(nA + nB)
std_err <- sqrt(pc*(1-pc)*((1/nA)+(1/nB)))
print(std_err)
z_statistic <- (pprime_A - pprime_B)/std_err
print(z_statistic)
# Lower-tailed test
p_value <- (pnorm(z_statistic,0,1))
#Upper-tailed test
#p_value <- (1 - pnorm(z_statistic,0,1))
# Two-tailed test
#p_value <- 2*(1-*pnorm(abs(z_statistic,0,1)))
print(p_value)
```

```
> print(std_err)
[1] 0.001438333
```

```
> print(z_statistic)
[1] -3.475269
```

```
> print(p_value)
[1] 0.0002551709
```

10.3 Comparing Two Independent Population Proportions

Pooled Proportion: $p_c = \dfrac{x_A + x_B}{n_A + n_B}$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right]$$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$.

Test Statistic (z-score): $z = \dfrac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$.

where

p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,

P_c is the pooled proportion, and n_A and n_B are the sample sizes.

Because p-value = 0, at a significance level of 0.05, the samples provide evidence that the proportion of obese southern women is loiwer than the proportion of obese southern men.

Matched or Paired Dependent Samples

- When using a hypothesis test for matched or paired samples, the following characteristics should be present:
  1. Two measurements (samples) are drawn from the same pair of individuals or objects. For example, measurements on a set of individuals, *before and after* a drug is taken.
  2. Differences in measurements are calculated from the matched or paired samples (i.e., for each individual it will be *after – before* or *before – after*).
  3. Simple random sampling is used.
  4. Sample sizes are often small.
  5. The differences form the sample that is used for the hypothesis test.
  6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

- In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, $\boldsymbol{\mu_d}$, is then tested using a Student's-t test for a single population mean with n–1 degrees of freedom, where n is the number of differences. The sample mean of differences is $\bar{X}_d$ and the sample standard deviation of differences is s_d.

- Null Hypothesis: $\mathbf{H_o}: \boldsymbol{\mu_d} = \mathbf{0}$; Alternate Hypothesis: $\mathbf{H_a}: \boldsymbol{\mu_d} <, \neq, > \ \mathbf{0}$

The test statistic (*t*-score) is:

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Practice with t-Tables

Practice with calculator

Matched or Paired Dependent Samples – Book Problem 10.115 – Page 601 (Two-sample.R)

- Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded in **Table 10.30**. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Table 10.30

The test statistic (*t*-score) is:

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

- The difference d is obtained as (Starting – Ending) and we want to test if this is significantly negative.

- Null Hypothesis: $\mathbf{H_o: \mu_d = 0}$; Alternate Hypothesis: $\mathbf{H_a: \mu_d > 0}$

Because p-value = 0.1353, at the 5% significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

```
# Book Problem 10.115 - Page 601
# Dependent samples - Matched or Paired samples - Paired t-test
#
start_chol <- c(140, 220, 110, 240, 200, 180, 190, 360, 280, 260)
end_chol <- c(140, 230, 120, 220, 190, 150, 200, 300, 300, 240)
xd <- start_chol - end_chol
x_bar_d <-mean(xd)
std_err <- sd(xd)/sqrt(length(xd))
print(x_bar_d)
print(std_err)
t_statistic <- x_bar_d/std_err
print(t_statistic)
deg_free <- length(xd) - 1
print(deg_free)
# Lower-tailed test
# p_value <- (pt(t_statistic,deg_free))
# Upper-tailed test
p_value <- (1 - pt(t_statistic,deg_free))
# Two-tailed test
#p_value <- 2*(1-pt(abs(t_statistic,deg_free)))
print(p_value)
#
t.test(start_chol,end_chol,paired=TRUE, alternative = "greater")
```

```
> # Book Problem 10.115 - Page 601
> # Dependent samples - Matched or Paired samples - Paired t-test
> #
> start_chol <- c(140, 220, 110, 240, 200, 180, 190, 360, 280, 260)
> end_chol <- c(140, 230, 120, 220, 190, 150, 200, 300, 300, 240)
> xd <- start_chol - end_chol
> x_bar_d <-mean(xd)
> std_err <- sd(xd)/sqrt(length(xd))
> print(x_bar_d)
[1] 9
> print(std_err)
[1] 7.666667
> t_statistic <- x_bar_d/std_err
> print(t_statistic)
[1] 1.173913
> deg_free <- length(xd) - 1
> print(deg_free)
[1] 9
> # Lower-tailed test
> # p_value <- (pt(t_statistic,deg_free))
> # Upper-tailed test
> p_value <- (1 - pt(t_statistic,deg_free))
> # Two-tailed test
> #p_value <- 2*(1-pt(abs(t_statistic,deg_free)))
> print(p_value)
[1] 0.1352787
> #
> t.test(start_chol,end_chol,paired=TRUE, alternative = "greater")

        Paired t-test

data:  start_chol and end_chol
t = 1.1739, df = 9, p-value = 0.1353
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -5.053866       Inf
sample estimates:
mean of the differences
                      9
```

Analysis of Variance - ANOVA

- Thus far, we saw how to test hypothesis involving the means or proportions of **two** groups.
- Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups (**more than two groups**).
  - For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.
- For hypothesis tests comparing averages between more than two groups, statisticians have developed a method called "Analysis of Variance" (abbreviated ANOVA).
- We will study ANOVA later in the semester.