Chapter 5 Section 5.1

Review of two-sample t-test
Analysis of Variance = ANOVA or AOV

In both cases:
• The response variable is quantitative.
• The explanatory variable is categorical
  – For a two-sample t-test, it has 2 categories.
  – For ANOVA, it has 2 or more categories.
  – However, when \( k = 2 \), ANOVA is equivalent to a two-sided two-sample t-test.

Some basic definitions

• A factor is a categorical explanatory variable.
• A level of a factor is one category.
• Categories are sometimes called groups.

Example

Does average time spent studying per week differ by type of major? Take random sample from each type of major, or one random sample and divide into the 3 majors.

• \( Y \) = time spent studying per week [response var.]
• Factor = Category of major (sciences, social sciences, humanities) [explanatory variable]
• The 3 levels of the factor (the 3 groups) are sciences, social sciences, humanities.

Two-sample t-test (Review)

Data: Independent samples from two groups
Summary statistics:
\[
\begin{array}{c|c|c}
& n_1 & \bar{Y}_1, s_1 \\
& n_2 & \bar{Y}_2, s_2 \\
\end{array}
\]

Conditions:
1. Normal populations (or large \( n \)'s)
2. Equal variances (sometimes)

Write as \( Y_k \sim N(\mu_k, \sigma) \), where
\( k = \text{group} (1 \text{ or } 2) \)
\( i = \text{ individual within group} = 1, 2, \ldots, n_k \)

Pooled Two-sample t-test (Review?)

Pooled variance:
\[
S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}
\]

Test statistic:
\[
t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

Reference distribution:
\[
t_{n_1 + n_2 - 2}
\]

Does Active Pulse Depend on Gender?

Explain why on white board.
Two-sample t-test ($R$)

```r
> t.test(Active~Gender, var.equal=TRUE)
Two Sample t-test

data:  Active by Gender
t = -2.7436, df = 230, p-value = 0.006556
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: -11.503416 -1.887046
sample estimates:
mean in group 0 mean in group 1
88.12295        94.81818
```

ANOVA for Means

```
> summary(aov(Active~Gender))
Df Sum Sq Mean Sq F value   Pr(>F)
Gender        1   2593 2592.96  7.5274 0.006556 **
Residuals   230  79228  344.47
---
> oneway.test(Active~Gender, var.equal=TRUE)
One-way analysis of means
data:  Active and Gender
F = 7.5274, num df = 1, denom df = 230, p-value = 0.006556
```

Conditions and assumptions

1. Normal populations (or large $n$ for each group)
2. Equal variances for all observations
3. All observations are independent, within and between groups.

Test: Are Group Means Equal (in the Population)?

<table>
<thead>
<tr>
<th>p-value</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.39</td>
<td>0.6</td>
</tr>
<tr>
<td>0.0015</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Some possible ways to get independent data

1. $K$ separate populations, take random sample from each.
   Ex: Groups = 4 regions of the US
   $Y_{ik}$ = time spent commuting to work

2. Take one random sample and measure response variable $Y$, and categorical explanatory variable $X$.
   Ex: Groups = type of major (Science, SocSci, Humanities)
   $Y_{ik}$ = time spent studying per week

3. Randomized experiment with $K$ treatments
   Ex: 30 cities available for experiment with 3 roadside billboards
   Randomly assign 10 cities to each type of billboard
   $Y_{ik}$ = Sales of product after 6 months in City $i$, with billboard $k$. 
Test: Are Group Means Equal (in the Population)?

What’s different?

Same n and means but smaller SDs

p-value = 0.39  
Effect size = 0.6

p-value = 0.0036  
Effect size = 1.5

Same (approx.) range among the means but larger n

p-value = 0.39  
Effect size = 0.57 to two decimal places

p-value = 0.002  
Effect size = 0.56

Summary of what decreases p-value and increases power of the test (easier to reject null hypothesis):

- Bigger difference between the means  
  - Increased effect size
- Smaller standard deviations  
  - Increased effect size
- Larger sample sizes  
  - Not an increase in effect size

Example: Random sample of \( n_k = 5 \) scores (Ys) from each of \( K = 4 \) exams (there are 4 levels)

<table>
<thead>
<tr>
<th>Exam #1</th>
<th>62, 94, 68, 86, 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam #2</td>
<td>87, 95, 93, 97, 63</td>
</tr>
<tr>
<td>Exam #3</td>
<td>74, 86, 82, 70, 28</td>
</tr>
<tr>
<td>Exam #4</td>
<td>77, 89, 73, 79, 47</td>
</tr>
<tr>
<td>Overall</td>
<td>20</td>
</tr>
</tbody>
</table>

Is there a difference in population mean score among the four exams?

Test:  
\[ H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \]

\[ H_1: \text{Some } \mu_k \neq \mu_j \]

Helpful R Command

```r
> means=tapply(X=Grade,INDEX=Exam,FUN=mean)  #FUNction = mean
> means
1 72  87  68  86
2  66  50
3  44  77  89  73  79  47
4  77  89  73  79  47

> sds=tapply(Grade,Exam,sd)  #we don’t have to state “X=”, etc.
> Sds  #standard deviations
1 17.88854
2 13.92839
3 23.23790
4 15.68439

> ns=tapply(Grade,Exam,length)  #length = sample size
> n
1 2 3 4
5 5 5 5

ANOVA (Means) Model

\[ Y = \mu_k + \varepsilon \]

Mean for group \( k \)

\[ \mu_k \]

\( N(0,\sigma) \) random error

Under \( H_0 \) (\( \mu_k \)'s all equal)  
\[ \bar{\mu_k} = \bar{Y} \]

Under \( H_1 \) (\( \mu_k \)'s differ)  
\[ \bar{\mu_k} = \bar{Y_k} \]

These are the least squares estimates for \( \mu_k \) for the two hypotheses.
“Predicting” in ANOVA Model
If the group means are the same ($H_0$):
\[ \bar{y} = \bar{Y} \] for all groups → residual = $Y - \bar{Y}$
If the group means can be different ($H_1$):
\[ \bar{y}_k = \bar{Y}_k \] for $k^{th}$ group → residual = $Y - \bar{Y}_k$
Do we do “significantly” better with separate means?
Compare sums of squared residuals…
\[ SSTotal = \sum (Y - \bar{Y})^2 \text{ vs. } SSE = \sum (Y - \bar{Y}_k)^2 \]

Partitioning Variability
\[ Data = Model + Error \]
\[ Y = \mu_k + \epsilon \]
TOTAL variation in response, $Y$
\[ \text{Variation explained by MODEL} + \text{Unexplained variation in RESIDUALS} \]

Key question: Does the MODEL explain a “significant” amount of the TOTAL variability?

Partitioning Variability ANOVA for Group Means
\[ Y = \mu_k + \epsilon \]
\[ (y - \bar{Y}) = (\bar{Y}_k - \bar{Y}) + (y - \bar{Y}_k) \]
\[ \sum (y - \bar{Y})^2 = \sum (\bar{Y}_k - \bar{Y})^2 + \sum (y - \bar{Y}_k)^2 \]
\[ SSTotal = SSGroups + SSE \]

Using familiar regression terminology
\[ \sum (y - \bar{Y})^2 = \sum (\bar{Y}_k - \bar{Y})^2 + \sum (y - \bar{Y}_k)^2 \]
Residuals if $H_0$ is true (same mean)

“Explained” by model with separate means
Still unexplained with separate means

\[ SSTotal = SSGroups + SSE \]

Decomposition: Four Exams

| Exam #1: 62, 94, 68, 86, 50 | Mean = 72.0 | $S_k$ = 17.89 |
| Exam #2: 87, 95, 93, 97, 63 | Mean = 87.0 | $S_k$ = 13.93 |
| Exam #3: 74, 86, 82, 70, 28 | Mean = 68.0 | $S_k$ = 23.24 |
| Exam #4: 77, 89, 73, 79, 47 | Mean = 73.0 | $S_k$ = 15.68 |
| Overall | Mean = 75.0 | $S_k$ = 18.11 |

\[ SSGroups = \sum (72 - 75)^2 + \sum (17 - 18)^2 + \sum (13 - 14)^2 + \sum (23 - 25)^2 = 1030 \]
\[ SSE = (62 - 72)^2 + (94 - 75)^2 + \cdots + (47 - 73)^2 = 5200 \]
\[ SSTotal = (62 - 75)^2 + (94 - 75)^2 + \cdots + (47 - 75)^2 = 6250 \]

Etc.
ANOVA Table (for $K$ Group Means)

$H_0$: $\mu_1 = \mu_2 = \ldots = \mu_K$

$H_1$: Some $\mu_k \neq \mu_j$

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>S.S.</th>
<th>M.S.</th>
<th>t.s.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>$K - 1$</td>
<td>SSGroups</td>
<td>$\frac{SSGroups}{K - 1}$</td>
<td>MGroups</td>
<td>use $F_{K - 1, n - K}$</td>
</tr>
<tr>
<td>Error</td>
<td>$n - K$</td>
<td>SSE</td>
<td>$\frac{s^2}{n - K}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>SSTotal</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Small p-value $\Rightarrow$ Reject $H_0$ $\Rightarrow$ There is evidence of a difference among the population means of the $K$ groups.

ANOVA Output in R

```r
> model=lm(Grade~as.factor(Exam))
> model

Call:
lm(formula = Grade ~ as.factor(Exam))

Residuals:
     Min      1Q  Median      3Q     Max

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.0000    17.8937   4.042 0.002077 **
as.factor(Exam)2  20.2000     17.8937   1.147 0.296263
as.factor(Exam)3  63.7000     23.2393   2.740 0.020275 *
as.factor(Exam)4  54.5000     23.2393   2.343 0.052240 .

Residual standard error: 18.02 on 16 degrees of freedom
Multiple R-squared: 0.3856, Adjusted R-squared: 0.3032
F-statistic: 2.621 on 3 and 16 DF, p-value: 0.08517
```

Note: $n$ = total sample size

Estimated effects may be unbalanced

```r
> summary(model)

Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(Exam)  3 1030.0   343.3  1.0564  0.3951
Residuals       16 5200.0   325.0

> 1-pf(1.0564,3,16) #if the P-value hadn't been given
[1] 0.3950020
```

After Installing Three Packages in R: gplots, gdata, gtools

```r
> plotmeans(Grade~Exam)
```

95% CI's for each group mean shown in blue. Notice the substantial overlap.

Partition Variability (different formulas) + df

Between groups: \((\text{d.f.} = K - 1)\)

\[
\text{SSGroups} = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + \cdots + n_K(\bar{y}_K - \bar{y})^2
\]

Within groups: \((\text{d.f.} = n - K)\)

\[
\text{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_K - 1)s_K^2
\]

Total: \((\text{d.f.} = n - 1)\)

\[
\text{SSTotal} = \sum (y - \bar{y})^2 = (n - 1)s^2_f
\]

\[
\text{SSTotal} = \text{SSGroups} + \text{SSE}
\]

Example: Four Exams

<table>
<thead>
<tr>
<th>Exam #1</th>
<th>62, 94, 68, 86, 50</th>
<th>$n_k$</th>
<th>Mean</th>
<th>$S_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam #2</td>
<td>87, 95, 93, 97, 63</td>
<td>5</td>
<td>72.0</td>
<td>17.89</td>
</tr>
<tr>
<td>Exam #3</td>
<td>74, 86, 82, 70, 28</td>
<td>5</td>
<td>68.0</td>
<td>23.24</td>
</tr>
<tr>
<td>Exam #4</td>
<td>77, 89, 73, 79, 47</td>
<td>5</td>
<td>73.0</td>
<td>15.68</td>
</tr>
<tr>
<td>Overall</td>
<td>20</td>
<td>75.0</td>
<td>18.11</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{SSGroups} = 5(72 - 75)^2 + 5(87 - 75)^2 + 5(93 - 75)^2 + 5(97 - 75)^2 = 1036
\]

\[
\text{SSE} = 4(17.89)^2 + 4(13.93)^2 + 4(23.24)^2 + 4(15.68)^2 = 5200
\]

\[
\text{SSTotal} = 19(18.11)^2 = 6230 \quad \text{(up to roundoff)}
\]

Alternate Form: ANOVA Model for Means

\[
Y = \mu + \alpha_k + \varepsilon
\]

Grand mean \quad Effect for $k^{th}$ group \quad Random error

\[
\hat{\alpha}_k = \bar{y}_k - \bar{y}
\]

Note: $\alpha_k$ sum to 0

\[
H_0: \mu_1 = \mu_2 = \ldots = \mu_K
\]

\[
H_1: \text{Some } \alpha_k \neq 0
\]

\[
H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_K = 0
\]

\[
H_1: \text{Some } \alpha_k \neq 0
\]
Estimating the common variance

\[ \varepsilon \sim N(0, \sigma_\varepsilon) \quad Y_{ik} \sim N(\mu_k, \sigma) \]

\[ SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_K - 1)s_K^2 \]

\[ \text{MSE} = \frac{SSE}{n - k} \]

MSE is an estimate of the (common) population variance \( \text{MSE} = \sigma^2 \)

Example: Four Exams

<table>
<thead>
<tr>
<th>Exam #1:</th>
<th>62, 94, 68, 86, 50</th>
<th>n_k</th>
<th>5</th>
<th>Mean</th>
<th>72.0</th>
<th>S_i</th>
<th>17.89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam #2:</td>
<td>87, 95, 93, 97, 63</td>
<td>5</td>
<td>87.0</td>
<td>13.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exam #3:</td>
<td>74, 86, 82, 70, 28</td>
<td>5</td>
<td>66.0</td>
<td>23.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exam #4:</td>
<td>77, 89, 73, 79, 47</td>
<td>5</td>
<td>75.0</td>
<td>15.68</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall

\[ \text{MSE} = \frac{5200}{16} = 325 = \text{estimate of popn variance} \]

\[ \sqrt{\text{MSE}} = \sqrt{325} = 18.03 \]

Section 5.2: Checking Conditions for ANOVA

\[ \varepsilon \sim N(0, \sigma_\varepsilon) \]

Check with residuals.

Zero mean: Always holds for sample residuals.

Constant variance:

- Plots and numerical checks:
  - Plot residuals vs. fits
  - Plot Y versus group, or boxplot for each group
  - Compare standard deviations of groups; check if largest is more than twice value of smallest.

Note: This is less crucial if the sample sizes are equal.

Checking Conditions, continued

Normality:

- Histogram of residuals
- Normal probability plot of residuals

Independence:

- Pay attention to data collection method. (See earlier slide.)

Plot of data and histogram of residuals

Section 5.3: Scope of Inference

Allocation of Units to Groups

Using Randomization

Random sample selected; units assigned randomly to treatment groups

Random samples selected from separate populations

Study units are found, then randomly assigned to treatment groups

Available units from separate populations are studied

Causal inferences can be drawn

Not using Randomization

Selection of Units

Not at Random

Available units from separate populations are studied

Inferences can be drawn to populations

Random sample selected; units assigned randomly to treatment groups

Random samples selected from separate populations

Study units are found, then randomly assigned to treatment groups

Available units from separate populations are studied

Causal inferences can be drawn

Not at Random

Available units from separate populations are studied

Inferences can be drawn to populations
Some Examples

Exercise 5.19 – Life spans  Not random
Exercise 5.28 – Fenthion  Random samples
Exercise 5.30 – Blood pressure  Random samples, cause/effect?
Example 5.1 – Fruit flies  Random allocation

Now do example of seat location.