HUMBEHV 3ST3
Evaluating hypotheses about 3 or more groups
(One-Way ANOVA)

Week 10
Prof. Patrick Bennett

Analysis of Variance (ANOVA)

- Concepts from previous lectures
  - comparisons of group means
  - null hypothesis significance testing
    - are differences among means unusual if H0 is true?
  - sampling distributions
- Extensions/generalizations of t tests
  - comparisons among 3 or more means
  - can measure effects of 2 or more independent variables
    - evaluate interactions between variables

Giancola & Corman (2007)

- effects of alcohol & attention on aggressive behaviour
- hypothesis: alcohol facilitates aggression by focusing attention on salient provocative cues
  - presenting Ss with a task should reduce attention on provocative cues & reduce aggression
  - but if task is too complex, Ss will not focus on task & aggressive behaviour will increase
- Ss were given alcohol to raise blood-alcohol level to 0.10%
- Different groups of Ss performed a memory task that varied in difficulty
- All Ss first received shocks from partner based on task performance
  - later, all Ss delivered shocks to partner
  - aggression measured as shock “level” (intensity & duration of shocks)
- What is/are the dependent & independent variables?

Giancola & Corman (2007)

- Subjects assigned randomly to 5 groups
  - groups varied in task difficulty (independent variable)
  - 12 Ss per group
- Shock “level” (intensity & duration) was dependent variable
- Compare mean shock levels in each group
  - Null Hypothesis (H0): population mean shock level is equal in all groups
  - Alternative Hypothesis (H1): population mean shock level not equal in all groups
  - [only source of variation between group means is sampling error]
- Are observed group differences large if H0 is true?
- Cannot evaluate H0 & H1 with a t test because we have 5 (not 2) groups
- Instead, evaluate H0 & H1 using the Analysis of Variance (ANOVA)
Why evaluate differences in group means with the analysis of variance?

By hypothesis, groups differ only in terms of mean aggression score

When null hypothesis is true, what causes variation within each group?
- population error variance $\sigma^2_e$
- variance within each group is an estimate of $\sigma^2_e$
- average within-group variance is best estimate of $\sigma^2_e$

$\sigma^2_e = s^2_j = s^2_1 + s^2_2 + s^2_3 + s^2_4 + s^2_5$

When $H_0$ is true, variance within and between groups is related to $\sigma^2_e$
- $\sigma^2_e$ is estimated by average of within-group variances $[MS_{within}, MS_{error}]$
- $\sigma^2_e$ also is estimated by product of $n$ and between-group variance $[MS_{group}]$
- Hence, when $H_0$ is true, within- & between-group variances provide 2 independent estimates of $\sigma^2_e$
- therefore, $MS_{group}$ & $MS_{error}$ should be similar

$\sigma^2_e = \frac{s^2_j \times n}{5}$

$\sigma^2_e = \frac{\text{average within-group variance}}{n}$

$\sigma^2_e = n \times \bar{s}^2_X$
Giancola & Corman (2007)

- When $H_0$ is false:
  - within-group variance still determined by $\sigma_e^2$
  - but variance between group means is related to $\sigma_e^2$ and to effects of the independent variable
  - hence, variance between group means should be greater than variance predicted solely by $[\sigma_e^2 + \sigma^2_{\text{effect}}]$ for grouping variable
- so $MS_{\text{group}}$ should be greater than $MS_{\text{error}}$

ANOVA (general strategy)

- compute measures of within-group & between-group variation:
  - within-group & between-group Mean Squares ($MS_W$ & $MS_B$)
- when $H_0$ is true, $MS_W$ & $MS_B$ are estimates of $\sigma_e^2$
  - so $MS_W$ & $MS_B$ should be similar
- when $H_0$ is false, $MS_B$ is the sum of $\sigma_e^2$ & effect of independent variable
  - so we expect $MS_B > MS_W$
- Therefore, we will evaluate $H_0$ by determining if the ratio $MS_B/MS_W$ is unusually large given that $H_0$ is true
ANOVA calculations

• Break overall variation into 2 parts: between- & within_group sums of squares
• Sums of Squares: sum of squared deviations from mean
  - $SS_{total} = \sum (X - \bar{X}_{gm})^2 \quad [\bar{X}_{gm} = \text{grand mean (mean of all scores)}]$  
  - $SS_{between\_group} = n \times [(\text{Mean Group}) - \bar{X}_{gm}]^2 \quad [n = \text{scores per group}]$  
  - $SS_{within\_group} = SS_{total} - SS_{between\_group}$  
  - $SS_{total} = SS_{between\_group} + SS_{within\_group}$
• degrees of freedom:  
  - $df_{total} = N - 1 \quad (N \text{ is the total number of observations})$  
  - $df_{between\_group} = k - 1 \quad (k \text{ is the number of groups})$  
  - $df_{within\_group} = df_{total} - df_{between\_group}$
• Mean Squares:  
  - $MS_{within\_group} = SS_{within\_group}/df_{within\_group} \quad [\text{estimate of } \sigma_v^2 \text{ when } H0 \text{ is true}]$  
  - $MS_{between\_group} = SS_{between\_group}/df_{between\_group} \quad [\text{estimate of } \sigma^2 \text{ when } H0 \text{ is true}]$  
• $F$ statistic: $MS_{between\_group}/MS_{within\_group}$

terms in ANOVA table

• Alternative terms for sums of squares:  
  - $SS_{between\_groups}$: $SS_{between}$, $SS_{group}$, $SS_{treatment}$  
  - $SS_{within\_groups}$: $SS_{within}$, $SS_{error}$, $SS_{residuals}$

• Alternative terms for Mean Squares:  
  - $MS_{between\_group}$: $MS_{between}$, $MS_{group}$, $MS_{treatment}$  
  - $MS_{within\_group}$: $MS_{within}$, $MS_{error}$, $MS_{residuals}$

ANOVA data...

![Figure 16.3](image)

Plot of Giancola and Corman's data on aggression as a function of level of distraction. Bars represent ±1 standard error.

<table>
<thead>
<tr>
<th>Table 16.4</th>
<th>Level of Shock Administered as a Function of Task Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D0</td>
</tr>
<tr>
<td>1.28</td>
<td>-1.18</td>
</tr>
<tr>
<td>1.35</td>
<td>0.15</td>
</tr>
<tr>
<td>3.33</td>
<td>1.16</td>
</tr>
<tr>
<td>3.36</td>
<td>2.61</td>
</tr>
<tr>
<td>3.59</td>
<td>0.66</td>
</tr>
<tr>
<td>3.25</td>
<td>1.32</td>
</tr>
<tr>
<td>2.98</td>
<td>0.73</td>
</tr>
<tr>
<td>1.53</td>
<td>-1.06</td>
</tr>
<tr>
<td>-2.68</td>
<td>0.24</td>
</tr>
<tr>
<td>2.64</td>
<td>0.27</td>
</tr>
<tr>
<td>1.26</td>
<td>0.72</td>
</tr>
<tr>
<td>1.06</td>
<td>2.28</td>
</tr>
</tbody>
</table>

Mean: 1.802  0.675  -0.912  0.544  1.889  0.880  0.000
St. Dev.: 1.656  1.140  0.515  1.180  2.370  1.800
Variance: 2.741  1.299  0.265  1.294  5.616  3.068
### ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4</td>
<td>62.460</td>
<td>15.615</td>
<td>6.90</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>124.458</td>
<td>2.363</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>186.918</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N = 5 x 12 = 60
df\text{Total} = N-1 = 59
df\text{Treatment} = 5-1 = 4
df\text{Error} = 59-5 = 55

### F distribution

- ratio of 2 independent estimates of the same variance follows F distribution
- F distribution is a theoretical distribution (like t distribution)
- family of curves defined by two parameters:
  - df in numerator & df in denominator
- large values of F are unusual when numerator & denominator are estimates of same population variance

### ANOVA Assumptions

- data in each group are distributed normally
- groups have the same variance
- observations are independent
- violations of assumptions mean p-value for F test will be inaccurate

### F test

- observed $F(4,55) = 6.90$
- critical $F(4,55) = 2.54$ ($\alpha=.05$)
- $F_{\text{observed}} > F_{\text{critical}}$
- reject $H_0$ (all means equal), $F(4,55)=6.90$, $p<.05$
- note that 2-tailed test does not make sense here because $H_1$ always predicts $MS_{\text{Group}} > MS_{\text{Error}}$
Non-normality & non-constant variance

• ANOVA reasonably robust to deviations from normality
  - if deviations are similar in all groups
  - robustness declines if n is not equal across groups
  - also declines if deviations differ across groups
    ‣ e.g., positive skew in 1 group, negative skew in others
• ANOVA is reasonably robust to 3-4 fold differences in variances
  - if scores are normally distributed and equal n per group

Common tests for non-normality

• Kolmogorov-Smirnov test
• Shapiro-Wilk’s test

```r
shapiro.test(residuals(mood.full))

## Shapiro-Wilk normality test
## data: residuals(mood.full)
## W = 0.85, p-value = 5e-04
```

Null hypothesis is that within-group scores are distributed normally.

Bartlett test for constant variance

```r
bartlett.test(mood.data$mood,mood.data$group)

## Bartlett test of homogeneity of variances
## data: mood.data$mood and mood.data$group
## Bartlett's K-squared = 2.6, df = 2, p-value = 0.2
```

Bartlett test is a common procedure for evaluating the null hypothesis that variance is constant across groups

When assumptions are violated

• perform ANOVA on transformed data
  - square-root, log, & inverse-sine transformations common
  - conclusions apply to transformed data
• Welch df correction for non-constant variance [assumes normality]
• Kruskal-Wallis test for group differences
  - does not assume normality or constant variance
  - KW test evaluates null hypothesis that means of ranked data are the same in each group
  - if we assume that distributions for each group have same shape (not necessarily normal), then KW test evaluates null hypothesis that group MEDIANS are equal
**Why measure association strength or effect size?**

- P-values are poor measures of the strength of an effect - depend **strongly** on sample size (n)
- They are expressed in units that are not easily related to independent and dependent variables
- Alternative indices of effect “strength”:
  - Association strength (proportion of variance accounted for)
  - Effect size (standardized differences among group means)

**Association strength**

- Association Strength:
  - **Proportion** of total variation in scores associated with grouping variable
  - Varies between 0 & 1
  - Similar to $R^2$ measure in linear regression
- **Eta-squared ($\eta^2$)**
  - $SS_{\text{group}}/SS_{\text{total}} = SS_{\text{group}}/(SS_{\text{group}} + SS_{\text{error}})$

**Association Strength (eta-squared)**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>3</td>
<td>33.8</td>
<td>11.26</td>
<td>2.39</td>
<td>0.098</td>
</tr>
<tr>
<td>error</td>
<td>20</td>
<td>93.9</td>
<td>4.69</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\eta^2 = \frac{SS_{\text{group}}}{SS_{\text{group}} + SS_{\text{error}}} = 0.26$

$r^2 = 0.504^2 = 0.25$

Eta-squared analogous to $r^2$ proportion of variance accounted for by grouping variable
Association strength

• Association Strength:
  - proportion of total variation in scores associated with grouping variable
  - varies between 0 & 1
  - similar to $R^2$ measure in linear regression
  - eta-squared ($\eta^2$)
    $SS_{group}/SS_{total} = SS_{group}/(SS_{group}+SS_{error})$
    biassed estimate of association score-group association in population
• omega-squared ($\omega^2$)
  - like $\eta^2$, $\omega^2$ measures association between scores & grouping variable
  - less biased than $\eta^2$
  - $\omega^2$ typically is less than $\eta^2$

Effect size

• Effect Size: standardized difference between group means
  - difference btwn group means relative to within-group SD
• Cohen's $d$: standardized difference between 2 group means
  - number of within-group SD between 2 group means (common with t tests)

Effect Size (Cohen’s $d$)

• Cohen’s $d$: common measure of effect size
• difference between means divided by within group standard deviation
• Cohen (1988) guidelines:
  - small effect: $d=0.2$
  - medium effect: $d=0.5$
  - large effect: $d=0.8$
• Unlike $p$ value, Cohen’s $d$ does not depend on $N$

$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$
estimate of Cohen's $d$

$d = \frac{\mu_1 - \mu_2}{\sigma}$
Effect of Sample Size on Cohen’s d

- \( \mu_1 = 100, \mu_2 = 102, \sigma = 15 \)
- 2 sample t test (medium size samples):
  - N=100 per group
  - \( t = 1.38, df = 198, p = 0.167 \)
  - \( d = 0.19 \) (small)
- 2 sample t test (large samples):
  - N=1000 per group
  - \( t = 3.48, df = 1998, p = 0.0015 \)
  - \( d = 0.14 \) (small)

\[
\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s_p}
\]

Effect size

- Effect Size: standardized difference between group means
  - difference between group means relative to within-group SD
- Cohen’s d: standardized difference between 2 group means
  - number of within-group SD between 2 group means (common with t tests)
- Cohen’s f: generalization of Cohen’s d to more than 2 groups
  - average standardized difference between group means & grand mean
  - similar to average Cohen’s d for each pair of groups

\[
Cohen's f = \frac{\text{standard deviation (group means)}}{\sqrt{\text{MS}_{error}}}
\]

low association strength & small effect size

- variation between groups is very small compared to variation within groups
- \( \eta^2 = 55/(55+4238) = 0.013 \)
  - \( \omega^2 \equiv 0 \)
- Cohen’s f \( \equiv 0 \)
  - average difference between group means and overall mean is approximately zero
high association strength & large effect size

- variation between groups is large compared to variation within groups
- \( \eta^2 = 3226/(3226+1067) = 0.751 \)
- \( \omega^2 \approx 0.73 \)
- Cohen's \( f \approx 1.65 \)
- on average, group means are 1.65\( \sigma_e \) away from overall mean

Giancola & Corman (2007) [Aggression Study]

- omnibus \( F=6.901, p<.01 \)
- reject null hypothesis of no group difference
  - \( \eta^2 = 62.46/(62.46+124.46) = 0.33 \)
  - \( \omega^2 = 0.28 \)
  - Cohen's \( f = 0.68 \)
  - big effect of group

Association strength & effect size

Guidelines from Cohen (1988)

<table>
<thead>
<tr>
<th>Magnitude of association or effect</th>
<th>Omega-squared</th>
<th>Cohen's f</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>medium</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>large</td>
<td>0.14</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Effect Size (Cohen's \( f \))

A measure of difference among means relative to within-group standard deviation

\[
\text{Cohen's } f = \frac{\text{standard deviation (group means)}}{\sqrt{\text{MS}_{\text{error}}}}
\]

Cohen's \( f = 0.47 \)

large effect that is not significant due to small N
Multiple Comparisons

Giancola & Corman (2007) [Aggression Study]

- significant omnibus F suggests H0 is false
  - reject H0 (group means are all equal) in favour of H1 (group means are not all equal)
- but how do groups differ?
- to answer this question, researchers often perform multiple comparisons/contrasts between groups
  - which pairwise group differences are significant?
  - does mean of g3 differ from mean of groups g1 & g5?

<table>
<thead>
<tr>
<th>Group (Distraction Condition)</th>
<th>Aggression Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>-1.0</td>
</tr>
<tr>
<td>g2</td>
<td>-0.5</td>
</tr>
<tr>
<td>g3</td>
<td>0.0</td>
</tr>
<tr>
<td>g4</td>
<td>0.5</td>
</tr>
<tr>
<td>g5</td>
<td>1.0</td>
</tr>
<tr>
<td>df</td>
<td>4</td>
</tr>
<tr>
<td>SS</td>
<td>62.46</td>
</tr>
<tr>
<td>MS</td>
<td>15.62</td>
</tr>
<tr>
<td>F</td>
<td>6.901</td>
</tr>
<tr>
<td>p</td>
<td>0.000142</td>
</tr>
</tbody>
</table>

df, SS, MS, F, p

Multiple Comparisons & Type I Error

- Consider case of evaluating each pairwise difference between 5 groups
  - total of 10 pairwise tests:
    » g1 vs g2, g1 vs g3, g1 vs g4, g1 vs g5
    » g2 vs g3, g2 vs g4, g2 vs g5
    » g3 vs g4, g3 vs g5
    » g4 vs g5
  - For each comparison, we use a t test with α = 0.05
  - Assume null hypothesis is true: all group means are the same
  - For our 10 tests, what is probability of making at least one Type I error?
    » p(at least 1 Type I error) = 1 - (1-.05)^10 = 0.40 [this value differs from textbook]
    » familywise α [α for entire set/family of tests] is 0.40, not .05
Giancola & Corman (2007) [Aggression Study]

Using Tukey HSD to evaluate all pairwise differences between group means:

<table>
<thead>
<tr>
<th>Group (Distraction Condition)</th>
<th>Aggression Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>-1.0</td>
</tr>
<tr>
<td>g2</td>
<td>-0.5</td>
</tr>
<tr>
<td>g3</td>
<td>0.0</td>
</tr>
<tr>
<td>g4</td>
<td>0.5</td>
</tr>
<tr>
<td>g5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Tukey multiple comparisons of means
95% family-wise confidence level

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p-adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>g2-g1</td>
<td>-2.860</td>
<td>0.605</td>
<td>0.364</td>
</tr>
<tr>
<td>g3-g1</td>
<td>-4.447</td>
<td>-0.983</td>
<td>0.000</td>
</tr>
<tr>
<td>g4-g1</td>
<td>-2.990</td>
<td>0.474</td>
<td>0.257</td>
</tr>
<tr>
<td>g5-g1</td>
<td>-1.645</td>
<td>1.819</td>
<td>1.000</td>
</tr>
<tr>
<td>g3-g2</td>
<td>-3.320</td>
<td>0.145</td>
<td>0.087</td>
</tr>
<tr>
<td>g4-g2</td>
<td>-1.863</td>
<td>1.601</td>
<td>1.000</td>
</tr>
<tr>
<td>g5-g2</td>
<td>-0.518</td>
<td>2.946</td>
<td>0.291</td>
</tr>
<tr>
<td>g3-g3</td>
<td>-2.755</td>
<td>0.189</td>
<td>0.139</td>
</tr>
<tr>
<td>g4-g3</td>
<td>-1.290</td>
<td>-2.990</td>
<td>0.474</td>
</tr>
<tr>
<td>g5-g3</td>
<td>-0.518</td>
<td>2.946</td>
<td>0.291</td>
</tr>
<tr>
<td>g4-g4</td>
<td>-1.275</td>
<td>-2.990</td>
<td>0.474</td>
</tr>
<tr>
<td>g5-g4</td>
<td>-0.518</td>
<td>2.946</td>
<td>0.291</td>
</tr>
</tbody>
</table>

Omnibus F & Multiple Comparisons

- Omnibus F and multiple comparison procedures often yield different results
  - e.g., omnibus F may be significant, but Tukey HSD may fail to find any significant pairwise differences
  - e.g., Tukey HSD may find significant pairwise differences but omnibus F may not be significant
  - these procedures evaluate different null hypotheses and therefore different results are not unexpected

- One exception: Omnibus F & Scheffe test are consistent:
  - if omnibus F is significant there is at least one linear contrast that is significant with Scheffe method
  - if omnibus F is not significant then no linear contrast will be significant with Scheffe method

One-way ANOVA summary

- One-way ANOVA: 1 independent variable with more than 2 groups
- Assumptions:
  - normality, homogeneity of variance, independent observations
  - Bartlett, Kolmogorov-Smirnov, & Shapiro-Wilks tests
  - alternative analyses: data transformations, Welch df correction, & Kruskal-Wallis test
- Sums-of-squares, degrees-of-freedom, Mean Squares
- omnibus F test:
  - $F = \frac{MS_{group}}{MS_{error}}$
  - F statistic follows F distribution
  - use F to evaluate null hypothesis of no difference among group means
- Association Strength & Effect Size: eta-squared, omega-squared, Cohen's $f$
- Multiple Comparisons & Linear Contrasts

Other procedures exist to do more complex comparisons:

- e.g., compare mean of g3 to the mean of g1 & g2 (combined)
- e.g., compare difference between g1 & g2 to difference between g5 & g4

- these comparisons are linear contrasts
  - sometimes called “trend analysis”
  - often more appropriate & powerful test of experimental hypothesis that omnibus F test

- important difference between planned vs post-hoc contrasts
  - post-hoc contrasts use Scheffé's method to control Type I error rate