STAT 3011 Discussion 015

Week 13: ANOVA and Tukey's HSD

Talha Hamza
University of Minnesota
College of Science and Engineering

Spring 2025

One-Way ANOVA Framework

Purpose

Compare means across more than two groups simultaneously

Notation

- g = number of groups
- n_i = sample size for group i
- N = total sample size
- $\bar{y}_i$ = group i mean
- $\bar{y}$ = grand mean
- s_i = group i standard deviation

Key Components

- Between-group variation (MSG)
- Within-group variation (MSE)
- F-test compares these variations

ANOVA Assumptions

Three Key Assumptions

1. Independent random samples from g populations
2. Normality: Each population is normally distributed
3. Equal Variance: All populations have same σ

Note

- F-test is robust to minor violations of normality and equal variance
- Check with boxplots and histograms

ANOVA Hypothesis Test

Hypotheses

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$
- H_a : At least two means differ

Test Statistic

$$F = \frac{MSG}{MSE} \sim F_{g-1,N-g} \text{ under } H_0$$

- MSG = Between-group variability
- MSE = Within-group variability

Decision Rule

Reject H_0 if p-value $< \alpha$ (typically 0.05)

ANOVA Table

Source	df	SS	MS	F
Group	$g-1$	SSG	MSG = SSG/(g-1)	MSG/MSE
Error	$N-g$	SSE	MSE = SSE/(N-g)	
Total	$N-1$	SST		

Key Calculations

- SSG $= \sum n_i(\bar{y}_i - \bar{y})^2$
- SSE $= \sum(n_i - 1)s_i^2$
- SST = SSG + SSE

Tukey's Honest Significant Difference (HSD) Test

What is Tukey's HSD?

- A post-hoc test after ANOVA to identify which specific group means differ.
- Controls the family-wise error rate (FWER) for multiple comparisons.

Key Concepts

1. Family-Wise Confidence Level (e.g., 95%)
   - Ensures 95% confidence that all pairwise comparisons are correct *simultaneously*.
   - More conservative than individual CIs (avoids false positives).
2. Adjusted P-values (`p adj`)
   - Accounts for multiple testing. Significant if `p adj < 0.05`.
3. Confidence Intervals (`lwr`, `upr`)
   - If CI excludes 0, the difference is significant.

R Commands for ANOVA & Tukey HSD

One-Way ANOVA Implementation

```
model <- aov(y ~ x, data = dataset)
```

- `y`: Quantitative response variable (e.g., hold time)
- `x`: Grouping/factor variable (e.g., message type)

```
summary(model)
```
 # Displays ANOVA table

Tukey's HSD Test

```
TukeyHSD(model, "x", conf.level = 0.95)
```

- `model`: Saved aov model object
- `"x"`: Exact name of your grouping variable (e.g., "therapy")

Lecture Example: Anorexia Treatments

Study Design

- 72 anorexic girls randomly assigned to:
  - Cognitive behavioral therapy
  - Family therapy
  - Control (no treatment)
- Response: Weight change after treatment

ANOVA Results

Source	df	SS	MS	F
Therapy	2	614.6	307.3	5.42
Error	69	3910.7	56.7	

- p-value $= 0.0065$ (significant at $\alpha = 0.05$)

R Output: Anorexia Example

```
> summary(aov(change ~ therapy))
            Df Sum Sq Mean Sq F value Pr(>F)
therapy      2  614.6  307.32   5.422 0.0065 **
Residuals   69 3910.7   56.68
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> TukeyHSD(aov1, "therapy", conf.level = 0.95)
$therapy
                    diff        lwr       upr     p adj
control-cog    -3.456897 -8.327276  1.413483 0.2124428
family-cog      4.257809 -1.250554  9.766173 0.1607461
family-control  7.714706  2.090124 13.339288 0.0045127
```

Interpretation

Only family vs. control shows significant difference (CI excludes 0, $p < 0.05$)

Questions?