Lecture12: ANOVA (Analysis of Variance)

Carolyn Anderson

Department of Educational Psychology

University of Illinois at Urbana-Champaign
The rest of the semester will all be hypothesis testing!

- Test of one proportion/mean

- Comparing two proportions/means
  - Independent samples
  - Matched samples

- Test of association between two categorical variables
  - Chi-square test
  - Fisher’s Exact test

- Comparing the means of multiple groups
  - One-way ANOVA
Topic 1: How Can We Compare Several Means? ------ ANOVA
Key Points

1. Motivation
2. Analysis of Variance
3. Hypotheses and Assumptions for the ANOVA Test
4. Variability Between Groups and Within Groups
5. ANOVA $F$-Test Statistic
6. The Variance Estimates and the ANOVA Table
7. The 5-Step ANOVA Test
8. Why Not Use Multiple $t$-tests?
Why Not Use Multiple t-tests?

- If separate $t$ tests are used, the significance level applies to each *individual* comparison, not the *overall* type I error rate for all the comparisons.
- When there are several groups, using the $F$ test instead of multiple $t$ tests allows us to control the probability of a type I error.
- However, the $F$ test does not tell us which groups differ or how different they are.
How Bad is Doing Multiple t-tests?

For a significance level equal .05, below are the minimum and maximum possible type I error rates.

- Let \( g \) = number of groups, so there are would \( g(g-1)/2 \) possible t-tests.

<table>
<thead>
<tr>
<th>( g )</th>
<th>number tests</th>
<th>P(at least 1 type I error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>.14 ( \leq ) .15</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>.26 ( \leq ) .30</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>.40 ( \leq ) .50</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>.54 ( \leq ) .75</td>
</tr>
</tbody>
</table>
Solution: Analysis of Variance

- The two independent sample t-test compares means of two groups.
- The analysis of variance (ANOVA) compares means of several groups.
  - Let \( g \) denote the number of groups (populations).
  - Each group has a corresponding population of subjects.
  - The means of the response variable for the \( g \) populations are denoted by \( \mu_1, \mu_2, \ldots, \mu_g \).
Hypotheses for the ANOVA Test

- The analysis of variance is a significance test of the null hypothesis of equal population means:
  - $H_0: \mu_1 = \mu_2 = \cdots = \mu_g$
- The alternative hypothesis is:
  - $H_a: \text{At least two of the population means are unequal.}$

The alternative is **NOT** $H_a: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_g$
Assumptions for the ANOVA Test

1. The population distributions of the response variable for the $g$ groups are normal.
   - You can construct box plots or dot plots for the sample data distributions to check for extreme violations of normality.
   - Moderate violations are not serious, especially when the sample sizes are large (greater than 8 in every group).
   - Misleading results may occur with the F-test if the distributions are highly skewed and the sample size $N$ is small.
Assumptions for the ANOVA Test

2. The population distributions of the response variable for the \( g \) groups have the same standard deviation for each group.
   - Moderate violations are also not serious.
   - As long as the largest group variance is no more than twice the smallest group variance, especially when sample sizes are equal across groups.
   - Misleading results may also occur with the \( F \)-test if the largest sample standard deviation is more than double the smallest one.
Assumptions for the ANOVA Test

3. Randomization:
   • In a survey sample, independent random samples are selected from each of the $g$ populations
   • In an experiment, subjects are randomly assigned separately to the $g$ different groups
Example: How Long Will You Tolerate Being Put on Hold?

- An airline has a toll-free telephone number for reservations
- The airline hopes a caller remains on hold until the call is answered, so as not to lose a potential customer
- The airline recently conducted a randomized experiment to analyze whether callers would remain on hold longer, on the average, if they heard:
  - An advertisement about the airline and its current promotion
  - Muzak (“elevator music”)
  - Classical music
Example: How Long Will You Tolerate Being Put on Hold?

- The company randomly selected one out of every 1000 calls in a week.
- For each call, they randomly selected one of the three recorded messages.
- They measured the number of minutes that the caller stayed on hold before hanging up (these calls were purposely not answered).
Example: How Long Will You Tolerate Being Put on Hold?

- Denote the means of holding time for the populations that these three random samples represent by:
  - $\mu_1 =$ mean for the advertisement
  - $\mu_2 =$ mean for the Muzak
  - $\mu_3 =$ mean for the classical music

- The hypotheses for the ANOVA test are:
  - $H_0: \mu_1 = \mu_2 = \mu_3$
  - $H_a: \text{at least two of the population means are different}$
**Example: How Long Will You Tolerate Being Put on Hold?**

<table>
<thead>
<tr>
<th>Recorded Message</th>
<th>Holding Time Observations</th>
<th>Sample Size</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertisement</td>
<td>5, 1, 11, 2, 8</td>
<td>5</td>
<td>5.4</td>
<td>4.2</td>
</tr>
<tr>
<td>Muzak</td>
<td>0, 1, 4, 6, 3</td>
<td>5</td>
<td>2.8</td>
<td>2.4</td>
</tr>
<tr>
<td>Classical</td>
<td>13, 9, 8, 15, 7</td>
<td>5</td>
<td>10.4</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Example: How Long Will You Tolerate Being Put on Hold?

- Here is a display of the sample means:

![Graph showing sample means for Muzak, Advertisement, and Classical music]

- This alone is not sufficient evidence to enable us to reject $H_0$
Another look at the distributions
Illustration of Ho and Some Ha

The Null hypothesis

One Possible Ha

Another Possible Ha

Using Data Mean & Std
Variability Between Groups and Within Groups

- The ANOVA method is used to compare population means.
- It is called analysis of variance because it uses evidence about two types of variability:
  - Between-group variability
  - Within-group variability
Variability Between Groups and Within Groups

- Two examples of data sets with equal means but unequal within-group variability.
- Which case do you think gives stronger evidence against $H_0: \mu_1 = \mu_2 = \mu_3$?

**FIGURE 14.2:** Data from Table 14.1 (Case a) and Hypothetical Data (Case b) that Have the Same Means but Less Variability within Groups. **Question:** Do the data in case b give stronger, or weaker, evidence against $H_0: \mu_1 = \mu_2 = \mu_3$ than the data in case a? Why?
Variability Between Groups and Within Groups

- In both cases the variability *between* pairs of means is the same.
- In ‘Case b’ the variability *within* each sample is much smaller than in ‘Case a’.
- The fact that ‘Case b’ has less variability within each sample gives stronger evidence against $H_0$. 
Why It’s Called ANOVA

We can estimate the common variance, $\sigma^2$, two different ways….assuming sample sizes are equal...

If the null is true, the between group variance estimator gives a good estimate of the variance:

$$\hat{\sigma}^2 = n\{(\bar{y}_1 - \bar{y})^2 + (\bar{y}_2 - \bar{y})^2 + \cdots + (\bar{y}_g - \bar{y})^2 \}/(g-1)$$

The within group variance estimator is always a good estimator, $\hat{\sigma}^2$

$$= (\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \cdots + \hat{\sigma}_g^2 ) / g$$

$$= ( s_1^2 + s_2^2 + \cdots + s_g^2 )/g$$
ANOVA F-Test Statistic

• The analysis of variance (ANOVA) F-test statistic is:

\[ F = \frac{\text{Between groups variability}}{\text{Within groups variability}} \]

• The larger the variability *between* groups relative to the variability *within* groups, the larger the F test statistic tends to be, and the stronger the evidence is against the null hypothesis of equal means

• P-value of the test statistic for comparing means is obtained from the *F*-distribution
The 5-Step ANOVA Test

1. **Assumptions:**
   - Independent random samples
   - Normal population distributions in each group
   - Equal standard deviations in each group

2. **Hypotheses:**
   - $H_0: \mu_1 = \mu_2 = \ldots = \mu_g$
   - $H_a$: at least two of the population means are different
The 5-Step ANOVA Test

3. Test statistic:

\[ F = \frac{\text{Between groups variability}}{\text{Within groups variability}} \]

- \( F \)-distribution has two degree of freedoms
  - \( df_1 = g - 1 = (\text{no. of groups} - 1) \)
  - \( df_2 = N - g = (\text{total sample size} - \text{no. of groups}) \)
The 5-Step ANOVA Test

4. **P-value:**
   Right-tail probability above the observed \( F \)-value.

5. **Conclusion:**
   - P-value \( \leq \) significance level (such as 0.05) \( \rightarrow \) Reject \( H_0 \) and conclude that the group means are not all equal
   - P-value > significance level (such as 0.05) \( \rightarrow \) Fail to reject \( H_0 \) and conclude that there is no sufficient evidence to show that group means are not all equal
The Variance Estimates and the ANOVA Table

When the sample sizes in each group are equal,

Within-groups variance estimate  \[ = \frac{s_1^2 + s_2^2 + \cdots + s_g^2}{g} \]

Between-groups variance estimate  \[ = \frac{n[(\bar{y}_1 - \bar{y})^2 + (\bar{y}_2 - \bar{y})^2 + \cdots + (\bar{y}_g - \bar{y})^2]}{g - 1} \]

where \( g \) denotes groups

\( s_g^2 \) denotes the variance of the response variable \( y \) in group \( g \) (that is the square of the standard deviation of \( y \) in group \( g \))

\( \bar{y}_g \) denotes the sample mean of \( y \) in group \( g \)

\( \bar{y} \) denotes the grand mean of \( y \) in all groups
Example:
ANOVA for Customers’ Telephone Holding Times

1: Assumptions

- Population distributions are normal?
  - For some group, the distribution does not look normal because the standard deviation is large compared to the mean
  - And the sample size is not large enough (< 8)
- These distributions have the same standard deviation
  - The largest standard deviation is lower than twice the smallest standard deviation (4.2 < 2.4*2 = 4.8)
- The data are independent random samples

<table>
<thead>
<tr>
<th>Recorded Message</th>
<th>Holding Time Observations</th>
<th>Sample Size</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertisement</td>
<td>5, 1, 11, 2, 8</td>
<td>5</td>
<td>5.4</td>
<td>4.2</td>
</tr>
<tr>
<td>Muzak</td>
<td>0, 1, 4, 6, 3</td>
<td>5</td>
<td>2.8</td>
<td>2.4</td>
</tr>
<tr>
<td>Classical</td>
<td>13, 9, 8, 15, 7</td>
<td>5</td>
<td>10.4</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Example: ANOVA for Customers’ Telephone Holding Times

2. Hypotheses
   - $H_0: \mu_1 = \mu_2 = \mu_3$
   - $H_a$: at least two of the population means are different

3. Test statistic & 4. P-value

![ANOVA Table for F Test Using Data from Table 14.1](image)
Example: ANOVA for Customers’ Telephone Holding Times

5. Conclusions

- Since P-value < 0.05, there is sufficient evidence to reject $H_0: \mu_1 = \mu_2 = \mu_3$

- We conclude that a difference exists among the three types of messages in the population mean amount of time that customers are willing to remain on hold
How to do ANOVA in R?

- `y1 <- c(5,1,11,2,8)`
- `y2 <- c(0,1,4,6,3)`
- `y3 <- c(13,9,8,15,7)`
- `y <- c(y1,y2,y3)`
- `n <- rep(5,3)`
- `group <- rep(1:3,n)`
- `boxplot(y~group)`
- `data <- data.frame(y = y, group = factor(group))`
- `result <- lm(y~group,data)`
- `anova(result)`
Key Points Revisited

1. Motivation
2. Analysis of Variance
3. Hypotheses and Assumptions for the ANOVA Test
4. Variability Between Groups and Within Groups
5. ANOVA F-Test Statistic
6. The Variance Estimates and the ANOVA Table
7. The 5-Step ANOVA Test
8. Why Not Use Multiple $t$-tests?
9. How to conduct ANOVA in R?
Topic 2: How Should We Follow Up an ANOVA F Test?
Follow Up to an ANOVA $F$-Test

- When an analysis of variance $F$-test has a small $P$-value, the test does not specify which means are different or how different they are.
- We can estimate differences between population means with confidence intervals.

And in R:

```r
pairwise.t.test(y, group, p.adj = "none")
```
Controlling Overall Confidence with Many Confidence Intervals

- How to do tests for pairs of means (or sets of means) so that the *entire set of intervals* rather than to each single interval has confidence level of 95%?
- Methods that control the probability that *all* confidence intervals will contain the true differences in means are called *multiple comparison methods*
  - Pooled SD Method: `pairwise.t.test(y, group)`
  - Bonferroni Method where $\text{Bonferroni} = \frac{a}{\# \text{comparisons}}$
    `pairwise.t.test(y, group, p.adj = "bonf")`
  - Tukey honest significant difference:
    - `TukeyHSD(aov(y ~ group, data))`
Exercise

Example: A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the following results:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Program 1</th>
<th>Program 2</th>
<th>Program 3</th>
<th>Program 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>6</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>9</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>8</td>
</tr>
</tbody>
</table>

\[ T_{A_j} = \sum y_{ij} \]

\[ \hat{\mu}_j = \frac{T_{A_j}}{N_j} \]

\[ \begin{array}{c|c|c|c|c}
11.8 & 8.8 & 12.2 & 8.6 \\
\end{array} \]

Estimate the treatment effects for the four programs.