ANOVA

Tom Bruning

2018-02-10

In this chapter, you learn:

- The basic concepts of experimental design
- How to use one-way analysis of variance to test for differences among the means of several groups
- When and how to use a randomized block design
- How to use two-way analysis of variance and interpret the interaction effect
- How to perform multiple comparisons in a one-way analysis of variance, a randomized block design, and a two-way analysis of variance

General ANOVA Setting

- Investigator controls one or more factors of interest
- Each factor contains two or more levels
- Levels can be numerical or categorical
- Different levels produce different groups
- Think of each group as a sample from a different population
- Observe effects on the dependent variable
- Are the groups the same?
- Experimental design: the plan used to collect the data

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to groups
  - Subjects are assumed homogeneous
- Only one factor or independent variable
  - With two or more levels
- Analyzed by one-factor analysis of variance (ANOVA)

One-Way Analysis of Variance

Evaluate the difference among the means of three or more groups
Examples:

- Number of accidents for 1st, 2nd, and 3rd shift
- Expected mileage for five brands of tires

Assumptions:

- Populations are normally distributed
- Populations have equal variances
- Samples are randomly and independently drawn

Hypotheses of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_c$
  - All population means are equal
  - i.e., no factor effect (no variation in means among groups)
- H_1 : Not all of the means are equal
  - At least one population mean is different
  - i.e., there is a factor effect
  - Does not mean that all population means are different (some pairs may be the same)
- When The Null Hypothesis is True All Means are the same
- When The Null Hypothesis is **NOT** true At least one of the means is different

Partitioning the Variation

Total variation can be split into two parts:

- SST = SSA + SSW
  Where:
  - SST = Total Sum of Squares
    (Total variation)
  - SSA = Sum of Squares Among Groups
    (Among-group variation)
  - SSW = Sum of Squares Within Groups
    (Within-group variation)
- *Total Variation* = the aggregate variation of the individual data values across the various factor levels (SST)
- *Among-Group Variation* = variation among the factor sample means (SSA)
- *Within-Group Variation* = variation that exists among the data values within a particular factor level (SSW)

  Partition of Total Variation

- Total Variation (SST) = Variation Due to Factor (SSA) + Variation Due to Random Error (SSW)

Total Sum of Squares

$$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \bar{\bar{X}})^2 \tag{1}$$

Where:

- SST = Total sum of squares
- c = number of groups or levels
- n_j = number of values in group j
- X_{ij} = ith observation from group j
- $\bar{\bar{X}}$ = grand mean (mean of all data values)

In lay terms: we are summing the squared difference between each individual value and the mean of all the values.

Among-Group Variation

SST = SSA + SSW

$$SSA = \sum_{j=1}^{c} (\bar{x}_j - \bar{\bar{X}})^2 \tag{2}$$

Where:

- SSA = Sum of squares among groups
- c = number of groups or levels
- n_j = number of values in group j
- X_j = sample mean from group j
- $\bar{\bar{X}}$ = grand mean (mean of all data values)

In lay terms we are summing the squared difference between the group means and the grand mean. This measures the variation between the groups.

- $MSA = \frac{SSA}{c-1}$ - the average difference between the group means and the grand mean.

Within-Group Variation

SST = SSA + SSW

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \tag{3}$$

Where:

- SSW = Sum of squares within groups
- c = number of groups

- n_j = sample size from group j
- X_j = sample mean from group j
- X_{ij} = ith observation in group j

In lay terms we are summing the variation within each group and then adding over all groups.

-Mean Square Within = $MSA = \frac{SSW}{n-c}$

Summary: Obtaining the Mean Squares

The Mean Squares are obtained by dividing the various sum of squares by their associated degrees of freedom

- Mean Square Among = $MSA = \frac{SSA}{c-1}$ (d.f. = c-1)
- Mean Square Within = $MSW = \frac{SSW}{n-c}$ (d.f. = n-c)
- Mean Square Total= $MST = \frac{SST}{n-1}$ (d.f. = n-1)

One-Way ANOVA

F Test Statistic

- $H_0 : \mu_1 = \mu_2 = ... = \mu_c$
- H_1 : At least two population means are different
- Test statistic:

  -$F_{stat} = \frac{MSA}{MSW}$

  MSA is mean squares *among* groups
  MSW is mean squares *within* groups
  Degrees of freedom

- $df_1 = c$–1 (c = number of groups)
- $df_2 = n$–c (n = sum of sample sizes from all populations)

Interpreting One-Way ANOVA

F Statistic:

- The F statistic is the ratio of the *among* estimate of variance and the *within* estimate of variance
- The ratio must always be positive
- $df_1 = c - 1$ will typically be small
- $df_2 = n - c$ will typically be large

Decision Rule:

- Reject H_0 if $F_{STAT} > F_\alpha$, otherwise do not reject H_0

One-Way ANOVA

F Test Example:

You want to see if three different golf clubs yield different distances. You randomly select five measurements from trials on an automated driving machine for each club. At the 0.05 significance level, is there a difference in mean distance?

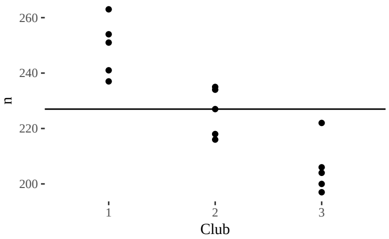
Table 1: Golf Ball Distances

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

One-Way ANOVA Example Computations

- Mean
  $X_1 = 249.2$
  $X_2 = 226.0$
  $X_3 = 205.8$
  $\bar{\bar{X}} = 227.0$

- Sample Size
  $n_1 = 5$
  $n_2 = 5$
  $n_3 = 5$
  n = 15
  c = 3



- $SSA = 5(249.2–227)^2 + 5(226–227)^2 + 5(205.8–227)^2 = 4716.4$
- $SSW = (254–249.2)^2 + (263–249.2)^2 + ... + (204–205.8)^2 = 1119.6$
- $MSA = \frac{4716.4}{(3-1)} = 2358.2$
- $MSW = \frac{1119.6}{(15-3)} = 93.3$
- $F_{STAT} = \frac{2358.2}{93.3} = 25.275$

One-Way ANOVA Example Solution

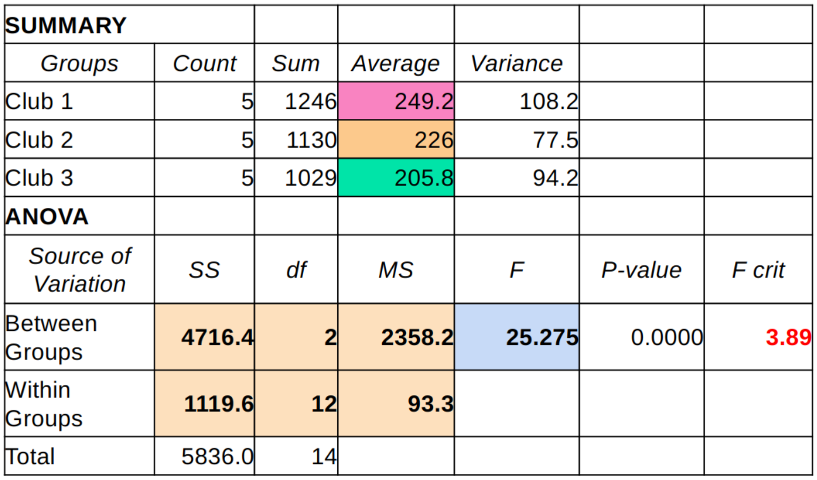
Values:

$H_0 : \mu_1 = \mu_2 = \mu 3$
$H_1 : \mu_j$ not all equal
$\alpha = 0.05$

$df_1 = 2$ $df_2 = 12$
Critical Value $F_\alpha = 3.89$
Test Statistic: $F_{STAT} = 25.275$

- Decision: Reject H_0 at $\alpha = .05$
- Conclusion: There is evidence that at least one μ_j differs from the rest

Excel Output



SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Club 1	5	1246	249.2	108.2		
Club 2	5	1130	226	77.5		
Club 3	5	1029	205.8	94.2		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**4716.4**	**2**	**2358.2**	**25.275**	0.0000	**3.89**
Within Groups	**1119.6**	**12**	**93.3**			
Total	5836.0	14				

Figure 1: Excel Output

After Determining that the means are not equal, now what?

The Tukey-Kramer Procedure

- Tells which population means are significantly different
  - e.g.: $\mu_1 = \mu_2 \neq \mu_3$
  - Done after rejection of equal means in ANOVA
- Allows paired comparisons
  - Compare absolute mean differences with critical range
- Critical Range $= Q_\alpha\sqrt{\frac{MSW}{2}(\frac{1}{n_j} + \frac{1}{nj'})}$
  where:
  - Q_α = Upper Tail Critical Value from Studentized Range Distribution with c and n - c degrees of freedom (see appendix E.7 table)

  - MSW = Mean Square Within
  - n_j and $n_{j'}$ = Sample sizes from groups j and j'

Table 2: Golf Ball Distances

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

Means:
- $x_1 = 249.2$
- $x_2 = 226.0$
- $X_3 = 205.8$

1. Compute absolute mean differences (AMD):
   $|x_1 - x_2| = |249.2 - 226.0| = 23.2$
   $|x_1 - x_3| = |249.2 - 205.8| = 43.4$
   $|x_2 - x_3| = |226.0 - 205.8| = 20.2$

2. Find the $Q\alpha$ value from the table in appendix E.7 with c = 3 and (n – c) = (15 – 3) = 12 degrees of freedom:
   $Q\alpha = 3.77$

3. Compute the Critical Range:
   $Q_\alpha\sqrt{\frac{MSW}{2}(\frac{1}{n_j} + \frac{1}{n_{j'})}} = 3.77\sqrt{\frac{93.2}{2}(\frac{1}{5} + \frac{1}{5})} = 16.285$

4. Compare the critical range to all three AMDs: 16.285: to 23.2, 43.4, 20.2

5. All of the absolute mean differences are greater than the critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

Thus, with 95% confidence we conclude that the mean distance for club 1 is greater than club 2 and 3, and club 2 is greater than club 3.

ANOVA Assumptions

- Randomness and Independence
  - Select random samples from the c groups (or randomly assign the levels)
- Normality

  - The sample values for each group are from a normal population
- Homogeneity of Variance
  - All populations sampled from have the same variance
  - Can be tested with Levene's Test