Analysis of Variance (ANOVA)

Your Name

March 15, 2023

Table of Contents

Introduction

- Analysis of variance (ANOVA) is a statistical technique used to determine if there are significant differences between the means of two or more groups.
- ANOVA decomposes the total variability in the data into variability between groups and variability within groups.
- The F statistic is used to test if the group means are equal.

Equations

- The total sum of squares (SST) is calculated as:
  $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$
- The between-group sum of squares (SSB) is calculated as:
  $SSB = \sum_{j=1}^{k} n_j(\overline{y_j} - \overline{y})^2$
- The within-group sum of squares (SSW) is calculated as:
  $SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j}(y_{ij} - \overline{y_j})^2$

Sum of Squares Decomposition

- The total sum of squares (SST) can be decomposed into the between-group sum of squares (SSB) and the within-group sum of squares (SSW).
- The equation for this decomposition is: $SST = SSB + SSW$
- The F statistic is calculated as the ratio of the mean square between groups (MSB) to the mean square within groups (MSW).
- The equation for the F statistic is: $F = \frac{MSB}{MSW}$, where $MSB = \frac{SSB}{k-1}$ and $MSW = \frac{SSW}{N-k}$

Relation with Linear Model

- ANOVA can be seen as a special case of the linear model.
- The linear model for ANOVA is: $y_{ij} = \mu + \alpha_j + \epsilon_{ij}$, where μ is the overall mean, α_j is the effect of group j, and ϵ_{ij} is the error term.
- The null hypothesis for ANOVA is that all group means are equal, which can be written as: $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$
- The alternative hypothesis is that at least one group mean is different from the others.

Numerical Example

- Suppose we have three groups with the following data:
  - Group 1: 6, 8, 7
  - Group 2: 9, 10, 11
  - Group 3: 12, 13, 14
- The overall mean is $\overline{y} = 10$
- The group means are $\overline{y_1} = 7$, $\overline{y_2} = 10$, and $\overline{y_3} = 13$
- The SST is calculated as:
  $SST = (6-10)^2 + (8-10)^2 + ... + (14-10)^2 = 60$
- The SSB is calculated as:
  $SSB = 3(7-10)^2 + 3(10-10)^2 + 3(13-10)^2 = 54$
- The SSW is calculated as:
  $SSW = (6-7)^2 + (8-7)^2 + ... + (14-13)^2 = 6$

Numerical Example (cont.)

- The F statistic is calculated as: $F = \frac{MSB}{MSW} = \frac{\frac{SSB}{k-1}}{\frac{SSW}{N-k}} = \frac{\frac{54}{3-1}}{\frac{6}{9-3}} = 9$
- The p-value is calculated by comparing the F statistic to an F distribution with $(k-1)$ and $(N-k)$ degrees of freedom.
- In this case, the p-value is very small, indicating that we can reject the null hypothesis and conclude that there are significant differences between the group means.

Interpretation of Results

- If the p-value is smaller than the significance level (usually 0.05), we reject the null hypothesis and conclude that there are significant differences between the group means.
- If the p-value is larger than the significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that there are significant differences between the group means.
- It is important to note that ANOVA only tells us if there are significant differences between the group means, but it does not tell us which specific means are different from each other.
- To determine which specific means are different from each other, we can use post-hoc tests such as Tukey's HSD test or Bonferroni's correction.