

Analysis of Variance

Analysis of Variance (ANOVA)

ANOVA is a statistical method used to compare the means of three or more groups to determine if at least one group mean is significantly different from the others. It helps in understanding whether the observed differences among sample means are due to actual differences in the population means or just random variation.

One-Way ANOVA

One-Way ANOVA is used when there is one independent variable with multiple levels (groups).

Hypothesis

(H_0): All group means are equal ($\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$).

(H_1): At least one group mean is different.

```
set.seed(123)
group <- factor(rep(1:3, each = 10))
values <- c(rnorm(10, mean = 5), rnorm(10, mean = 6),
            rnorm(10, mean = 7))
```

```
# Perform One-Way ANOVA
anova_result <- aov(values ~ group)
summary(anova_result)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## group         2   12.24    6.122    6.435 0.00518 **
## Residuals    27   25.68    0.951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way ANOVA

Two-Way ANOVA is used when there are two independent variables.

Hypothesis:

(H₀): There is no interaction effect between the two factors, and the means are equal across the levels of each factor.

(H₁): There is an interaction effect, or the means are not equal across the levels of each factor.

```
# Example dataset
data <- data.frame(
  Factor1 = rep(c("A", "B"), each = 6),
  Factor2 = rep(c("X", "Y", "Z"), times = 4),
  Response = c(4.2, 4.8, 3.6, 5.0, 4.9, 3.7, 5.1, 5.3, 4.2, 5.6, 5.4, 4.8)
)

# Perform two-way ANOVA
anova_result <- aov(Response ~ Factor1 * Factor2, data = data)

# Display results
summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Factor1         1  1.470   1.4700   13.781 0.00994 **
## Factor2         2  2.502   1.2508   11.727 0.00845 **
## Factor1:Factor2  2  0.065   0.0325    0.305 0.74812
## Residuals       6  0.640   0.1067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Table

An ANOVA table summarizes results of the ANOVA test. It includes the following features:

1. **Source of Variation:** Identifies the different sources of variation (e.g. between groups, within groups).
2. **Sum of Squares (SS):** Measures the total variation.
3. **SS Between (SSB):** Variation due to the differences between group means.
4. **SS Within (SSW):** Variation within each group.
5. **Degrees of Freedom (df):** Number of independent values that can vary. df Between: (k - 1) (where (k) is the number of groups). df Within: (N - k) (where (N) is the total number of observations).
6. **Mean Square (MS):** Average variation (SS divided by df).
7. **MS Between:** (SSB / df_between).
8. **MS Within:** (SSW / df_within).
9. **F-Statistic:** Ratio of MS Between to MS Within. $F = (MS_{\text{between}}) / (MS_{\text{within}})$.
10. **P-Value:** Probability of observing the data that null hypothesis is true.

Making Decisions After ANOVA

Check the P-Value: Compare the p-value to the significance level (alpha), typically 0.05. If (p < alpha), reject the null hypothesis (indicating significant differences between group means). If (p > alpha), fail to reject the null hypothesis (indicating no significant differences).

Post-Hoc Tests

If the null hypothesis is rejected, perform post-hoc tests (e.g. Tukey's HSD) to determine which specific groups are different.

```
# Perform Tukey's HSD test
tukey_result <- TukeyHSD(anova_result)
print(tukey_result)

##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = Response ~ Factor1 * Factor2, data = data)
##
## $Factor1
##      diff      lwr      upr      p adj
## B-A  0.7 0.2386059 1.161394 0.0099424
##
## $Factor2
##      diff      lwr      upr      p adj
## Y-X  0.125 -0.5835876 0.8335876 0.8545292
## Z-X -0.900 -1.6085876 -0.1914124 0.0187826
## Z-Y -1.025 -1.7335876 -0.3164124 0.0104074
##
## $'Factor1:Factor2'
##      diff      lwr      upr      p adj
## B:X-A:X  7.500000e-01 -0.5498125 2.04981251 0.3194679
## A:Y-A:X  2.500000e-01 -1.0498125 1.54981251 0.9641786
## B:Y-A:X  7.500000e-01 -0.5498125 2.04981251 0.3194679
## A:Z-A:X -9.500000e-01 -2.2498125 0.34981251 0.1623455
## B:Z-A:X -1.000000e-01 -1.3998125 1.19981251 0.9994241
## A:Y-B:X -5.000000e-01 -1.7998125 0.79981251 0.6614328
## B:Y-B:X  1.776357e-15 -1.2998125 1.29981251 1.0000000
## A:Z-B:X -1.700000e+00 -2.9998125 -0.40018749 0.0147887
## B:Z-B:X -8.500000e-01 -2.1498125 0.44981251 0.2286203
## B:Y-A:Y  5.000000e-01 -0.7998125 1.79981251 0.6614328
## A:Z-A:Y -1.200000e+00 -2.4998125 0.09981251 0.0694030
## B:Z-A:Y -3.500000e-01 -1.6498125 0.94981251 0.8765888
## A:Z-B:Y -1.700000e+00 -2.9998125 -0.40018749 0.0147887
## B:Z-B:Y -8.500000e-01 -2.1498125 0.44981251 0.2286203
## B:Z-A:Z  8.500000e-01 -0.4498125 2.14981251 0.2286203
```

Kruskal-Wallis test

The Kruskal-Wallis test is a non-parametric alternative to one-way ANOVA. It is used when the assumptions of ANOVA (e.g. normality, homogeneity of variances) are not met.

Hypothesis:

(H_0): The distributions of the groups are the same.

(H₁): At least one group distribution is different.

```
set.seed(123)
group <- factor(rep(1:3, each = 10))
values <- c(rnorm(10, mean = 5), rnorm(10, mean = 6),
            rnorm(10, mean = 7))

# Perform Kruskal-Wallis test
kruskal_result <- kruskal.test(values ~ group)
print(kruskal_result)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  values by group
## Kruskal-Wallis chi-squared = 8.1858, df = 2, p-value = 0.01669
```