Statistical Inference

Analysis of Variance

Contents

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among more than two group means developed by statistician and evolutionary biologist **Ronald Fisher**.

Example: There are 20 plots of wheat and 5 fertilizers applied to four different plots. The yield of wheat is recorded for each of the 20 plots.

ANOVA can be used to find out whether the effect of these fertilisers on yields is equal or significantly different.

ANOVA

- Note that although the name is 'Analysis of Variance', the method is used to analyze the differences among group means.
- Variation in the variable is inherent in nature. In general, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.
- The total variance in any variable is due to a number of causes which may be classified “assignable causes (which can be detected and measured)” and “chance causes (which are beyond human control and cannot be traced separately)”.
- Hence, ANOVA is the separation of variance ascribable to one group of causes from the variance ascribable to another.

ANOVA assumptions

The assumptions of ANOVA are listed below:

- The samples drawn are random samples.

-The populations from which samples are drawn have equal & unknown variances.

-The populations follow a normal distribution.

Testing Normality assumption

- An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.

  - **Graphical approach**

    - Box-Whisker plot (used to assess symmetry rather than normality)
    - Quantile-Quantile plot (Q-Q plot).

  - **Statistical approach**

    - Shapiro-Wilk test
    - Kolmogorov-Smirnov test

Normality tests are covered in the Parametric test session.

One Way ANOVA

- One Way ANOVA can be considered as an extension of the t test for independent samples.
- One Way ANOVA is used to test the equality of K population means.
  (when K=2, t test can be used.)
- For two levels (K=2),the t test and One Way ANOVA provide identical results.
- The **Mathematical model** is :

  $X_{ij} = \mu_i + \varepsilon_{ij}$

  Where X_{ij} is the jth observation due to ith level of a factor. μ_i is the effect of ith level of a factor. ε_{ij} is the error term. i=1,2,..,k ; j=1,2,...,n_i
- The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = ... = \mu_K = \mu$$

Partitioning Total Variance

- Total variation is partitioned into two parts:
  Total SS= Between Groups SS + Within Groups SS
  where, SS stands for sum of squares

$$SS_{total} = \sum_i \sum_j \left(X_{ij} - \bar{X}_{..} \right)^2$$

$$SS_{between} = \sum_i n_i \left(\bar{X}_{i.} - \bar{X}_{..} \right)^2$$

$$SS_{error} = \sum_i \sum_j \left(X_{ij} - \bar{X}_{i.} \right)^2$$

Total variation (Total SS)

Variation due to Assignable causes (Between SS)

Variation due to Chance causes (Within SS)

- Total SS is calculated using squared deviations of each value from overall mean.
- Between SS is calculated using squared deviation of each group mean from overall mean.
- Within Group SS can be obtained by subtracting Between SS from Total SS

Case Study - 2

Background

A large company is assessing the difference in the 'Satisfaction Index' of employees in it's Finance, Marketing and Client-Servicing departments.

Objective

To test whether the **mean satisfaction indices** for employees in three departments (CS, Marketing, Finance) are equal.

Sample Size

Sample size: 37
Variables: satindex, dept

Data Snapshot

One way anova

Variables

Observations

satindex	dept
75	FINANCE
56	FINANCE
72	FINANCE
59	FINANCE
66	FINANCE
58	FINANCE
58	MARKETING
63	MARKETING
51	MARKETING
64	MARKETING
55	MARKETING
72	CS
69	CS

Columns	Description	Type	Measurement	Possible values
satindex	Satisfaction Index	Numeric		Positive Values
dept	Department	Character	MARKETING, CS, FINANCE	3

One Way ANOVA

Testing equality of means in one factor with more than two levels.

Objective	**To test whether the mean satisfaction indices for employees in three departments (CS, Marketing, Finance) are equal.**

Null Hypothesis (H_0): Mean satisfaction index for 3 departments are equal i.e. $\mu1 = \mu2 = \mu3$

Alternate Hypothesis (H_1): Mean satisfaction index for 3 departments are not equal

Test Statistic	**The test statistic is denoted as F and is based on F distribution.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Calculation

Overall Mean	65.59	n=37
Mean for Finance	64.42	n1=12
Mean for Marketing	63.25	n2=12
Mean for CS	68.85	n3=13

Total SS = (75-65.59)^2+(56-65.59)^2+……+(65-65.59)^2+(76-65.59)^2
= 1840.92

Between Groups SS = 12*(64.42-65.59)^2+12*(63.25-65.59)^2+13*(68.85-65.59)^2
= 220.0599

Within Groups SS = Total SS – Between SS **= 1620.86**

One Way ANOVA table

Sources of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MS=SS/df)	F-Value
Between groups	K-1=3-1 =2	SSA=220.0599	MSA=110.03	F=2.3080
Within groups (error)	n-k=37-3 =34	SSE=1620.86	MSE=47.6724	
TOTAL	n-1=37-1 =36	TSS=1840.92		

One Way ANOVA in R

```
# Import data
data<-read.csv("One way anova.csv",header=TRUE)

# ANOVA table
anovatable<-aov(formula=satindex~dept, data=data)
summary(anovatable)
```

- *'aov' is the R function for ANOVA .*
- *formula specifies 'satindex' as analysis (dependent) variable and 'dept' as factor (independent) variable.*
- *anovatable is user defined object name created to store output.*
- *summary function displays the ANOVA table output.*

```
# Output:
            Df Sum Sq Mean Sq F value Pr(>F)
dept         2  220.1  110.03   2.308  0.115
Residuals   34 1620.9   47.67
```

Interpretation :

- *Since p-value is >0.05, do not reject H0. There is no significant difference in satisfaction index among 3 different departments.*

Quick Recap

ANOVA	• Analysis of variance (ANOVA) is a collection of statistical models used to analyze the difference among more than two group means developed by statistician and evolutionary biologist Ronald Fisher.
Partitioning the variance	• The total variance in any variable is due to a number of causes which may be classified as "assignable causes (which can be detected and measured)" and "chance causes (which is beyond human control and cannot be traced separately)".
One Way ANOVA	• Comparing several means of different levels of one factor. $H_0 : \mu_1 = \mu_2 = ... = \mu_K = \mu$