BUSINESS ANALYTICS

Analysis Of Variance (ANOVA)

- Compares - multiple independent variable with a continuous variable
  - Dependent Vs Independent Variables
  - Continuous Vs Category (2 or more groups)
  - Continuous Vs Category + Continuous+….(many Independent variable)

- Extension of the independent t-tests

Example - One Way ANOVA

- Marks obtained in the same subject by 3 students belonging to three different schools are given below
- Does the data suggest any association between schools and marks?

SCHOOL	A	B	C
Marks	82	83	38
	83	78	59
	97	68	55

Steps

1. Calculate the means
   - School A : mean(82,83,97) = 87.3
   - School B : mean(83,78,68) = 76.3
   - School C : mean(38,59,55) = 50.6

2. Calculate the grand mean
   - Grand mean $\bar{X}$= mean(82,83,97,83,78,68,39,59,55)/9 = 71.4

- Calculating the variations
  - Sum of Squared Deviations about the grand mean, across all observed values

    ie. $(X-\bar{X})^2 = (82-71.44)^2+(83-71.44)^2+\ldots$

    $SS_{Total} = 2630.2$
  - Sum of Squared Deviations of group mean about the grand mean – three group means against the grand mean

    ie. $n(X-\bar{X})^2 = 3\{(87.3-71.4)^2+(76.3-71.4)^2+(50.6-71.4)^2\}$

    $SS_{Between} = 2124.2$
  - Sum of Squared Deviations of observations within a group about their group mean, added across all groups

    ie. $SS_{Within} = 506$

- Calculate the degrees of freedom for every variance
  - df_{Total} = Number of observations – 1 = 9 -1 = 8
  - $df_{Between}$ = Number of groups -1 = 3-1 = 2
  - df_{Within} = Number of observations – number of groups = 9-3 = 6

- Calculate the Mean Squared Variances
  - Mean Squared variance between groups

    $MS_{Between} = SS_{Between} / df_{Between} = 2124.2/2 = 1062.1$
  - Mean Squared variance within groups

    $MS_{Within} = SS_{Within} / df_{Within} = 506/6 = 84.3$

- Calculate the f-statistic
  - F-value: $MS_{Between}/MS_{Within} = 1062.1/84.3 = 12.59$

- Calculate the p-value from the F-table
  - p-value for given f-value 12.59 and degrees of freedom 2 and 6 is 0.007

Type of ANOVA

One way ANOVA

- Compare the means of two or more independent (unrelated) groups
- E.g. Is there a difference in student’s scores based on the row he is seated – front/middle/back?

Two way ANOVA

- Examines the influence of two category independent variables on one continuous dependent variable
- E.g. Does the race and gender affect a person’s yearly income?

Note: Can use ANOVA, when there is more than 2 groups in Category variable and sample size is more than 30

Chi-Square Test (X^2)

- Determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more category variables
- Dependent Vs Independent
- Category (multiple groups) Vs Category (multiple groups)
- Hypothesis:

- Null hypothesis states Variable A and Variable B are independent
- Alternate hypothesis states Variable A and Variable B are not independent

Note: If both the variables are category variable, can use chi-square test

Example on Chi-Square Test

- Ice-cream flavours survey taken based on gender

	Choco	Vanilla	Strawberry	Total
Men	100	120	60	280
Women	350	200	90	640
Total	450	320	150	920

- Proportion of population Men = 280/920 = 0.3043
- Proportion of population Women = 640/920 = 0.6957

Thus, expected values:

Population with Choco = 450

Choco Men: 450 * 0.3043 = 136.935

Choco Women: 450 * 0.6957 = 313.065

Population with Vanilla = 320

Vanilla Men: 320 * 0.3043 = 97.376

Vanilla Women: 320 * 0.6957 = 222.624

Population with Strawberry = 150

Strawberry Men: 150 * 0.3043 = 45.645

Strawberry Women: 150 * 0.6957 = 104.355

Calculate the Chi-squared statistic

- $X^2 = \sum \frac{(observerd\ frequency\ - expected\ frequency)^2}{expected\ frequency}$

$$= \frac{(100-136.935)^2}{136.935} + \frac{(350-313.065)^2}{313.065} + \frac{(120-97.376)^2}{97.376} + \frac{(200-222.624)^2}{222.624} + \frac{(60-45.645)^2}{45.645} + \frac{(90-104.355)^2}{104.355}$$

- $X^2 = 28.362$

Note: From Chi-square table corresponding p-value will be calculated

- ANOVA and Chi-Square analysis in R & SAS
- CASE STUDY – Internet Survey in R & SAS