5691 Seminar Report

Analysis of Variance

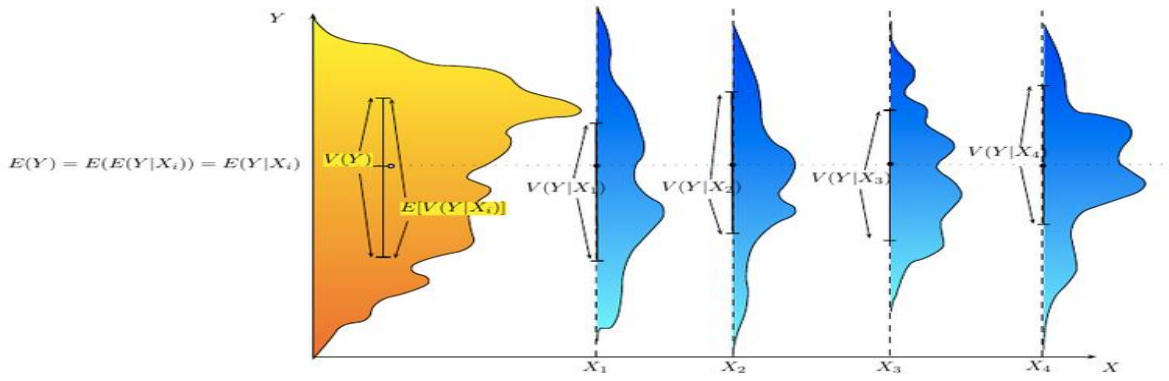

Figure 2: ANOVA : No fit

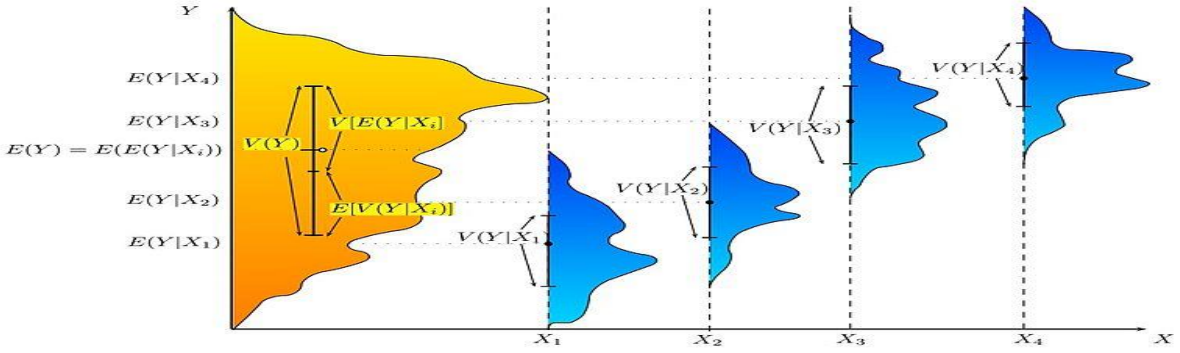

Figure 1: ANOVA : Fair fit

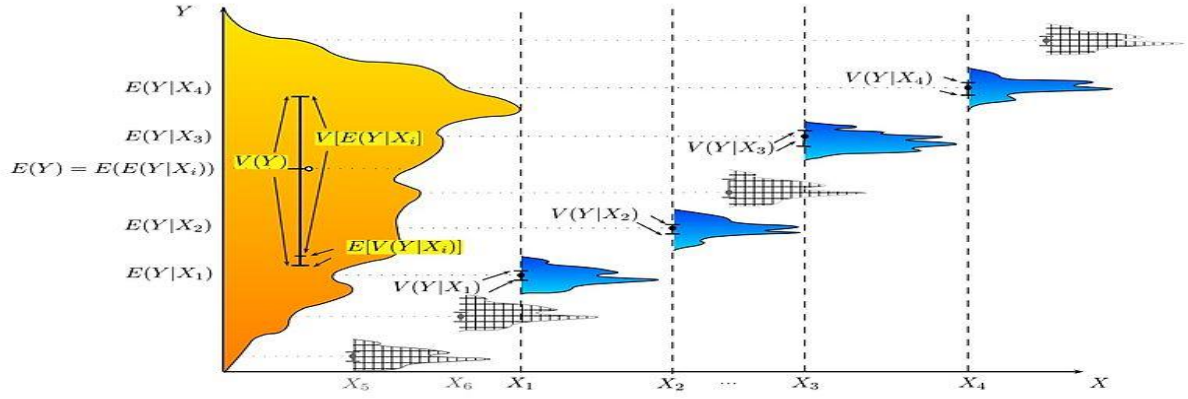

Figure 3: ANOVA : very good fit

Contents:

1. Introduction

ANOVA (a.k.a. ANalysis Of VAriance) is a statistical procedure that computes the significance of an experiment where the experiment consists of various groups and their comparison. This test is also called the Fisher analysis of variance.

In Anova, the inference is made by comparing means of various groups by analyzing their variances. The null hypotheses in case of n groups is that mean of all the groups are same, as follows:

$$\mu_1 = \mu_2 = \mu_3 = \dots \dots \dots \dots = \mu_n \qquad \{\text{Omnibus Null Hypothesis}\}$$

In a regression study, the variance test is useful to analyze the impact of independent variables on dependent variables. The shortlisted variables are the major factors of a dataset. The results can be put in an f-test to align with the proposed model.

2. Motivation

Pioneering work in experimental design was done by Ronald A. Fisher, a British statistician and biologist. One such experiment was Lady Tasting Tea.

Lady Tasting Tea: A randomized experiment devised, is quoted in his book "Design of Experiments". The experiment originated in a tea party which Sir Ronald Fisher happened to attend.

A tea can be made by adding either of tea or milk first. A lady in the party claimed that she could tell which has been added first just by taking the sip of tea. Then and there, Ronald Fisher devised an experiment in which the lady was given eight random cups of tea – four of each kind. She was asked to select four cups in which tea had been added first.

The Null Hypothesis was that lady selecting the right cups was a mere chance. Thus, she had 50-50 chance of making the right choice for each pair. The probability chart would be a normal distribution in this case. Fisher did not consider alternative hypothesis. Failure of null hypothesis would mean that the lady did not possess the ability to detect the differences between cups.

The technique demonstrates how a restricted number of experiments can be sufficient to device general laws considering several variables at the same time.

A normal distribution proves that it would have been difficult for the lady to choose all the correct answers if she were only guessing.

Using the combination formula, there are 70 possibilities.

$${}^{8}C_{4} = 70$$ (choosing 4 correct from 8 cups)

Success count	Permutation of selection	Permutations
0	Oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	Xxxx	1 × 1 = 1
Total		70

The result came out that the lady guessed all the cups right. There next question was repetition of the experiment and then comparing the results, which further enhanced his idea of experimental design and thus leading to formulation of Anova.

3. Background and Terminologies

Hypothesis testing and variance estimation was performed in late 18th century. The term variance was, though, later coined in 19th century by Sir Fisher.

Anova extends the concept of *t*-tests and the *z*-tests which have the limitations of only analyzing two groups.

$$t - score = \frac{difference\ between\ various\ groups}{difference\ within\ the\ group} = t$$

t => $groups\ are\ t\ times\ as\ different\ from\ each\ other\ as\ they\ are\ within\ each\ other.$

For multiple groups, the solution was to run multiple t-tests, in which case the total error would amplify, hence the Anova. As it analyses H_0(null hypothesis) of multiple pairs in one go, it results in less type I error. Thus, it is the conservative form of multiple two-sample t-tests.

Decision	**H0 = True**	**H0 = False**
H0= reject	Type 1 error, (false positive) Denoted as α Aka “significance level of a test”	Correct inference (True Positive)
H0 = Fail to reject	Correct inference (True Negative)	Type II error (False Negative) Denoted as β

The era when it was introduced, the ease of its usage with computers broadcasted it further. Tables of the F function were easily supplied.

4. Anova designs

In ANOVA computation, the variance of a variable is attributed to different variation sources, and thus significance is calibrated by considering the interaction between two categories. The use of ANOVA depends on the research design. In its simplest form, ANOVAs are used in three ways: one-way ANOVA, two-way ANOVA, and N-way ANOVA. Here, we will see:

- One-way Anova,
- Factorial Anova,
- Repeated Measures Anova,
- M-AN-O-VA

One way or two way refers to the number of independent variables in variance test.

Designs

A. One-Way ANOVA
   One-way ANOVA has one factor or independent variable and one dependent variable. For example, tea test to answer one question (what is poured first – milk or tea)
B. Factorial Anova
   Anova with more than two factors come under Factorial Anova.
   Two-Way ANOVA
   Two-way ANOVA refers to an ANOVA using two independent variables. The above example can be expanded as examining the tea taste at various levels
   N-Way ANOVA
   Anova with n factors.
C. Repeated Measures Anova
   Testing is done repeatedly on the same sample. Extending the above example: Tea is being poured to the same people multiple times and results are being calibrated for the same set of variables.
D. M-AN-O-VA
   Multivariate Analysis of Variance is application of Anova on multiple independent variables as well as multiple dependent variables. The above examples can be expanded as: tea being tested at various levels (sugar, content, spices, etc.) for various factors (taste, color)

5. Computing F Score and Anova Table

The calculation requires below steps:

Total Sum of Squares, SST a measure of all variations in the dependent variable

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$$

Mean Square Total, MST is the SST divided by the degree of freedom which is total number of variables minus one

$$MST = \frac{SST}{df(SST)} = \frac{SST}{N-1},$$

Sum of Squares Treatment, SSC

$$SSC = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x})^2$$

Mean Square Treatment, MSC is the SSC divided by the total number of groups minus one

$$MSC = \frac{SSC}{df(SSC)} = \frac{SSC}{k-1}.$$

Sum of Squares of Errors,

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

Mean Square Error, MSE is the SSE divided by the degree of freedom, total variables minus groups

$$MSE = \frac{SSE}{df(SSE)} = \frac{SSE}{N-k},$$

Ideally, Sum of Squares Total is the sum of Sum of Squares Treatment and Sum of Squares Error

$$SST = SSC + SSE$$

Thus, F-score is computed as the ratio of MSC and MSE

$$F = \frac{MSC}{MSE}.$$

If F-score is greater than the critical value from the F-statistics table, then we reject the null Hypotheses

$$F > F_{1-\alpha, k-1, N-k},$$

The table is as follows:

Variance source	Sum of squares *SS*	Degrees of freedom *df*	Mean square *MS*	*F*-statistic
Between	SSC	$k-1$	MSC	MSC/MSE
Within	SSE	$N-k$	MSE	—
Total	SST	$N-1$	—	—

6. Conclusion

This report describes the various designs of ANOVA and its effectiveness in case of non-applicability of t-tests. With large datasets, with various dependent and independent variables, Anova gives faster results as it converges numerous variables to few relevant variables. The technique can be used in regression modelling as the starting point.

7. References

I. Fisher, Ronald A. (1971) [1935]. The Design of Experiments (9th ed.). Macmillan. ISBN 0-02-844690-9.
II. R. A. Fisher and the Design of Experiments, ISSN: 00031305
III. Design of Experiment - An Integration of Fisher, Taguchi and Shainin DOE Methodology, 2015, ISSN: 16627482
IV. Analysis of Variance – Why is it more important than ever by Andrew Gelman,
V. Statistical Methods for Research Workers, Ronald A. Fisher, 1970
VI. Methodology and Application of One-way Anova by Eva Ostertagová, Oskar Ostertag, November 2013
VII. Onlinestatbook.com developed by David Lane
VIII. Tuning MPI Runtime Parameter Setting for High Performance Computing by Simone Pellegrini, Radu Prodan, Thomas Fahringer, 2012 (IEEE International Conference on Cluster Computing Workshops)