Application of One-Way ANOVA

Repaka Rama Rao[1], Kalidindi Narayana Raju[2], P L Suresh[3]

[1]Associate Professor, Department of Computer Science, B V Raju College, Bhimavaram, Andhra Pradesh, India
Email: *ramarao.r[at]bvricedegree.edu.in*

[2]Associate Professor, Department of Statistics, B V Raju College, Bhimavaram, Andhra Pradesh, India
Email: knraju20[at]gmail.com

[3]Associate Professor, Department of Mathematics, B V Raju College, Bhimavaram, Andhra Pradesh, India
Email:p.arunasuresh[at]gmail.com

Abstract: ***This paper describes the dominant statistical method one-way ANOVA that can be used in many engineering and manufacturing applications and presents its application. This method is proposed to analyze variability in data in order to deduce the inequality among population means.***

Keywords: one-way ANOVA test, normality tests

1. Introduction

Analysis of variance (ANOVA) is a statistical procedure concerned with comparing means of several samples. It can be thought of as an extension of the t-test for two independent samples to more than two groups. The purpose is to test for significant differences between class means, and this is done by analysis the variances. The ANOVA test of the hypothesis is based on a comparison of two independent estimates of the population variance. When performing an ANOVA procedure the following assumptions are required: The observations are independent of one another. The observations in each group come from a normal distribution. The population variances in each group are the same . ANOVA is the most commonly quoted advanced research method in the professional business and economic literature. This technique is very useful in revealing important information particularly in interpreting experimental outcomes and in determining the influence of some factors on other processing parameters. The original ideas of analysis of variance were developed by the English statistician Sir Ronald A. Fisher (1890-1962) in his book “Statistical Methods for Research Workers” (1925). Much of the early work in this area dealt with agricultural experiments.

One-way ANOVA
Test Procedure The simplest case is one-way ANOVA. A one-way analysis of variance is used when the data are divided into groups according to only one factor.

Assume that the data $x_{11}, x_{12}, x_{13}, \dots x_{1n_1}$ are the samples from the population 1 and $x_{21}, x_{22}, x_{23}, \dots x_{2n_2}$ are the samples from the population 2 …. $x_{k1}, x_{k2}, x_{k3}, \dots x_{kn_k}$are the samples from the population k. Let x_{ij} denote the data from the ith group (level) and jth observation.

We have values of independent normal random variables $x_{ij\,,}$ $\text{wherei} = 1,2,3,\dots k$ and $j = 1,2,3,..n_i$ with mean μ_i and constant standard deviation σ, $x_{ij} \sim N(\mu_i, \sigma)$. Alternatively, each $x_{ij} = \mu_i + \varepsilon_{i\,j}$ wherev are normally distributed independent random errors, $\varepsilon_{i\,j} \sim N(0, \sigma)$.Let N $= n_1 + n_2 + \cdots n_k$ is the total number of observations where n_i is the sample size of the ith group.

The parameters of this model are the population means $\mu_1, \mu_2, \mu_3, \dots, \mu_k$ k and the common standard deviation σ.Using many separate two-sample t-tests to compare many pairs of means is a bad idea because we don’t get a p-value or a confidence level for the complete set of comparisons together. We will be interested in testing the null hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k \tag{1}$$

against the alternative hypothesis

$$H_1: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k \tag{2}$$

(there is at least one pair with unequal means).

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{3}$$

Where $\bar{x}$ represent the grand mean, the mean of all the data points

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} \tag{4}$$

s_i^2 represent the sample variance:

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \tag{5}$$

and $s^2 = \text{MSE}$ is an estimate of the variance σ^2 common to all k populations,

$$s^2 = \frac{1}{N-k}\sum_{i=1}^{k} \quad (n_i - 1)s_i^2 \tag{6}$$

ANOVA is centered around the idea to compare the variation between groups (levels) and the variation within samples by analyzing their variances. Define the total sum of squares SST, sum of squares for error (or within groups) SSE, and the sum of squares for treatments (or between groups) SSC:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}\ \ )^2 \qquad (7)$$

$$SSB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \qquad (8)$$

$$SSB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x}\ \ )^2 \qquad (9)$$

Consider the deviation from an observation to the grand mean written in the following way

$$SST = SSW + SSB \qquad (10)$$

The total mean sum of squares MST, the mean sums of squares for error MSE, and the mean sums of squares for treatment MSC are:

$$MST = \frac{SST}{N-1}$$

$$MSW = \frac{SSW}{N-k}$$

$$MSB = \frac{SSB}{k-1}$$

The one-way ANOVA, assuming the test conditions are satisfied, uses the following test statistic:

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

Under H_0 this statistic has Fisher's distribution F (k-1,N-k).In case it holds for the test criteria

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

Where $F_{1-\alpha,k-1,N-k}$ is $(1-\alpha)$ quantile of F distribution with $k-1, N-k$ degrees of freedom, then H_0 hypothesis is rejected on significance level α ,The results of the computations that lead to the F-statistic are presented in an ANOVA table, the form of which is shown below.

One Way Anova Table

Variance Source	Sum of squares SS	Degrees of freedom df	Mean square MS	F-statistic
Between	SSC	$k-1$	MSC	$F = \frac{MSC}{MSE}$
Within	SSE	$N-k$	MSE	
Total	SST	$N-1$		

Example

One important factor in selecting software for word processing and database management systems is the time required to learn how to use a particular system. In order to evaluate three database management systems, a firm devised a test to see how many training hours were needed for six of its word processing operators to become proficient in each of three systems . The data from this experiment are in the below table. Using a 5 % significance level, is there any difference between the training time needed for the three systems?

Experiment data in hours

Observation	System 1	System 2	System 3	System 4
1	8	12	18	13
2	10	11	12	9
3	12	9	16	12
4	8	14	6	16
5	7	4	8	15

Solution:

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
∑A=45	∑B =50	∑C =60	∑D =65

A	B	C	D
64	144	324	169
100	121	144	81
144	81	256	144
64	196	36	256
49	16	64	225
$\sum A^2$=421	$\sum B^2$ =558	$\sum C^2$ =824	$\sum D^2$= 875

Data table

Group	A	B	C	D	Total
N	n_1 =5	n_2 =5	n_3 =5	n_4 =5	N =20
$\sum x_i$	T_1 = $\sum x_1$=45	T_1 = $\sum x_2$= 50	T_1 = $\sum x_3$= 60	T_1 = $\sum x_4$=65	$\sum x$ = 220
$\sum x_i^2$	$\sum x_1^2$=421	$\sum x_2^2$ =558	$\sum x_3^2$=824	$\sum x_4^2$ =875	$\sum x^2$ =2678
Mean $\bar{x}_i$	$\bar{x}_1$=9	$\bar{x}_2$ =10	$\bar{x}_3$=12	$\bar{x}_4$ =13	Overall $\bar{x}$=11
Standard deviationS_i	S_1 =2	S_2 =3.8079	S_3=5.099	S_4 =2.7386	

Let k = the number of different samples

$$n = n_1 + n_2 + n_3 + n_4 = 5 + 5 + 5 + 5 = 20$$

$$\bar{x} = \frac{220}{20}$$

$$\frac{(\sum x)^2}{n} = 2420$$

$$\frac{\sum T^2}{n_i} = 2470$$

$$\sum x_i^2 = \sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2$$
$$= 421 + 558 + 824 + \; +875 = 2678$$

sum of squares between samples

$$SSB = \frac{\sum T^2}{n_i} - \frac{(\sum x)^2}{n} = 2470 - 2420 = 50$$

sum of squares within samples

$$SSW = \sum x^2 \; - \frac{\sum T^2}{n_i} = 2678 - 2470 = 208$$

Total sum of squares

$$SST = SSB+SSW$$
$$= 50+208 = 258$$

variance between samples

$$MSB = \frac{SSB}{k-1}$$
$$= \frac{50}{3} = 16.6$$

variance within samples

$$MSW = \frac{SSW}{n-k}$$
$$= \frac{208}{20-4} = 13$$

test statistic F for one way ANOVA test

$$F = \frac{MSB}{MSW} = \frac{16.6}{13} = 1.28$$

The degrees of frees between samples= k-1 =3

The degrees of frees within samples = n-k= 16

One Way Anova Table

Variance Source	Sum of squares SS	Degrees of freedom df	Mean square MS	F-statistic
Between	SSB=50	k – 1=3	MSB=16.6	$F = \frac{MSB}{MSW}$=1.28
Within	SSW=208	N – k= 16	MSW= 13	
Total	SST=258	N – 1=19		

Null Hypothesis H_0: There is no significance difference between samples

Alternative Hypothesis H_1: There is a significance difference between samples

Level of Significance: At $\alpha = o.o5$, $F_{0.05}(3,16) = 3.23$

Conclusion: Calculated Value F = 1.28

Table Value $F_\alpha = 3.23$

$F < F_\alpha$

So We accept the Null hypothesis

That is there is no significance difference between samples

2. Conclusion

In many statistical applications in business administration, psychology, social science, and the natural sciences we need to compare more than two groups. For hypothesis testing more than two population means scientists have developed ANOVA method. The ANOVA test procedure compares the variation in observations between samples (sum of squares for groups, SSC) to the variation within samples (sum of squares for error, SSE). The ANOVA F-test rejects the null hypothesis that the mean responses are equal in all groups if SSC is large relative to SSE. The analysis of variance assumes that the observations are normally and independently distributed with the same variance for each treatment or factor level.

References

[1] Aczel, A.D., Complete Business Statistics, Irwin, 1989.

[2] Brown, M., Forsythe, A., "Robust tests for the equality of variances," Journal of the American Statistical Association, 364- 367. 1974.

[3] Montgomery, D.C., Runger, G.C., Applied Statistics and Probability for Engineers, John Wiley & Sons, 2003.

[4] Ostertagová, E., Applied Statistic (in Slovak), Elfa, Košice, 2011.

[5] Parra-Frutos, I., "The behaviour of the modified Levene's test when data are not normally distributed," Comput Stat, Springer, 671-693. 2009.

[6] Rafter, J.A., Abell, M.L., Braselton, J.P., "Multiple Comparison Methods for Means," SIAM Review, 44 (2). 259-278. 2002.

[7] Rykov, V.V., Balakrishnan, N., Nikulin, M.S., Mathematical and Statistical Models and Methods in Reliability, Springer, 2010.

[8] Stephens, L.J., Advanced Statistics demystified, McGraw-Hill, 2004.

[9] Taylor, S., Business Statistics.www.palgrave.com.