International Journal of Social Science and Human Research

ISSN (print): 2644-0679, ISSN (online): 2644-0695

Volume 08 Issue 06 June 2025

DOI: 10.47191/ijsshr/v8-i6-05, Impact factor- 8.007

Page No: 3971-3976

Comments and Observations on The Analysis of Variance

Donald L. Buresh, Ph.D., Esq.
Touro University Worldwide

ABSTRACT: What is the purpose of conducting an analysis of variance? This is an interesting question that demands an answer. From a prima facia perspective, an analysis of variance helps a researcher evaluate how data is distributed. Does it bunch together, or are there wide gaps between data points? An analysis of variance can stand by itself, but it is particularly beneficial to appreciate before conducting a regression analysis. Thus, given that a researcher has previously determined the independent or explanatory variables and the dependent variable, this paper aims to help researchers understand the significance of an analysis of variance before they embark on deciding what regression technique to employ.

KEYWORDS: Analysis of Variance, Anova, Krushal-Wallis Test, Mann-Whitney Test

DONALD L. BURESH BIOGRAPHY

Donald L. Buresh earned his Ph.D. in engineering and technology management from Northcentral University. His dissertation assessed customer satisfaction for both agile-driven and plan-driven software development projects. Dr. Buresh earned a J.D. from The John Marshall Law School in Chicago, Illinois, focusing on cyber law and intellectual property. He also earned an LL.M. in intellectual property from the University of Illinois Chicago Law School (formerly, The John Marshall Law School) and an LL.M. in cybersecurity and privacy from Albany Law School, graduating summa cum laude. Dr. Buresh received an M.P.S. in cybersecurity policy and an M.S. in cybersecurity, concentrating in cyber intelligence, both from Utica College. He has an M.B.A. from the University of Massachusetts Lowell, focusing on operations management, an M.A. in economics from Boston College, and a B.S. from the University of Illinois-Chicago, majoring in mathematics and philosophy. Dr. Buresh is a member of Delta Mu Delta, Sigma Iota Epsilon, Epsilon Pi Tau, Phi Delta Phi, Phi Alpha Delta, and Phi Theta Kappa. He has over 25 years of paid professional experience in information technology and has taught economics, project management, negotiation, managerial ethics, cybersecurity, business law, and quality management at several universities. Dr. Buresh is an avid Chicago White Sox fan and is active in fencing épée and foil at a local fencing club. Dr. Buresh is a member of the Florida Bar.

MISCELLANEOUS CONSIDERATIONS

I thank Leizza Buresh for her tireless editorial efforts. She is helpful beyond measure. Any other errors that remain in this article are mine.

INTRODUCTION

What is the purpose of conducting an analysis of variance? This is an interesting question that demands an answer. From a prima facia perspective, an analysis of variance helps a researcher evaluate how data is distributed. Does it bunch together, or are there wide gaps between data points? An analysis of variance can stand by itself, but it is particularly beneficial to appreciate before conducting a regression analysis. Thus, given that a researcher has previously determined the independent or explanatory variables and the dependent variable, this paper aims to help researchers understand the significance of an analysis of variance before they embark on deciding what regression technique to employ.

THE PURPOSE OF CONDUCTING AN ANALYSIS OF VARIANCE

In this section, the purpose of employing an analysis of variance is discussed. The first subsection describes the purpose of an analysis of variance. The second subsection addresses various data models.

Comments and Observations on The Analysis of Variance

Purpose of an Analysis of Variance

The purpose of analyzing variance is to test whether several means are equal to each other.[1] Analysis of variance is derived from partitioning the total variance into its component parts.[2] The point is that when there are several subsamples whose differences need to be tested, a practical approach is to use an analysis of variance, also known as ANOVA.[3] According to Aczel, the technique involves analyzing the different types of variances associated with the random sample under consideration.[4]

The original ideas for an analysis of variance are attributed to Sir Ronald A. Fisher during the beginning of the 20th Century.[5] The early work focused on agricultural experiments, where crops were subjected to various fertilization treatments with the aim of increasing yield. According to Aczel, the researchers wanted to discover if all of the treatments were equally effective or if some treatments were better than others.[6] Because of this history, the words *treatments* and *populations* are used interchangeably. For example, if the mean income for several different communities is compared, then the populations of each community can be considered a treatment.

Models of the Data

It is always helpful to describe the observations from an experiment with a model. One way to characterize this model is:

$$y_{ij} = \mu_i + \varepsilon_{ij} \text{ where } i = 1, \ldots, m \text{ and } j = 1,\ldots, n$$

where y_{ij} is the *ij*th observation, μ_i is the mean of the ith factor, and ε_{ij} is the random error of all sources of variability in the experiment.[7] This model is known as the *means model*. If μ is the mean that is common to all of the treatments, and τ_i is the parameter that is unique to the ith treatment, then

$$\mu_i = \mu + \tau_i \text{ where } i = 1, \ldots, m$$

If this equation is substituted into the equation above,

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \text{ where } i = 1, \ldots, m \text{ and } j = 1,\ldots, n$$

This equation is known as the *effects model* because it has a common mean, representing the effects of the various treatments, and accounts for the random error due to all sources of variability.

Now, this equation is called a one-way or *single-factor* analysis of variance because only one variable is under investigation.[8] An experiment with one factor demands that it be performed in random order so that the environment where the treatments are applied is as uniform as possible.[9] This means that the design of the experiment is completely randomized, where the objective is to test the appropriate hypothesis about the treatment means, as well as to estimate them. In other words, assuming that the variance σ^2 is known, then $y_{ij} \sim N(\mu + \tau_i, \sigma^2)$, meaning that the observations are mutually independent. An analysis of variance is derived from partitioning the variation into its component parts.[10] If the total sum of squares is:

$$SS_{Total} = \sum_{i=0}^{m}\sum_{j=0}^{n}(y_{ij} - y - bar)^2$$

where y-bar is the overall mean of y_{ij}, then adding and subtracting:

$$y_i - bar = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

where y_i-bar is the ith mean of y_{ij}. Because the cross term is equal to zero, it is straightforward to derive that:

$$\sum_{i=0}^{m}\sum_{j=0}^{n}(y_{ij} - y - bar)^2 = n\sum_{i=1}^{m}(y_i - bar - y - bar)^2 + \sum_{i=0}^{m}\sum_{j=0}^{n}(y_{ij} - y_i - bar)^2$$

---

[1] DOUGLAS C. MONTGOMERY, DESIGN AND ANALYSIS OF EXPERIMENTS, (John Wiley & Sons. Inc. 10th ed. Jun. 2019).
[2] *Id.*
[3] TERRY G. VARVA, IMPROVING YOUR MEASUREMENT OF CUSTOMER SATISFACTION: A GUIDE TO CREATING, CONDUCTING, ANALYZING AND REPORTING CUSTOMER SATISFACTION MEASUREMENT PROGRAMS (American Society for Quality 1997).
[4] *Id.*
[5] AMIR D. ACZEL, COMPLETE BUSINESS STATISTICS (McGraw-Hill, Inc. 8th ed. Jan. 2012).
[6] *Id.*
[7] Douglas C. Montgomery, *supra*, note 1.
[8] *Id.*
[9] *Id.*
[10] *Id.*

or $SS_{Total} = SS_{Treatments} + SS_{Error}$, where $SS_{Treatments}$ is the sum of the squares due to the treatments, and SS_{Error} is the sum of the squares due to the errors. Because N equals m times n total observations, SS_{Total} has $N - 1$ degrees of freedom, while $SS_{Treatments}$ has $m - 1$ degrees of freedom, and SS_{Error} has $n - 1$ degrees of freedom. Since m sample variances can be combined as a weighted average to provide one estimate of the population variance, it can be shown that $SS_{Error} / (N - m)$ is a pooled estimate of the common variance within each of the m treatments.[11]

Furthermore, if there are no differences between the m treatment means, it can be shown that $SS_{Treatments} / (N - m)$ is an estimate of σ^2. Thus, there are two estimates of σ^2, one based on the variability within the treatments and the other on the variability between treatments.[12] If there is no difference in the treatment means, then the two estimates are approximately the same. If the two estimates are significantly different from each other, then it is apparent that there are differences in the treatment means. The hypothesis under consideration can be written as:[13]

$$H_0: \tau_1 = \ldots = \tau_m = 0$$
$$H_1: \tau_i \neq 0 \text{ for at least one i}$$

More commonly, this is written as:

$$H_0: \mu_1 = \ldots = \mu_m$$
$$H_1: \mu_i \neq \mu_j \text{ for some } i \neq j$$

According to Cochran's Theorem, $SS_{Treatments} / \sigma^2$ and SS_{Error} / σ^2 are independently distributed chi-square random variables.[14] Thus, if the null hypothesis is true, then the ratio:

$$F - Statistic = \frac{SS_{Treatments} / (m - 1)}{SS_{Error} / (N - m)}$$

is distributed as an *F-statistic* with $m - 1$ and $N - m$ degrees of freedom and is the test statistic for the null hypothesis stated above.

REQUIRED ASSUMPTIONS OF AN ANALYSIS OF VARIANCE

According to Berger et al., the following subsections outline procedures for determining whether data from three or more populations created by the associated treatments in an experiment measuring a single factor indicate that the population means are distinct.[15] The first subsection deals with a one-way analysis of variance. The second subsection considers a two-way analysis of variance with equal variances. The final subsection talks about a two-way analysis of variance with unequal variances.

Procedure for One-Way Analysis of Variance

The assumptions for a one-way analysis of variance are:

- The samples are independent;
- The populations are normal; and
- The standard deviations are equal (i.e., homogeneity of variances).

The hypotheses are $H_0: \mu_1 = \ldots = \mu_m$ and $H_1: \mu_i \neq \mu j$ for some $i \neq j$. This is always a right-tail test Next, the α value must be specified. With this information, one can calculate SS_{Total} with $N - 1$ degrees of freedom, $SS_{Treatments}$ with $m - 1$ degrees of freedom, and SS_{Error} with $N - m$ degrees of freedom. This information is needed to derive the F-statistic defined above. In the fifth step, the critical value in an F-table with $m - 1$ degrees of freedom in the numerator and $N - m$ degrees of freedom in the denominator is found. In the sixth step, a researcher determines whether the null hypothesis should be rejected. In other words, the calculated F-statistic is greater than or equal to the F-statistic found in the table. In this case, the null hypothesis is rejected, and the alternate hypothesis is accepted. On the other hand, if the calculated F-statistic is strictly less than the F-statistic from the table, then the null hypothesis cannot be rejected. Finally, the conclusion is stated in terms of the original problem.

[11] *Id.*
[12] *Id.*
[13] *Id.*
[14] *Id.*
[15] ROGER W. BERGER, DONALD W. BENBOW, AHMAD K. ELSHENNAWY, & H. FRED WALKER (EDS), THE CERTIFIED QUALITY ENGINEER HANDBOOK (American Society of Quality Press.2nd ed. 2006).

Comments and Observations on The Analysis of Variance

Procedure for Two-Way Analysis of Variance with Equal Variances

Berger et al. also discussed the procedure for conducting a two-way analysis of variance, where there are two factors to be analyzed, such as A and B.[16] In this case, there are five sources of variation:

- Factor A
- Factor B
- Interaction or intersection of A and B (A x B)
- Errors
- Total

In this case, if *SS* denotes the sums of the square, then SS_A has $(n_A - 1)$ degrees of freedom, SS_B has $(n_B - 1)$ degrees of freedom, SS_{AB} has $(n_A - 1)(n_B - 1)$ degrees of freedom, SS_{Error} has $N - n_A n_B$ degrees of freedom, and SS_{Total} has $N - 1$ degrees of freedom, where N is the total number of readings, n_A is the number of levels for factor A, and n_B is the number of levels of factor B.[17] Three F-statistics should be calculated according to the following formulas:

$$F_A = [SS_A / (n_A - 1)] / [SSE / (N - n_A n_B)]$$
$$F_B = [SS_B / (n_B - 1)] / [SSE / (N - n_A n_B)]$$
$$F_{AB} = [SS_{AB} / (n_B - 1)] / [SSE / (N - n_A n_B)]$$

The three F-statistics are then compared to their associated critical values found in an F-statistics table. If any of the calculated F-statistics above are greater than their associated value in the F-statistics table, then that indicates that there is a statistically significant effect due to the relevant factor. It should be noted that for both one-way and two-way analysis of variance tests, the populations, treatments, or factors are normally distributed, with equal standard deviations.

Procedure for Two-Way Analysis of Variance with Unequal Variances

Another form of a two-way analysis of variance is a test for the population standard deviation, which is a chi-square test (the Greek letter χ).[18] First, it is assumed that the population is normally distributed. The hypotheses to be tested are:

$H_0: \sigma = \sigma_0$

$H_1: \sigma \neq \sigma_0$ or $\sigma < \sigma_0$ or $\sigma > \sigma_0$ (two-tail, left tail and right tail tests respectively)

The α value is then specified. The critical values are obtained from a χ^2 table, similar to how a t-table is employed. The degrees of freedom are $n - 1$, where n is the number of observations.

- For a two-tail test, the critical values are $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$.
- For a left tail test, the critical value is $\chi^2_{1-\alpha}$
- For a right tail test, the critical value is χ^2_{α}

The formula for the test statistic is $\chi^2 = [(n - 1) s^2] / \sigma^2_0$, where s is the sample standard deviation σ_0 is the standard deviation being used in the comparison. A researcher should reject the null hypothesis if the value of the test statistic is in the reject region. Otherwise, the null hypothesis should not be rejected. Finally, an individual should state the conclusion in terms of the problem.

OBSERVATIONS ABOUT THE ANALYSIS OF VARIANCE

This section is divided into general observations and a description of the Kruskal-Wallis test. Each subsection is discussed in turn.

General Observations

In all three analyses of variance tests considered above, the populations were assumed to be independent and normal. In the first two tests, the standard deviations of the relevant population were assumed to be equal. Aczel restated the independence assumption that independent *random* sampling is employed for each of the populations under consideration.[19] The reason that these assumptions are in force is to ensure that the three ratios listed above have an *F* distribution when the null hypothesis is true. Aczel observed that

[16] *Id.*
[17] *Id.*
[18] *Id.*
[19] Amir D. Aczel, *supra*, note 5,

if the populations do not exactly possess a normal distribution, but are close, the methods listed above still yield good results.[20] If the populations are skewed in one direction or another, or if the population variances are not approximately equal, then it is recommended that the analysis of methodology not be used.[21] In this latter case, the appropriate non-parametric technique to be used is the Krushal-Wallis test.[22]

Krushal-Wallis Test

The Krushal-Wallis test is a non-parametric analysis of variance test that employs the ranks of the observations, rather than the actual data itself. The assumption behind this test is that the observations are on an interval scale.[23] It should be noted that the Krushal-Wallis test is identical to the Mann-Whitney U test, also known as the Wilcoxon rank-sum test,[24] when there are only two populations under consideration.[25] This means that the power of the Krushal-Wallis lies where there are *k* populations, or treatments, under consideration, and $k > 2$.

The null hypothesis is that the *k* populations have the same distribution, while the alternative hypothesis is that at least two of the populations are different. It is important to note that the Kruskal-Wallis test does not specify the underlying distribution. It only assumes that the distributions for the populations are the same. Thus, the null and alternative hypotheses are:

H_0: All *k* populations have the same distribution.
H_1: Not all *k* populations have the same distribution.

Aczel observed that this test is quite sensitive to differences in the locations of the populations.[26] Another assumption for the Krushal-Wallis test is that the *k* samples are randomly drawn from their respective populations.[27]

To use this test, all of the data are ranked from smallest to largest, regardless of their original population. The next step is to sum all of the ranks from each of the samples. Let n_1 be the number of data items from the first sample, n_2 be the number of data items from the second sample, etc., and $N = n_1 + n_2 + \ldots + n_k$. If R_1 is the sum of the ranks from the first sample, R_2 is the sum of the ranks from the second sample, etc., then the Krushal-Wallis test statistic H is defined by the equation:[28]

$$H = \frac{12}{N(N-1)}\left[\sum_{j=1}^{k} R_j^2 / n_j\right] - 3(N+1)$$

Aczel stated that for sample population sizes less than *5*, the exact distribution for H can be found in books on non-parametric statistics.[29] On the other hand, for sample population sizes equal to *5* or more, the test statistic for H is approximately equal to the chi-square distribution with $k - 1$ degrees of freedom.[30] The null hypothesis is rejected when the value of the H statistic is greater than $\chi^2_{(k-1)}$ for the specified level of significance α.

[20] *Id.*
[21] *Id.*
[22] *Id.*
[23] *Id.*
[24] The Mann-Whitney U test, is a non-parametric statistical test employed to compare two independent groups when the dependent variable is ordinal or continuous and not normally distributed. It is a non-parametric alternative to the independent samples t-test. The test determines if there's a statistically significant difference between the medians (or overall distributions) of two independent groups. It should be used when the assumptions of the independent samples t-test (normally distributed data) are not satisfied, or when the data is already ranked. The test ranks all data points from both groups together and then calculates the sum of the ranks for each group. A large difference between the rank sums indicates a statistically significant difference. The Mann-Whiteny U test assume at (1) the dependent variable should be ordinal or continuous, (2) the independent variable should be categorical with two groups, and (3) the data are independent.
[25] *Id.*
[26] *Id.*
[27] *Id.*
[28] *Id.*
[29] *Id.*
[30] *Id.*

CONCLUSION

In conclusion, this article sought to explain how to analyze data variance. Although an analysis of variance can be viewed as a standalone statistical analysis, in the author's opinion, it is particularly beneficial when used as a prelude to regression analysis. Granted that the researcher has already identified the independent or explanatory variables and the dependent variables, an analysis of variance can be employed to show the variance, divergence, or disparity in the data. With this knowledge, a researcher is better equipped to select which regression technique to employ. Given the various data interactions, such knowledge is essential when effectively analyzing data.

REFERENCES

1) Amir D. Aczel, Complete Business Statistics (Mcgraw-Hill, Inc. 8th Ed. Jan. 2012).
2) Douglas C. Montgomery, Design And Analysis Of Experiments, (John Wiley & Sons. Inc. 10th Ed. Jun. 2019).
3) Roger W. Berger, Donald W. Benbow, Ahmad K. Elshennawy, & H. Fred Walker (Eds), The Certified Quality Engineer Handbook (American Society Of Quality Press 2nd Ed. 2006).
4) Terry G. Varva, Improving Your Measurement Of Customer Satisfaction: A Guide To Creating, Conducting, Analyzing And Reporting Customer Satisfaction Measurement Programs (American Society For Quality 1997).