UYEMURA International Corporation

The Analysis of Variance: Drawing Conclusions from Data That Are Correct, Unambiguous, and Defensible

by: Pat Valentine, PhD
Technical Department, Uyemura USA, Southington, CT USA

Abstract

Statistical tests are powerful tools that help process engineers make better decisions on process improvement projects. Drawing conclusions from an improvement project's data that are correct, unambiguous, and defensible is crucial for the process engineer. One of the most common parameters of interest with improvement projects is the mean. The purpose of this paper is to go over the appropriate steps for using the analysis of variance with multiple mean responses. The analysis of variance is reviewed, along with model validation and the key data acceptance criteria required. A worked example is provided using pulse acid copper throwing power.

Keywords: ANOVA, means, model validation, data acceptance, pulse plating

Introduction

The two-sample t-test is used to determine if two population means are equal. A typical application tests if a new process or treatment is superior to a current one. But what if we have three or more means we want to test? The t-test is inappropriate for this analysis.

For example, a young engineer is testing the mean brightener concentration in her four acid copper pulse plating tanks (A, B, C, D). There are six pairwise comparisons: AB, AC, AD, BC, BD, CD. Using the t-test, if the probability of correctly accepting the null hypothesis for each test is $1 - \alpha = 0.95$, then the probability of correctly accepting the null hypothesis for all six tests is $(0.95)^6 = 0.74$, or 74%. In other words, $1 - 0.74 = 26\%$ chance of committing a Type I error. Recall that a Type I error occurs when we reject a true null hypothesis (no statistical difference) and claim that there is a statistical difference. The multiple comparisons cause a significant increase in Type I errors. The appropriate procedure for testing the equality of several means is the analysis of variance [1].

The Analysis of Variance

The analysis of variance (ANOVA) was invented by British statistician R.A. Fisher in 1918. While the t-test was limited to comparisons between two groups, the ANOVA was designed to allow for comparisons between multiple groups using a single test. The ANOVA gained popularity after being included in Fisher's text, Statistical Methods for Research Workers, in 1925.

Today, the ANOVA is the most useful technique in the field of statistical inference. The ANOVA is a general linear statistical model technique used to test the hypothesis that the means of two or more groups are equal. The linear function refers to the mathematical relationship between the model parameters and the dependent variable (y). Specifically,

the response variable (y) is a linear function of the model parameters (the average outcome is linearly related to each term in the model) [2].

There are two types of assumptions with the ANOVA model. The first assumption is about the form of the model. These initial assumptions pertain to choosing the correct predictors (they are related to the response variable), and the average outcome is linearly related to each term in the model [2, 3].

The second assumption is about the distribution of the errors (residuals). It is generally assumed that the sampled populations are approximately normally distributed, the observations are independent, the variances are equal across groups (homogeneity), and the observations have been randomly sampled. The ANOVA technique is robust to minor deviations from normality, independence, and homogeneity. You can get clues about whether most of these assumptions will be met before building the model. But we typically build the model first and then verify the assumptions. Suppose you've done the foundational work in the early steps. In that case, testing assumptions is about looking for minor deviations, not major transgressions [2, 3].

The ANOVA tests the null hypothesis (H0) that two or more population means are equal versus the alternative hypothesis (H1) that at least one mean is different. Using the formal notation of statistical hypotheses, for k means we write:

H_0: $\mu 1 = \mu 2 = ... = \mu k$

H_1: At least one mean is not equal to the others

In statistics, the alternative hypothesis can be either one-tailed or two-tailed. The one-tailed tests are for either inferiority or superiority, while the two-tailed tests are for parity (not equal). The ANOVA is a bit more complex.

With ANOVA, we test "not all means are equal." Suppose we are comparing three groups; the alternative hypothesis says that at least one of the following is true:

Mean 1 is not equal to mean 2.

Mean 1 is not equal to mean 3.

Mean 2 is not equal to mean 3.

As implied, the ANOVA analyzes variances to test means. But why analyze variances to derive conclusions about the means? Remember that "means are different." And the larger the differences between the means, the more variation there is present. The ANOVA assesses the amount of variability between the group means in the context of the variation within groups to determine whether the mean differences are statistically significant. When the ANOVA signals statistically significant results (p-value < 0.05), indicating that not all means are equal, you'll need to use post hoc tests to complete pairwise comparisons.

Let's look at how the ANOVA works by using an example. Table 1 shows three factors (A, B, C), with three measured responses per factor, along with descriptive statistics. The data is fictitious and is presented for explanatory purposes only.

	A	B	C
	1	4	7
	2	5	8
	3	6	9
Mean:	2	5	8
Std Dev:	1.0	1.0	1.0

Table 1. Three-factor data set.

A raw ANOVA table is shown in Table 2, followed by the detailed ANOVA calculations. Finally, the completed ANOVA is shown in Table 3.

Source of Variation	Degrees of Freedom	Adj Sum of Squares	Adj Mean Square	F-Value	P-Value
Factor	df_{factor}	SS_{factor}	MS_{factor}	F	p
Error	df_{error}	SS_{error}	MS_{error}		
Total	df_{total}	SS_{total}			

Table 2. Raw ANOVA table.

Descriptive Statistics

Descriptive statistics, such as the mean and standard deviation, summarize a set of data [4, 5].

Mean of A: 1 + 2 + 3 / 3 = 2

Mean of B: 4 + 5 + 6 / 3 = 5

Mean of C: 7 + 8 + 9 / 3 = 8

Grand Mean: 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 / 9 = 5

Degrees of Freedom

Degrees of freedom (n – 1) are the number of independent values that a statistical analysis can estimate; more specifically, they define how many units within a set can be selected without constraints. Let's say we have three numbers that add up to 12. There are two degrees of freedom (3 – 1 = 2). After picking the first two numbers, there is no freedom to choose the last number; it is "determined" by the other two numbers. The first and second numbers can be any positive or negative numbers. For example, if the first number is 3, the second number is 7, the third number must be 2 [4, 5].

Factor: 3 – 1 = 2

Error: 8 – 2 = 6

Total: 9 – 1 = 8

Sum of Squares

The sum of the squared deviations of scores from their mean. The total sum of squares helps express the total variation that can be attributed to various factors. The adjusted sum of squares is the unique portion of the sum of squares explained by a factor, given all other factors in the model, regardless of the order they were entered into the model [4, 5].

Factor (between the factors): $3 * [(2 – 5)^2 + (5 – 5)^2 + (8 – 5)^2] = 54$. (Note: "3" is the number of levels within the factors, not the number of factors, and "5" is the grand mean.)

Error (within the factors):

SS of A: $(1 – 2)^2 + (2 – 2)^2 + (3 - 2)^2 = 2$

SS of B: $(4 – 5)^2 + (5 – 5)^2 + (6 – 5)^2 = 2$

SS of C: $(7 – 8)^2 + (8 – 8)^2 + (9 – 8)^2 = 2$

Error: 2 + 2 + 2 = 6

Total: 54 + 6 = 60

Mean Squares

A term used in the analysis of variance to refer to the variance in the data due to a particular source of variation. Converting the sum of squares into mean squares by dividing by the degrees of freedom lets you compare these ratios and determine whether there is a

significant difference. The larger this ratio is, the greater the factor's impact on the outcome [4, 5].

Factor: 54 / 2 = 27

Error: 6 / 6 = 1

F-value

Calculated by dividing the factor mean square by the error mean square. As an alternative to calculating the p-value, F-critical can be used. The F-critical is found in the F-table, using the degrees of freedom for the factor and error, F(2, 6). An F-value greater than F-critical indicates statistical significance [4, 5].

F-value: 27 / 1 = 27

F-critical: 5.14

P-value

The P-value indicates the probability of observing the given F-value (or a more extreme value) under the assumption that the null hypothesis is true. It is calculated from the F-distribution, F(2, 6), using the F-value [4, 5].

F-value: 27
Probability (p-value) of X ≥ 27, F(2, 6) = 0.001

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	2	54	27	27	0.001
Error	6	6	1		
Total:	8	60			

Table 3. Completed ANOVA table.

The ANOVA signals statistically significant results (P-value < 0.05), indicating that not all means are equal. But before action is taken, the model needs to be validated by examining the residuals. If all looks good, a post hoc test needs to be conducted for all pairwise comparisons. Finally, a review of the five requirements for data acceptance is required.

Model Validation

The ANOVA work does not stop when the model is fit. As discussed previously, the second assumption is about the distribution of the residuals. If your model is not adequate, it will incorrectly represent your data. For example, incorrect F- and P-values. Models can be adversely affected by as few as one or two points [4].

To validate the model, the assumptions about the distribution of the residuals must be met. These assumptions include that the residuals are normally distributed, have independence of observations (no autocorrelation), and have homogeneity of variances (equal variances across groups). Residuals are elements of variation unexplained by the model. Since they are a form of error, the same general principles apply to the group of residuals as would apply to errors in general: one expects them to be normal and independently distributed (NID) with a mean of zero and constant variance NID(0, σ^2). Departures from these assumptions usually mean that the residuals contain unaccounted-for information. Validating the model helps ensure the conclusions drawn are correct, unambiguous, and defensible [1, 3].

Normality

Virtually any graph suitable for displaying the distribution of a set of data is ideal for judging the normality of the distribution of a group of residuals. The two most common plots and graphs are the normal probability plot and the histogram [3, 4].

Interpretation: The normal probability plot of the residuals should approximately follow a straight line, see Figure 1. The histogram helps identify whether the data are skewed or contain outliers, as shown in Figure 2. With histograms, it's best to have at least 50 data points (n ≥ 50) to make interpretation robust [4].

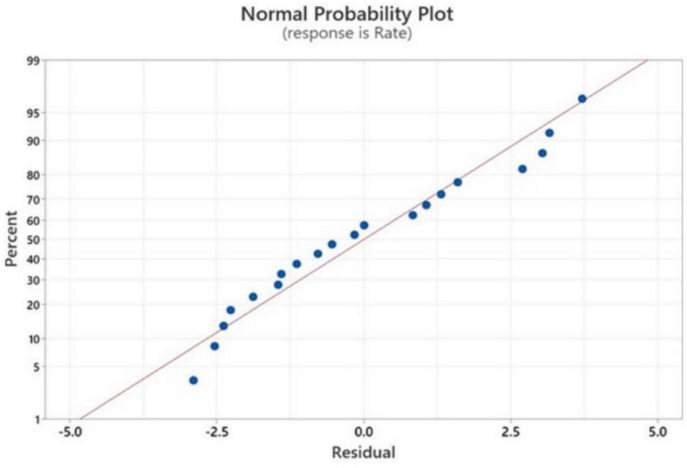

Figure 1. A reasonable probability plot.

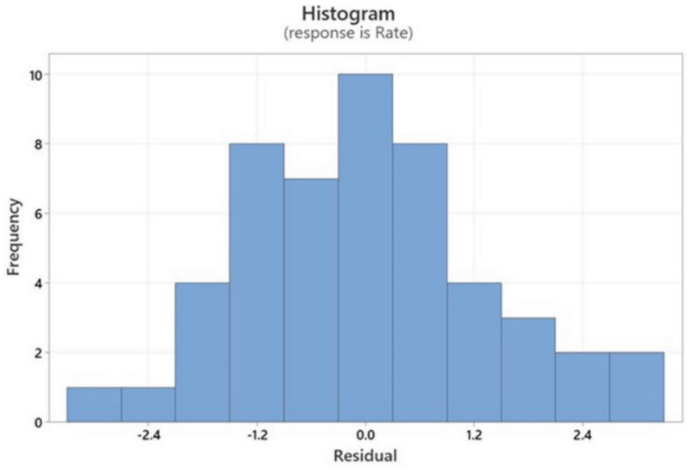

Figure 2. A reasonable histogram.

Independence

Suppose the order of the observations in a data table represents the order of execution of each test. In that case, a plot of the residuals of those observations versus the time order of the observations will test for lack of independence. For example, drift in equipment will produce models with autocorrelation. [3, 4].

Interpretation: Independent residuals show no trends or patterns when displayed in time order. Patterns in the data points indicate that residuals near each other may be correlated and thus not independent. The residuals on the plot should fall randomly around the center line with a mean of zero and constant variance NID(0, σ^2) with no recognizable patterns or trends in the points, see Figure 3 [4].

Homogeneity

Plotting residuals versus the value of a fitted response should produce a distribution of points scattered randomly about zero, NID (0, σ^2), regardless of the size of the fitted value. Quite commonly, however, residual values may increase as the size of the fitted value increases. When this happens, the residual cloud becomes "funnel-shaped" with the larger end toward larger fitted values; that is, the residuals have larger and larger scatter as the value of the response increases [3, 4].

Interpretation: Ideally, the points should fall randomly around the center line with a mean of zero and constant variance NID(0, σ^2) with no recognizable patterns, trends, or outliers in the points, see Figure 4 [4].

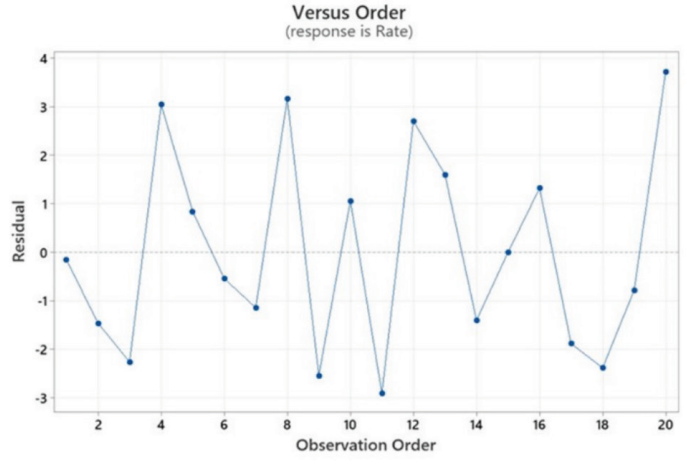

Figure 3. A reasonable residuals versus time order plot.

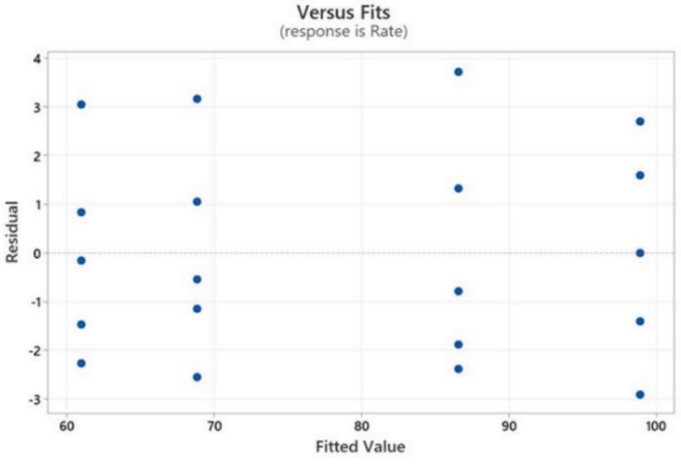

Figure 4. A reasonable residuals versus fits plot.

Post-hoc Testing

Suppose the ANOVA indicates a statistical difference (p-value < 0.05), and the model assumptions have been validated. In that case, post-hoc tests are used to identify which specific groups differ from each other. Standard post-hoc tests include Tukey, Fisher, Dunnett, and Hsu MCB. The Tukey and Fisher tests compare all pairs of groups. The Dunnett test compares the treatment groups to a control group. In contrast, the Hsu MCB test compares each group to the group with either the largest or the smallest mean (chosen by the process engineer). The process engineer must consider individual and family error rates in conjunction with post-hoc testing [4].

The individual error rate is the maximum probability that one or more comparisons will incorrectly conclude that the observed difference is significantly different from the null hypothesis. It is equivalent to the alpha level selected (typically 0.05) for the hypothesis test. The family error rate is the maximum probability that a procedure consisting of more than one comparison will incorrectly conclude that at least one of the observed differences is significantly different from the null hypothesis. The family error rate is based on both the individual error rate and the number of comparisons. It is essential to consider the family error rate when making multiple comparisons because your chances of committing a Type I error for a series of comparisons are greater than the error rate for any one comparison alone [4].

The Tukey test is a robust, widely used, and popular post-hoc test. It compares all pairs of groups while controlling the simultaneous confidence level (SCL). The SCL is the percentage of times that a group of confidence intervals will all include the true population parameters or true differences between factor levels if the study were repeated multiple times. The SCL level is based on both the individual confidence level and the number of confidence intervals. The Tukey family error rate is typically controlled at 0.05 (5%). The trade-off with Tukey's is the less precise confidence intervals and hypothesis tests that are less powerful than either Dunnett's or Hsu's MCB [4, 6].

Data Acceptance

There are five requirements if conclusions drawn from data analysis are to be correct, unambiguous, and defensible. These five requirements are an equitable sample, stability, statistical significance, practical significance, and truth. Each of these is discussed below.

Equitable Sample: The sample is representative of the population. Free from bias and confounding. Sample size is sufficient, or confirmation runs have been done.

Stability: No unusual conditions when the data was collected. No outliers, trends, shifts, or non-random patterns.

Statistically Significant: p-values are real, not noise, typically $\alpha < 0.05$, and residuals are normal.

Practical Significance: Is the magnitude of difference worthwhile? Does anybody care?

Truth: Can you explain why it is true? Do you have a theory? Does the conclusion fit with the subject matter knowledge?

A Worked Example

Process characterization is an integral part of any continuous improvement program. There are many steps in that program for which process characterization is required. These include instances when we introduce a new process or tool for use, as well as when we bring a tool or process back online after scheduled/unscheduled maintenance, when we want to compare tools or processes, when we want to check the health of our process during the monitoring phase, when we are troubleshooting a bad process, or when we need to improve a process [3].

A young process engineer is completing a process improvement project on her acid copper pulse plating tanks, looking to improve throwing power. She conducts an experiment looking at three different pulse recipes. The first pulse recipe (P1) is the control (current wave), while recipes P2 and P3 are experimental. The test vehicle is an 18" x 24" panel with 20:1 aspect ratio holes. The engineer plates four panels with each of the three pulse recipes and measures the throwing power (Note: the runs are randomized to protect against noise variables). The throwing power percentages, along with descriptive statistics, are shown in Table 4.

	Recipe P1	Recipe P2	Recipe P3
	56	69	80
	48	66	85
	47	74	91
	52	75	88
Mean:	50.8	71.0	86.0
Std Dev:	4.1	4.2	4.7

Table 4. Throwing power percentages and descripive statistics.

The process engineer analyzes the throwing power data using an ANOVA. The Recipe p-value is less than 0.05, indicating that not all means are equal, see Table 5.

Source	DF	Adj SS	ADj MS	F-Value	P-Value
Recipe	2	2503.5	1251.75	65.98	0.000
Error	9	170.7	8.97		
Total	11	2647.2			

Table 5. Pulse recipe ANOVA.

Next, the engineer validates the model by examining the residuals. The probability plot of the residuals approximately follows a straight line. The histogram is ignored due to the presence of fewer than 50 data points, making interpretation difficult. The residuals versus order points fall randomly around the center line with a mean of zero and constant variance NID(0, σ^2) with no recognizable patterns or trends in the points. The residuals versus fit points fall randomly around the center line with a mean of zero and constant variance NID(0, σ^2) with no recognizable patterns, trends, or outliers in the points. All four plots can be seen in Figure 5. The model has been validated. The process engineer now needs to use a post hoctest to complete pairwise comparisons.

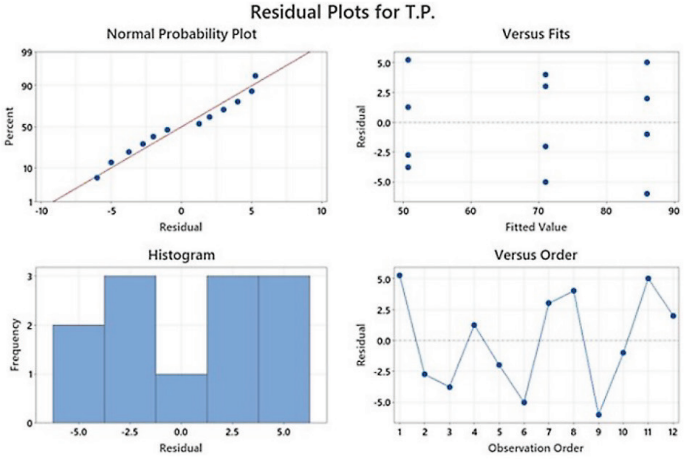

Figure 5. Four-in-one residual plot.

The process engineer decides to use the Tukey post hoc test. She uses the grouping information table to quickly determine whether the mean difference between any pair of groups is statistically significant. Groups that do not share a letter are significantly different. In these results, Table 6 shows that group A contains Recipe P3, group B contains Recipe P2, and group C contains Recipe P1.

Recipe	N	Mean	Grouping
P3	4	86	A
P2	4	71	B
P1	4	50.75	C

Table 6. Tukey post hoc test.

Means that do not show a layer are signifigantly different.

Discussion: The ANOVA model has been built, validated, and a post hoc test completed. The process engineer concludes that all three Recipe means are statistically different; the results in the data are unlikely to be explained by chance alone. The data acceptance criteria has been met: Equitable Sample (18" x 24" panel, 20:1 aspect ratio, four test panels), Stability (all parameters were in range during the testing), Statistical Significance (P-value < 0.05, and residuals are normal and independently distributed with a mean of zero and constant variance NID(0, σ^2)), Practical Significance (36% improvement in throwing power), and Truth (significant modifications to the pulse waves improve throwing power). Recipe P3 has been statistically proven to improve throwing power over Recipe P1 by an average of 36% (86% – 50%). The process engineer concludes her improvement project's data are correct, unambiguous, and defensible. She can confidently implement the process change.

Conclusions

The analysis of variance (ANOVA) is over 100 years old. Today, the ANOVA is the most useful technique in the field of statistical inference. The ANOVA is designed to allow for comparisons between multiple groups using a single test. The ANOVA work does not stop when the model is fit; the model must be validated. Validation is accomplished by verifying that the residuals are normally distributed, have independence of observations, and have homogeneity of variances. When the ANOVA indicates a statistical difference, and the model assumptions have been validated, a post-hoc test is used to identify which specific groups differ from each other. The Tukey test is a robust, widely used, and popular post-hoc test. Finally, data acceptance is based on five requirements: equitable sample, stability, statistical significance, practical significance, and truth. Drawing conclusions from an improvement project's data that are correct, unambiguous, and defensible is crucial for the process engineer.

References

[1] Montgomery, D. (2001). *Design and Analysis of Experiments, 5th Ed.* United States: John Wiley & Sons.

[2] The Analysis Factor: When to Check Model Assumptions. Available from: https://www.theanalysisfactor.com/when-to-check-model-assumptions/ (accessed 2 June 2025).

[3] NIST Engineering Statistics Handbook. (2012). http://www.itl.nist.gov/div898/handbook/

[4] Minitab software.

[5] Hinton, P. (2004). *Statistics Explained, 2nd Ed.* London, England: Routledge.

[6] Nanda, A., et al. (2021). Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. International Journal of Statistics and Applied Mathematics; 6(1): 59-65.

Biography

Patrick Valentine Technical and Lean Six Sigma Manager for Uyemura USA (uyemura.com). He holds a Doctorate Degree in Quality Systems Management from Cambridge College, a Six Sigma Master Black Belt certification from Arizona State University, and ASQ certifications as a Six Sigma Black Belt and Reliability Engineer.

uyemura.com

Corporate Headquarters:
(909) 466-5635

Tech Center:
(860) 793-4011

PValentine@uyemura.com