

Statistics and ARM Overview

- Part 1: Introduction and Data Confirmation
- Part 2: Analysis of Variance (AOV) and Data Review
- Part 3: Mean Comparison Tests

In this video, we introduce statistics for research – specifically geared towards the tools and analysis offered within ARM.

This is the first of a three-part presentation covering: analysis of variance, mean comparison tests, and the assumptions of AOV to watch out for. This lays the foundation for the various data management and analysis features that can be used in ARM.

Research Basics

- Determine if treatments of interest perform the same
 - Treatments: Products, formulations, rates
 - Perform: effectiveness, crop safety, cost
- Research to gather observations
- Are observed differences REAL or due to random CHANCE?
 - Statistics are used because we can't test *everywhere*



The goal of research is often to determine whether treatments of interest perform differently. The choice of treatments could be different products, different formulations, or different rates of application. We want to know if those treatments can be distinguished from one another, in their effectiveness, safety to the crop, or cost (sometimes the hope is that performance is indistinguishable, if your product is cheaper!)

Research trials are then performed to obtain observations of treatment performance (we call them assessments or ratings). But the question is, are the observed differences due to the treatments, or due to random chance? If we could test in every field across the world, we would have a definitive answer. Of course that is impossible, so statistics are used to provide an idea of how confident we can be in conclusions drawn from the results

You should not immediately jump to the analysis after recording data. Data Confirmation is important to catch outliers and inaccuracies with the assessment data, before analysis.

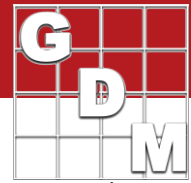
Before you analyze data...

- Analysis is dependent on data accuracy
 - Review rating **immediately**, at trial site
 - **Outliers** can exist, typos should not!
 - Decide **acceptable** range of values (before)



The best way to ensure accuracy is by reviewing the assessment immediately afterwards, while still at the trial site. This way, any irregular values can be investigated and resolved. If you wait to input and review the data back in the office, these irregularities can only be ignored in the analysis – causing a loss in statistical power!

The issues we are looking for may sometimes just be a typo from data entry. Or it could be an outlier, which is an observation outside the normal or expected distribution. But just because a value is 'unusual', does not mean it is 'not usable'! Some discernment is needed to determine if the data should be kept or thrown out.



Another tip for data accuracy: identify an acceptable range of values before the assessment is taken. Extreme values may then stand out during notetaking, cutting down on surprises at the end review. Learn about the tools in ARM for this process in the Data Confirmation tutorial.

Statistical variance is a measure of how “spread out” the data is about its mean. The goal of AOV is to dissect the variability in the recorded assessments, to understand whether the differences we have observed between treatments is due to the treatment itself, or just by random chance.


There is controlled variability that the researcher had a hand in, from the treatment and blocking structure. Then everything we cannot control is called the experimental error.

The overall approach is to consider that the observed response has three components: the “baseline” response if nothing had occurred in the trial, the effect that the treatments and blocking specifically had, and the effect of random influences too.

In order to draw conclusions about the treatment effects, we want to estimate what the random or uncontrolled effects are. To do this, we’re going to need to make some assumptions about their behavior.

Analysis of Variance

- Variance – how *spread out* the data is
- Break down observed variance into causes:
 - Controlled (treatments, blocking)
 - Uncontrolled (experimental error)
- Observed response = overall mean + controlled effects + uncontrolled effects




Assumptions of AOV

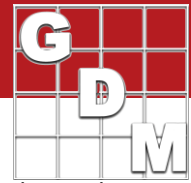
It turns out that these assumptions are quite important, because if they are not met, then the results of the analysis cannot be relied upon! So these “Assumptions of AOV” become a key focus in data review. Let’s take a closer look at them.

Assumptions of AOV

- Homogeneity of variance
 - Test: Levene’s/Bartlett’s & box-whisker
- Normal distribution
 - Test: Skewness, Kurtosis, Shapiro-Wilks
- Additivity of effects
- Independence of errors
 - Tool: Assessment heat map




The first is **homogeneity of variance**, that the variances of the different treatments are approximately equal. If a treatment is much more variable (or much less variable) than the others, the analysis will not work. So to determine if this is the case, we can look at a box-whisker graph and run a test for homogeneity of variance – Levene’s is generally chosen but Bartlett’s is also an option.



The next assumption is that the distribution of errors is **normal**, or bell-shaped. If we have this and equal variances, then the math can “do its thing” because we can calculate a reasonable guess for the amount of error based on the standard bell-shaped curve. Tests for this include Skewness, Kurtosis, and Shapiro-Wilks.

Assumptions of AOV

- Homogeneity of variance
 - Test: Levene's/Bartlett's & box-whisker
- Normal distribution
 - Test: Skewness, Kurtosis, Shapiro-Wilks
- Additivity of effects
- Independence of errors
 - Tool: Assessment heat map



The other two assumptions are harder to test for, but should still be considered especially during trial design. The first one of these is "additivity of treatment and block effects". For example, the effects of insecticide treatments are likely to be multiplicative on the population count, affecting a proportion of the insect population. So a more suitable assessment to analyze efficacy would be a “percent of untreated” calculation instead of raw counts.

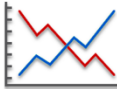
Finally, the last assumption is "independence of errors", that one rating does not directly impact the next or previous value. The key is to not introduce bias in the assessment, like edge effects or subjective ratings performed inconsistently. The assessment map may be useful to detect some of these potential issues.

What happens if these assumptions are not met? We cannot rely on the results, unless we take another action.

There are **data correction transformations**, designed to correct the underlying distribution of the data (and thus the error) without affecting the results. The Logarithm transformation is used with counts of things that occur in clusters (a log-normal distribution). The Square Root transformation is for counts of things that occur at random or rarely (following a Poisson distribution). The “Arcsine Square Root Percent” transform works with proper percentages that often follow a binomial distribution.

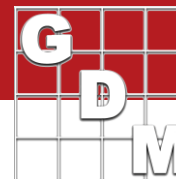
Fail to meet assumptions?

- Data correction **transformations**
 - Log, Square Root, Arcsine
- **Spatial** models (when blocking fails)
 - Trend, Nearest neighbor
- **Non-parametric** statistics
 - Rank-based instead of means



Spatial models try to recover information about hidden variables across the field. There are 3 Trend and 4 Nearest Neighbor models that can be used in ARM. This is commonly used when the blocking design fails, or when blocking is not a realistic option. **Non-parametric statistics** throw out AOV and its assumptions, and instead analyze the data by ranking the assessment values. A mean comparison test can still be run on the mean of those ranks.

Learn about the Column Diagnostic tool in ARM that is used to resolve these issues in the Data Review tutorial.



Example: 5 treatments, 4 replicates, harvested yield.

- **Degrees of Freedom (DF)** = # independent comparisons can be made
 - $DF = n - 1$
- **Sum of Squares (SS)** = total amount of variation in the data
- **Mean Square (MS)** = amount of variation + degree to which it can occur
 - $MS = SS / DF$

Source	Degrees of Freedom	Sum of Squares	Mean Square	F	Prob (F)
Between blocks (Rep)	3	305.8	101.93	0.626	0.6121
Between treatments	4	2268.8	564.7	3.466	0.0421
Error	12	1955.2	162.93		
Total	19	4519.8			

After reviewing the data, we are now ready to perform the analysis. Here is an example output of the AOV analysis, from a trial with 5 treatments and 4 reps.

Degrees of Freedom is a bit tricky to define, but is the number of values in a final calculation that are free to vary. With variance, DF equals $N-1$, so for 5 treatments you have 4 degrees of freedom. The **Sum of Squares** represents the amount of variation in the data, split up into each source listed in the table. The **Mean Square** is the Sum of Squares divided by Degrees of Freedom; we are taking the amount of variation and dividing it by the degree to which it can occur.



- **F ratio** = indicates level of variation compared to random variation
 - $F = MS / \text{Error MS}$
 - Large F \rightarrow variance may NOT be due to random chance
- **Prob(F)** = probability that variability occurred by chance
 - Low p \rightarrow confident that real differences exist

Source	Degrees of Freedom	Sum of Squares	Mean Square	F	Prob (F)
Between blocks (Rep)	3	305.8	101.93	0.626	0.6121
Between treatments	4	2268.8	564.7	3.466	0.0421
Error	12	1955.2	162.93		
Total	19	4519.8			

Then we calculate an **F ratio** which indicates the level of variation from that source, compared to random variation. The larger this ratio is, the more likely that variance from this source is NOT due to just random chance. To quantify this, we calculate a **p-value**: the probability that observed differences in treatment (or block) means could have arisen by chance.

A low p-value (typically $< 5\%$) indicates we can be confident that there are real differences among treatments. Whereas a high p-value (like between blocks) indicates that there is no evidence of a real difference (among blocks in this case). This could be because there is no real difference or that the differences are too small to detect.



A long story short...

- ANOVA breaks apart variance into sources
- Calculates a P-value (= probability)
 - $p = \%$ confidence that no differences exist
 - $p = 0.04 \rightarrow 96\%$ sure of differences
- But we are not done yet!

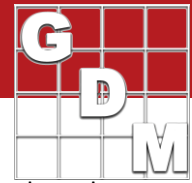


To summarize, the AOV analysis breaks apart the variance of observations into sources, then calculates a p-value as its end-result. Remember, a p-value is a probability!

The AOV p-value is the probability that no differences in treatments exist. So a p-value of .04 means we are 96% confident that there are significant differences between treatments. But, we are not done yet! We need to know which observed differences between treatments are significant, and which are not!


Mean Comparison Tests

The AOV analysis does not provide an answer about which observed differences in treatment means are significant, just whether there is evidence that a significant difference does exist. So the next step of the analysis is to run a mean comparison test.



Mean Comparison Test

- Compare treatment means for significant differences
- Different tests for different situations:
 - LSD
 - Duncan's
 - SNK
 - Tukey's HSD
- Significance level (or alpha, α)



Also known as “Multiple Comparison Test” or “Mean Separation Test”, these compare treatment means to determine what differences are significant. ARM has a variety of mean comparison tests to choose from, for different situations and research types. The most common ones are LSD, Duncan’s, Student-Newman-Keuls, and Tukey’s HSD.


We must also choose a significance level (or alpha) to run these tests. This is similar to the p-value for "confidence" in AOV results, but sets the threshold for significant differences for the test. Thus an alpha of (point)05 implies that we only accept differences that have a probability of 95% or higher of being true.

This selection should be made *prior* to running any analysis, as the choice should not be biased by desired outcomes.

Tests can be classified by how liberal or conservative they are. This means that tests can be more likely to report a false positive (Type I error) or a false negative (Type II error), respectively. The first 4 tests listed in ARM are arranged from liberal to conservative. The liberal LSD is the most likely to report differences but suffers from higher false-positive rates, while you can be most confident in differences that the conservative Tukey's HSD reports, even though separation is less likely to be found.

What to choose?

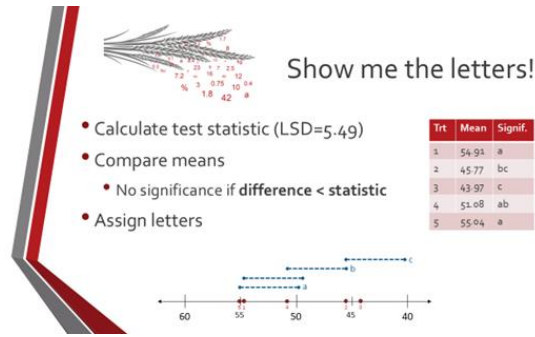
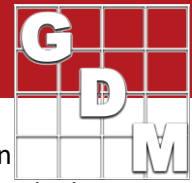
- Liberal vs. conservative
- Type I vs. Type II error
- Industry or sponsor standards
- Purpose of study
- Penalty of failure



When deciding which to use, following industry- or discipline-standards is a good starting point. Or the study sponsor may request a specific test and alpha level.

The purpose of the study should also be considered. When trial is used for registration, then a conservative test is appropriate. But a screening trial is all about potential, and so a more liberal test may be preferred, because any false-positives will be "weeded out" in future studies.

Finally, consider the penalty of failure of the treatments. For example, the failure of a crop protection product will likely have an impact on yield, and so an alpha level of 5% is common. But for a biological product for plant health, the penalty of failure is just the cost of the product and so has a much lower penalty of failure. So a 15 or 20% significance level can be appropriate.



Here's an example of how a mean comparison test works. The specifics vary for each test, but the basic concepts are the same.

First, a test statistic is calculated to perform the comparison of means. In our example we will use LSD at 5% significance level, which is 5.49 for this data.

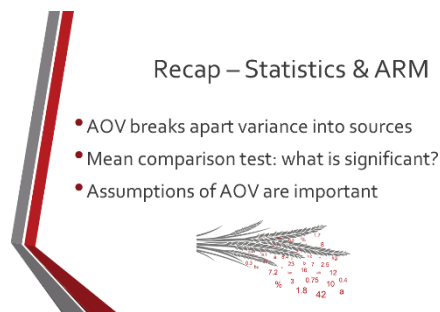
Then, compare the largest treatment mean to each of the other means, using the test statistic. If the difference between means is smaller than the LSD value, then they are not considered significantly different.

To visualize this process, we can use a number line. The LSD value creates a range around each mean, and anything outside that range is significantly different according to the mean comparison test.

Let's start with the largest mean, and extend to the lower bound. All treatments within this range are not significantly different, so we assign the letter 'a' to treatments 5, 1, and 4.

Repeat this process for each mean. Because there are no new means in this next range, we don't need a letter here. The next range does include treatment 2, and so treatments 4 and 2 are assigned letter 'b'. The next range includes treatments 2 and 3 and is assigned 'c'. There are no more means without a letter, so we are finished. The letters are typically reported to the right of the means.

Learn about the reporting tools in ARM used to perform and present this analysis, in the AOV Report tutorial.



In summary, the AOV analysis breaks apart the variance of observations into sources, then calculates a p-value probability.

Then we perform a mean comparison test to determine which treatment mean differences are significant.

But we must be cautious with this analysis – if the data does not fit the assumptions, we cannot rely on the results.