

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

We will discuss 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.

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 other, 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? This is where statistics come in, namely Analysis of Variance.

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

But how do we avoid these problems? By keeping detailed records of the site and assessment and confirming the data as soon as it is taken (both are fundamental benefits of the TDCx add-in). And by identifying an acceptable range of values before the assessment is taken. We will focus on this Data Confirmation step in another video.

Agriculture research is always subject to variation from a variety of sources, so we perform an Analysis of Variance on the data. Statistical variance is a measure of how “spread out” the data is about its mean.

Also called ANOVA or AOV, the analysis of variance dissects the variability in the recorded assessments. There is controlled variability, from the treatments and blocking structure, and the experimental error is uncontrolled. In essence, we consider “if the observed differences are due to chance, then the variance **between** treatments must agree with the variance **within** treatments”

Research Basics

- Determine if treatments of interest perform the same
 - Treatments: Products, formulations, rates
 - Perform: effectiveness, crop safety, cost
- Research to gather observations
 - aka Assessments or Ratings
- Are observed differences REAL or due to random CHANCE?



Before you analyze data...



- Rogue values/outliers can occur
- Analysis is dependent on data accuracy
 - Keep detailed records (TDCx)
 - Decide acceptable range of values (before)

Analysis of Variance

- Variability in observations is analyzed
- Variance – how *spread out* the data is
- Break down variance into causes:
 - Controlled (treatments, blocking)
 - Uncontrolled (experimental error)
- Variance **within** treatments vs. **between** treatments





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	0.628	0.6121
Between treatments	4	2258.8	564.7	3.466	0.0421
Error	12	1955.2	162.93		
Total	19	4519.8			

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	0.628	0.6121
Between treatments	4	2258.8	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

- aka Multiple Comparison or Mean Separation
- Tests treatment means for significant differences
- Different tests for different situations:
 - LSD
 - Duncan's
 - SNK
 - Tukey's HSD



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.

Which test to run?

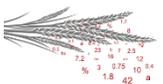
- Liberal vs. conservative
 - Type I vs. Type II error
- Penalty of failure



Tests can be classified by how liberal or conservative they are. This is not a political discussion, but rather this means that tests can be more likely to report a false positive (**Type I** error) or a false negative (**Type II** error), respectively.

Consider the penalty of failure for the study. When trial is used for registration, then conservative is appropriate. But for a screening trial, it is all about potential, and so a more liberal test may be preferred.

Show me the letters!



- Calculate test statistic (LSD=5.49)
- Compare means
 - No significance if $\text{diff} < \text{statistic}$
- Assign letters

Trt	Mean	Signif.
1	54.93	A
2	45.77	BC
3	43.97	C
4	51.08	AB
5	55.04	A

Here’s an example of how a mean comparison test works. The specifics vary for each test, 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 which is 5.49 for this data.

Then compare means using the statistic: if the two treatments do not differ by at least the LSD amount, they are not considered significantly different.

Letters are assigned such that means with no letter in common are significantly different according to the test. Thus here we can say that there is evidence that treatments 2 and 5 differ (because there is no letter in common) but there is no evidence of a difference between treatments 4 and 5.

A warning about LSD: this particular test is intended only for specific comparisons chosen before the data is collected. When used to make more than one comparison (like across a whole trial shown in this example), the true significance level becomes much larger as the number of treatments increase (because there are more pairwise comparisons being made).



Assumptions of ANOVA

One big caveat to all of this analysis, which we alluded to earlier, is that ANOVA has some requirements of the data that it is run on. If these requirements are not met, then the results of the analysis can NOT be relied upon!

Assumptions of ANOVA

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



So what are these assumptions? The first is **homogeneity of variance**, that the variances within the treatments are approximately equal. If a treatment is much more variable (or much less variable) than the others, the ANOVA 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 requirement is that the data is **normally distributed**. If your data is not bell-shaped then you can't use the same p-values in the analysis. "Its just math" as Peter our statistician would say! Tests for this include Skewness, Kurtosis, and Shapiro-Wilks.

Finally, the last two assumptions are **independence of errors** and **additivity of treatment and block effects**. These are harder to test, but generally are not an issue. The biggest thing here is to not introduce bias in the assessment, like edge effects or subjective ratings being influenced, etc. The assessment map may be useful to detect some of these things.

We will demonstrate how to review your data to test for these assumptions, in another video.

What happens if these assumptions are not met? There are data correction transformations, designed to correct the underlying distribution of the data without affecting the results.

The Log transform is useful for counts of things that occur in clusters, which results in a log-normal distribution.

The square root transformation is used on counts of things that occur at random (or rarely) and that results in a Poisson distribution.

Finally, the arcsine square root percent transformation is useful for percentage data which many times becomes a binomial distribution.

Fail to meet assumptions?

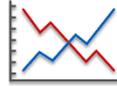
- Data correction **transformations**
 - Log, Square Root, Arcsine
- **Spatial models**
 - Trend, Nearest neighbor
- **Non-parametric statistics**
 - Rank-based instead of means





Fail to meet assumptions?

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

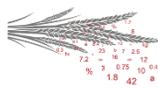


Spatial models try to recover information about hidden variables across the field. There are 3 Trend and 4 Nearest Neighbor models that can be used in ARM, but are typically only used when the blocking design fails.

Finally, non-parametric statistics throw out AOV & 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.

Recap – Statistics & ARM

- AOV breaks apart variance into sources
- Mean comparison test: what is significant?
- Assumptions of AOV are important



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.