Statistics Overview with ARM

Research is designed to answer a question. When we pose a research question, we want to know if the outcome is from:

1. The treatments (independent variables), or
2. Random chance (meaning tested treatments are probably not effective)

Many experiments seek to show equivalence of treatments, that they are not perform differently.
Inferential statistics are used to make generalizations from a sample to a population - making an estimation of the population effect. The reason for calculating an inferential statistic is to get a p value (p = probability).

For crop research, Analysis of Variance is commonly used for the initial statistical analysis to obtain a p value.

Randomized Complete Block (RCB) AOV

Source	DF	Sum of Squares	Mean Square	F	Prob(F)
Total	19	4519.80			
Replicate	3	305.80	101.933333	0.626	0.6121
Treatment	4	2258.80	564.700000	3.466	0.0421
Error	12	1955.20	162.933333		

- RCB has replicates which should be designed in such a way that the amount of variation, background variation within a replicate should be less than that which is across the entire experimental area.
- **DF** (Degrees of Freedom) - is the number of independent things you can measure. If you have 4 replicates, you have an average of the replicates and 3 other bits of information.
- **Sum of Squares** - how much variation is there between the individual experimental units (plots). The stats try to determine how much is due to the replicate, treatment, or unexplained variation.
- **Prob(F)** – the probability that the Null Hypothesis tested is true.
  The Null Hypothesis is that every treatment is the same. If we reject the Null, we are stating that there is evidence that at least one treatment is different. In the example, we are roughly 40% certain (p=.61) that a replicate might be different. Some explainable variation due to replicate affect. If we did a really good job of laying out the replicates, we would have a smaller Prob(F) which correlates to a larger percent of certainty that the replicate is explaining the variation.
  We are roughly 96% certain (p=.04) that at least one treatment is different.
- **Error** - if you can do a better job with the experimentation, you can reduce the error term and by getting a smaller Sum of Squares - we should have better precision. Really need 12 DF of error as the minimum. EPPO guidelines states this for trials requesting registration.
- Make sure that you have enough replicates to statistically distinguish an expected treatment difference.

ARM statistical tools

Power and Efficiency Planner - planning tool to design trials with sufficient treatments and replicates to detect expected differences based on chosen statistical parameters.

Randomization Quality – improve a trial's precision by optimizing the plot randomization in the field. Recommended block size, randomize all replicates, avoid edge effect, and balance treatment dispersion.

Assessment map – a 'heat map' of plot values to identify spatial differences in the assessment results.

Column Diagnostics panel – review data to find outliers and resolve failed assumptions of AOV by applying transformations, spatial models, or non-parametric statistics.

AOV Means Table report – Run comparison test on treatment means and calculate summary statistics for assessments.

Treatment mean comparisons

Type I error: **Accept** that a difference exists, when it does not (false positive).

α = Prob(Type I error)

Type II error: **Reject** that a difference exists, when there really is one (false negative).

β = Prob(Type II error)

The Power of a statistical test is its ability to avoid Type II errors. Power = 1 – β

Alpha (α) is the statistical significance level, which we default to 5%. Meaning, the treatment Prob(F) needs to be .05 or less. Basis came from Fisher in 1930's, statistician for Ag statistics : if there is a 1 in 20 chance that a difference between treatments is due to random chance, then that is a reasonable risk.

- Use the alpha level which is appropriate for what you are testing.
- Choose the alpha level based upon the severity or **penalty of failure** of a treatment which we believe is statistically better but could be due to random chance.
  - If the failure of a treatment causes a yield loss or quality loss, then should choose a more stringent alpha level - high penalty of failure.
  - BioStimulants, BioHealth, and MicroNutrients are good examples of have a low penalty of failure and thus could use a higher alpha level. Penalty would be the cost of the product/application.
    Case study: if you can claim your product gives an 80% chance that a $10 application could result in a $20 profit from increased yield, would a grower do it?

Selecting a mean comparison test

The first four mean comparisons are listed in ARM are in order by most liberal to most conservative.

- Liberal = more likely to report a false positive (Type I error)
- Conservative = more likely to report a false negative (Type II error)

The selection of a mean comparison test should occur before the data is collected (or at least without looking at the results to determine which test provides the "right" or anticipated results). When the trial is to register a product, a more conservative test would be appropriate, to be certain about the results (and avoid false positives). But when doing a screening trial, a more liberal test would be better to avoid overlooking a potential product.

- **Least Significant Difference (LSD) test**
  - Sensitive to number of comparisons. Type I error becomes very high as number of treatments increase.
  - Use for a specific pair of treatments to be compared (should identify this before the experiment)
- **Duncan's New Multiple-Range (MRT) test**
  - Powerful, low type II error. Type 1 error increases with number of treatments
  - Cannot be used with unequal samples sizes (i.e. missing data)
  - Mainly used in US.
- **Student-Newman-Keuls (SNK) test**
  - Most acceptable method as a general letter test
  - Can be used with unequal sample sizes (i.e. missing data)
  - Mainly used in Europe.
- **Tukey's Honestly Significant Difference (HSD) test**
  - Very conservative, low type 1 error
  - Can be used with unequal sample sizes (i.e. missing data)