STATISTICS FOR AG RESEARCH:

Easy Button, Cookbook, or Decision Tree?

Leslie Fuquay

NAICC 2017

Some good quotes:

"Statistics is the science of extracting information from data. It is thus through statistics that we understand the world, make better decisions, and improve the human condition."

– David Hand, British Statistician

"It is possible to conduct an investigation without statistics, but impossible to do so without subject-matter knowledge. However, by using statistical methods, convergence to a solution is speeded and a good investigator becomes an even better one."

– G.E.P. Box, J.S. Hunter, W.G. Hunter

"The best time to plan an experiment is after you've done it."

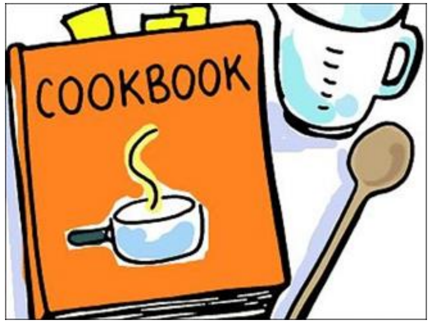
– R.A. Fisher

"It is not unusual for a well-designed experiment to analyze itself."

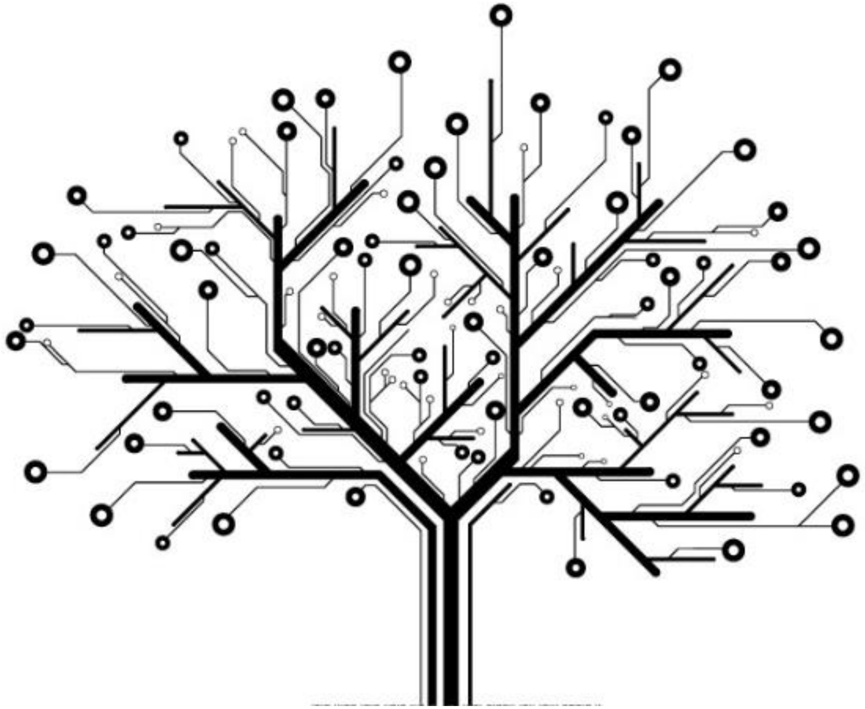
– G. E. P. Box

So . . . When it comes to analyzing your data, how many of you have one of these?

- I am not a statistician.
- I am an experienced **statistical practitioner**, with **subject matter expertise** in the biological sciences.
- If you aren't already, you can be that too.
- But not with one of these:

- And not really with one of these: 

What you really need is . . .



A comprehensive decision tree!

Key Question:

- Are you the subject matter expert, the data analyst, or both?
- Both have very important roles and neither role should be taken for granted.
- Logistics <<<<<<< — >>>>>>> Statistics
- If you are acting as both, you must ask yourself questions and challenge your answers along the way!

Experimental Planning & Design

- What are your input variables (treatment factors)?
  - 1, 2, or more?
  - How many levels of each?
  - Do any result in constraints?
- Will you compare with a control or a standard?
- Write these out in detail!

Experimental Planning & Design

- What are your output variables (responses)?
  - 1, 2, or many?
  - Avoid index scales – attempt to measure vs. estimating
- How will you evaluate them (method, precision, timing, frequency)?
- How variable do you expect them to be?
  - SME experience or prior data helps here
- What size of a difference do you need/want/expect to detect?

Experimental Planning & Design

- What question(s) do you need to answer about each response?
  - Difference? Direction of difference?
  - Magnitude of difference?
  - Equivalence?
  - Maximize?
  - Minimize?

Experimental Planning & Design

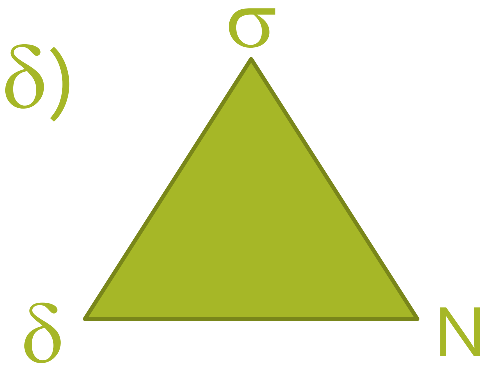
- Will experimental factors interact to affect the response(s)?
- What statistical method/model will you employ?
  - Regression
  - Analysis of variance
  - Analysis of covariance
  - A nonparametric method
- Does the trial have multiple objectives? More than one analysis may be needed.

Experimental Planning & Design

- What about replication?
- The more natural variability in experimental units, the more replicates needed.
- Will there be multiple locations (or years)?
  - As additional replication or to test outcomes in different environments?

Experimental Planning & Design

- Power – Level of certainty that that you will detect a “true” difference
- 3-way tensioning between
  1. expected variation (estimate of σ)
  2. size of meaningful difference (estimate of δ)
  3. number of experimental units (N)
- For a given alpha-level
- N is trts x reps (affects degrees of freedom)

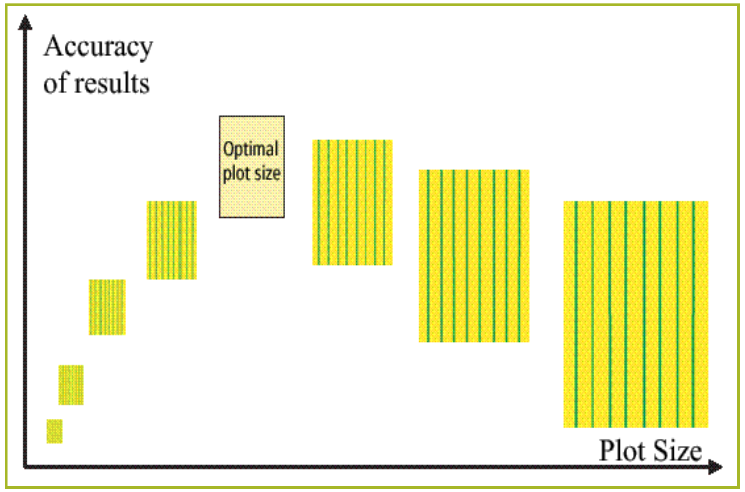


Experimental Planning & Design

- What about blocking and randomization?
- Block what you can (should) and randomize what you can't!
- Block against a known or suspected gradient. For some objectives, don't block at all.
- Randomization is your insurance against bias . . . and all experiments are subject to bias!

Experimental Planning & Design

- Experimental unit (plot) size
- Depends on the variability and sensitivity of the response(s) of interest.
- Bigger is not necessarily better – there is usually a "sweet-spot".
- It is often crop-dependent.



Experimental Planning & Design

- Size of guard-areas (between plots) and alleyways (between blocks)
  - Protect the assumption of independence of plots!
  - Avoid spray/vapor drift, competition, pest migration from plots of lesser control
- Do everything possible to assure all plots (and samples) are treated alike, except for the experimental treatment.

Experimental Error

FACT:

As agricultural researchers, we work in some of the most variable experimental conditions in science.

And we work with large, expensive, difficult to manage experimental units.

Experimental Error

- Beware variation in a response caused by something other than the experimental treatments.
- IT IS ALWAYS PRESENT!!!
- Identify as many potential sources of error as possible.
- What is the cost (vs benefit) of controlling each source?
- When might an error source actually confound the response?

Experimental Error

- Control what you can during the experiment, partition what you should during analysis, and the rest is experimental error.
- Experimental error is the ‘noise’ and it obscures the treatment ‘signal’ if it is ‘loud’.

If Background Noise **>** Treatment Difference

Then

- ➢ **fail to detect a true difference or**
- ➢ **draw an incorrect conclusion**

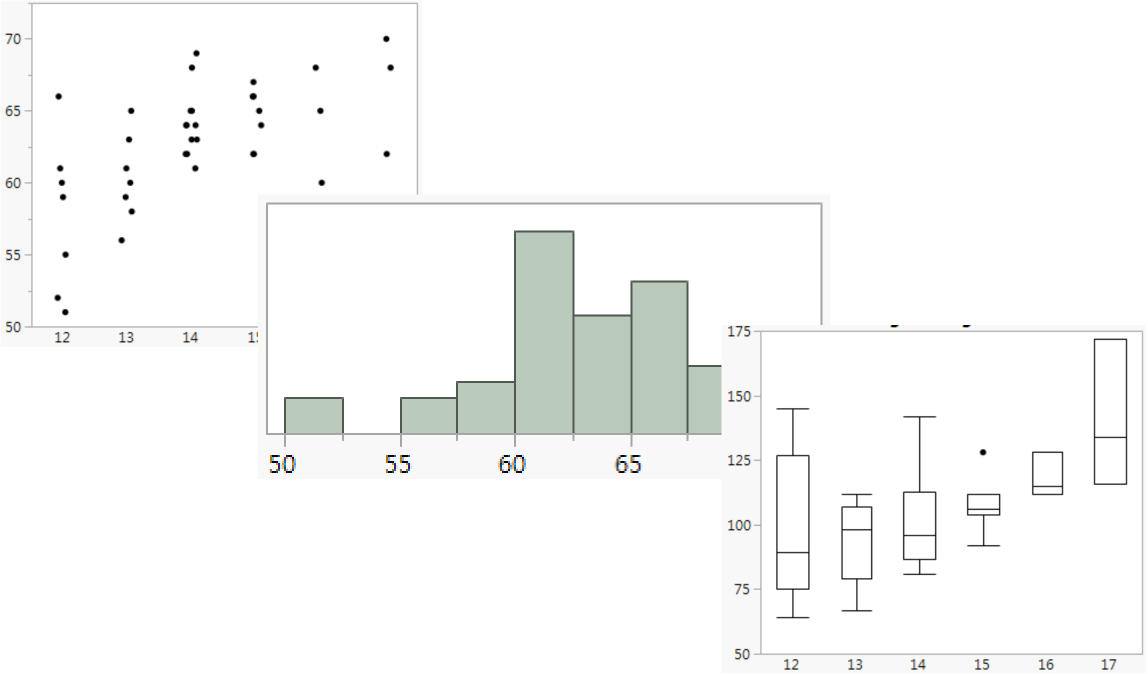
Statistical Assumptions (for most models)

- Normal distribution (of responses and errors)
- Homogeneity of error variances
- Independence of effects
- Additivity of effects

What are the potential effects of ignoring?

The Analysis – Know Thy Data!

- Graph it and check applicable assumptions
- Look for anomalies & look at the variation
- Examine it multiple ways
  - Scattergrams
  - Distribution histograms
  - Box plots



The Analysis

- ANOVA is robust to small departures from a normal distribution.
- But not to heterogeneous error variances among the treatments.
- There are formal statistical tests, but you can often see this in scattergrams of the raw data or by plotting the residuals.
- **DON'T** ignore it!
- A data transformation may improve or correct both.

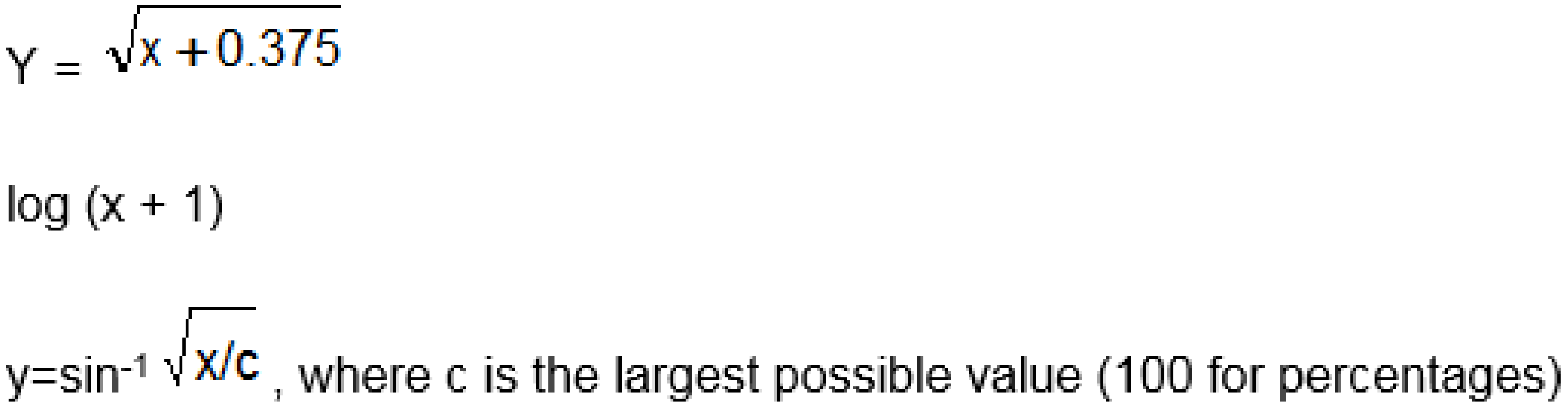
The Analysis

- Transformation rules of thumb:

Assessment	Transformation
Counts of things which tend to occur at random	Square root
Counts of things which tend to occur in colonies or clusters	Log
Proper percentages	Arcsine square root
Height, yield	Untransformed

The Analysis

- Common transformation equations with constants to avoid issues with 0-values in your data



$$Y = \sqrt{x + 0.375}$$

$$\log (x + 1)$$

$y = \sin^{-1} \sqrt{x/c}$, where c is the largest possible value (100 for percentages)

The Analysis

- Usually best practice to leave all treatments in the analysis for a robust estimate of variation.
- Consider eliminating treatments that result in no variation.
  - All 100% (pest control)
  - All 0% (crop injury)
- Do not contribute to variance estimation.
- Lead to heterogeneity of variance that no transform can fix.
- Make it difficult to find "true" difference among other treatments.

The Analysis – What about “outliers”?

- Should be evident in pre-analysis graphs, but you can also examine the residuals.
- My “rule of thumb” is >3 SD within a location & >4 SD across multiple locations . . . but it depends.
- Initiate an investigation: Is it real?
  - A ‘real’ mistake? Correct or replace with missing value.
  - Perhaps it is a valid response in the tail of the distribution.

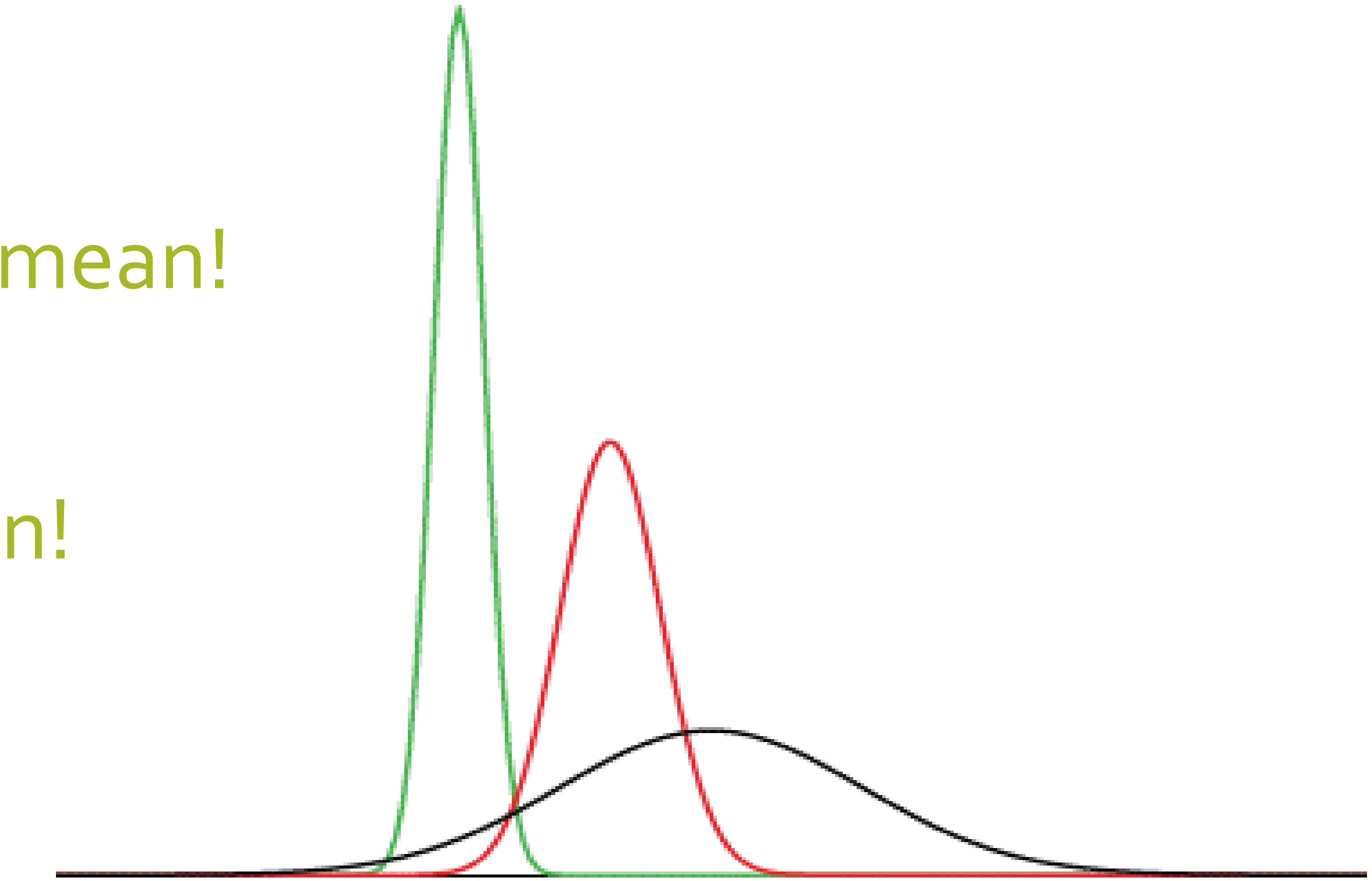
Analysis Watch Outs!

- Avoid confusing correlation with causation.
- Just because two treatments are not significantly different does not mean that they are “equal”.
- The ‘p’ in p-value is for **probability**. Statistical outcomes are inferences based on probability.
- That means IT DEPENDS! On a lot!

Watch Outs!

- It's not all about the mean!

- It's the variation, Vern!



Watch Outs!

- Assuming you are pushing the right buttons or have typed the correct code, most statistical computing software will give you answers!
- Just because it runs doesn’t mean it’s done.
- It is up to you to know if the answer is correct or even real.
- And . . . is it biologically or economically relevant!

There is no easy button, cookbook, or even computer software that can answer most of these questions!

You need good SMEs, well-thought-out experimental techniques, and decision-based data analysis.

Thanks!

leslie.fuquay@syngenta.com