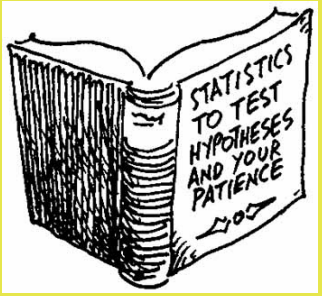
Statistical issues in horticulture: common issues and some fixes

Matt Kramer, USDA Agricultural Research Service
August 2020, matt.kramer@usda.gov



J. Amer. Soc. Hort. Sci. 141(5):400–406. 2016. doi: 10.21273/JASHS03747-16

Statistics in a Horticultural Journal: Problems and Solutions

Matthew H. Kramer[1]
U.S. Department of Agriculture, Agricultural Research Service, Statistics Group, Building 005, Room 130, 10300 Baltimore Avenue, Beltsville, MD 20705

Ellen T. Paparozzi
Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583-0724

Walter W. Stroup
Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583-0963

HortScience 54(9):1605–1609. 2019. https://doi.org/10.21273/HORTSCI13952-19

Best Practices for Presenting Statistical Information in a Research Article

Matthew H. Kramer[1]
U.S. Department of Agriculture, Agricultural Research Service, Statistics Group, Building 003, Room 330, 10300 Baltimore Avenue, Beltsville, MD 20705

Ellen T. Paparozzi
Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68583-0724

Walter W. Stroup
Department of Statistics, University of Nebraska, Lincoln, NE 68583-0963

86 articles from JASHS (all of 2014 and 1st issue of 2015) were examined for statistical issues.

http://influentialpoints.com---gives results for use and misuse of statistics in biology (neuroscience emphasis)

Review of 513 biological and medical articles in 5 top-ranking journals, 78 used the correct procedure, 79 an incorrect procedure (assume that the rest could not be judged?)

In an article aimed at biologists, the main message should be that the observed ‘effect’ is biologically, economically, or scientifically consequential, not that a *P* value is statistically significant.

“Most people use statistics like a drunk man uses a lamppost; more for support than illumination”
— Andrew Lang

The purpose of the statistics section in a research paper

(1) The design, data collection, method of analysis, and software used must be described with sufficient clarity to demonstrate that the study is capable of addressing the primary objectives of the research; the statistical analysis must be reproducible.

(2) Authors must provide sufficient documentation to create confidence that the data have been analyzed appropriately.

(3) Data and their analyses must be presented coherently.

(4) Readers should not have to guess which scientific questions the analysis answers. Effects which are statistically significant must be biologically important.

(5) Readers should be able to use information in the statistical reporting section as a resource for planning future experiments.

Summary of statistical problems

Problem	Count
Need experiment-wise control/multiple dependent variables	30
Incorrect analysis	24
Means separation	20
Missing information	10
Miscellaneous	8

Multiple Dependent Variables

Problem:

When each plant is measured on several characteristics, the measures are correlated through the plants. However, each characteristic is analyzed as if it were measured on an independent group of plants, with significance set at $\alpha = 0.05$. Experiment-wise error rate is not controlled.

Solution:

Control experiment-wise error rate to account for the correlation, e.g. by adjusting *p* values (one method to do this is by using FDR, false discovery rate). This allows different characteristics to be analyzed in different ways (e.g. some assuming a normal distribution, some assuming a binomial distribution).

Give the correlations of the dependent variables.

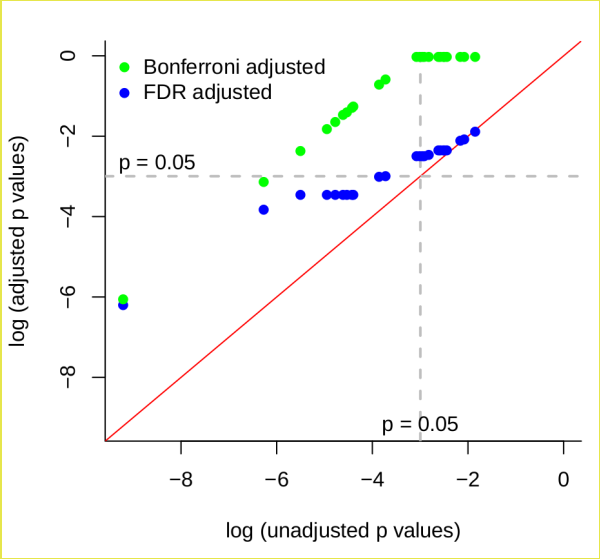
Example

You have analyzed 8 plant characteristics from each plant in a CRD with two factors, variety and treatment, each factor has 3 levels (3 varieties, 3 treatments).

Six of those characteristics are assumed to be normally distributed (some had to be transformed) but one is a count variable (number of leaves) and one is binomial (proportion of ripe fruit). The count variable was analyzed assuming a negative binomial distribution. The binomial variable was analyzed assuming a quasi-binomial distribution (binomial, but allowing for over-dispersion).

There are 8 x 3 (2 main effects + interaction) *p* values that need adjusting. If one also uses multiple comparisons to look at treatment combination means, those, too, should be adjusted.

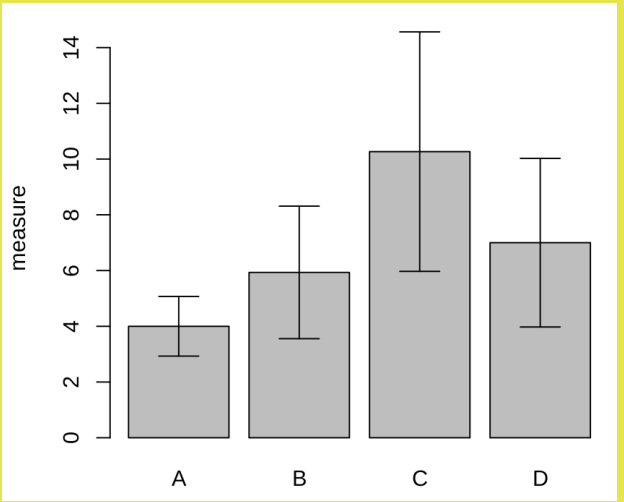
Results for example. Unadjusted *p* values on x axis and adjusted *p* values on y axis, both on log scale. Number of significant (at 0.05) *p* values: unadjusted: 13, FDR adjusted: 9, Bonferroni adjusted: 2.



Incorrect analysis

Problem	Count
Variance a function of mean	11
Random effect treated as fixed or ignored	7
Ignored spatial variability	1
Repeated measures ignored	1
Wrong repeated measures covariance structure	1
Pooled different treatments	1
Ignored censoring	1
Regression with 3 observations	1

Variance is a function of the mean



Problem:

ANOVA assumption is that the variance is approximately the same for all treatment combinations (technically, the variances are samples from the same chi-square distribution).

Solutions:

OK solution: Transform the data so that the variances are independent of the means. The Box-Cox family of power transformations are a good starting point. Proportions (percents) can usually be transformed as logits or probits.

Better solution: Use a statistical model based on the appropriate sampling distribution (generalized linear models framework), e.g. negative binomial for count data (says that data are samples from an over-dispersed Poisson distribution).

Random effect not treated correctly

Whether an effect is treated as fixed or random can have large consequences on hypotheses tests (and on inferences).

Random effects allow for a broader inference space because you are saying that the levels of the random factor is a random sample from some larger population of levels. So, your inference space is to the entire population of levels, so all blocks that might have been used in your experiment, or all greenhouses that might have been used.

You 'pay' for this with larger standard errors on fixed effect means and more conservative *p* values.

Some effects are clearly fixed (e.g. treatments), some are clearly random (e.g. blocks), but for others there may not be a clear categorization. Also, if there are just a few levels of the random effect, you have to ask yourself if you are really capturing the representative variability in that random effect.

Means separation

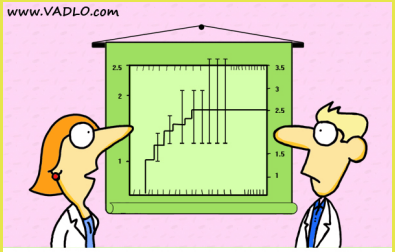
Problem	Count
Duncan's used for means separation	8
Undisclosed means separation technique	5
No adjustment for multiple comparisons (e.g. used *t*-tests)	4
Means comparisons without prior ANOVA	2
Used non-overlapping confidence intervals as means comparison	1

Missing information

Problem	Count
Missing necessary statistical information	7
Not clear what stat. software was used for	1
Undisclosed tests	1
PCA results not explained well	1

Miscellaneous

Problem	Count
Sample sizes not given	3
Measure of variability not reported	2
Stepwise variable selection with proc mixed	1
Show just fitted curves	1
Figure issues	1

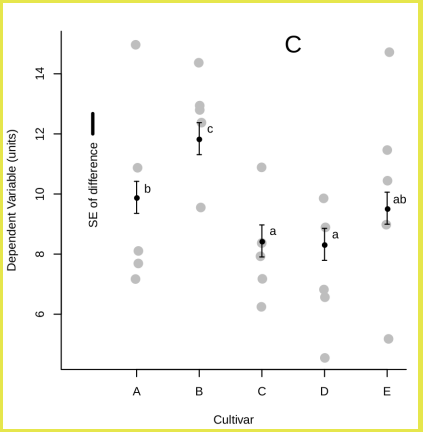


"Did you really have to show the error bars?"

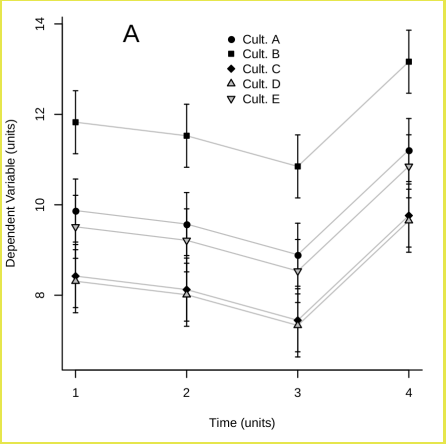
Figures

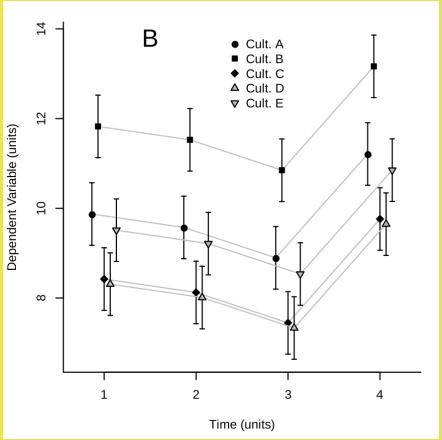
(1) Individual means with their standard errors are not that useful if you are interested in comparing means, in which case you are interested in the **difference of the means and their standard errors**. There is not yet an established graphic for depicting that, it might be better presented as a table/matrix, with the upper right triangle giving mean differences, their standard errors, and grouping letters. Along the diagonal are the means and their standard errors.

	A	B	C	D
A	3.1 (0.7)	-0.1 (1.3)	-0.3 (1.1)	0.3 (1.1)
B		3.2 (0.9)	0.2 (1.1)	0.4 (1.2)
C			3.4 (0.3)	0.6 (0.9)
D				2.8 (0.5)



(2) Do not overlap standard error bars in figures. Separate groups with a little horizontal space.





(3) In figures, make clear what are data and what are model results.

Data are the observations (but may also include summary statistics, such as treatment combination means and standard deviations).

Model results include model means (expected marginal means or least squares means), standard errors coming from a model fit, other model parameters, regression lines.

If possible, put data points in gray in the background when showing model results, so that the reader can visually gauge how well the model fits the data.

Resources available to biologists

A large and diverse number of statistical books aimed at biologists (Amazon in 2016 brought up 3,785 results for “statistics biology”)

Different emphases, but many have some material on mixed models, means separation, and experiment-wise control---common issues in horticultural science

Why are these mistakes being made?

Is there a problem with how statistics is taught (in general)? Is this kind of statistical material not taught or emphasized?

Is it forgotten/ignored by the time biologists become researchers? Many times researchers simply copy what others are doing in their field.

Where is the balance between what a biologist should know and knowing when it is time to consult with a statistician?

“If your experiment needs a statistician, you need a better experiment.”
— Ernest Rutherford

The End

Thanks for listening!

Matt Kramer
matt.kramer@usda.gov