Kansas State University Libraries

New Prairie Press

THOU SHALL NOT BRUSH YOUR TEETH WHILE EATING BREAKFAST: A 7- STEP PROGRAM FOR RESEARCHERS PREVIOUSLY HURT IN DATA ANALYSIS

Edzard van Santen

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

Recommended Citation

THOU SHALL NOT BRUSH YOUR TEETH WHILE EATING BREAKFAST: A 7-STEP PROGRAM FOR RESEARCHERS PREVIOUSLY HURT IN DATA ANALYSIS

Edzard van Santen

Edzard van Santen, Ph.D., Dept. of Agronomy and Soils, Auburn University, AL 36849-6248. USA. Email: vanedza@auburn.edu

ABSTRACT

After years of providing statistical advice to fellow faculty members and graduate students, I have come to realize that it is not necessarily the big issues, but lack of knowledge of basic data analysis principles that get my clients into trouble. My claim is that if researchers and students internalized two basic definitions they would not have any problems analyzing most of their experiments. The definitions of Experimental Unit (EU) as the smallest physical unit to which a treatment may be applied and Experimental Error (Exp. Err.) as the variation among EUs treated alike are the basis for successful data analysis of experiments. I follow a seven-step data analysis program for my graduate student and faculty clients: (1) Understanding the experiment; (2) Checking the data; (3) Getting a feel for the data; (4) Checking underlying assumptions; (5) Testing; (6) Estimating; and (7) Interpreting results. Clients who have adhered to the program generally have had fewer problems than clients who, for some reason or another, did not get on board of the program. I will also touch on the implications for teaching experimental design and data analysis to non-statistics majors.

The relationship between client and consulting statistician can be a complicated one. Issues range from client's ignorance about basic design principles to failure to communicate. Like in any human relationship, big problems can be solved with some effort but it is the accumulation of little annoyances that lead to a breakdown of the collaborative effort. Over the last 15 years I have developed a protocol for dealing with graduate student and faculty clients. This seven-step program consists of the following elements:

1. Understand the experiment
2. Check the data
3. Get a feel for the data
4. Check underlying assumptions
5. Testing, i.e. variance analysis
6. Estimation, i.e. establish treatment differences
7. Interpretation

Such an approach may appear trivial to many readers but personal experience suggests that it is not trivial. Skipping to step 6 or 7, which is the real interest of the researcher, more often than not creates a host of problems than can be avoided by following a systematic approach. In fact, once the first four steps have been navigated successfully, the remaining steps are often anticlimactic. While I will be using SAS in my examples because of familiarity and SAS' traditional ties to agronomy and related fields, software implementation can also be found in R (http://cran.r-project.org/) and Genstat (http://www.vsni.co.uk/software/genstat/).
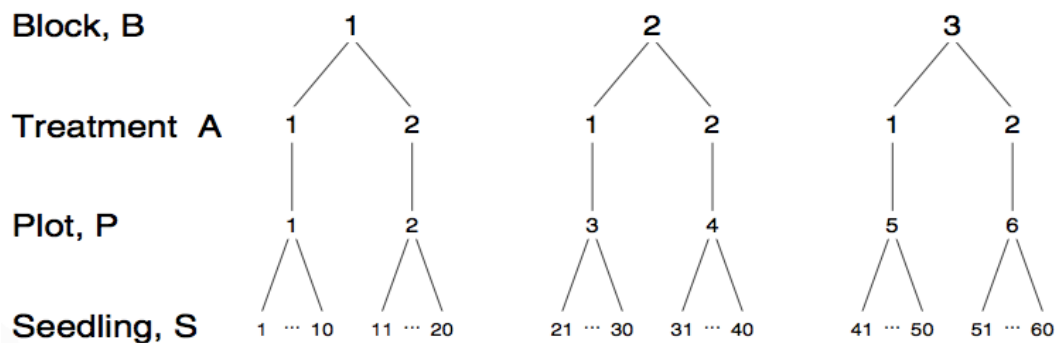

Fig. 1 Unit structure for a randomized block design with subsampling (from Bergerud, 1996b)

Understanding one's own or some else's experiment is critically important. The most intuitive approach I have found is the Factor Relationship Diagram described by(Bergerud, 1996a; Bergerud, 1996b). It will lead almost effortlessly into the appropriate linear model and programming statements if done thoroughly and correctly. Other approaches to describe an experiment such as Nelder'sshorthand operators (dot, product term, nesting, crossing), demonstrated by (Piepho et al., 2003), by themselves work well for a consulting statistician with the necessary "technical" know-how, but are less useful to the researcher, who simply wants to analyze the data rather than completing a statistics course. Graduate students have found it to be helpful to complete the Materials and Methods section of their thesis/dissertation at this stage.

The assumption that the data have been entered correctly is seldom warranted. The need for caution as a necessary requirement was emphasized by (Portman and Keata, 1997) and by Fox et al. (1997). In fact Cochran(1947) in Table 1 of his publication provided an early example of the importance of checking the data, linking the error to lack of attention to detail by the researcher. But as Cochran(1947) continued, "Habitual rejection of outlying values leads to a marked underestimation of errors." Thus a common practice used by nematologists to automatically exclude the highest and lowest observation is of questionable value. President Reagan's adage of dealing with the opposing superpower comes to mind, "Trust but verify." In fact, it usually takes me no longer than a minute to spot a likely error in the data submitted for analysis using my favorite tool,PROC TABULATE(van Santen, 2010).

Numerous authors have encouraged what I call "**getting a feel for the data**," e.g.Yandell (1997). This is not to be confused with "data-snooping," the practice of letting the data determine the type of treatment comparisons to be conducted. The intent is to look at the plausibility of the data and, in conjunction with step 2, eliminate errors in the dataset. This preliminary assessment can also give the analyst an idea of the distributional properties of the data set. This step naturally transitions the researcher into the next, **checking the assumptions** underlying the analysis, viz.

normality, homogeneity of variances, and independence.The normality assumption is that all observations (Y_{ij}) in an experiment have a joint multivariate normal distribution (Eisenhart, 1947). As Eisenhart (1947) pointed out, these assumptions are only needed when inferring population behavior from a sample; if ANOVA is used solely to summarize data no assumptions are needed.An example of simply summarizing data would be the cultivar performance trials conducted by many Agricultural Experiment Stations as a service to farmers in their state.

Non-homogeneity of error variances will lead to a loss of efficiency (Cochran, 1947), yet agronomists for the last six decades have continued to blithely ignore this problem, even though the theoretical basis for addressing this problem has existed since Cochran's article was published. Ever since PROC MIXED became widely available, enabling the analyst to subdivide the error variance into homogeneous groups as suggested by Cochran (1947), there really is no reason to ignore this issue, as the computational difficulties mentioned by the author have all but disappeared.Cochran furthermore mentioned a common form of heterogeneity of variances observed in situations where the treatment variance is some function of the treatment means, as would be the case for counts following a Poisson distribution, where the variance equals the mean. These situations were originally addressed through empirically derived transformations such as inBartlett (1947) or Box and Cox (1964). More recently, generalized linear models and generalized linear mixed models based procedures (as implemented in PROC LOGISTIC, PROC GENMOD, and PROC GLIMMIX) have enabled researchers to analyze response data in an appropriate manner. However, there are quite a number of unresolved issues with the more complex models.

It is interesting to note in passing thatBox and Cox (1964) indicated that the last assumption, independence, cannot be addressed or "fixed" by transformation. Cochran (1947) also pointed out over 60 years ago that the error structure in real field experiments would likely be more intricate than could be captured in models, bringing to mind Box's contention that "all

models are wrong but some are useful"(Box and Draper, 1987). The independence assumption comes into play in repeated measures situations, where multiple observations are taken on the same experimental unit. Older standard textbooks such asSteel et al. (1997)often treata repeated measures design as a variant of the split-plot restriction on randomization, called a split-plot in time, noting that a researcher would do well by seeking advice from a professional. The problem with the split-plot approach is that it doesn't reflect reality, because it assumes homogeneity of variances among time points and zero correlation among successive time points; neither one is a realistic assumption. Classic examples are experiments with perennial forage, fruit, or nut crops. A similar situation exists in soil science when sampling by depth, where neither assumption is valid. The situation in soil science is somewhat more complicated, as the traditional approach to divide a soil core by uneven depth-increments precludes certain "canned" covariance models that rely on equidistant points.

Only after steps 1-4 have been dealt with in a satisfactory manner does step 5,**hypothesis testing**, make any sense. "**Analyze them as you randomized them**" has long been a battle cry of consulting biometricians. While the origins of this expression are lost in the fog of history, the basic approach still holds true. Randomization is one of the key elements of statistical design and analysis, providing the justification for proper statistical analysis. The other basic element is proper replication of treatments. My claim is that if researchers internalized two basic definitions they would not have any problems analyzing most of their experiments. The definitions of Experimental Unit (EU) as the smallest physical unit to which a treatment may be applied and Experimental Error (Exp. Err.) as the variation among EUs treated alike are the basis for successful data analysis of combined experiments, e.g., an agronomic experiment conducted at multiple locations and/or multiple years. Some researchers consider treatment structure as an inconvenience during data analysis. They use a treatment structure in the design phase because they learned in graduate school that it is an important tool in designing an experiment, but forget

and/or ignore that the treatment design affects the analysis. An example would be a factorial treatment design, where the factorial structure is ignored during analysis and a single LSD value is deemed to suffice to compare all treatment combinations. It is the responsibility of the consulting statistician to "persuade" the client to see the opportunity and advantages of a proper analysis in factorial treatment designs. The analysis of augmented factorials (Lentner and Bishop, 1993) in particular, often confuses researchers. This need not be viewed as a difficulty, because linear contrasts allow us to handle such situations very efficiently. A recent publication offers some insights on how to evaluate and deal with interaction (Crossa et al., 2013).

However, in most cases, the goal is not hypothesis testing but the **estimation of treatment differences** (step 6). Many researchers confuse the two and get bogged down in testing when they really should place more emphasis on estimation(Pearce, 1988). In the regression example with three rates,Pearce demonstrates that the response for the intermediate rate has to be outside the range (30% larger than the range) of the response for the lowest and highest rates in order for the deviation from linearity contrast to be significant. Reading that article and working through the two examples given would be a worthwhile and eye-opening exercise for any researcher. In an ideal world, researchers would have clearly stated research objectives and customized hypotheses that could be addressed through appropriate linear combination of means. But more often than not, estimation of multiple treatment differences from the same body of data are involved, leading to an increased Type I error rate. There is plenty of advice in the literature on how to deal with such situations(Carmer et al., 1989; Carmer and Swanson, 1971; Milliken and Johnson, 2009; Saville, 2013; Stroup, 2013), but little general consensus beyond the realization that some kind of protocol or procedure is needed to deal with such situations.Openness is the key; tell the reviewer/reader about the procedures that have been used If the reviewer is not convinced, he/she has the opportunity to challenge assumptions or procedures.

Lastly, the researcher needs to remember that **interpretation** of trial data is his/her responsibility and absolute prerogative. An experiment is not complete without a thorough and thoughtful interpretation. The exhortation "Treasure your Exceptions" has been attributed to William Bateson (1861-1926), one of the leading geneticists of the early 20th century. Researchers would be well advised to take that advice to heart when interpreting their own research results, keeping in mind that "spectacular breakthroughs" in agronomic research may well be the result of errors caused by inadequate data checking.

This brings up the question about what statistics courses graduate students in agriculture should take and what content should be offered. Relevancy is an important consideration. Using the famous (infamous) example of teacher evaluations to convey the idea of hierarchical (nested) models might work for students in rural sociology or economics but doesn't really pique the interest of a student in agronomy, crop science, or soil science. However, there is value in using data from outside the immediate area of interest of the students, because it forces students to deal with the issues rather than relying on familiarity and rote. Wright (2013) compiled data from uniformity trials, animal trials, tree data and horticultural and field crops into a single source, which is an excellent teaching and training tool. I would argue that non-statistics majors need a 2-semester sequence, preferably early in their graduate student career, where the first course would concentrate on the fundamentals such as probability, population *versus* sampling concepts, sampling distributions, etc. The choice of textbook is highly personal and the absolute prerogative of the instructor. I believe the book by Whitlock and Schluter, (2008) deserves attention for the fundamentals. It offers mostly single-topic chapters with numerous examples and problems as well as superb graphics.

The second course should be in experiment design and analysis. It is in my strong opinion that design and analysis be linked; students will need to learn that design has consequences. Stroup's "what would Fisher do" approach is one way to accomplish this (Stroup,

2013). It is not necessary that students have a working knowledge of all possible designs.If they get the "big picture" (Stroup, 2013) combined with basic designs such as CRD, RCB, Latin square, and split-plot and split-block restrictions on randomization, they should have the basic tools to analyze most experiments. It is critical, however, to teach design and analysis as a creative process(Casler, 2013) rather than a set of rules that should be blindly followed. While I use well-worn canned examples such as are found in Steel et al. (1997) I much prefer real life data that have inherent problems. The old Roman motto 'per aspera ad astra' exemplifies my teaching approach.

Finally, everyone teaching statistics and experimental design is forced to confront the issue that time is the most limiting resource, particularly as universities begin to penalize graduate students for taking too many course credits. Beloved old content may have to make room so information on more relevant design and analysis topics can be presented. The goal is to train non-statistics majors so they can handle most of their data analysis projects but also teach them to recognize their limits. Once they approach a limit we want them not to be afraid or embarrassed to seek out external expertise. What we absolutely do not want are students who analyze their own data in a manner consistent with their advisors' statistical training rather than in a manner consistent with current statistical thinking.

LITERATURE CITED

Bartlett M.S. (1947) The Use of Transformations. Biometrics 3:39-52.

Bergerud W.A. (1996a) Displaying Factor Relationships in Experiments. The American Statistician 50:228-233. DOI: 10.1080/00031305.1996.10474385.

Bergerud W.A. (1996b) Displaying Factor Relationships in Experiments, Ministry of Forests and Range. pp. 1-12.

Box G.E.P., Cox D.R. (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26:211-252.

Box G.E.P., Draper N.R. (1987) Empirical Model-Building and Response Surfaces, Wiley.

Carmer S.G., Nyquist W.E., Walker W.M. (1989) Least Significant Differences for Combined Analyses of Experiments With Two- or Three-factor Treatment Designs. Agron. J. 81:665-672. DOI: 10.2134/agronj1989.00021962008100040021x.

Carmer S.G., Swanson M.R. (1971) Detection of Differences Between Means: A Monte Carlo Study of Five Pairwise Multiple Comparison Procedures1. Agron. J. 63:940-945. DOI: 10.2134/agronj1971.00021962006300060036x.

Casler M.D. (2013) Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments. Agron. J. 0:-. DOI: 10.2134/agronj2013.0114.

Cochran W.G. (1947) Some Consequences When the Assumptions for the Analysis of Variance are not Satisfied. Biometrics 3:22-38.

Crossa J., Vargas M., Cossani C.M., Alvarado G., Burgueño J., Mathews K.L., Reynolds M.P. (2013) Evaluation and Interpretation of Interactions. Agron. J. 0:-. DOI: 10.2134/agronj2012.0491.

Eisenhart C. (1947) The Assumptions Underlying the Analysis of Variance. Biometrics 3:1-21.

Fox P.N., Mead R., Talbot, M, Corbett J.D. (1997) Data Management and Validation, in: R. A. Kempton, et al. (Eds.), Statistical Methods for Plant Variety Evaluation, Chapman & Hall, London. pp. 19-39.

Lentner M., Bishop T. (1993) Experimental design and analysis. 2nd ed. Valley Book Company,, Blacksburg, VA.

Milliken G.A., Johnson D.E. (2009) Analysis of messy data. 2nd ed. CRC Press, Boca Raton.

Pearce S.C. (1988) Analysis of data from agricultural experiments. Tropical Agriculture (Trinidad) 65:2-5.

Piepho H.P., Büchse A., Emrich K. (2003) A Hitchhiker's Guide to Mixed Models for Randomized Experiments. Journal of Agronomy and Crop Science 189:310-322. DOI: 10.1046/j.1439-037X.2003.00049.x.

Portman P., Keata H. (1997) Field Plot Technique, in: R. A. Kempton, et al. (Eds.), Statistical Methods for Plant Variety Evaluation, Chapman & Hall, London. pp. 9-18.

Saville D.J. (2013) Multiple Comparison Procedures—Cutting the Gordian Knot. Agron. J. 0:-. DOI: 10.2134/agronj2012.0394.

Steel R.G.D., Torrie J.H., Dickey D.A. (1997) Principles and Procedures of Statistics : a Biometrical Approach, Mcgraw-Hill Series in Probability and Statistics; Variation: Mcgraw-Hill Series in Probability and Statistics, McGraw-Hill.

Stroup W.W. (2013) Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press, Boca Raton, FLorida.

van Santen E. (2010) Data checking with SAS PROC TABULATE. Communications in Biometry and Crop Science 5:1-3.

Whitlock M., Schluter D. (2008) The Analysis of Biological Data Roberts and Company Publishers, Greenwood Village, Colorado.

Wright K. (2013) agridat: Agricultural datasets. R package (http://cran.r-project.org/web/packages/agridat/index.html).

Yandell B.S. (1997) Practical Data Analysis for Designed Experiments. 1st ed. Chapman & Hall, London [etc.].