Ten common misuses of statistics in agronomic research and reporting

Larry A. Nelson and John O. Rawlings

ABSTRACT

Ten common misuses of statistics in agronomic research and reporting are discussed. Some of these are a result of changes in statistical philosophy over the years to which biologists in general, and agronomists in particular, have not responded in terms of their data analytic and interpretational techniques. Others have been created by an overdependence on computers without careful study of the basic data patterns or without careful consideration of the calculations which the computer is performing. The importance of planning experiments properly with a view toward subsequent analysis is emphasized. Careful, well-controlled experimental technique is also recommended. Proper planning usually assures that logical comparisons can be made without resorting to the use of mechanical procedures such as multiple comparisons. The matter of misuse of multiple comparison procedures such as Duncan's Multiple Range Test is also discussed. It is pointed out that in cases where logical structure doesn't exist in the treatments (a rare event), the use of multiple comparison procedures is valid.

Additional index words: Statistics, Research, Accuracy.

MODERN applied statistics has been utilized extensively in agronomic research in the United States and elsewhere for the past 40 years. During this period, some of our concepts have changed. For example, there now is much less emphasis on hypothesis testing and far more emphasis on estimation of the magnitudes of differences and other effects. No longer are we expecting an experiment to provide the last word based upon the result of some mechanical process such as a hypothesis test at a stated level of significance. Now we are looking for indications of effects and we rely on the data to provide us estimates of the magnitudes of these effects. There has also been some revision in our thinking about some concepts as a result of extensive application of statistical techniques to biological problems. For example, formerly we recommended the use of large plots in field experiments because the variance of large plots is small. Now we recommend the use of small plots with a compensating increase in number of replications to use the available resources.

Misuse I—Failing to Involve Statistical Considerations at the Planning Stage of the Experiment

Statistical considerations should be involved at the conceptual stage of the experiment. This perhaps is secondary only to the need for a good researchable idea. The reason for involving statistical principles early is to insure that one obtains quality data which bear upon the problem being studied. There are sampling considerations with respect to the populations to which the results of the experiment will be extrapolated; the populations of environmental conditions, of experimental material, and of treatments. There are experimental design considerations such as which experimental design should be used, what treatments should be studied, and how many replications are needed. There are a number of practical considerations relating to the orientation, size, and shape of plots and blocks. Involvement of a statistician

The ready availability of efficient electronic computers has had both positive and negative effects. On the positive side, we can process data very efficiently and accurately at low cost. There are analyses which can be run which would have been impossible before the advent of modern computing techniques. On the other hand, ready access to statistical packages by researchers with limited statistical background has increased the misuse of statistical procedures.

Today there is abundant evidence of the misuse of statistics by research agronomists. A current issue of any one of the plant science journals will provide ample cases in point. *Agronomy Journal* is attempting to improve the situation but authors are asking for more assistance in deciding which statistical techniques they should use and how they should be used.

The purpose of this paper is to point out and discuss 10 common misuses of statistics in agronomic applications. Some of these are more a lack of use of statistics rather than misuses.

**MISUSES IN PLANNING EXPERIMENTS**

**Fig. 1. Example of an aerial spraying experiment for plant disease control in a South American country.**

{\textsuperscript{1}}Contribution from the Inst. of Statistics, North Carolina State Univ., Raleigh, NC 27650.

{\textsuperscript{2}}Professors of statistics, North Carolina State Univ., Raleigh, NC 27650.
at an early stage in the planning of an experiment, particularly if the researcher is not well trained in statistics, is helpful in focusing on the specific questions to be answered and the relevant statistical methods for estimation and/or hypothesis testing. He also can provide assistance in choosing treatments in such a way that the treatment comparisons can be made efficiently at the data analysis stage.

Misuse 2—Using an Improper Experimental Design or Misusing a Proper Design

The most popular design in agronomic research is the randomized complete block design. It is a simple design and is reasonably efficient in most cases if the blocking has been constructed appropriately. The second most popular design is some version of the split-plot design. There are many situations in which a split-plot design is used where clearly a randomized complete block design with factorial arrangement of treatments would be preferred. In cases where there are only a few levels of the whole-plot factor or there are few replications, the whole-plot error in the split-plot analysis of variance will be estimated with low precision and there will not be a good test on the whole-plot factor. A factorial arrangement of treatments within the framework of a randomized complete block design will provide equal (and adequate) precision for all effects, i.e., both sets of main effects and interactions.

There are many instances in which a split-plot design arises out of an inappropriate handling of what was intended to be a randomized complete block design. At some point in the conduct of the experiment, certain subsets of the treatments are handled in groups such as in data collection, exposing to treatment factors, harvesting by maturity, etc. Such nonrandom handling of the experimental units may introduce positive interplot correlations among the units handled as groups, generating a “split-plot” design. Failure to recognize this can lead to inappropriate analyses and erroneous experimental error variance estimates.

A proper experimental design can be destroyed by failing to recognize what constitutes the experimental unit. For example, an individual may try to provide replication by subdividing larger plots to which treatments have been applied (Fig. 1). In an aerial spraying experiment for disease control, three chemicals were applied in long strips and then the strips were subdivided to provide what the researcher considered were replications. Figure 1 is not a randomized complete block design which the researcher had intended to use. Each strip is one experimental unit and the subdivisions are samples, not replicates. A randomized complete block design (Fig. 2) was then provided the researcher as an appropriate alternative. In this case there are nine experimental units in three blocks of three each. Randomization was carried out as required in the randomized complete block design and the aerial spraying was performed according to this revised plan. The danger that chemical effect estimates would be confounded with

Misuse 3—Failing to Use Proper Randomization Procedures

Randomization is used to insure that we will have unbiased estimates of treatment effects and experimental error. Failure to use proper randomization techniques could cause certain treatments to be favored or hampered due to the position in which they are placed in the experimental area and cause differences in degrees of precision for different comparisons. In our extensive consulting with plant science researchers during the past 20 years, we have encountered numerous examples in which the researcher did not take the need for randomization seriously and consequently compromised the quality of the experiment and the results therefrom. Methods of proper randomization are discussed extensively in statistical methods texts. There are also computing routines available for this purpose in some of the statistical computing packages.

It is common for randomization to be considered relevant only at the time of assignment of the treatment to the experimental units. It is important that the researcher take care during all phases of the research that spurious correlations are not introduced among experimental units receiving the same treatments or any other correlations not accounted for by the design (see misuse 2). Such might occur, for example, if potted plants from different replicates are grouped for easy administration of a daily nutrient supplement.
When experiments are conducted in series (over sites and/or years) it is necessary to provide different randomizations for each experiment. This will reduce biases which might result from two adjacent treatments interacting and will tend to equalize the precision of all comparisons.

There are also experiments in which the entire experimental process is a chain of individual steps. For example, plants might be grown according to an experimental design in the greenhouse and then transferred to the field and used in a second stage experiment under field environmental conditions. Or, more likely, plants may be grown in a haphazard arrangement in the greenhouse for part of their growth cycle and then put into a designed experiment at a specified stage of growth. Proper randomization at that stage would avoid biasing treatment effects and experimental error (arising from environmental variation during the earlier stage of growth) but a more efficient experimental design would incorporate provisions for error control at all stages of plant growth.

**Misuse 4—Using an Improper Size of An Experiment**

It is important to use an appropriate number of replications in an experiment. Under-replication could result in very imprecise estimates, whereas over-replication can be costly. Agronomists probably err on the side of under-replication more often than on the other side. One still sees self-contained field experiments where the researcher is attempting to achieve adequate precision with only two replications. There are very few field situations where this number of replications would be adequate.

Table 2.1 in Cochran and Cox (1957) provides a useful guide to the determination of the proper number of replications if the approximate variability of the experimental material (coefficient of variation) is known and the researcher is willing to estimate differences between means of a given percent. It is also necessary to assume a probability level for the test (α) and a probability of rejecting a false null hypothesis.

Equally important in agronomic research is for the replication to adequately sample the reference population of environmental conditions. This is usually accomplished by growing the test over several years at several locations within the geographical area of interest. It is very unlikely that a test at one site in 1 year will provide a reliable inference to any except the most restricted reference population of environments.

**Misuse 5—Using Improper Experimental Technique**

The precision of data from an experiment depends to a large degree upon the experimental technique used. Because statistics deal with variability and methods for dealing with it, experimental technique does fall within the realm of a statistician's concern. In fact, statisticians have perhaps made one of their more important contributions to agronomic research in asking questions of the researcher about his experimental technique with a view to its improvement.

We find that many agronomists do not know the meaning of effective blocking. In many cases, they are blocking just to provide replication, not error control. Others attempt to run experiments before they have adequately become familiar with the experimental process (perhaps through small pilot studies) and so they do not use the best technique. Some lack care in controlling variation. We say that their experimental technique is out-of-control. Others do not oversee the experimental process adequately or they do not take note of unusual events which took place at the experimental site. Consequently, when these unusual events have generated "outliers" in the experimental data, there is no basis for rejection of the extreme datum points from the data set.

Overall precision can be increased by using a uniform experimental technique throughout the series of experiments. Some ways of standardizing technique are to: 1) write out procedures for conducting various phases of the experiments and a time schedule for their execution; 2) make all personnel dealing with the treatments, plots and data aware of the various sources of error and the need for good technique; 3) apply the treatments uniformly; 4) exercise sufficient control over external influences so that every treatment produces its effect under controlled, comparable conditions; 5) devise suitable unbiased measures of the effects of treatments; and 6) prevent gross errors.

There are many aspects to technique such as choice of proper plot size and shape, choice of dosages in quantitative controlled variable experiments and proper timing of operations. It is very important from a statistical point of view that individuals who lay out the experiments are trained in the subject matter discipline as well as in field plot technique. Otherwise it is impossible to provide credentials to imprecise data once they have been collected under dubious sets of circumstances.

**MISUSES IN ANALYSIS AND INTERPRETATION OF DATA**

**Misuse 6—Using Inappropriate Error Terms for Testing or for Calculating Standard Errors**

Use of an inappropriate error term for testing or for providing standard errors has been a problem for many years but has increased recently with the common use of statistical computing packages which use a default error term. By this is meant that all terms not included in the linear model spelled out in the instructions to the computer are pooled into an error term which the computer uses for tests and estimates of precision. In a very large proportion of the cases, the tests of significance automatically provided by computer packages are incorrect. Each user of a computer package is responsible for the correctness of his or her analysis.

The analysis of variance of data from a randomized complete block design in which each plot has been sampled (Table 1) has a sampling error in addition to the
Table 1. An analysis of variance of data from a randomized complete
block experiment (four blocks and 10 treatments) in which the entire
plot was not harvested, but instead six samples were taken within
each plot and analyzed for level of the response
variable being studied.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>3</td>
<td>9000</td>
</tr>
<tr>
<td>Treatments</td>
<td>9</td>
<td>4000</td>
</tr>
</tbody>
</table>
| Error      | 27 | 2000| $F = 2$ NS
| Sampling error | 200 | 400 |

usual experimental error. The appropriate error term for
testing treatments is experimental error, not sam-
pling error. The $F$ of 2 based upon forming the ratio of
Treatment MS to Error MS with 9 and 27 degrees of
freedom is not significant at the 0.05 level. If one placed
only Blocks and Treatments in the linear model used in
fitting by the computer, the pooled error would be $[(27 \times 2000) + (200 \times 400)]/227 = 590$. Using this inap-
propriate pooled error, the computer would use a de-
nominator of 590 rather than 2000 in the $F$ ratio and the
denominator degrees of freedom would be 227 rather
than the correct value of 27. The resulting inference
would be incorrect, i.e., treatments would now be sig-
nificant at the 0.05 level.

Or if the model is specified so that “error” is
separated from “sampling error,” some computer
packages will use the sampling error in testing the treat-
ments category resulting in a large upward bias in the
Treatments $F$ ratio.

There are other cases where the researcher desires to
use the appropriate error term but it is difficult due to
the nature of the design constraints. A common ex-
ample is the design of growth chamber experiments. It is
difficult to provide enough chambers for replication of
the temperature-humidity conditions. The chambers are
often shared by a number of researchers and also it is
difficult to change the temperature-moisture settings for
a given chamber. Replication of the factor-combinations
within the chamber is achieved more readily but the error
term resulting from this second type of replication is not appropriate for testing temperature-
humidity treatments. The answer to this problem usually
is to provide multiple runs of the experiment using the
same temperature settings within a chamber from run to
run but with new randomizations of the factor-com-
binations within a chamber for the various runs. The
run then serves as a block in a randomized complete
block whole-plot arrangement of temperatures and the
plots within the chambers within a run are considered as
sub-plots.

In short, the researcher needs to be sure that the ap-
propriate error term is being used whether the analysis
of variance is being conducted by a desk calculator
under his personal supervision or whether by an elec-
tronic computer. The computer is in no position to de-
termine the appropriate error term with which to test so
one should not blindly assume that significance tests or
error estimates obtained from the computer were ob-
tained using the proper error. The procedure of writing
out expectations of mean squares which is described in
statistics methods texts is useful in guiding one to the ap-
propriate error term for testing, especially in situations
where there are several plot sizes or levels of error within
the same experiment.

Misuse 7—Failing to Study Patterns in Data

With modern computers, there is a tendency to rou-
tinely process data through standard data analysis rout-
ines without careful study of the patterns of variation
in the data. As a result, researchers are much less
familiar with the behavior of their experimental data
than when analyses were done by hand and it is easy for
badly behaved data to escape detection. The presence of
a single outlier can grossly inflate variances without
being detected if, for example, only routine analyses of
variance are run. Heterogeneous variances, inade-
quacies of the model, and model assumptions will seldom
be detected without a careful study of the data. If
the data set is too large for a careful study by hand, vari-
ous computer programs for editing, residual analysis,
tests of normality, etc., are available for assisting a
complete analysis of the data. If causes for the outliers
can be identified, it is possible to replace them by
missing plot estimates. In some cases, entire sections of
the data (or even entire data sets) are rendered invalid by
effects of uncontrolled factors or by improper experi-
mental design or layout. It is important that such cases
be recognized and dealt with appropriately.

Detection of portions of the data where the variance
differs from that of other parts of the data may be ac-
complished by comparing the ranges among the
replications for each treatment or by analysis of
residuals. There are several courses of action once it has
been determined that the errors in the data set are
heterogeneous. In some cases, a transformation such as
the log-transformation may be used. Another approach
is to partition the data into sets which have homoge-
nous variance and conduct separate analyses of
variance for each set.

A careful study of the data patterns will also help to
determine if the a priori biological model is adequate or
if the patterns show that some other model would be
more appropriate.

Misuse 8—Depending Excessively on One Class of
Statistical Analyses

In agronomic research, the analysis of variance has
almost become the universal method of analysis. This is
the statistical procedure emphasized in all basic stati-
sical methods courses and it is familiar to most agrono-
mists. The analysis of variance is a powerful technique
for understanding variational patterns in the data and
deserves to be a primary tool. However, excessive de-
pendence on the analysis of variance, or any other
single statistical method of analysis, is a handicap to the
researcher. There are two problems associated with this
dependence. First, the particular statistical analysis
simply may be inappropriate for the problem either be-
cause basic assumptions required for the analysis are not satisfied or because other more powerful methods may be available. Secondly, other methods of analysis may be more revealing of the basic structure of the data. For example, a principal component analysis or other multivariate methods can be very informative of the correlational structure among several variables. The researcher should always be critical of the standard method of analysis and seek statistical assistance if he suspects his method is inadequate in any sense.

**MISUSES IN REPORTING EXPERIMENTAL RESULTS**

**Misuse 9—Misapplying Multiple Comparison Procedures Such as Duncan's New Multiple Range Test**

Perhaps the most frequently occurring of the misuses in agronomic reporting is misuse of multiple comparison procedures. There has been a recent tendency to overuse and misuse mechanical comparison procedures such as Duncan's New Multiple Range Test in interpreting and reporting research results. The plant science journals have faced this problem for a number of years. Unfortunately, the misuse of these procedures is recorded permanently in some of the papers of these journals.

In approaching the problem of drawing inferences about treatment effects, it is important to review the plans and the major goals set forth by the researcher before the experiment was initiated. It is also important to review the experimental and treatment designs and an account of what has taken place at the research site throughout the period of the experiment. The treatments are then studied in detail to see what comparisons are logical from the structure of the treatments. As an example, consider the following set of five treatments: the 2 x 2 factorial set of treatments consisting of varieties A and B with and without fertilization plus a check treatment consisting of the standard variety and the standard fertilization. The comparisons which would be logical for these five treatments are:

- Ck vs. other four treatments,
- Var A vs. Var B,
- Fertilizer vs. No fertilizer, and
- (Var A vs. Var B) x (Fertilizer vs. No fertilizer).

One should consult a statistics methods text for the methodology to construct contrasts and to calculate the sum of squares associated with each comparison. It should be pointed out that multiple comparison procedures should not be used for comparing the five means because there are specific comparisons suggested by the factorial structure of the data. We want the power of the tests to be focused on these particular comparisons.

For situations in which the experiment involves a quantitative controlled factor (e.g., an N-rate experiment involving treatments of 0, 50, 100, 150, and 200 kg/ha N) the appropriate statistical analysis is to fit a response curve. This may be reported in a graph showing the response relationship together with the equation and measures of precision. The preceding approach may be extended to more than one fertilizer variable; a response surface in two or more variables would be fitted to the yield data. Adoption of a response curve (or surface) to represent the behavior of the data completes the tests of significance of effects of changes in the independent variables. It would be inappropriate, for example, to ask, "Is the change in Y from X = 10 to X = 20 statistically significant?" Or, "How much does X have to be increased (above X = 10) before a statistically significant increase in Y is achieved?" Questions of this type reflect a general over-emphasis on hypothesis testing at the expense of estimation. The statistical test of significance is an aid for deciding whether changes in Y are real or are an artifact of the random "noise" in the experiment. Adoption of the response curve, presumably using appropriate tests of significance, implies that we have rejected random noise as the explanation for the changes in Y. Having adopted the response curve, the above questions should be rephrased as estimation questions; e.g., "What is the estimated change in Y as X is changed from 10 to 20, and what is the standard error of the estimated change?" or "How much does X have to change to produce a (biologically) meaningful increase in Y of S units, and what is the standard error of this estimated change in X?"

It is difficult to lay down hard and fast rules for interpreting data from factorial experiments. Each experiment will present a different interpretational pattern. Usually if interaction(s) are negligible it is possible to use the techniques described above on the main effect means. If interaction is sizeable, the above techniques are used for one factor and each level of the other factor(s).

Where treatments have structure, the inclusion of a brief analysis of variance table in the RESULTS section of a report or paper will often be quite useful to the reader (Table 2). The check mean was different from the mean of all of the other treatments. Also the Var A mean was different from the Var B mean. One can then report the check mean, the average of the other four means and the Var A and Var B means. This gives the information which exists in this set of data.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4</td>
<td>1450</td>
<td>4.8*</td>
</tr>
<tr>
<td>Ck vs others</td>
<td>1</td>
<td>2225</td>
<td>7.5*</td>
</tr>
<tr>
<td>Variety (V)</td>
<td>1</td>
<td>1500</td>
<td>5.0*</td>
</tr>
<tr>
<td>N-rate (N)</td>
<td>1</td>
<td>1250</td>
<td>4.2</td>
</tr>
<tr>
<td>V x N</td>
<td>1</td>
<td>825</td>
<td>2.8</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

* Significant at the 0.05 level.
Waller-Duncan k-Ratio Procedure (Steel and Torrie, 1980) seems appropriate. However, it should be pointed out that in most experiments there is at least some structure in the treatments.

**Misuse 10—Failing to Report in the MATERIALS AND METHODS Section of the Research Report the Experimental Design and Statistical Procedures Used**

One of the common failures in reporting of research results is to inadequately describe in the MATERIALS AND METHODS section of the research report or paper the experimental design and statistical procedures used. This information is very essential to the reader's proper interpretation of the results reported in the paper. A detailed account of what procedures were used will also allay any fears on the part of the reviewer and the editor of the journal and the reader that improper design or statistical procedures were used.

An example of a statement which would be used to describe the experimental design and statistical procedures is as follows: The experiment was conducted according to a split-plot design with a randomized complete block arrangement of the whole-plot factor (Varieties). There were four blocks. The sub-plot factor was N fertilizer rate (0, 50, 100, 150 kg/ha of N supplied as anhydrous ammonia). Because there was a significant Variety x N-rate interaction, separate quadratic N-rate response curves were fitted for each of the varieties. Economic optimal rates were computed according to the methods of Heady et al. (1955) assuming a cost/price ratio of 0.0244 (N fertilizer in kg, corn yield in quintals).

**CONCLUSIONS**

Misuses of statistics in agronomic applications are far more prevalent than most agronomists realize. The statisticians' views of statistical concepts have changed considerably since the early days of statistical application (e.g., when Duncan's New Multiple Range Test was in vogue). Agronomists apparently are not fully aware of these changes. Widespread use of computers, together with the ready availability of "user friendly" software packages have also resulted in a number of misuses of statistics.

Agronomists should recognize their need for statistical assistance in planning experiments and in analyzing and interpreting experimental data. The responsibility for improving the statistical practice in agronomic research rests jointly with the agronomist, the statistician, the agronomic journal reviewers and editors. Such improvement should result in well-planned agronomic experiments which focus upon the problem(s) being researched, the results from which are clear-cut and the conclusions from which are scientifically valid and relevant to the underlying problem.

**LITERATURE CITED**