January 1965

STATISTICS IN THE DESIGN AND ANALYSIS OF PHYSIOLOGY EXPERIMENTS

L. Dale Van Vleck
University of Nebraska-Lincoln, dvan-vleck1@unl.edu

C. R. Henderson
Cornell University

Follow this and additional works at: http://digitalcommons.unl.edu/animalscifacpub

Part of the Animal Sciences Commons

http://digitalcommons.unl.edu/animalscifacpub/322

This Article is brought to you for free and open access by the Animal Science Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Papers and Publications in Animal Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Abstract: Statistics should be considered as a tool by the animal experimenter in much the same way that a chemical analysis or a radiation counter is used as a tool. Too often statistics is considered as something magical which can restore order out of chaos and perhaps absolve the experimenter from mistakes in logic and procedure. It is true of course that the statistician through his knowledge of statistical procedures can sometimes help salvage some results from an otherwise hopelessly muddled experiment. One point must be made clear at the outset—the statistician is not a tool of the experimenter although statistics is such a tool. What then is the proper role of the statistician in biological experimentation? In this time of specialization, most professional men and women are highly trained in one aspect of one field of interest. The old cliche that continued education is simply “Learning more and more about less and less” is nevertheless cogently descriptive of our age.
STATISTICS should be considered as a tool by the animal experimenter in much the same way that a chemical analysis or a radiation counter is used as a tool. Too often statistics is considered as something magical which can restore order out of chaos and perhaps absolve the experimenter from mistakes in logic and procedure. It is true of course that the statistician through his knowledge of statistical procedures can sometimes help salvage some results from an otherwise hopelessly muddled experiment.

One point must be made clear at the outset—the statistician is not a tool of the experimenter although statistics is such a tool.

What then is the proper role of the statistician in biological experimentation? In this time of specialization, most professional men and women are highly trained in one aspect of one field of interest. The old cliché that continued education is simply "Learning more and more about less and less" is nevertheless cogently descriptive of our age.

Many biological problems, however, overlap various areas of specialization. A team consisting of many specialists may be needed to solve such problems. A not unusual team in the area of hormone secretion and function may include an endocrinologist, a biochemist, a physiologist, a physicist, and should we not include, also, either a biomathematician or a statistician? An example of such a team is the group which postulated the structure of genetic material.

In this sense the statistician should be considered as a partner of the team, if not a full partner, at least a junior partner. He should not be in the position of a lawyer who wants in at the beginning of a case, not just after the client has been convicted.

This does not imply that the biochemist or physiologist needs no training in mathematics or statistics. An appreciation of the problems and techniques of statistics is important for all experimenters. This knowledge may be no deeper than the survey courses biologists take in the humanities or social sciences. An expert knowledge is not required. In order to talk intelligently to the statistician, however, at least a knowledge of the rudiments of statistics is necessary.

A reasonable statistical requirement for research biologists would be the completion of a 1-year course at the level of Snedecor (1956) or Steel and Torrie (1960). Many of the underlying principles and most of the mathematical derivations obviously would not be learned. Yet the experimenter would become more aware of the limitations as well as the potentials of statistics as a tool. Most biologists who have not had such formal training can acquire an equivalent understanding through independent study.

It is desirable for the statistician on the other hand, as a member of the team, to have some acquaintance with the field of the biologist. Many biometricians who have come up through biologically oriented curricula, with later specialization in statistics, will have this acquaintanceship. Statisticians with formal mathematical training, on the contrary, do not often have a biological background. A general observation, however, is that it is much easier for a mathematician to gain a general knowledge of biology than for a biologist to gain a general knowledge of mathematics.

What then can the statistician offer as a member of the team?

All statisticians will not agree, since Fisher has remarked that variance is a term describing the attitude of one statistician to another. Practicing biometricians, however, find the following statements by Cochran and Cox (1957) to be descriptive of the problems they face everyday:

"Statisticians are often asked for advice in making inferences from the results of experiments. Since the inferences that can be made depend on the way in which the experiment was carried out, the statistician should request a detailed description of the experiment and its objectives. It may then become evident..."
that no inferences can be made or that those which can be made do not answer the questions to which the experimenter had hoped to find answers. In these unhappy circumstances, about all that can be done is to indicate, if possible, how to avoid this outcome in future experiments. Consequently, it has come to be realized that the time to think about statistical inference, or to seek advice, is when the experiment is being planned.

"Participation in the initial stages of experiments in different areas of research leads to a strong conviction that too little time and effort is put into the planning of experiments. The statistician who expects that his contribution to the planning will involve some technical matter in statistical theory finds repeatedly that he makes a much more valuable contribution simply by getting the investigator to explain clearly why he is doing the experiment, to justify the experimental treatments whose effects he proposes to compare, and to defend his claim that the completed experiment will enable its objectives to be realized. For this reason the remainder of this chapter is devoted to some elementary comments on the subject of planning. These comments are offered with diffidence, because they concern questions on which the statistician has, or should have, no special authority, and because some of the advice is so trite that it would be unnecessary if it were not so often overlooked.

"It is good practice to make a written draft of the proposals for any experiment. This draft will in general have three parts: (1) a statement of the objectives; (2) a description of the experiment, covering such matters as the experimental treatments, the size of the experiment, and the experimental material; and (3) an outline of the method of analysis of the results."

Along with the statements of objectives, hypotheses to be tested, and effects to be estimated there should be a literature review of similar or related experiments. Such a review will often point out some of the potential experimental difficulties, e.g., death losses resulting in unequal replication, various effects which should be removed by blocking, interactions among blocks and treatments and among other variables and treatments. Estimates of the size of possible differences among treatments, or the size of differences which are to be detected, and the amount of sampling variation inherent in the experimental material will aid in deciding whether the experiment has a chance of detecting important differences among treatments if such differences actually exist. It should be common knowledge that statistically significant differences can nearly always be found in biological experimentation, if the numbers of replicates are large enough or if the significance level for rejection of the hypothesis of no differences is set at a sufficiently high level. This aspect of significance levels will be discussed more fully a little later.

Kempthorne (1952) and Federer (1955) both suggest in their books a similar set of general principles of experimentation. Kempthorne (1952) states that "a statistically designed investigation may be said to consist of the following steps:

1. Statement of the problem.
2. Formulation of hypotheses.
3. Devising of experimental techniques and design.
4. Examination of possible outcomes and reference back to the reasons for the inquiry to be sure the experiment provides the required information to an adequate extent.
5. Consideration of the possible results from the point of view of the statistical procedures which will be applied to them, to ensure that the conditions necessary for these procedures to be valid are satisfied.
7. Application of statistical techniques to experimental results.
8. Drawing conclusions with measures of the reliability of estimates of any quantities that are evaluated, careful consideration being given to the validity of the conclusions for the population of objects or events to which they are to apply.
9. Evaluation of the whole investigation, particularly with other investigations on the same or similar problems."

Experienced and outstanding investigators usually will require little statistical assistance with points 1 and 2. Most of us, however, do not fall into that category. As attested by Cochran and Cox (1957), the statement of the problem may often be unclear and actually confusing. In such a situation, formation of exact hypotheses is next to impossible. So with most experimenters, the statistician will begin to show his worth as a member of the team at the beginning. A written statement of objectives and hypotheses will allow the statistician to make a judgment as to the logic of the experiment. Also, if the experimenter is re-
required to explain to the statistician what it is he wants to learn and how he proposes to do it, the usual result is that the purpose of the experiment becomes clearer to the experimenter himself. Do not assume that the statistician understands your problem at the outset. He may be intelligent and a good statistician, but he cannot be expected to be an expert in your field. Therefore be patient, clear and thorough in your briefing of the biometrician. Similarly the statistician won’t expect you to know much about the experimental designs which he will devise based upon your statement of the problem, review of literature, and available material. Here again, do not be awed by his vast array of designs. Make sure that he understands the potential hazards of the experiment. The statistical design cannot be any better than the biological model which underlies it.

This makes the statistician an important partner through points 1 to 5. He, however, will contribute nothing to the physical performance of the experiment. The statistician is again likely to be active at point 7 in the application of statistical techniques to experimental results. If the experiment was performed as designed, this help may be mostly passive. With the widespread availability of electronic computers the statistical analyses of such problems should be programmed beforehand, so that the analyses can be completed as soon as possible after termination of the experiment. The statistician may or may not make the necessary arrangements. The advantages of immediate statistical analysis are obvious to any who have ever been delayed in the completion of this aspect of experimentation.

Experienced investigators will need little support at point 8. Most of us, however, will at least want the reassurance of the statistician in making inferences and in drawing conclusions from the analyses.

Point 9 will be largely the responsibility of the biologist—the integration of the current results with previous similar investigations. But here again the statistician may be able to devise procedures for statistical pooling of various investigations.

Thus, we see that only in exceptional cases is the statistician excluded from most phases of the statistically designed investigation.

Now, however, point 8 should be considered in more detail. As is well known, two types of errors can be made in accepting or rejecting an experimental hypothesis. This is shown in the diagram.

<table>
<thead>
<tr>
<th>When hypothesis is actually:</th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
<td>Type II Error</td>
</tr>
<tr>
<td>False</td>
<td>Type I Error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

If we say the null hypothesis (the hypothesis of no or very small differences) is false when there are no differences, the error is said to be an error of the first kind, the so-called type I error. We can also say that there are no differences when there really are differences. This error is said to be an error of the second kind, the so-called type II error.

What is usually considered by experimenters is the type I error. Texts, journals, legislation, statistical tables have all led to common use of 5% and 1% probability levels of errors of this kind in tests of the null hypothesis, i.e., 1 or 5 times out of 100 we will say there is a difference when there really is no difference. Usually not remembered, however, is the fact that we should in some cases be more interested in the type II error than in the type I error. If an economic value could be assigned to type I and type II errors (the decision function approach which consists of minimizing the expectation of risk and loss), then we would want a decision rule which would minimize our loss.

For example, two drugs may be tested against a type of cancer. Drug A, the standard, has only a slight effect against the disease. If drug B has a much greater effect, the loss because of type I error may be very small as compared to the loss due to a type II error. Perhaps in such a case we would want to reduce the probability of a type II error, even if the probability of a type I error is raised above the mystical values of 1 and 5%—perhaps to 25 or 40%. Naturally the risks of side effects must also be considered in formulation of the loss functions.

Too often we forget that the usual tests of significance are very conservative and tend to maintain the status quo. When we test the hypothesis of no or little difference among treatments, the only concern is that we want to make the error of rejecting a chance effect as a chance effect (the type I error) no more often than 5 times in 100.
Although significance levels appear to have little to do with publication policies, we shall see now that tests of significance may play a surprising role in what we read in scientific journals.

Sterling (1959) points out that in some fields many experiments appear not to be reported. The first selection of reports to be published is done by the experimenter. In such cases the experimenter decides which results warrant the effort required for publication. In addition, the experimenter may make observations on many variables during the course of an experiment, but may report only those which support his conclusions. The final selection occurs at the journal level where editorial policy further reduces the number of papers which are finally published.

Since editorial policy usually requires some evidence of statistical analysis, Sterling (1959) asks if the probability levels for rejection of the null hypothesis stated in the papers published which survive the two-stage selection barrier are, in fact, what they are stated to be.

"Tests of hypotheses" or "tests of significance" are commonly presented. These tests are based on the probability of rejecting the null hypothesis when it is really true—the type I error.

The frequency of such tests is shown for several biological journals in table 1 adapted from Sterling's (1960) paper. He also points out that the significance level used is almost always 5%.

Unfortunately, but usually, the probability of a type II error is ignored. The right side of table 1 adapted from table 3 of Sterling (1960) illustrates the frequency of use of the 5% significance level. His tabulation shows that 43 of 44 tests were made at the 5% probability level.

Sterling (1960) also shows what has happened in one scientific field where apparently only reports which reject the null hypothesis at the 5% level are published (table 2 adapted from table 4 of Sterling, 1960). This is an extreme case, but he believes biologists are following a similar path. He states, "Whenever a test of significance is used, a difference is found. The criterion point for deciding on a difference is the 5% level of significance."

A survey of the May 1964 issue of the Journal of Animal Science reveals a pattern similar to that shown in tables 1 and 2. This tabulation is shown in table 3.

Most reports (41 of 42) tabulated made use of tests of significance, an even higher rate than for the journals tallied by Sterling. All except three reports rejected at least one null hypothesis. Evidently either researchers are hesitant to submit articles to the Journal of Animal Science in which "non-significant" or "negative" results are reported, or the editorial board of the Journal rejects articles which present negative results (the 1963 rate of rejection by the editorial board does not support the idea that this is a policy of the editorial board). Nearly all applied statistics or experimental design texts stress the fact that there are no "negative" results. Showing that the null hypothesis may be true is just as positive as showing that the alternate hypothesis is probably true. The question of whether or not "negative" results are reported ultimately rests with you as experimenters and with the editorial boards of your journals. These tabu-

<table>
<thead>
<tr>
<th>Journal</th>
<th>Number of research reports</th>
<th>Number using tests of significance</th>
<th>Number of articles rejecting null hypothesis for major hypotheses with $p &gt; 0.05$</th>
<th>Number of articles rejecting null hypothesis for major hypotheses with $p &lt; 0.05$</th>
<th>Number of articles failing to find significant differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Clin. Invest. Vol. 39, No. 1</td>
<td>26</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Amer. J. Physiol. Vol. 198, No. 1</td>
<td>45</td>
<td>24</td>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>J. Biol. Chem. Vol. 235, No. 1*</td>
<td>48</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>J. Pharm. Exp. Therap. Vol. 128, No. 1</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>44</td>
<td>1</td>
<td>42</td>
<td>1</td>
</tr>
</tbody>
</table>

* Adapted from tables 1 and 3 of Sterling (1960). All articles contained in the January 1960 issue were used.

* The Journal of Biological Chemistry was included in the tally, because it is often used as a source of reference. Actually, the articles contained in its pages seldom investigate hypotheses. Usually they are concerned with new chemical methods or describe chemical activity. It would be unlikely that statistical tests would be found in large numbers. However, many of the descriptions and response curves could be quantified by appropriate probability functions. It is interesting that not one example of this valuable use of statistics occurred.
TABLE 2. OUTCOMES OF TESTS OF SIGNIFICANCE FOR FOUR PSYCHOLOGY RESEARCH JOURNALS

<table>
<thead>
<tr>
<th>Journals: All issues from January to December</th>
<th>Total number of research reports</th>
<th>Number of research reports that reject $H_0$ with $P &lt; .05$</th>
<th>Number of research reports that fail to reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Psychology (1955)</td>
<td>124</td>
<td>106</td>
<td>105</td>
</tr>
<tr>
<td>Comparative and Physiological Psychology (1956)</td>
<td>118</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>Clinical Psychology (1955)</td>
<td>81</td>
<td>62</td>
<td>59</td>
</tr>
<tr>
<td>Social Psychology (1955)</td>
<td>39</td>
<td>32</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>362</td>
<td>294</td>
<td>286</td>
</tr>
</tbody>
</table>

*a Adapted from table 4 of Sterling (1960).

lations do not, of course, give a basis for decision as to publication philosophies or policies of authors or editorial boards.

If it is true that only experiments are conducted which provide statistically significant results at the 5% level of probability, then the logical conclusion is that all these experiments are designed in such a way that treatment differences are expected and are detected. Replication is thus sufficient in each experiment, and the differences to be detected are satisfactory to the experimenter. From actual experience one would doubt the conclusion that only experiments which have a good chance of rejecting the null hypothesis are conducted. Far too many seminar reports end with the plaintive note that no differences were statistically significant, because of the small number of replicates involved or else that not too much reliance should be placed on the results because of the small numbers of animals involved in the experiment.

Another possibility is that, if enough variables are observed, then some of them will call for a rejection of the null hypothesis. Again the possibility that about half of all variables tested (see middle of table 3) would call for the null hypothesis to be rejected does not seem very probable. This ratio of about one-half is true for both null hypotheses tested by analyses of variance or correlation techniques.

Sterling (1960) makes these comments which warrant inclusion in this discussion. "It appears then very definitely that the probability with which an experiment may be made part of the scientific record is conditional on the probability with which it can reject, either by statistical or by logical test of significance, the hypothesis that these same effects are due to chance alone. This may have one or more of the following consequences:

1. An experimental design for which treatment effects are non-existent or small may have a high frequency of replication by investigators who do not know that this particular comparison has been made previously and that previous experimental tests have failed to reject the hypothesis of chance differences between test and control groups.
   a. The number of replications is probably inversely related to the magnitude of the actual difference engendered by the treatment effect.
   b. The chain of replications may be terminated at the point at which one of the investigators finds he can reject the hypothesis of chance differences between experimental and control groups.
   c. When use is made of statistical tests of significance, the actual incidence of false positives may be increased out of proportion. The introduction of tests of significance, coupled with the fixed level of 5 in 100 necessary for rejection, decreases the average number of replications needed for publication of a false positive to 1 in 20.

2. From the point of view of publishing false positives, it actually makes little difference if an experiment is replicated 20 times or if 20 incorrect hypotheses are examined (Tullock, 1959).
   a. Since the number of investigators is so much larger than the number of papers
TABLE 3. SUMMARY OF STATISTICAL TECHNIQUES USED FOR REPORTS IN THE
MAY 1964 ISSUE OF THE JOURNAL OF ANIMAL SCIENCE

<table>
<thead>
<tr>
<th>Item</th>
<th>Number</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of research reports</td>
<td>42</td>
<td>This excludes genetic and symposium papers and one for which tests were unintelligible.</td>
</tr>
<tr>
<td>Number of research reports using tests of significance</td>
<td>41</td>
<td>This includes two reports that must have used an analysis of variance, but did not say so.</td>
</tr>
<tr>
<td>Number using analysis of variance</td>
<td>38</td>
<td>Three others reported standard errors of treatment means.</td>
</tr>
<tr>
<td>Number reporting error mean square</td>
<td>9</td>
<td>Two of these were obviously using 5% level, but did not say so.</td>
</tr>
<tr>
<td>Number rejecting null hypotheses at p&lt;.05</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Number rejecting null hypotheses at p&lt;.10</td>
<td>2</td>
<td>All were at P=.05 level although two of these were obviously using 5% level, but did not say so.</td>
</tr>
<tr>
<td>Number rejecting null hypotheses at more than one level of probability</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Number not rejecting at least one null hypothesis</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Number of variables for which null hypotheses were tested</td>
<td>565</td>
<td></td>
</tr>
<tr>
<td>Number of variables for which null hypotheses were rejected</td>
<td>265</td>
<td>Some reports included both correlation estimates and analyses of variance.</td>
</tr>
<tr>
<td>Number of reports using Duncan’s multiple range test</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Number using other multiple range tests</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Number reporting correlation coefficients</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Number reporting standard errors of variables</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Number rejecting hypotheses or t=0 at p&lt;.05</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Number of correlation coefficients reported</td>
<td>703</td>
<td></td>
</tr>
<tr>
<td>Number for which null hypotheses were rejected</td>
<td>355</td>
<td></td>
</tr>
<tr>
<td>Number of reports citing Snedecor’s text</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Number of reports citing Steel and Torrie’s text</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Number of reports citing other texts</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

that can be published, it means that it is impossible to estimate the number of times an incorrect hypothesis is tested. An incorrect hypothesis may be defined as one that tests for the positive effect of a treatment where the treatment effect is actually zero or close to it. With selection of positive results for publication, an appreciable number of these false hypotheses may be published as indicating treatment effects.

b. The use of statistical tests and the fixed level of significance insures that about 5% of all investigations of incorrect hypotheses will be considered suitable for publication. However, this does not mean that only 5% of the articles that make use of tests of significance are false positives. The number of false positives is unknown. It may be assumed to be large, since the determining factor in the publication of false positives is not the number of articles published but the number of investigations.

3. The selective nature of experiments themselves increases the probability that they will ascribe treatment effects erroneously.

a. The investigator may use a sequential method of correlating treatment effects with a number of possible antecedents and stop when he has found a significant association. This method is almost sure to unearth a spurious association.

b. The investigator may have started with a large number of variables or increase this number as he sees fit. The probability is high that some of the observed differences will be due to chance. When tests of significance are used with the fixed level of significance, the investigator examining large numbers of vari-
ables is almost bound to unearth occasional spurious treatment effects.’’

Thus, following the argument of Sterling (1960) we see that the probability of a type I error for a reader of a journal may well be closer to 100% than to the “safe and sane” 5%. On the other hand the chance of the reader’s making a type II error may be near zero. Thus, we will have, in the limit, gone the full circle from a conservative type I error rate and a relatively unknown or high type II error rate to a radically high type I error rate and a low type II error rate.

The article by Tullock (1959) cited by Sterling (1960) points out that, in most areas of research if on the average one of 20 investigators by chance alone rejects the null hypothesis at the 5% level of probability and, therefore, presumably qualifies his research for publication, then most of the other 19 researchers may hurriedly write letters to the editor or attempt to publish their “negative” results. How frequent this may be is unknown. More likely, most experiments are not exactly alike in view of the genetic and environmental variability of experimenters. Therefore, Tullock (1959), as well as Sterling (1960), reasons that on the average out of 20 such unlike experiments at least one will qualify for publication on the grounds of rejecting a null hypothesis, even when rejection is due to sampling variation.

Similarly if 20 variables are measured for which the null hypothesis is true, one of these on the average will lead to rejection of the null hypothesis. How many experimenters fail to report the results for the other 19 variables is also open to question. Hopefully, most would not. Yet nearly half of all variables in the May 1964 issue of the Journal of Animal Science were reported to be associated with a rejection of the null hypothesis. Particularly open to question is the testing of correlation coefficients. If the experimenter has the patience, sooner or later a variable can be found which has a significant correlation with another variable no matter how spurious this correlation may be.

This fact raises the question as to how valuable are tests of significance for the null hypothesis that a correlation coefficient is different from zero, the usual null hypothesis for tests of significance of correlation coefficients. With enough observations true correlations as low as 0.01 can be found statistically significant at any reasonable probability level. Yet, what is the biological importance of such a correlation? Certainly not very great if one is attempting to predict one variable when the other is known. Perhaps a correlation of 0.95 which is not statistically different from zero is more important than a correlation of 0.05 which is.

However, point estimates seem to be in disrepute among investigators in most fields reporting in the Journal of Animal Science. The all powerful cult of the significance test appears to be in complete command. Animal geneticists on the other hand have long been forced to use point estimates of heritabilities and genetic correlations even though the statistical confidence limits are not very satisfactory.

Certainly some knowledge of the magnitude of differences among treatment effects is more valuable than a test of the null hypothesis at some probability level. In such cases confidence intervals for differences among point estimates of treatment effects appear preferable to tests of significance which yield little knowledge of the quantitative and, hence, economic value of the differences among treatments. This conclusion applies equally as well to estimates of correlation coefficients. Most statisticians would probably concur. Evidently many non-statisticians do not, since most of the reports tabulated in table 3 relied on an analysis of variance, many with multiple range tests, for tests of significance. Also, only 9 of 38 reports gave an estimate of residual mean square which is essential for a reader to calculate even the simplest confidence intervals.

Statisticians will, in theory, quarrel with another point shown in table 3. The probability level for rejection of the null hypothesis should also be stated before the experiment is performed. Yet in 25 of 38 reports more than one level of probability was stated as the significance level. Perhaps experimenters attempt to quantitate their results on the basis of the significance level. Additional replication would also increase the significance level if a true difference actually exists. The evident doubt as to what significance level to report suggests that experimenters do not set the probability level for the type I error until after the statistical analyses are complete. Only three of the 38 reports gave the impression that the 5% level for tests of significance was established before the data were analyzed.

Even though the 5% probability level is stated for tests of significance based on the analysis of variance, for Duncan’s (1955)
multiple range test, or for other range tests, the true error rates are not usually the same. In other words, on a comparable basis Duncan's test may have a higher error rate than the orthogonal comparisons tested by analysis of variance. Harter (1957), Federer (1955), and Steel and Torrie (1960) discuss the relative error rates for several methods of testing for differences among treatment means. Many designs imply such specific questions that any multiple range test is inappropriate. These points do not seem to be well understood by many biological experimenters.

What can we conclude from the apparent dependency of publication of experimental results on the finding of significant differences?

1. Unless investigators report results according to the merits of the problem and of the experiment, the reader is forced to judge reports on the basis of what he knows about the research group making the report. This reputation will, of course, partially depend on their past publication policy. Most readers will be unable to judge reports adequately on this basis. Sterling (1960) goes even farther and suggests that papers should be submitted for relevance even before the experiment is performed. Knowledge of the editorial policy of the journal and of its potential contributors would also be required. Somehow this counter strategy would seem to be ineffective for most readers.

2. Channels for introducing so-called "negative" results into technical literature should perhaps be found. Such results would not require much journal space. Perhaps they could be in the form of technical notes stating in full detail the conditions of the experiment, the variables measured, the point estimates of treatment effects, and the appropriate sampling variances. Abstracts are of doubtful value in establishing this necessary detail.

3. Much more reliance should be placed on the biological considerations studied in the investigations and less on significance levels found for treatment differences. This statement does not imply that statistical analyses should be discarded, but rather that they should be placed in their proper perspective—as means and not ends in themselves. Point estimates of treatment differences and their confidence intervals are to be preferred to tests of significance. The confidence interval approach includes a test of significance, but the test of significance does not yield the confidence interval. Possible economic losses due to type I and type II errors should be considered in establishing the type I error rate. The single and double star levels are certainly convenient, but they should be understood to be only that. Other probability levels may prove more useful. Modern computing equipment can easily compute statistical tables for use with other probability levels which in the past were accomplished with much desk calculator work for the arbitrarily selected probability levels of 5 and 1%.

4. This problem of publication policy should be faced by editorial boards of biological journals. New policies may be required in view of the findings of Sterling (1960). These boards would find essential the assistance of competent statisticians for the formulation of effective policies which would eliminate the selective nature of results which are published.

Thus, we have seen that the statistician will be an important member of the experimental team at all stages of the experimental investigation with the exception of the actual performance of the experiment. Blind faith in statistical tests of significance by biological experimenters in conjunction with the natural tendency for all experimenters to prefer differences to no differences may have led to selection of results which are published—the final stage of the investigation. This selective nature of publication, in turn, can prejudice the value of reported literature to an incalculable degree. Statisticians may have helped create this problem and working together with biologists are jointly responsible for solving this problem of selective publication through more judicious use of statistical techniques, especially the downgrading of tests of significance, and through a greater understanding of the relationship of "positive" and "negative" results in statistically designed investigations.

**Literature Cited**


