Are Nonsignificant Differences Really Not Significant?

Richard P. Marini

Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0327

Many researchers and reviewers hesitate to publish data when treatments are not "significantly" different. Most of us were taught to use a particular probability value (usually 5% or 1%) as a critical value to test hypotheses. The term "significant" has come to be synonymous with the phrase "significant at the 5% level." Using relatively conservative critical values prevents readers from evaluating, or even seeing, some data that should be published. When using 5% as a critical value, we would conclude that two treatments are different when the P -value is 0.05, but not when it is 0.051. Reviewers often suggest that data not differing at the 5% level should not be included in a publication, but the author should state that the difference was not significant. Is this an intelligent use of statistical tests? Researchers use such tests to help interpret experimental data. Probability values allow us to determine the probability of incorrectly declaring that responses are different when the difference is actually due to chance. The use of critical levels to test hypotheses has led to a black or white (reject vs. fail to reject) view of data when the real world is mostly gray.

When a probability value of 5% is used as a critical value, there is a 5% chance that the null hypothesis will incorrectly be rejected. In other words, 5% of the time responses are declared different when the difference is really due to chance rather than to treatment. This type of mistake is known as a Type I error. The probability of a Type I error is α . Most researchers do not consider a Type II error, which is the probability of incorrectly concluding that treatment differences are due to chance rather than to experimental treatments. The probability of a Type II error, denoted as β , is determined by the choice of α and the difference between two treatment means. Power ($1 - \beta$) is the ability to detect differences that are due to treatment. Ideally we would like to have a low probability of making both types of errors, but Type I and Type II errors are inversely related, and decrease only as sample size increases. The concept of Type II error is especially important in determining the sample size needed to detect a difference of a specified magnitude. When both types of error are considered in the case of a fixed sample size, a reduction in Type I error must be accompanied by an increase in Type II error. To decrease Type I error without increasing Type II error requires either increasing the sample size or reducing the magnitude of the treatment differences we are able to declare as being significant.

Received for publication 14 Oct. 1998. Accepted for publication 26 Jan. 1999. The cost of publishing this paper was defrayed in part by the payment of page charges. Under postal regulations, this paper therefore must be hereby marked *advertisement* solely to indicate this fact.

The concept of critical P -values is presented in many statistics texts, is used by many statisticians, and is accepted in most scientific journals. The choice of any particular P -value as a "critical value" is purely arbitrary. Historically, obtaining exact probability values was difficult. Therefore, most statistical texts included tables with values of the probability of obtaining larger z , t , and F statistics. Although some texts had slightly more extensive tables, many contained values for only the 1% and 5% levels because of the expense of large tables.

The concept of always using the same probability value as a "cut-off" level is not only arbitrary, but also oversimplifies a complex issue. Researchers should choose a level of probability based both upon the seriousness of failing to reject the hypothesis when it is false and the probability of rejecting it when it is true (Steele and Torrie, 1980). For example, when comparing the efficacy of two similarly priced plant growth regulators, a liberal test (possibly the 20% level of significance) may be desirable because the consequences of incorrectly declaring one product superior to the other are not serious. However, if one product were four times more expensive, a more conservative P -value would reduce the likelihood of incorrectly declaring the more expensive product superior to the less expensive product. One may also want to consider the sample size and the amount of variation inherent in some types of data. The 5% probability level is probably too conservative for experiments

where large sample sizes are not practical or are too expensive. Some types of data are quite variable and would require extraordinarily large sample sizes to detect differences at the 5% level of significance. For example, >30 peach trees per treatment are required to detect a 15% difference in yield at the 5% level of significance (Marini, 1983). Therefore, if a 10% difference in yield is considered economically important, using a more liberal P -value may be justified.

Several years ago, Wehner and Shaw (1994) suggested that authors be allowed to present comprehensive analysis of variance (ANOVA) tables in ASHS publications. I agree with their recommendations and I would like to suggest two additional changes in the way the data are presented.

1) Rather than using asterisks to indicate the arbitrarily chosen probability level at which the null hypothesis is rejected, present the exact P -values. While perusing the last six issues of *HortScience* published in 1997 (Vol. 32), I noticed that P -values were presented in only five of 98 articles. Modern statistical software packages provide these values by default. Presentation of P -values allows readers to develop their own interpretation of the data. I am not a gambler, but as an extension specialist I am willing to recommend a new inexpensive practice that increases yield by 15% even when there is a 20% ($P = 0.20$) probability that the yield increase was not due to the new practice.

2) Publish data that are significant at some

Table 1. The influence of outliers on P -values for data sets with small and large sample sizes.

Treatment	Rep	Experiment			
		1	2	3	4
A	1	14	14	13	13
	2	15	15	13	13
	3	15	15	14	14
	4	16	25	14	14
	5	---	---	15	15
	6	---	---	15	15
	7	---	---	16	16
	8	---	---	16	16
	9	---	---	17	17
	10	---	---	17	43
B	1	21	20	20	20
	2	22	22	20	20
	3	22	22	21	21
	4	23	24	21	21
	5	---	---	22	22
	6	---	---	22	22
	7	---	---	23	23
	8	---	---	23	23
	9	---	---	24	24
	10	---	---	24	50
P -value		0.001	0.105	0.001	0.099
Mean A		15.0	17.2	15.0	17.6
Mean B		22.0	24.2	22.0	24.6
Range		2	13	4	30
MSE		0.67	26.9	2.22	81.4

reasonable probability level. These decisions must be determined on a case-by-case basis. While deciding on an appropriate significance level, authors and reviewers should consider the economic implications of the data, the sample size used for the experiment, the inherent variability in the type of data being reported, and the seriousness of failing to detect differences that are not due to chance.

To demonstrate how *P*-values can be influenced by a single outlier in experiments with small and large sample sizes, four hypothetical data sets along with results from the ANOVAs are presented in Table 1. The data sets were analyzed as completely randomized designs. For each experiment, the difference between treatment means is 7.0 and the value for treatment B is $\approx 46\%$ greater than for treatment A. Experiments 1 and 2 have only four replicates, whereas Experiments 3 and 4 have

10. There is little variability in the data for Experiments 1 and 3, indicated by the small ranges and MSEs from the ANOVA, and the *P*-values are 0.001. Experiment 2 has one outlier in each treatment, resulting in a range of 13, an MSE of 26.9, and a *P*-value of only 0.105. Experiment 4 also has one outlier per treatment and a *P*-value of 0.099, but the range is 30.0 and the MSE is 81.4. These analyses demonstrate that *P*-values are affected to a greater extent by high-influence observations or variable data in experiments with few replications. Presentation of scatter plots, rather than treatment means, would also allow readers to observe the nature of the variation in the data. In other words, is the variation causing the inflated *P*-value due to a single outlier or to a highly variable data set? While evaluating *P*-values one should consider the magnitude of the difference between treatment means,

sample size, and the variability inherent in the types of data being reported.

The changes suggested above would undoubtedly increase the length of some manuscripts, but they will also allow readers to see data now unavailable to them. This will increase our knowledge base on many topics of investigation and allow readers and authors to interpret data slightly differently.

Literature Cited

- Marini, R.P. 1985. Sample size estimates for peach tree growth and yield experiments. *J. Amer. Soc. Hort. Sci.* 110:604–608.
- Steele, R.E. and J.H. Torrie. 1980. *Principles and procedures of statistics—A biometrical approach*. 2nd ed. McGraw-Hill, N.Y.
- Wehner, T.C. and D.V. Shaw. 1994. Presentation of analysis of variance results and graphical data. *HortScience* 29:608.