Letter to the Editor

Use of Statistical Tests of Equivalence (Bioequivalence Tests) in Plant Pathology

K. A. Garrett

Department of Botany and Plant Pathology, 2082 Cordley Hall, Oregon State University, Corvallis 97331-2902, and Centro Internacional de la Papa, Casilla 17-21-1977, Quito, Ecuador.
Accepted for publication 16 January 1997.

Hypothesis tests currently used in plant pathology are almost always based on a null hypothesis of equal means. In this framework, the experimenter determines whether or not there is evidence that the means are, in fact, different. This framework makes sense for many common questions such as whether a new management technique gives an increase in yield over existing management techniques. But suppose, for example, that a disease management technique is so effective that an experimenter is interested in whether its use in the presence of disease achieves the same yield as in the absence of disease. In this case, a more appropriate null hypothesis would be that mean yields are different. Examples of questions in plant pathology for which a null hypothesis of equal treatment means is not suitable include (corresponding phrasing for one-sided questions is in parentheses as appropriate):

(i) Is an engineered organism equivalent to the original organism in all relevant characteristics except the intended change?

(ii) Is disease severity the same for a cheaper or safer management strategy as for a standard strategy? (Is it at least as low?)

(iii) Does a cultivar with potential for higher yield have the same level of disease resistance as a proven resistant cultivar? (Is it at least as high?)

(iv) Is the level of pesticide on a plant surface the same for different application techniques?

(v) Does the yield reach that of pathogen-free or disease-free plants when

- a biocontrol agent is used?
- a pesticide is applied?
- a resistant cultivar is grown? (Is it at least as high?)

(vi) In general, does a new approach with certain advantages perform as well as a standard "best" approach? (Does it perform at least as well?)

For the last question, in the standard hypothesis testing framework, a test would be made to determine whether there is evidence that performance of the new approach is different from performance of the standard. Based on a null hypothesis of equal means, if there is not evidence for this difference, the experimenter may fail to reject the null hypothesis, but cannot, therefore, conclude that performance is equal. Failure to find evidence for a difference might result simply because the experiment has low statistical power (see below). The equivalence null hypothesis framework offers techniques for testing whether means are functionally equivalent using a null hypothesis of different means (4,5,7). Equivalence tests are perhaps the second most useful general class of hypothesis tests after the standard hypothesis testing framework. Following is a discussion of the standard hypothesis framework, the concept of statistical power, how the equivalence framework differs from the standard framework, an example of an equivalence test, and a summary of some of the available equivalence test techniques.

Corresponding author: K. A. Garrett; E-mail address: garrettk@bcc.orst.edu

Publication no. P-1997-0220-02O

Standard hypothesis framework. The standard hypothesis testing framework is based on the idea that what constitutes "important news" is evidence of a difference between treatment means. This is, of course, true for many questions. Traditionally, greater emphasis has been placed on protecting against type I errors (concluding there is evidence of a difference when a difference does not exist) than on protecting against type II errors (concluding there is not evidence of a difference when a difference does exist) (2). This has been part of a general notion that the burden of proof should be on researchers to demonstrate that they have important news if their work is to be published and considered. There has generally been little concern for whether negative results are also published and given consideration, which has the potential to result in the "file drawer" problem (13). In other words, positive results may be published, while negative results are filed and a literature-wide bias may result.

When the standard hypothesis framework is used for questions in which interest lies in whether means are equivalent, the results are often inconclusive. If the difference between means is relatively large and there is adequate power, there may be evidence to reject a null hypothesis of equivalent means. If the means are similar or there is low power, the typical emphasis on protecting against type I errors may mean that there will not be evidence to reject the null hypothesis of equivalent means. This does not indicate that the means are equivalent, however. The experiment may simply have had low power because of a small sample size or high variation. Thus, an experiment that is too small may be more likely to result in a lack of evidence for a difference, regardless of what the means actually are.

Statistical power. Power is the probability that the null hypothesis will be rejected if it is not true, or one minus the probability of a type II error (3). The power of a test increases as the sample size increases and as the level of unexplained variability decreases. For a null hypothesis of equal means, power increases as the actual difference between means increases. High power also results in narrower confidence intervals around parameter estimates. Low power may make it difficult to demonstrate a real difference between treatments, especially if the difference, or effect size, is small in magnitude. A power analysis can be an important part of planning an experiment, allowing an experimenter to pick an appropriate sample size for an estimate of the variance and the effect size (3).

Equivalence test hypothesis framework. The equivalence test hypothesis framework employs a null hypothesis of unequal means. In this context, "important news" is evidence that means are equivalent. Because two population means will never be truly identical, the null hypothesis used in practice is that the difference

between means is greater than some tolerance defined by a researcher prior to experimentation. The tolerance level used could be based on an arbitrary level of similarity such as allowing for a 5% difference, or it might be based on knowledge of how much leeway is tolerable in the response being measured. The researcher determines whether there is evidence that the difference between means is, in fact, less than this tolerance. This framework puts the burden of evidence on the experimenter to demonstrate that the means are equivalent within a reasonable tolerance.

The need for establishing such an a priori tolerance may seem like a shortcoming of this approach, but such a tolerance is, in a sense, understood in the standard hypothesis framework. If a standard framework hypothesis test is performed when there is very high power, a small difference between means may be statistically significant when a researcher is skeptical of the difference being biologically significant. However, researchers can often be assured of having high enough variability and, thus, low enough statistical power, that they are not forced to consider whether a small difference is biologically important!

It might seem that determining the power of an experiment in the standard framework would give the same information as an equivalence test. But such a power analysis gives information on how small a difference is likely to be statistically significant for a given sample size and level of variability (3,15). Using the fact that a small difference would have a large chance of being detected, yet was not detected, is a roundabout approach to determining that there is evidence for treatments being equivalent. But power is an issue in equivalence tests as in the standard framework, because adequate power is needed to reject the null hypothesis.

Examples of equivalence tests. Hypothesis tests can be viewed as a subset of confidence interval construction. If testing whether there is evidence to reject a standard null hypothesis of equal means, the researcher could determine whether or not the confidence interval includes zero. If not, then there is evidence to reject the null hypothesis of equal means at the confidence level used to construct the interval. For a simple equivalence test, the same confidence interval can be used; the researcher determines whether the interval includes the predetermined tolerance (11). If not, then there is evidence to reject the null hypothesis of unequal means (means more different than the tolerance.) As an example, suppose that an experiment is performed to compare a new management technique with a standard technique for controlling powdery mildew of roses. The experimenter, planning to measure percent infection at the end of the experiment, might determine a prior that a difference of 5 in percent infection is insignificant from a practical standpoint. If the mean percent infection of 10 replicates of the new technique is 12 and of the standard technique, 9, the observed difference between means is 3. Using a standard null hypothesis of equal means, the experimenter would construct a confidence interval around this estimate and determine whether it includes zero. Suppose the data are approximately normally distributed, so that an interval based on a t distribution can be used (2). If the pooled standard deviation is 6.1, then a 95% confidence interval around the estimated difference is $3 \pm 2.1 \times 6.1 \times 0.45$, or 3 ± 5.8, in which 2.1 is the critical t value for 18 degrees of freedom (df) and $0.45 = \sqrt{1/10 + 1/10}$. This interval includes zero, so a null hypothesis of equal means is not rejected. To test the equivalence null hypothesis of different means, the researcher determines whether any of the confidence interval lies outside the predetermined tolerance, –5 to 5. The confidence interval extends above the upper tolerance of 5, so the equivalence null hypothesis is not rejected either. For this experiment, a larger sample size or lower variance would be needed to determine conclusively whether the mean effects of the techniques are different or equivalent at the 95% confidence level.

A series of possible outcomes of hypothesis tests are illustrated in Figure 1. In the first example (Fig. 1A), the confidence interval includes both zero and the upper tolerance; there is so much variability that the null hypothesis cannot be rejected in either the standard hypothesis framework or in the equivalence framework. Note that if the observed difference falls outside the tolerance, it is not possible to reject the equivalence framework null hypothesis. In the second example (Fig. 1B), the confidence interval is narrower and does not include zero; the null hypothesis of equal means can be rejected in the standard hypothesis framework. Next, suppose the observed difference between means does fall within the tolerance around zero. For the third example, though the observed difference is near zero, the confidence interval extends beyond the tolerance; the equivalence test null hypothesis of means more different than the tolerance is not rejected (Fig. 1C). For the same observed difference with a narrower confidence region that falls inside the tolerance (Fig. 1D), the equivalence test null hypothesis of means more different than the tolerance can be rejected. For this case, there is evidence that the means are equivalent as defined by the tolerance. For the fifth example (Fig. 1E), the confidence interval is so small that both the standard null hypothesis and the equivalence test null hypothesis are rejected. For this case, there is evidence that the means are different, but also that their difference is less than the tolerance. Such an example might arise if an experiment has very high power or the tolerance is very wide.

Since interest in equivalence tests has often focused on comparisons of levels of pharmaceutical compounds in blood, or bioequivalence tests, the emphasis has been on two-sided tests. But, for many cases in plant pathology, a one-sided test might be more appropriate. It is not usually considered problematic if yield is actually higher than, or disease incidence is actually lower than, a desirable standard. In either case, the results of standard or equivalence tests can be reported in greater generality by including the P value for the observed difference rather than simply reporting whether or not the null hypothesis is rejected at a particular confidence level. Statistical computing packages typically output P values for standard hypothesis tests. The P value for the

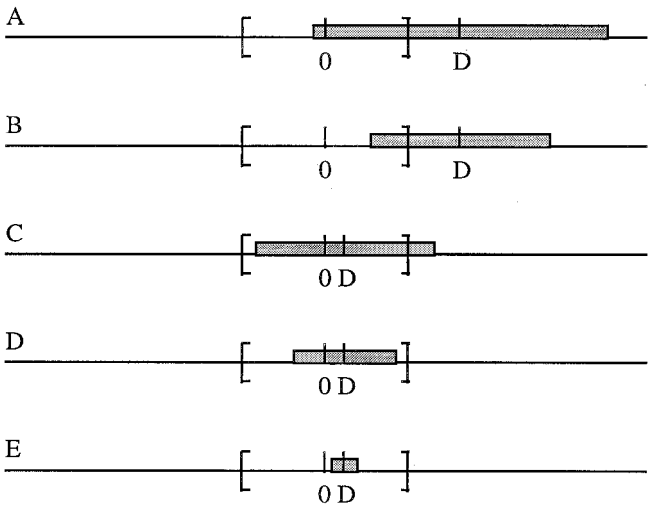

Fig. 1. Examples of hypothesis testing scenarios. Brackets indicate the equivalence tolerance around zero; the shaded region indicates a confidence interval around the observed difference in treatment means (D). **A,** The observed difference falls outside the tolerance; neither the null hypothesis of equal means nor the null hypothesis of a nontrivial difference in means is rejected. **B,** The observed difference falls outside the tolerance; the null hypothesis of equal means is rejected, while the null hypothesis of a nontrivial difference in means is not. **C,** The observed difference falls within the tolerance; neither null hypothesis is rejected. **D,** The observed difference falls within the tolerance; the null hypothesis of equal means is not rejected, while the null hypothesis of a nontrivial difference in means is rejected. **E,** The observed difference falls within the tolerance; both the null hypothesis of equal means and the null hypothesis of a nontrivial difference in means are rejected.

upper tolerance of the equivalence test example given above can be calculated as follows. The difference between the observed difference in means, 3, and the upper tolerance, 5, is 2. This difference is scaled by the standard error: $2/(6.1 \times 0.45) = 0.73$. For a *t* distribution with 18 df, 0.73 is at percentile 0.76, giving a *P* value of $1 - 0.76 = 0.24$. The percentile can be found using a statistical program, for example, the function pt in S-plus (StatSci, Seattle), or from a table of the *t* distribution (2). The null hypothesis of means more different than 5 would not be rejected at the 95% confidence level, as shown previously, but would be rejected at the 75% confidence level for a one-sided test. Reporting the actual *P* value summarizes results for tests at all confidence levels.

Other types of equivalence tests. Several variations exist on the simple equivalence test described in the example based on an assumption of normality (1,8,16). Rather than the absolute difference between treatment means, it may be the proportion that is of interest. Erickson and McDonald (7) describe a test using a null hypothesis in which a treatment mean is equal to a proportion of the standard mean. Erickson and McDonald (7) also report the sample size, coefficient of variation, and tolerance combinations required to yield a desired level of power. Hauschke et al. (9) have suggested a distribution-free procedure based on use of a nonparametric Mann-Whitney-Wilcoxon test. A response may be binomial, such as the presence/absence of disease. Dunnett and Gent (6) have described equivalence tests for the case of binomial samples. The general idea of constructing a null hypothesis of difference greater than some tolerance and determining whether there is evidence of equivalence can, of course, be adapted to tests for other parameters, as well as to different distributional assumptions. Because tests of bioequivalence have become a standard type requested by the U.S. Food and Drug Administration, these tests are beginning to be included in statistical packages. For example, the nonparametric statistics package Testimate (SciTech, Chicago) includes one- and two-sided equivalence tests. Erickson and McDonald (7) discuss how a series of comparisons with a standard might be made with an adjustment for the number of tests such as a Bonferroni's correction (12). However, when comparisons are based on several levels of a continuous treatment, a regression analysis would be more appropriate (10,14).

As Box et al. stated (2), "Significance testing in general has been a greatly overworked procedure, and in many cases where significance statements have been made it would have been better to provide an interval within which the value of the parameter would be expected to lie." By expressing the results of many hypothesis tests simultaneously, an interval estimate may be the most informative product of an experiment. While a researcher would still be advised to decide upon an acceptable tolerance prior to experimentation and report the *P* value for that tolerance, a confidence interval supplies information to allow readers to test their own a priori tolerance level when evaluating the research results.

To summarize, when a hypothesis test of whether treatments are equivalent is desired, equivalence tests offer a more appropriate framework than the standard null hypothesis. By using a null hypothesis of treatment differences, they place the burden of proof on the experimenter to demonstrate that the treatments are equivalent.

ACKNOWLEDGMENTS

This work was supported, in part, by the UDSA STEEP Program. This is Oregon State University Extension and Experiment Station Communications Series paper 11101. I thank G. Forbes, D. Gross, K. Johnson, C. Mundt, M. Powelson, F. Ramsey, and two anonymous reviewers for comments that improved this manuscript and P. Dixon for introducing me to bioequivalence tests.

LITERATURE CITED

1. Anderson, S., and Hauck, W. W. 1983. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Commun. Stat. Theory Methods 12:2663-2692.
2. Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building. John Wiley & Sons, New York.
3. Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. LEA Press, Hillsdale, NJ.
4. Dixon, P. M. Assessing effect and no effect with equivalence tests. In: Risk Assessment: Logic and Measurement. M. C. Newman and C. L. Strojan, eds. Ann Arbor Press, Ann Arbor, MI. In press.
5. Dixon, P. M., and Garrett, K. A. 1994. Statistical issues for field experimenters. Pages 439-450 in: Wildlife and Population Modeling: Integrated Studies of Agroecosystems. R. J. Kendall and T. E. Lacher, eds. CRC Press, Boca Raton, FL.
6. Dunnett, C. W., and Gent, M. 1977. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. Biometrics 33:593-602.
7. Erickson, W. P., and McDonald, L. L. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. Environ. Toxicol. Chem. 14:1247-1256.
8. Hauck, W. W., and Anderson, S. 1984. A new statistical procedure for testing equivalence in two-group comparative bioavailability studies. J. Pharmacokinet. Biopharm. 12:83-91.
9. Hauschke, D., Steinijans, V. W., and Diletti, E. 1990. A distribution-free procedure for the statistical analysis of bioequivalence studies. Int. J. Clin. Pharmacol. Ther. Toxicol. 30(suppl. 1):S37-S43.
10. Madden, L. V., Knoke, J. K., and Louie, R. 1982. Considerations for the use of multiple comparison procedures in phytopathological investigations. Phytopathology 72:1015-1017.
11. Metzler, C. M. 1974. Bioavailability—A problem in equivalence. Biometrics 30:309-317.
12. Milliken, G. A., and Johnson, D. E. 1984. Analysis of Messy Data. Vol. 1: Designed Experiments. Van Nostrand Reinhold, New York.
13. Rosenthal, R. 1979. The 'file drawer' problem and tolerance for null results. Psychol. Bull. 86:638-641.
14. Swallow, W. H. 1984. Those overworked and oft-misused mean separation procedures—Duncan's, LSD, etc. Plant Dis. 68:919-921.
15. Toft, C. A., and Shea, P. J. 1983. Detecting community-wide patterns: Estimating power strengthens statistical inference. Am. Nat. 122:618-625.
16. Westlake, W. J. 1976. Symmetrical confidence intervals for bioequivalence trials. Biometrics 32:741-744.