THE STATISTICAL ANALYSIS OF ACTIVE CONTROL EQUIVALENCE STUDIES

Armando Garsd

Susana F. Delfino

The statistical analysis of active control equivalence studies

by Armando Garsd and Susana F. Delfino

Department of Statistics, School of Agricultural Engineering,
University of Buenos Aires, Argentina

Abstract

In experimental work, the notion of equivalence falls short of the idea of equality. Thus, the effects of two treatments, while not identical, may still be regarded as equivalent if their difference is negligible in a certain sense. This simple distinction raises not only technical difficulties, since of necessity it results in special statistical procedures, but also deeper conceptual issues, since one has to ask why two treatments should be equivalent but not equal, more specifically, whether their being merely equivalent has any bearing on the practical questions posed by the data. In this paper we present examples, drawn from agricultural experiments, to address the statistical analysis of studies intended to show equivalence of effects. We formalize two notions of equivalence in the context of a horticultural experiment conducted on witloof chicory plants, to compare the efficacy of two treatments to prevent root infection. We then extend the work to include the concept of multivariate equivalence for the specific case of two simultaneous endpoints, seed implantation and germination, as the key features to accept that two corn planters are equivalent. We address this type of equivalence via nominal α level adjustments for multiple endpoints. Finally, we discuss these approaches and suggest areas for further research. Among these, we entertain the broader concept of equivalent performance under a defined range of experimental conditions.

1. Introduction

In experimental work, the notion of equivalence falls short of the idea of equality. Thus, the effects of two treatments, while not identical, may still be regarded as equivalent if their difference is negligible in a certain sense. This simple distinction raises not only technical difficulties, since of necessity it results in special statistical procedures, but also deeper conceptual issues, since one has to ask why two treatments should be equivalent but not equal, more specifically, whether their being merely equivalent has any bearing on the practical questions posed by the data.

In this paper we present examples, drawn from agricultural experiments, to address the statistical analysis of studies intended to show equivalence of effects. In section 2 we formalize

two notions of equivalence in the context of a horticultural experiment. In section 3 we extend the work to include the concept of multivariate equivalence. In section 4 we discuss these approaches and suggest areas for further research. Among these, we entertain the concept of performance equivalence.

2. Two notions of equivalence

Consider the case of a study conducted on witloof chicory plants, to compare the efficacy of two treatments to prevent root infection (watery soft rot) by soil fungus *Sclerotinia sclerotiorum*. A new and ecologically sound experimental method based on biological control (by *Trichoderma harzianum*) is thought to be about as effective as the standard chemical method based on the fungicide iprodione. Everything being equal, one would be inclined to recommend the new biological method. To summarize results, Table 1 shows the root infection status of each of 400 chicory plants at the end of the field study. Clearly, both methods perform very similarly.

Table 1
Infection status at the end of field study of roots from 400 chicory plants on one of two treatments

Treatment	Uninfected	Infected	Total
Standard	135 (67.5%)	65	200
Experimental	134 (67.0%)	66	200
	269	131	400

Let π_s and π_e denote the true proportion of plants with roots protected by the standard and experimental treatment, respectively, and p_s and p_e their sample counterparts. The usual set of hypotheses

$$H_o: \pi_s=\pi_e \text{ versus } H_1: \pi_s\neq\pi_e, \quad (1)$$

is clearly inadequate to show equivalence; one should actually disprove H_1, not H_o. This point has long been recognized in the biometric literature (Metzler, 1974). It is a logical difficulty that propagates to all levels of statistical work on the subject, including design (Westlake, 1974), estimation of sample size (Dilletti, Hauschke and Steinijans, 1991), and data analysis (Chow and Liu, 1992).

It has also been noted that a two-step statistical procedure based

on rejection of the null hypothesis H_o, or its acceptance if attended by high statistical power for a targeted alternative— the so called power approach, is flawed, since it cannot control the procedure size at the nominal level. For the chicory plants example, H_o is readily accepted and the power to detect a 10 percent difference (67.5% vs 57.5%) is 54.2%, that is to say, below the standard 80% power. Thus, these results could be construed as constituting lack of evidence in favor of equality, an unsettling state of affairs in view of the numbers in Table 1.

One approach to address equivalence stems from the idea that the experimental therapy may be as effective as the standard therapy but within certain limits. This is the concept of broad sense, or δ-equivalence. In statistical terminology:

$$H_o: |\pi_s - \pi_e| \geq \delta, \quad H_1: |\pi_s - \pi_e| < \delta, \quad (2).$$

This setup is also known as the role reversal approach because H_o is an interval hypothesis of inequivalence, eventually to be rejected in favor of equivalence. The value assigned to δ is arbitrary but should in principle be sensitive to the intuitive notion of equivalence.

Various tests and procedures have been proposed to test this type of equivalence. The Schuirmann procedure is often the standard choice in bioquivalence studies (Schuirmann, 1987). It is the procedure that we adopted here, extending the formulae to handle binary outcomes. In essence this is a two one-sided t tests based on H_o in (2).

For the chicory plants example, the Schuirmann procedure is significant ($p<0.02$). This is tantamount to rejecting the inequivalence between the two treatments in favor of their equivalence within 10 percent of the reference value (67.5%).

On the other hand, Blackwelder (1982) has argued that for efficacy, it is often the case that the question of interest is whether the new treatment is as effective, but perhaps no more effective, than the standard treatment. It may then be more meaningful to test for quasi-equivalence (also known as the δ-no-worse-than approach):

$$H_o: \pi_s - \pi_e \geq \delta, \quad H_1: \pi_s - \pi_e < \delta, \quad (3) \text{ (no absolute values involved).}$$

Not surprisingly, for the case at hand H_o in (3) is also rejected at $p<0.02$.

3. Multivariate equivalence

The ideas underlying δ-equivalence for one response carry over to situations involving two or more variables. The rationale for these multivariate extensions can be justified by the following example.

Consider an experiment involving a comparison of the simulated characteristics of two corn sowing machines under no-tillage soil conditions. The aim of the study was to compare the perfomance of a conventional machine (A) with a new machine (B) presumably as effective to sow, but less abrasive on soil physical properties. Certain basic aspects of sowing performance would render the machines equivalent for practical purposes. Assume that implantation of the seed and germination are regarded as the key features to accept that the machines are equivalent. Based on 500 seeds per system and after various intervening estimations, the resulting (hypothetical) data from implantation and germination can be summarized in the following Table.

Table 2
Performance of two sowing machines

		System				
	n	A	B	σ	z	p
%implantation	500	72.20	68.43	0.10	-2.16	<0.02
%germination	500	94.25	93.25	0.05	-2.61	<0.01

z: standard normal score from Blackwelder's test. p: p-value.

In this case, the desired notion of equivalence calls for testing both endpoints in a multivariate setting. The endpoints, however, may represent different scales of measurement.

One way of addressing this type of equivalence is via nominal α level adjustments for multiple endpoints. Recall that when each of two variables must show a statistically significant difference, then the nominal levels that shoud be used must be between 0.05 and 0.2236, inclusive, depending on whether they are perfectly correlated or independent, respectively (Offen and Helterbrand, 1996). This adjustment is therefore the inverse of the Bonferroni.

For the corn experiment, Table 2 shows that under hypotheses (3) and global α<0.02 system B is quasi-equivalent to A.

4. Discussion

The precise correspondence between classical hypothesis testing and the role reversal approach is still a matter of debate (Ng, 1996). What is beyond a doubt to the practitioner is the need for a statistical framework that accomodates equivalence as a concept different from efficacy. The vitality of equivalence as a topic of statistical interest is rooted in concrete problems, the bioequivalence of two medical drugs being one of the best known cases (Chow and Liu, 1992).

We presented some simple examples showing the relevance of equivalence to agricultural testing problems. There are of course many situations and extensions of these basic ideas that we did not cover. For example, our definitions of equivalence do not account for covariate adjustment.

Consider first the case of a study with multiple binary endpoints to show the therapeutical equivalence of lindane and ivermectin for the treatment of sheep scabies. Eight days after treatment application, one of the symptoms (A), not always present at baseline, showed the following evolution

Table 3
Status of symptom A

		Ivermectin	Lindane	*z*	*p*
Baseline	Sample Size	25	26		
	Symptom A	8 (32%)	10 (38.46%)	0.48	0.64
8 days	Sample Size	19	24		
	Symptom A	1 (5.26%)	1 (4.17%)	-0.17	0.87

z: standard normal score for the difference of proportions,
p: p-value.

It seems from the Table that the lindane treatment "equaled" the performance of ivermectin, but it did so from a more adverse status at baseline. A more meaningful notion of equivalence would take this difference into account by applying to a wider range of baseline conditions.

Consider now the case of a continuous dependent response Y, a covariate x with average value m within a designed range R of experimental conditions, and a treatment indicator I. In standard notation (see, for example, Neter et al, 1990).

$$Y_{ij} = \beta_0 + \beta_1 I_i + \beta_2 x_i + \epsilon_{ij}.$$

For positive δ, β_0, β_1 and m, and setting $I_i=1$ for the standard

treatment, the hypothesis of inequivalence then becomes

$$H_o: \beta_o + \beta_1 + \beta_2 m \geq (\beta_o + \beta_2 m) + \delta(\beta_o + \beta_1 + \beta_2 m), \text{ or alternatively}$$

$$H_o: \beta_1 - \delta(\beta_o + \beta_1 + \beta_2 m) \geq 0.$$

Rejection of H_o would indicate that the two treatments show equivalent performance within the scope of R.

Finally, a word of caution may be in order. Equivalence is conceptually different from equality. This holds true for the statistical analysis as well as the specifics of the subject matter under study. Furthermore, one should resist the temptation of appealing to equivalence procedures as a watered-down alternative to unsuccessful efficacy tests. The positive magnitude of the δ value may imply an admission that a treatment is to a certain extent inferior to the standard one. Statisticians might do well to alert experimenters about the implications of statistical equivalence for sizable values of δ.

References

_ Blackwelder, W. C. 1982. Proving the null hyphothesis in clinical trials. *Controlled Clinical Trials* 3: 345-353.

_ Chow, S. C., and Liu, J. P. 1992. *Design and analysis of bioavailability and equivalence studies*. Marcel Dekker, pp. 19-21.

_ Dilletti, E.; Hauschke, D.; and Steinijans, V. W. 1991. Sample size determination for bioequivalence assessment by means of confidence interval. *International Journal of Clinical Pharmacology, Therapy and Toxicology,* 29:1-8.

_ Metzler, C. M. 1974. Bioavailability: A problem in equivalence. *Biometrics*: 30, 309-317.

_ Neter, J.; Wasserman , W.; and Kutner, M. H. 1990. *Applied linear statistical models.* Third Edition. Irwin, pp 861-898.

_ Ng. T-H, 1996. Reply to letter to the editor. *Controlled Clinical Trials*, 17: 349-350.

_ Offen W. W., and Helterbrand, J. D. 1996. Strategies for nominal alpha level adjustments for multiple endpoints. *Biopharmaceutical Report*, 4: 11.

_ Schuirmann, D. J. 1987. A comparison of the two one-sided test procedures and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics,* 15: 657-680.

_ Westlake, W. J. 1974. The use of balanced incomplete block designs in comparative bioavailability trials. *Biometrics,* 30: 319 - 327.