Conference on Applied Statistics in Agriculture 1997 - 9th Annual Conference Proceedings

EQUIVALENCE TESTING IN AGRICULTURE EXPERIMENTS

Brian J. Fergen

Equivalence Testing in Agriculture Experiments

Brian J. Fergen
Pfizer Animal Health Group

Abstract

Equivalence testing is a relatively new area of research in statistics. It's development has been motivated in large part by the need for statistical methods for determining if generic drugs are bioequivalent to their name brand counterparts. The application of equivalence testing methods to data resulting from experiments and surveys unrelated to drug development, and in particular agriculture-related experiments, is infrequent and possibly non-existent. These methods provide useful alternatives to the analysis methods currently being used. In this paper, an overview of the philosophy of equivalence testing and a review of equivalence testing methods are presented. Additionally, experimental situations for which equivalence testing would be appropriate are discussed. Examples that illustrate the application of the philosphy of equivalence testing to experimental designs commonly used in agriculture research are also presented.

Keywords: Equivalence Testing, Completely Randomized Design, Split-plot Design

1 Introduction

The analysis of data resulting from experimental designs commonly used in agriculture typically proceed according to the hypothesis testing structure corresponding to the Analysis of Variance (ANOVA), General Linear Mixed Models (GLMM) or Generalized Linear Models (GLM) analysis methods. In this paper, the application of any of these methods and their related hypothesis testing structure is referred to as a **traditional analysis**. In many situations, a traditional analysis is appropriate, and the results from the hypothesis tests provide information that is relevant to the objectives of the study. However, for this to be the case, the objectives of the study must be in agreement with the hypothesis testing structure related to the analysis method used. If the objectives of the study are not in agreement with the hypothesis tests conducted, the conclusions are likely incorrect.

A common situation where the hypothesis testing structure is not appropriate is when the objectives involve demonstrating the comparability or equivalence of parameters. In this situation neither the ANOVA, GLMM or GLM hypothesis testing approaches are appropriate. It is this situation that is commonly referred to as equivalence testing.

The topics discussed in this paper: what is equivalence testing; why use equivalence testing; current equivalence testing methods, and examples of equivalence testing; provide a brief history of equivalence testing, the motivation for using equivalence testing, and a broad overview of

equivalence testing methods. In addition, the examples provide a bridge between the philosophy of equivalence testing and its application to agriculture experiments.

2 What is Equivalence Testing

As an introduction to equivalence testing, consider the application of equivalence testing that has motivated most of the research in equivalence testing - bioequivalence. Bioequivalence is the assessment of the comparability (the words equivalence and comparability are used interchangeably throughout this paper) of the bioavailabilty of two formulations of a drug, such as Name Brand and Generic formulations. The interest in bioequivalence reflects the need for appropriate statistical methods to assess the equivalence of parameters indicative of the therapeutic effect of a drug. While the literature for equivalence testing dates back at least as far as Bondy (1969), Westlake (1972) and Metzler (1974) introduce the concept of equivalence testing in the setting of bioavailablilty - bioequivalence. Indeed, the majority of publications dealing with equivalence testing are motivated by and relate to bioequivalence. Thus it is appropriate to introduce equivalence testing by considering bioequivalence.

Let μ_G = average response for a Generic drug, and μ_{NB} = average response for the Name Brand drug, for a response considered indicative of the therapeutic effect of the drug. To conclude the drugs are bioequivalent, it must be demonstrated that μ_G and μ_{NB} are comparable. The Food and Drug Administration has determined that two formulations can be declared bioequivalent if it can be demonstrated that the ratio of Generic to Name Brand average responses is greater than 0.8 and less than 1.25. This requirement can be expressed via null and alternative hypotheses as

H_0: $0.8 \geq \frac{\mu_G}{\mu_{NB}}$ or $\frac{\mu_G}{\mu_{NB}} \geq 1.25$ vs H_A: $0.8 < \frac{\mu_G}{\mu_{NB}} < 1.25$. For the analysis of data from a bioequivalence trial, it is assumed that the null hypothesis is true. The objective of the study is then to demonstrate the alternative hypothesis is true, i.e. the two formulations are bioequivalent. This is the typical manner in which hypothesis testing is conducted, assume the null hypothesis is true, and demonstrate the research objective by demonstrating the alternative hypothesis is true. To model the physiological process involved in the assessment of bioavailabilty, the analysis is conducted on the log transformed data. Under the assumption of equal formulation variances, this is equivalent to testing

H_0: $-\Delta \geq \delta_G - \delta_{NB}$ or $\delta_G - \delta_{NB} \geq \Delta$ vs H_A: $-\Delta < \delta_G - \delta_{NB} < \Delta$, where $\delta_i = \ln(\mu_i)$, and

$\Delta = \ln(1.25)$. The key to a bioequivalence test, and what differentiates it from hypothesis tests conducted in traditional analyses, is that the **alternative** hypothesis indicates the comparability of μ_G and μ_{NB}.

It is also reasonable to expect that the concept of equivalence is appropriate in some agriculture research. As an example, consider a Canada Thistle Study to determine if two eradication methods result in comparable canada thistle counts two years after application. The methods

compared are the Standard Method (SM) - the application of a common herbicide in conjunction with tilling, and a New Method (NM) - mowing in conjunction with tilling. The study design has a One-Way Treatment Structure, where the treatments are the New and Standard Methods of eradication, with a Randomized Complete Block Design Structure, where the blocks are various locations in fields. The experimental unit is a plot of constant size and the response of interest is the number of thistles on a plot 2 years after the application of an eradication method.

An appropriate model is $y_{ij} = \mu_i + b_j + e_{ij}$, i = SM, NM; j = 1, 2, ..., k, where μ_i = average count for eradication method i, b_j = random effect due to location j ($E(b_j) = 0$, $V(b_j) = \sigma_b^2$), and e_{ij} = random error associated with treatment i in block j ($E(e_{ij}) = 0$, $V(e_{ij}) = \sigma_i^2$). Typically, it is assumed that the random components of the model are independent (all b_j and e_{ij} are independent).

The objective of the study is to determine if the eradication methods result in comparable thistle counts. It might be appropriate to assess this objective by

testing $H_0: \Delta_L \geq \frac{\mu_{NM}}{\mu_{SM}}$ or $\frac{\mu_{NM}}{\mu_{SM}} \geq \Delta_U$ vs $H_A: \Delta_L < \frac{\mu_{NM}}{\mu_{SM}} < \Delta_U$ or alternatively by

testing $H_0: \Delta_L^* \geq \mu_{NM} - \mu_{SM}$ or $\mu_{NM} - \mu_{SM} \geq \Delta_U^*$ vs $H_A: \Delta_L^* < \mu_{NM} - \mu_{SM} < \Delta_U^*$. If the assessment of the comparability of the variability of the eradication methods is important,

$H_0: \Delta_L^+ \geq \frac{\sigma_{NM}^2}{\sigma_{SM}^2}$ or $\frac{\sigma_{NM}^2}{\sigma_{SM}^2} \geq \Delta_U^+$ vs $H_A: \Delta_L^+ < \frac{\sigma_{NM}^2}{\sigma_{SM}^2} < \Delta_U^+$ could also be tested. It may also be reasonable to assess the objectives of the study by determining the comparability of both the means and variances of the eradication methods. Regardless of the hypothesis test(s) used, it is appropriate to specify the comparability of the parameters in terms of a lower and upper limit (Δ_L and Δ_U) and, through the hypothesis testing process, attempt to demonstrate that the alternative hypothesis is true and the eradication methods are comparable.

It is important to note that the objective is not to demonstrate the superiority of NM to SM, but the comparability of the methods. Interest in this objective may be related to cost of NM vs SM, an ecological benefit associated with eliminating the use of the herbicide, and/or a variety of other reasons. Whatever the underlying reasons might be, establishing that the the eradication methods are comparable provides a basis from which the underlying reason can be put forth by the researcher as the justification for switching to the New Method. If the traditional approach had been used, failure to reject the null hypothesis does not demonstrate that the methods are comparable.

At this point, equivalence testing is formally defined. Equivalence testing is the statistical assessment of the comparability of functions of parameters related to the distribution of random

variables, or of properties of the distribution of probabilities related to the comparability of random variables. This assessment requires a guideline (e.g., Δ_L and Δ_U) to be used in the determination of equivalence where the interval (Δ_L, Δ_U) is referred to as the equivalence interval.

3 Why Use Equivalence Testing

The primary reason to use equivalence testing is that it is correctly assessing the objectives of the experiment. This is reflected in the performance of the decision rules corresponding to the hypothesis tests for an equivalence analysis and a traditional analysis, as indicated in Schuirmann (1987). This performance is now illustrated for the Canada Thistle Study previously described.

Recall the model for the Canada Thistle Study: $y_{ij} = \mu_i + b_j + e_{ij}$, i = SM, NM; j = 1, 2, ..., k. Under the assumption that e_{ij} is normally distributed, $\mu_{SM} = 100$, k = 10 (10 blocks), $\alpha = 0.05$, $\sigma^2_{SM} = \sigma^2_{NM}$, $\Delta_L = -10$ and $\Delta_U = 10$, the performance of the decision rules corresponding to the traditional and equivalence analyses can be graphically illustrated. These assumptions are not neccessary for the relationship exhibited to hold, but are made to allow for the graphical representation that follows.

First, consider the performance of the decision rule appropriate for the equivalence analysis,

H_0: $\Delta_L \geq \mu_{NM} - \mu_{SM}$ or $\mu_{NM} - \mu_{SM} \geq \Delta_U$ vs H_A: $\Delta_L < \mu_{NM} - \mu_{SM} < \Delta_U$, where

H_A: $\Delta_L < \mu_{NM} - \mu_{SM} < \Delta_U$ indicates equivalence. This is illustrated in Figure 1 for combinations of the estimated difference between μ_{NM} and μ_{SM} and the estimated standard error of the difference between μ_{NM} and μ_{SM}. There are two properties to note. First, the conclusion of equivalence never occurs if the estimated difference in μ_{NM} and μ_{SM} is either greater than 10 (Δ_U) or less than -10 (Δ_L). Second, for a given estimated difference falling between 10 and -10, as the estimated standard error of the difference increases from 0, the decision rule indicates that the hypothesis testing result will change from a conclusion of equivalence to a conclusion of inequivalence. Both of these properties are appropriate, the first as a consequence of the objective of the trial being comparability, and the second as an illustration of the role variability plays in making decisions.

Next, consider the performance of the decision rule appropriate for the traditional analysis, H_0: $\mu_{NM} = \mu_{SM}$ vs H_A: $\mu_{NM} \neq \mu_{SM}$, where H_0: $\mu_{NM} = \mu_{SM}$ indicates equivalence. (Clearly, by the definition of equivalence, neither of the hypotheses from the traditional analysis are appropriate when the study objectives involve demonstrating μ_{NM} and μ_{SM} are comparable. But, if the traditional analysis is used, the null hypothesis is more indicative of equivalence than the alternative hypothesis.) Consider the two properties mentioned for the equivalence analysis, as illustrated in Figure 2 for combinations of the estimated difference of μ_{NM} and μ_{SM} and the estimated standard error of the difference between μ_{NM} and μ_{SM}. First, the conclusion of

equivalence can occur for *any* estimated difference in μ_{NM} and μ_{SM}! Obviously, this is not a desirable property for the hypothesis test as it in no way indicates that the eradication means are comparable. Second, for a given estimated difference, as the estimated standard error of the difference increases from 0, the decision rule indicates that the hypothesis testing result will eventually change from a conclusion of inequivalence to a conclusion of equivalence. This is the exact opposite of the equivalence analysis and is also not a desirable property. The performance of the traditional analysis should not be a suprise, as it is not addressing the objectives of the study.

4 Current Equivalence Testing Methods

Next, the framework needed to categorize and describe the current equivalence testing methods is presented. Appropriate modifications have been made to generalize the categories beyond the scope of bioequivalence Hauck (1997) presented.

To begin, specify a general linear mixed model similar to the model presented by Henderson (1984). Let $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{U} + \epsilon$, where $\mathbf{y}$ is a vector of measured responses, $\mathbf{X}$ is a known matrix, β is a vector of unknown fixed effects, $\mathbf{Z}$ is a known matrix, $\mathbf{U}$ is a vector of random effects with $E(\mathbf{U}) = \mathbf{0}$, ϵ is a random vector with $E(\epsilon) = \mathbf{0}$ and $\mathrm{Var}\begin{bmatrix}\mathbf{U}\\ \epsilon\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$, where $\mathbf{R}$ and $\mathbf{G}$ are positive definite covariance matrices. Under these assumptions, $E(\mathbf{y}) = \mathbf{X}\beta$ and $V(\mathbf{y}) = \Sigma = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, where the covariance parameters in Σ are denoted by θ. (Readers interested in a more detailed description are referred to Henderson.) Using this model for the Canada Thistle Study, results in

$$\beta = \begin{bmatrix}\mu_{NM}\\ \mu_{SM}\end{bmatrix} \text{ and } \theta = \begin{bmatrix}\sigma_b^2\\ \sigma_{SM}^2\\ \sigma_{NM}^2\end{bmatrix}.$$

Equivalence tests can be classified as either moment-based or probability-based. Moment-based equivalence tests test hypotheses which are a function of the parameters (moments) of the distribution of $\mathbf{y}$. Moment-based equivalence tests can be further classified into aggregate and disaggregate methods.

An aggregate moment-based equivalence test involves a single hypothesis test which encompasses all parameters (from β and θ) of interest. The test of equivalence is dependent upon this hypothesis test, which can be expressed as

$$H_0: \Delta_L \geq \mathrm{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} \text{ or } \mathrm{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} \geq \Delta_U \quad \text{vs} \quad H_A: \Delta_L < \mathrm{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} < \Delta_U.$$

An example of an aggregate moment-based test presented previously is

H_0: $\Delta_L \geq \frac{\mu_{NM}}{\mu_{SM}}$ or $\frac{\mu_{NM}}{\mu_{SM}} \geq \Delta_U$ vs H_A: $\Delta_L < \frac{\mu_{NM}}{\mu_{SM}} < \Delta_U$. A second example is

H_0: $\Delta_L \geq \lambda$ or $\lambda \geq \Delta_U$ vs H_A: $\Delta_L < \lambda < \Delta_U$, where

$\lambda = \frac{(\mu_{NM} - \mu_{SM})^2 + (\sigma^2_{NM} + \sigma^2_{SM})}{2\sigma^2_{SM}}$. In this situation, $\lambda = \frac{E[(y_{NMj} - y_{SMj})^2]}{2V^*(y_{SMj})}$

where $V^*(y_{SMj})$ does not include block-to-block variability. For this hypothesis test, a lower bound is not necessary, but rather than create futher classifications, Δ_L can by definition be set to 0, thus continuing with the use of an equivalence interval.

A disaggregate moment-based equivalence test involves multiple hypothesis tests, where each hypothesis test involves subsets of the parameters (from β and θ) of interest. Thus, the equivalence testing process is dependent upon k hypothesis tests ($k > 1$), which can be expressed as

$$H_{01}: \Delta_{L1} \geq \text{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} \text{ or } \text{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} \geq \Delta_{U1} \quad \text{vs} \quad H_{A1}: \Delta_{L1} < \text{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} < \Delta_{U1}$$

$$\vdots$$

$$H_{0K}: \Delta_{LK} \geq \text{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} \text{ or } \text{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} \geq \Delta_{UK} \quad \text{vs} \quad H_{AK}: \Delta_{LK} < \text{fnc}\begin{pmatrix}\beta\\ \theta\end{pmatrix} < \Delta_{UK}.$$

Equivalence is concluded if all null hypotheses are rejected and partial equivalence can be concluded if at least one null hypothesis is rejected.

An example of a disaggregate moment-based equivalence test is

$$H_{01}: \Delta_{L1} \geq \mu_{NM} - \mu_{SM} \text{ or } \mu_{NM} - \mu_{SM} \geq \Delta_{U1} \quad \text{vs} \quad H_{A1}: \Delta_{L1} < \mu_{NM} - \mu_{SM} < \Delta_{U1}, \text{ and}$$

$$H_{02}: \Delta_{L2} \geq \frac{\sigma^2_{NM}}{\sigma^2_{SM}} \text{ or } \frac{\sigma^2_{NM}}{\sigma^2_{SM}} \geq \Delta_{U2} \quad \text{vs} \quad H_{A2}: \Delta_{L2} < \frac{\sigma^2_{NM}}{\sigma^2_{SM}} < \Delta_{U2}.$$

Note that if the block effects (b_j) and random errors (e_{ij}) for the randomized complete block design are normally distributed, a conclusion of equivalence for the disaggregate moment-based equivalence test example indicates a conclusion of equivalence in the distributions of the two eradication methods as the normal distribution is completely specifed by its mean and variance.

Probability-based equivalence tests involve demonstrating with a sufficiently high probability

that measured responses for treatments are comparable. For example, consider P_j = P(thistle count for the Standard and New Methods satisfy a 'comparability criteria' at location j), e.g. P_j=P($\Delta_L < y_{NMj} - y_{SMj} < \Delta_U$). We can then further classify probability-based equivalence tests as either tolerance interval or expectation methods.

The expectation probability-based methods involve demonstrating $E(P_j) > \tau$. If $\mu_P = E(P_j)$, then an appropriate hypothesis test is H_0: $\mu_P \leq \tau$ versus H_A: $\mu_P > \tau$.

The tolerance interval probability-based methods involve demonstrating $P(P_j > \tau`) > \omega$. If $P_{\tau`} = Pr(P_j > \tau`)$, then an appropriate hypothesis test is $H_0$: $P_{\tau`} \leq \omega$ versus H_A: $P_{\tau`} > \omega$.

5 Examples

Having given an overview of equivalence testing in the previous sections, it is appropriate to address the question of extending the concept of equivalence testing to various treatment and design structures. The literature does not address this question for designs other than the two-period crossover design and the four-period, two-treatment replicate designs such as those described by Hauck(1997). An obvious place to begin this discussion is with the most basic design, a one-way treatment structure in a completely randomized design structure. The second example is a two-way treatment structure in a split-plot design with completely randomized design structure for the whole plot and sub plot experimental units. The philosophy and ideas illustrated in these two examples can be generalized to most common experimental designs used by agriculturists.

One-Way Treatment Structure in a Completely Randomized Design Structure

Consider the means model described in Milliken and Johnson (1992, Chapter 1): $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $E(\mathbf{y}) = \mathbf{X}\beta$, $V(\mathbf{y}) = \Sigma = \sigma^2\mathbf{I}$, $\beta' = [\mu_1\ \mu_2 \ldots \mu_t]$ and $\theta = [\sigma^2]$. Though there are typically several hypotheses related to the objectives of the study that are of interest, frequently the null hypothesis of equality of treatment or population means is the initial hypothesis tested when the objective of the study is to detect differences among the t treatment means (the traditional analysis). Here, assume the objective of the study is to assess the comparability of the t treatment means and to detect treatment equivalences, i.e., the situation where two or more treatment means are comparable.

In trying to relate this objective to a hypothesis test, it is illustrative to consider the initial hypothesis test in the traditional analysis, the test of equality of treatment means. Here, the null and alternative hypotheses can be expressed as H_0: $\mu_1=\mu_2=\ldots=\mu_t$ vs H_A: $\mu_i \neq \mu_i$ for some $i \neq j$. The process of testing this hypothesis involves the use of a t-1 by t matrix of contrasts, such as

$$\mathbf{H} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & -2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & -3 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & \vdots \\ 1 & 1 & 1 & 1 & . & . & . & (t-1) \end{bmatrix}. \text{ Then } \mathbf{H\beta} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 + \mu_2 - 2\mu_3 \\ \mu_1 + \mu_2 + \mu_3 - 3\mu_4 \\ \vdots \\ \mu_1 + \ldots \mu_{t-1} - (t-1)\mu_t \end{bmatrix} \text{ and}$$

the previous hypotheses can be expressed as H_0: $\mathbf{H\beta} = \mathbf{0}$ vs H_A: $\mathbf{H\beta} \neq 0$

The first step in an equivalence analysis is the correct specification of the null and alternative hypotheses. In the traditional analysis, the objective is to demonstrate there is a least one pair of population means that are different, thus the null hypothesis specifies that all population means are equal. For an equivalence analysis, the objective is to demonstrate that at least one pair of the population means is equivalent, thus the null hypothesis specifies that all pairs of population means are inequivalent. The correct formulation for the null and alternative hypotheses is then H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$ vs H_A: $\Delta_L < \mu_i - \mu_j < \Delta_U$ for at least one $i \neq j$.

For simplicity (and because it is reasonable in many equivalence testing situations), assume $\Delta_L = -\Delta_U$. Intuitively, it appears reasonable to reformulate the null and alternative hypotheses for the equivalence analysis using the contrast matrix $\mathbf{H}$ previously defined and vectors $\mathbf{\Delta_L}$ and $\mathbf{\Delta_U}$, where the elements of $\mathbf{\Delta_L}$ and $\mathbf{\Delta_U}$ are appropriate upper and lower equivalence intervals for the linear contrasts represented by the elements of $\mathbf{H\beta}$. $\mathbf{\Delta_L}$ and $\mathbf{\Delta_U}$ are the lower and upper equivalence vectors, respectively. Then, the previous set of hypotheses can be expressed as H_0: $\mathbf{\Delta_L} \geq \mathbf{H\beta}$ or $\mathbf{H\beta} \geq \mathbf{\Delta_U}$ vs H_A: $\mathbf{\Delta_L} < \mathbf{H\beta} < \mathbf{\Delta_U}$, where H_A: $\mathbf{\Delta_L} < \mathbf{H\beta} < \mathbf{\Delta_U}$ indicates that at least one of the elements of $\mathbf{H\beta}$ is within the corresponding equivalence interval from $\mathbf{\Delta_L}$ and $\mathbf{\Delta_U}$. The selection of the equivalence vectors should provide a mapping from H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$ to H_0: $\mathbf{\Delta_L} \geq \mathbf{H\beta}$ or $\mathbf{H\beta} \geq \mathbf{\Delta_U}$. In particular, one might

select $\mathbf{\Delta_L} = \Delta_L \begin{bmatrix} (t-1) \\ (t-1)+(t-2) \\ \vdots \\ \sum_{i=1}^{t} (i-1) \end{bmatrix}$ and $\mathbf{\Delta_U} = \Delta_U \begin{bmatrix} (t-1) \\ (t-1)+(t-2) \\ \vdots \\ \sum_{i=1}^{t} (i-1) \end{bmatrix}$ and assess this mapping. (The

elements of these vectors can be derived by considering the relationship among the population means for all combinations of inequivalence under the null hypothesis and relating those restrictions to the linear contrasts in $\mathbf{H\beta}$.)

Unfortunately this mapping is not one-to-one except in the trivial situation where t=2. This occurs because the constraints represented by H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$ can not be replicated by the t-1 degrees of freedom (t-1 constraints) corresponding to the rank of $\mathbf{H}$. This result can be clearly illustrated using an example with a completely randomized design structure, a one-way treatment structure with t=3 populations, and $\Delta_U = 5$.

Using the matrix of contrasts defined previously, the two elements of $\mathbf{H}\beta$ are $\mu_1 - \mu_2$ and $\mu_1 + \mu_2 - 2\mu_3$, and the elements of Δ_U are 10 and 15 (the elements of Δ_L are -10 and -15). If $\mu_1 < \mu_3 < \mu_2$, and each mean differs by at least 5 (Δ_U), then $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$. This situation illustrate why it is appropriate to consider ± 10 ($2*\Delta_U$) for the equivalence interval on $\mu_1 - \mu_2$ from $\mathbf{H}\beta$.

Next, consider $\mu_1 = 55$, $\mu_2 = 65$ and $\mu_3 = 60$, here again $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$, but $\mu_1 + \mu_2 - 2\mu_3 = 0$ is <15 ($3*\Delta_U$) and > -15 ($3*\Delta_L$) . Hence H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$ incorrectly maps into H_A: $\Delta_L < \mathbf{H}\beta < \Delta_U$. Similarly, consider $\mu_1 = 30$, $\mu_2 = 60$ and $\mu_3 = 64$, then $\Delta_L < \mu_i - \mu_j < \Delta_U$ for at least one $i \neq j$ ($\mu_2 - \mu_3 = 4$), but $\mu_1 - \mu_2 = 30 > 10$ and $\mu_1 + \mu_2 - 2\mu_3 = 38 > 15$, thus $\mathbf{H}\beta \geq \Delta_U$, and H_A: $\Delta_L < \mu_i - \mu_j < \Delta_U$ for at least one $i \neq j$ incorrectly maps into H_0: $\Delta_L \geq \mathbf{H}\beta$ or $\mathbf{H}\beta \geq \Delta_U$. The problem is related to the order of the population means and the magnitude of their differences, e.g. when $\mu_1 < \mu_3 < \mu_2$ a one-to-one mapping from H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$ to H_0: $\Delta_L \geq \mathbf{H}\beta$ or $\mathbf{H}\beta \geq \Delta_U$ doesn't not exist for all possible combinations of μ_i's satisfying H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$.

Thus the extension of the use of equivalence intervals for contrasts of the parameters ($\mathbf{H}\beta$) is not, in general, acceptable for assessing H_0: $\Delta_L \geq \mu_i - \mu_j$ or $\mu_i - \mu_j \geq \Delta_U$ for all $i \neq j$ vs H_A: $\Delta_L < \mu_i - \mu_j < \Delta_U$ for at least one $i \neq j$. However, under the assumption that $\Delta_L = -\Delta_U$, the null and alternative hypothesis are equivalent to H_0: $|\mu_i - \mu_j| \geq \Delta_U$ for all $i \neq j$ vs H_A: $|\mu_i - \mu_j| < \Delta_U$ for at least one $i \neq j$, which is equivalent to H_0: $\min(|\mu_i - \mu_j|) \geq \Delta_U$ vs H_A: $\min(|\mu_i - \mu_j|) < \Delta_U$, where $\min(|\mu_i - \mu_j|)$ is the smallest absolute difference among all possible unique pairs of the t treatment means. Hence, an assessment of the comparability of the t treatments can be conducted under the hypothesis test H_0: $\min(|\mu_i - \mu_j|) \geq \Delta_U$ for all $i \neq j$ vs H_A: $\min(|\mu_i - \mu_j|) < \Delta_U$ for all $i \neq j$.

Two-Way Treatment Structure in a Split-Plot Design Structure

Next, consider a two-way treatment structure in a split-plot design using the means model as described in Milliken and Johnson (Chapter 5, Section 2). For this example, consider an extension of the Canada Thistle Study previously introduced. For this study, the 3 whole-plot treatments are the following eradication methods:

Treatment 1: Application of a Common Herbicide in Conjunction with Tilling
Treatment 2: Mowing in Conjunction with Tilling
Treatment 3: Application of a Common Herbicide

The 3 sub-plot treatments are:

Treatment 1: Intermediate Wheatgrass
Treatment 2: Western Wheatgrass
Treatment 3: Control

The intermediate and western wheatgrass treatments are the seeding of wheatgrass in addition to whatever treatment is applied to the whole-plot experimental unit, and the control treatment is equivalent to doing nothing in addition to whatever treatment is applied to the whole-plot

experimental unit. The response for the experiment is the thistle count two years after the attempted eradication. It was hypothesized that the wheatgrasses would provide additional supression of the thistles. The primary study objectives are to assess equivalence among eradication methods in combination with the two wheatgrass varieties and to assess if these treatment combinations are superior to the eradication method - control treatment combinations.

An appropriate model is $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{U} + \epsilon$, where $\beta' = \left[\mu_{11}\ \mu_{12}\ \mu_{13}\ \mu_{21} \cdots \mu_{33}\right]$

and $\theta' = \left[\sigma^2_{wp}\ \sigma^2_{sp}\right]$. Here μ_{ij} = average count for whole-plot treatment i and subplot treatment j,

σ^2_{wp} is the variability among whole-plot experimental units, and σ^2_{sp} is the variability among sub-plot experimental units. Further description of **X**, **Z**, **U** and ϵ is provided by Milliken and Johnson(1992).

One approach to the analysis of the data from this study is to first determine if the wheatgrass variety - eradication method treatment combinations are superior to the eradication method - control treatment combinations and to then determine if there is equivalence among eradication method - wheatgrass variety treatment combinations. Using this analysis approach, the first step in the analysis is to determine an appropriate hypothesis test for testing the superiority of the wheatgrass varieties versus control. To accomplish this step, H_0: $\min(\mu_{ij}, i=1,2,3; j=1,2) \leq \max(\mu_{13}, \mu_{23}, \mu_{33})$ vs H_A: $\min(\mu_{ij}, i=1,2,3; j=1,2) > \max(\mu_{13}, \mu_{23}, \mu_{33})$ could be tested. This hypothesis test assesses if intermediate wheatgrass ($\mu_{11,}\ \mu_{21}, \mu_{31}$) and western wheatgrass ($\mu_{12,}$ μ_{22}, μ_{32}) provide superior thistle suppression to control ($\mu_{13,}\ \mu_{23}, \mu_{33}$). This would be a reasonable hypothesis to test under the assumption that the population means for the eradication method - wheatgrass variety treatment combinations are equivalent. However, this may not be a reasonable assumption.

If one is not willing to assume the population means for the eradication method - grass variety treatment combinations are equivalent or if the null hypothesis in the previous hypothesis test is not rejected, the next step in the analysis could be to test for the superiority of the wheatgrass varieties versus control for each whole-plot treatment, i.e. test H_{0i}: $\min(\mu_{i1}, \mu_{i2}) \leq \mu_{i3}$ vs H_{ai}: $\min(\mu_{i1}, \mu_{i2}) > \mu_{i3}$ for i=1,2,3.

Having addressed the objective of superiority of the wheatgrass varieties to control, it is appropriate to assess equivalence among the eradication and wheatgrass variety treatment combinations. As a starting point in determining what an appropriate first step might be, consider the traditional analysis approach. The first step in the the traditional analysis is a test for interaction. (For this experiment, this test would not involve the control - eradication treatment combinations as there is no interest in including them in the assessment of equivalence.) The interaction hypothesis from a traditional analysis can be expressed as H_0: $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$ for all $i \neq i'$ and $j \neq j'$ vs H_A: $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \neq 0$ for at least one

$i \neq i'$ and $j \neq j'$. For this example, if $\mathbf{H} = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & -1 & 0 & -2 & 2 & 0 \end{bmatrix}$,

then $\mathbf{H\beta} = \begin{bmatrix} (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) \\ (\mu_{11} - \mu_{12} + \mu_{21} - \mu_{22}) - 2(\mu_{31} - \mu_{32}) \end{bmatrix}$, and the previous hypotheses can be expressed as

H_0: $\mathbf{H\beta} = \mathbf{0}$ versus H_A: $\mathbf{H\beta} \neq 0$.

An analagous first step for an equivalence analysis is to test H_0: $\Delta_L < \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} < \Delta_U$ for all $i \neq i'$ and $j \neq j'$ vs H_A: $\Delta_L \geq \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}$ or $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \geq \Delta_U$ for at least one $i \neq i'$ and $j \neq j'$. These hypotheses arise naturally from the combination of the equivalence testing and interaction concepts, thus, the related hypothesis test is refferred to as a test for equivalence interaction. In the traditional analysis for interaction, no interaction is assumed (H_0: $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$ for all $i \neq i'$ and $j \neq j'$) and a hypothesis test is conducted based on this assumption. In the equivalence analysis for equivalence interaction, no equivalence interaction (H_0: $\Delta_L < \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} < \Delta_U$ for all $i \neq i'$ and $j \neq j'$) is assumed, and a hypothesis test based on this assumption is conducted.

As with the previous example, when the rank of **H** is greater than 1, it is not possible to construct a hypothesis test addressing equivalence interaction using the contrasts in $\mathbf{H\beta}$ that provides a one-to-one mapping with H_0: $\Delta_L < \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} < \Delta_U$. However, note that $\Delta_L < \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} < \Delta_U$ for all $i \neq i'$ and $j \neq j'$ if and only if $\max(\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}) < \Delta_U$ and $\Delta_L < \min(\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'})$, where $\max(\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'})$ is the maximum value of $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}$ when considering all appropriate $i \neq i'$ and $j \neq j'$ and $\min(\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'})$ is the minimum value of $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}$ when considering all appropriate $i \neq i'$ and $j \neq j'$. Here again, in most situations it is reasonable to assume $\Delta_L = -\Delta_U$, and under this assumption, $\max(\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}) < \Delta_U$ and $\Delta_L < \min(\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'})$ if and only if $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|) < \Delta_U$ where $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|)$ is the maximum value of $|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|$ when considering all appropriate $i \neq i'$ and $j \neq j'$, thus the equivalence interaction hypotheses can be restated as H_0: $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|) < \Delta_U$ vs H_A: $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|) \geq \Delta_U$, where $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|)$ is the maximum value of $|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|$ when considering all appropriate $i \neq i'$ and $j \neq j'$.

In most experimental designs involving a two-way treatment structure, the researcher would like (if appropriate) to compare one set of treatments after averaging over the second set of treatments, e.g., for a split-plot design the researcher would like to compare the sub-plot treatments after averaging over the whole-plot treatments. If H_0: $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|) < \Delta_U$ is not rejected, it is appropriate to assess the equivalence of the whole-plot and sub-plot treatments (seperately) using the equivalence testing procedure presented for the one-way treatment structure in a completely randomized design structure. Hypothesis tests to assess the comparability of the whole-plot treatment means in order to detect whole-plot treatment equivalences, and to assess the comparability of the sub-plot treatment means in order to detect

sub-plot treatment equivalences can be conducted and interpretations can be made independently. If H_0: $\max(|\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}|) < \Delta_U$ is rejected, the equivalences among one set of treatments may depend on the level of the second set of treatments. Hypothesis tests to assess the comparability of whole-plot treatments and the comparability of sub-plot treatments can be conducted, but the interpretation of the results of these hypothesis tests may not be unrelated (or even meaningful).

6 Summary

The use of equivalence testing, though not commonplace if occurring at all, is warranted in some experimental situations involving research in agriculture. There exists a wide variety of equivalence testing methods, but few (if any) have been generalized for use in designs other than the two-period crossover design and the four-period, two-treatment replicate designs commonly used in bioequivalence trials. This paper develops the philosophy and illustrates the application of that philosophy to two common experimental designs. These examples can be generalized to most common experimental designs used in agriculture experiments. However, there are several areas of research related to equivalence testing that need to be addressed before equivalence testing can successfully be used for the wide variety of designs employed in agricultural research.

First, equivalence testing methodology based on distributional assumptions for $\mathbf{y}$, or appropriate nonparametric methods must be developed. At the same time, methodology to address the overall or experimentwise error rate associated with the multiplicity of pairwise equivalence assessments (the equivalence analysis counterpart to pairwise comparisons in a traditional analysis) must be developed.

Additionally, guidelines need to be developed to assist researchers in understanding and choosing appropriate Δ_L and Δ_U and appropriate equivalence testing methods as well as to delineate how to apply equivalence testing methods when interested in various inference spaces such as discussed by McLean *et all* (1991).

After these initial research efforts, extensions of equivalence testing methods to all GLMMs, GLMs, group sequential tests and the field of covariance analysis will broaden the application of equivalence testing.

The development of equivalence testing methods that are applicable to agricultural research and the simultaneous production of examples, guidelines and explanations to assist the researchers involved in this research will greatly benefit the efficiency and effectiveness of this research.

Equivalence and Inequivalence Regions for Equivalence and Traditional Analyses
SM Mean= 100, Number Blocks= 10, Type I Error Rate= 0.05

Figure 1
Equivalence Test

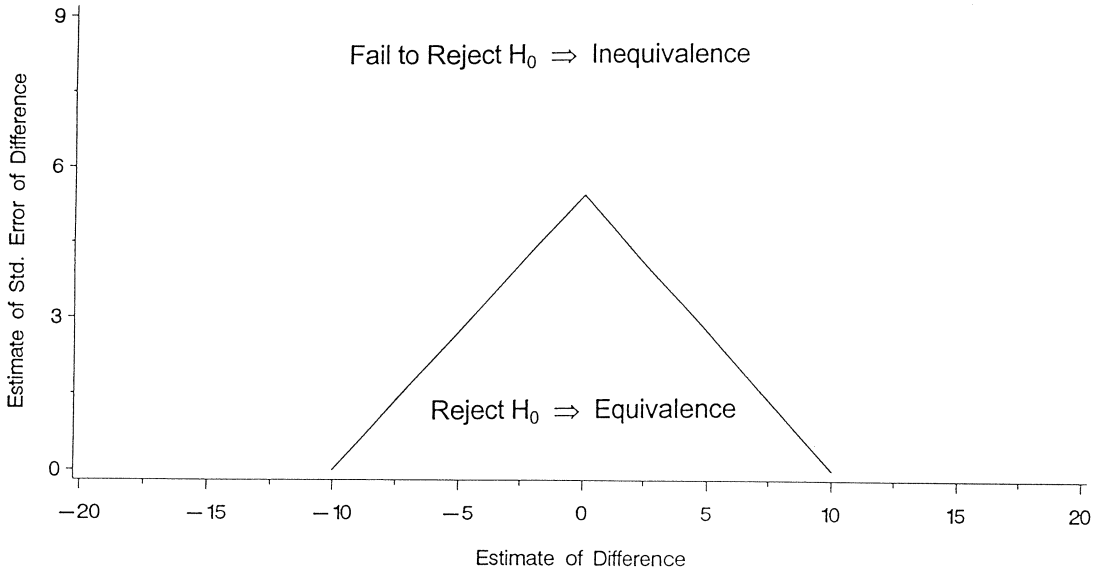
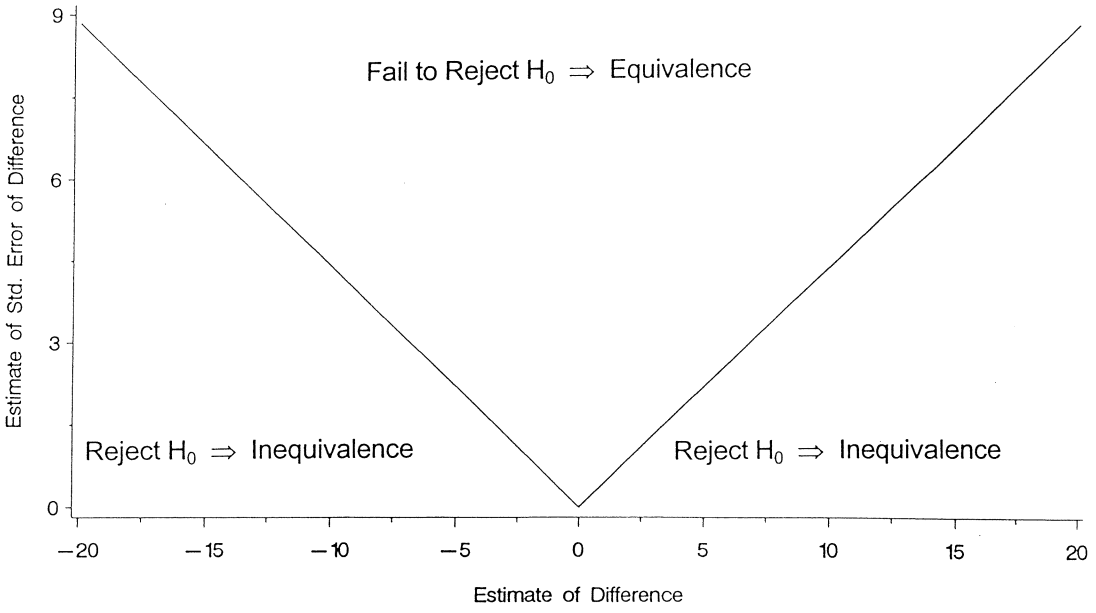


Figure 2
Traditional Test



References

Bondy, Warren H. A Test of an Experimental Hypothesis of Negligible Difference between Means. *The American Statistician*, **23(5)**, 28-30 (1969).

Hauck, Walter W. Individual and Population Bioequivalence. Short Course presented at the 1997 Spring ENAR Meeting in Memphis Tenessee, March 23, 1997.

Henderson, Charles H. Applications of Linear Models in Animal Breeding. *University of Guelph* (1984).

Metzler, C. M. Bioavailability - a Problem in Equivalence. *Biometrics*. **30**, 309-317 (1974).

McLean, R. A., Sanders, W. L. And Stroup, W. W. A Unified Approach to Mixed Linear Models. *American Statistician*. **45**, 54-63 (1991).

Milliken, George A. and Johnson, Dallas E. *Analysis of Messy Data, Volume 1: Designed Experiments*. New York: Chapman Hall (1992).

Schuirmann, Donald J. A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetic and Biopharmaceutics*. 15(6), 657-680 (1987).

Westlake, W. J. Use of Confidence Intervals in Analysis of Comparative Bioavailabilty Trials. *Journal of Pharmaceutical Sciences*, **61**, 1340-1341 (1972).