生物安全学报 2016, 25(2): 77-91
JOURNAL OF BIOSAFETY http://www.jbscn.org

DOI: 10.3969/j.issn.2095-1787.2016.02.002

Equivalence tests to support environmental biosafety decisions: theory and examples

David A. ANDOW[1*], Gábor L. LÖVEI[2], Débora Pires PAULA[3]

[1]*Department of Entomology, University of Minnesota, St. Paul, MN* 55108, *USA*; [2]*Department of Agroecology, Aarhus University, Flakkebjerg Research Centre, DK*-4200 *Slagelse, Denmark*; [3]*Embrapa Genetic Resources and Biotechnology, Parque Estação Biológica, W*5 *Norte, P.O. Box* 02372, *Brasília, DF*, 70770-917, *Brazil*

Abstract: A major role of ecological risk assessment (ERA) has been to provide scientific guidance on whether a future human activity will cause ecological harm, including such activities as release of a genetically modified organism (GMO), exotic species, or chemical pollutant into the environment. This requires the determination of the likelihoods that the activity: would cause a harm, and would not cause a harm. In the first case, the focus is on demonstrating the presence of a harm and developing appropriate management to mitigate such harm. This is usually evaluated using standard hypothesis analysis. In the second case, the focus is on demonstrating the absence of a harm and supporting a decision of biosafety. While most ERA researchers have focused on finding presence of harm, and some have wrongly associated the lack of detection of harm with biosafety, a novel approach in ERA would be to focus on demonstrating directly the safety of the activity. Although, some researchers have suggested that retrospective power analysis can be used to infer absence of harm, it actually provides inaccurate information about biosafety. A decision of biosafety can only be supported in a statistically sound manner by equivalence tests, described here. Using a 20% ecological equivalence standard in GMO examples, we illustrated the use of equivalence tests for two-samples with normal or binomial data and multi-sample normal data, and provided a spreadsheet calculator for each. In six of the eight examples, the effects of Cry toxins on a non-target organism were equivalent to a control, supporting a decision of biosafety. These examples also showed that demonstration of equivalence does not require large sample sizes. Although more relevant ecological equivalence standards should be developed to enable equivalence tests to become the main method to support biosafety decision making, we advocate their use for evaluating biosafety for non-target organisms because of their direct and accurate inference regarding safety.

Key words: GM crops; average bioequivalence; environmental impact; ERA; statistical methods

Introduction

A major role of ecological risk assessment (ERA) has been to provide scientific guidance on whether a future human activity will cause ecological harm (Suter, 2006). These activities may include land use changes, such as rural and suburban development, agricultural expansion, or deforestation, the release of organic and inorganic chemicals, such as CO_2, NO_X, pesticides and other toxic chemicals, and the release of biological organisms, such as biological control organisms, exotic species, and genetically modified organisms (GMOs). The logic of all of these ERAs directs the assessments to determine the likelihood that the activity would cause an adverse environmental change, and at the same time the likelihood that the activity will not cause an adverse environmental effect. A risk assessor is interested in both the probability that there will be an adverse effect(s) and the probability that there will be no adverse effect(s). The first of these can be determined by standard statistical hypothesis tests. In the case of "no harm", the assessor must be able to conclude that a test treatment has similar risk or effect as the control treatment, i.e. the two are equivalent, and this requires the use of equivalence tests, described here.

Under standard experimental design and statistical hypothesis testing, the null hypothesis is that the

收稿日期(Received): 2015-12-15 **接受日期(Accepted)**: 2016-02-29

* 通讯作者(Author for correspondence), E-mail: dandow@umn.edu

responses to the test and control treatments are the same. Rejecting the null hypothesis (that they are the same) will lead to the conclusion that they are different. In other words, it allows a conclusion that there is a difference between the two treatments, but it does not allow a conclusion that they are the same or similar. Equating the lack of statistical significance with "no difference" or biosafety is a serious logical flaw, because the lack of significance can be related to low replication and/or high error variation, not because there is truly no effect. Inability to reject the null hypothesis can lead to a Type Ⅱ error (= not rejecting the null hypothesis when in fact the treatments are different). For any given estimated difference between the treatments, as the estimated standard error of the difference increases from 0, the result will eventually change from an inference of significant difference to one of equivalence, which is the opposite of what is desired for equivalence testing. For the most part, this problem — that the null hypothesis cannot be proved with standard hypothesis testing — is recognized, but because alternatives are not recognized it is largely ignored (e.g., Raybould,2010).

In recent decades, however, a new branch of statistical theory, equivalence testing, has been developed to address these problems. An equivalence test inverts the null and alternative hypotheses, so that the null hypothesis is that the treatments are different and the alternative hypothesis is that they are equivalent. Thus the rejection of the null hypothesis enables a sound statistical inference that the treatments are equivalent. Equivalence tests have gained widespread use for supporting regulatory decisions about new generic drugs, and there are now textbooks for conducting such tests (e.g., Patterson & Jones,2005). In this paper, we summarize the statistical theory underlying equivalence tests, compare this approach with standard hypothesis testing and power analysis, illustrate how to conduct these statistical tests with examples from ecological risk analysis experiments for testing the safety of GM crops, and suggest that equivalence testing is superior to standard hypothesis testing for assessing ecological safety. We use GM crops because we have conducted research in this area. A spreadsheet calculator for the equivalence tests described in this paper is provided in the supplementary material. Even though the examples are solely related to GMOs, the potential scope of application of equivalence tests in ecological risk assessment and environmental policy is quite broad (Diamond *et al.*, 2012; Hanson,2011; Kristofersson & Navrud,2005).

Statistical theory for equivalence tests

There are three kinds of equivalence tests: average equivalence, population equivalence and individual equivalence (Liu & Chow,1996). Average equivalence evaluates the similarity in the average response between a test and control treatment. Population equivalence evaluates the similarity in the entire statistical distribution (average, variance, skew, kurtosis, etc) of the responses to the treatments. Population equivalence is a more rigorous similarity standard than average equivalence because the average, variance, and possibly higher statistical moments all must be similar. Individual equivalence tests examine the similarity in the responses to the treatments within the same individuals. This last is often used in drug testing, where each individual is exposed to both treatments with a suitable re-equilibration period between treatments (a so-called two-period crossover design), and also includes other designs, such as repeated measures, and paired designs. From such a design it is possible to evaluate equivalence individual by individual.

For ecological risk assessment, average equivalence is the more generally applicable of the three. Individual equivalence testing will be limited because it is often be difficult to expose individuals to more than one treatment, especially in a toxicity assay. Population equivalence testing will also be limited because it is usually not possible to have sufficiently high replication to test for equivalence in variance, skewness, etc.

Equivalences tests can be understood by contrasting them with standard hypothesis tests (Fig. 1). If the average response of some biological entity to a test

treatment i is denoted μ_i and the average response to the control treatment (negative control) is μ_c, a standard hypothesis for normally distributed data is

$$H_0: \mu_i/\mu_c = 1 \qquad H_a: \mu_i/\mu_c < 1 \text{ or } \mu_i/\mu_c > 1 \quad [1]$$

where H_0 is the null hypothesis and H_a is the alternative hypothesis. The null hypothesis is that the two populations have the same mean, and the alternative hypothesis is that they do not.

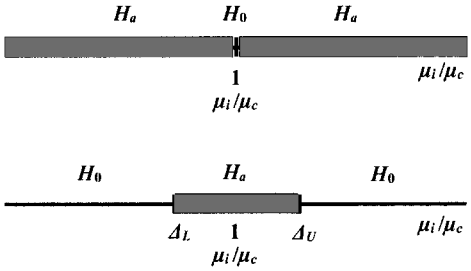

Fig.1　Comparison of standard hypothesis tests and equivalence tests for hypotheses [1] and [2]

Upper bar shows standard hypothesis test [1] with $H_0 = 1$ and H_a otherwise (gray region). Lower bar shows equivalence test [2] with H_0 less than Δ_L and greater than Δ_U, and H_a between Δ_L and Δ_U (gray region).

An equivalence hypothesis reverses the null and alternative hypotheses. Using the same notation, the analogous equivalence hypothesis is

$$H_0: \Delta_L \geqslant \frac{\mu_i}{\mu_c} \text{ or } \frac{\mu_i}{\mu_c} \geqslant \Delta_U \qquad H_a: \Delta_L < \frac{\mu_i}{\mu_c} \text{ and } \frac{\mu_i}{\mu_c} < \Delta_U \quad [2]$$

where the values Δ_L and Δ_U are equivalence standards set according to regulatory, statistical and biological considerations that define how close the means must be to be considered equivalent. The null hypothesis states that the ratio of the averages is either less than the lower or greater than the upper equivalence standard, and the alternative hypothesis is that the value of the ratio is between the two standards. Note that the hypothesis that the averages are the same, is now a part of the alternative hypothesis. The equations in [2] are typically reformulated as:

$$H_{01}: 0 \geqslant \mu_i - \Delta_L \mu_c \text{ versus } H_{a1}: 0 < \mu_i - \Delta_L \mu_c$$
$$H_{02}: 0 \leqslant \mu_i - \Delta_U \mu_c \text{ versus } H_{a2}: 0 > \mu_i - \Delta_U \mu_c \quad [3]$$

The null hypothesis in [2] is rejected and the means are equivalent if and only if both null hypotheses [3] are rejected. For some regulatory procedures, ±20% is a commonly used equivalence standard ($\Delta_L = 0.80$ and $\Delta_U = 1.25$).

When equation [2] is log-transformed, linear hypotheses are produced:

$$H_0: \theta_L \geqslant \eta_i - \eta_c \text{ or } \eta_i - \eta_c \geqslant \theta_U$$
$$H_a: \theta_L < \eta_i - \eta_c \text{ and } \eta_i - \eta_c < \theta_U \quad [4]$$

where $\eta = \ln(\mu)$ and $\theta = \ln(\Delta)$. In the US and Europe, the value $\theta_U = -\theta_L = 0.223144$ is required for most generic drug tests, which is the same as ±20% on the untransformed scale.

Hypotheses [2] ~ [4] are called intersection-union hypotheses (Berger & Hsu, 1996a). The null hypothesis is the intersection of two one-sided hypotheses, and the alternative hypothesis is the union of two one-sided hypotheses. A test of an intersection-union hypothesis is called an intersection-union test (IUT) and is often formulated as a test of two one-sided hypotheses, which is called a two one-sided test (TOST). The formulation of an equivalence test as an IUT allows the application of some general mathematical theorems to determine the Type Ⅰ error rate for the test (Berger, 1982; Berger & Hsu, 1996a). Although it might be thought that the Type Ⅰ error rate for the two tests would need to be adjusted because there are multiple tests with the same data, the theorems prove that such corrections are not needed for any of the tests discussed in this paper (Berger, 1982; Berger & Hsu, 1996a).

Another theorem (Theorem 4, Berger & Hsu, 1996a) provides conditions for constructing confidence intervals (or regions) on the statistical parameter(s) so that confidence intervals (or regions) can be used to test equivalence in lieu of hypothesis testing. If and only if an IUT rejects the null hypothesis with a Type Ⅰ error of 0.05, the 95% confidence interval around $\eta_i - \eta_r$ will be entirely contained in the interval $[\theta_L, \theta_U]$, which is called the equivalence region or interval. This demonstrates the identity between hypothesis testing and interpretation of confidence intervals and regions. We will use this theorem to test the equivalence of multiple test treatments to a single control treatment.

Comparison with retrospective (observed) power

A retrospective analysis of the statistical power of an experiment has been proposed to address the problem of Type Ⅱ error in GMO ecological risk assessment (e.g., Romeis *et al.*, 2011). The power of an experimental design is an estimate of the probability of not making a Type Ⅱ error (not rejecting the null hypothesis, when in fact it should have been rejected). There are two kinds of power analysis: prospective and retrospective. Prospective power analysis uses information from previous experiments to optimize the design of experiments yet to be conducted, and is a legitimate and useful statistical tool (Hoenig & Heisley, 2001). It can also be used to optimize equivalence tests.

Hoenig & Heisley (2001) provide a deep critique of retrospective power analysis. Retrospective power analysis aims to provide an independent estimate of the probability of not making a Type Ⅱ error based on the design and data of an experiment that has already been completed, and relies on a statistic called "observed power". Advocates for retrospective power analysis argue that high observed power indicates a low Type Ⅱ error rate and therefore the null hypothesis is more likely to be true when it is not rejected and there is high observed power (e.g., Romeis *et al.*, 2011). These arguments and inferences are logically flawed because retrospective power analysis does not provide an independent estimate of the probability of not making a Type Ⅱ error (Brosi & Biber, 2009; Nakagawa & Foster, 2004; Perry *et al.*, 2009).

Hoenig & Heisley (2001) provided a specific example to illustrate this serious logical flaw in the use of observed power. Suppose two similar experiments are conducted, and neither rejects the null hypothesis, but the observed power in the first experiment was larger than the observed power in the second one. Advocates of the use of observed power may wish to infer that the first experiment gives stronger support favoring the null hypothesis than the second. However, this leads to a fatal logical contradiction. Suppose the experiments were tested with a one-sided t-test. Let t_{p1} and t_{p2} be the observed test statistics from the respective experiments. Because the observed power was higher in the first experiment, this implies that $t_{p1} > t_{p2}$, because observed power is an increasing function of the t_p statistic. But if $t_{p1} > t_{p2}$, then the p-values from the experiments would have $p_1 < p_2$, because a higher test statistic has a smaller p-value. In other words, experiment 1 has a smaller p-value and higher observed power. Thus, by usual inferential standards based on the p-value, experiment 1 gives stronger support against the null hypothesis, because it has the lower probability of error (p-value), while advocates of observed power would wish to infer that experiment 1 gives stronger support favoring the null hypothesis. This has been called "the paradox of power" (Hoenig & Heisley, 2001). Inference based on observed power leads to this serious logical error, and observed power should not be used to infer support for or against the null hypothesis. Observed power is not independent of the observed p-value, because the p-value completely determines the observed power. Reporting both the p-value and the observed power is, in effect, reporting the p-value twice. Instead, the better solution is to structure the null hypothesis as an equivalence test that allows sound statistical inference (Perry *et al.*, 2009). Therefore, retrospective power analysis does not address the problem of Type Ⅱ error, and cannot replace equivalence testing (van der Voet *et al.*, 2011).

Conducting equivalence tests

Equivalence tests can be conducted for many different experimental designs, and one area of active research is extending their use for more complex designs. Here we provide a basic introduction to equivalence tests for some common and simple experimental designs: two independent treatments and normal data, two independent treatments and binomial data, multiple independent treatments and normal data, and replication of experiments with multiple independent treatments and normal data. We use examples for GMO biosafety testing because we have been conducting research in this area and can use real data to illustrate

the use of equivalence tests. These examples include only ones where no significant difference was detected using standard hypothesis tests, and are used to illustrate when it is possible to conclude that there is statistical equivalence supporting a biosafety decision and when this is not possible.

Normal data

Sasabuchi (1980) first proposed a standard TOST of normal data, and Westlake (1981) and Schuirmann (1987) proposed the standard TOST of log-normal data. Here we provide an example for normal data. Let $\overline{X}$, $\overline{C}$ be the two sample means (test and control treatments with m and n samples, respectively), and S^2 be the pooled estimate of σ^2. The null hypotheses [3] are both rejected if

$$T_L > t_{\alpha,\nu} \text{ and } T_U < -t_{\alpha,\nu} \qquad [5]$$

$$\text{where } T_L = \frac{\overline{X} - \Delta_L \overline{C}}{S\sqrt{1/m + \Delta_L^2/n}} \text{ and } T_U = \frac{\overline{X} - \Delta_U \overline{C}}{S\sqrt{1/m + \Delta_U^2/n}} \qquad [6]$$

These have a Student's t-distribution with $\nu = m + n - 2$ degrees of freedom. The TOST [3] is conducted using the ordinary, $\alpha = 0.05$, one-sided t-test based on T_L for the one-sided hypothesis [3 upper] and the ordinary, $\alpha = 0.05$, one-sided t-test based on T_U for the one-sided hypothesis [3 lower]. A numerical example is provided in Box 1 and the supporting information.

Binomial data

For binomial data there are several alternatives for constructing equivalence intervals and designating equivalence standards, based on the binomial parameter, π. They can be modeled on the arithmetic difference between the test (π_i) and control (π_c) parameters ($\pi_i - \pi_c$), the proportional difference in the parameters (π_i/π_c), or the proportional difference in the odds of a response ($\pi_i(1-\pi_c))/((1-\pi_i)\pi_c$), which is based on the odds ratios in each treatment. We calculated the equivalence interval for the three models (arithmetic, proportional, and odds ratio), and expressed it as the interval of the test treatment (π_i) as a function of the control treatment (π_c). The proportional difference model, which was ideal for normal data [2], is asymmetrical across the range of π_c (Fig. 2), which is problematic because equivalence will depend on which response was chosen as the focal response. The others are symmetric (Fig. 2), but differ near $\pi_{r1} = 0$ or $\pi_{r1} = 1$. Both can be justified, depending on whether the absolute differences or the difference in odds is critical. Here we provide equivalence tests for arithmetic differences with sufficiently large samples (m, $n \geqslant 50$), because these have a stable Type Ⅰ error rate (Chen *et al.*, 2000). For small samples, exact methods are required (Agresti, 1996).

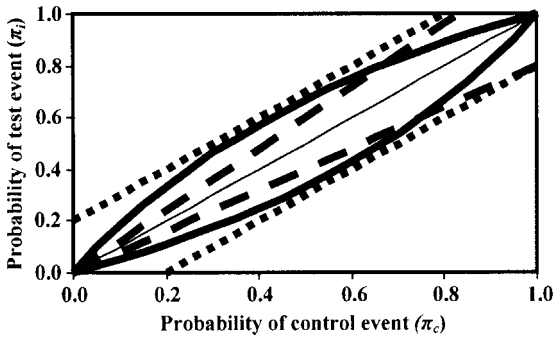

Fig.2　Equivalence intervals for binomial data as a function of the binomial parameter, π_c, in the control (c) treatment for three different ways of designating the interval

Intervals are given for the range of the test treatment, π_i, that would be equivalent. Solid line: proportional odds ratios $(\pi_i(1-\pi_c))/((1-\pi_i)\pi_c)$; dashed line: proportional difference in test to control treatment (π_i/π_c); dotted line: arithmetic difference $(\pi_i - \pi_c)$. All intervals are calculated for ±20% of the control value, π_c.

Let $X_1, \cdots, X_m$ denote the independent binomial responses (m, π_i) to the test treatment and $C_1, \cdots, C_n$ denote the independent binomial responses (n, π_c) to the control treatment, where π_i and π_c are the true response probabilities for the test and control treatments, respectively, and m and n are the number of independent observations for each. In addition, let y_c be the total number of observed "positive" responses in the control treatment and x_i be the number of "positive" responses in the test treatment, so that $n - y_c$ and $m - x_i$, respectively, become the number of "negative" responses in the two treatments.

For the arithmetic difference in response probabilities, the equivalence hypothesis is

$$H_0: \Delta_L \geqslant \pi_i - \pi_c \text{ or } \pi_i - \pi_c \geqslant \Delta_U$$
$$H_a: \Delta_L < \pi_i - \pi_c \text{ and } \pi_i - \pi_c < \Delta_U \qquad [7]$$

where Δ_L and Δ_U are equivalence standards determining how close π_i and π_c must be to be considered equivalent. For drug testing, the standards used vary between 10%~20%, but here we set the standards as $\Delta_L=-0.2$ and $\Delta_U=0.2$, with π_i and π_c bounded by the interval [0, 1]. More generally, the equivalence intervals can be adjusted based on the observed values of π_i and π_c.

A TOST for [7] comes from the asymptotic test statistic for the difference between two binomial parameters, $\pi_i-\pi_c$, and is based on the following two statistics with a standard error estimated by maximum likelihood (Farrington & Manning, 1990),

$$z_{FM,L}=\frac{[\hat{\pi}_i-\hat{\pi}_c-\Delta_L]^2}{\tilde{\sigma}_{FM}^2} \text{ and } z_{FM,U}=\frac{[\hat{\pi}_i-\hat{\pi}_c-\Delta_U]^2}{\tilde{\sigma}_{FM}^2}$$

$$\hat{\pi}_i=\frac{x_i}{m}$$

$$\hat{\pi}_c=\frac{y_c}{n}$$

$$\tilde{\sigma}_{FM}^2=\frac{\tilde{\pi}(1-\tilde{\pi})}{n}+\frac{(\tilde{\pi}+\Delta_j)(1-\tilde{\pi}-\Delta_j)}{m} \qquad [8]$$

where $\tilde{\sigma}_{FM}^2$ is calculated separately for each one-sided test substituting either Δ_L or Δ_U for Δ_j, and $\tilde{\pi}$ is the unique solution of a restricted maximum likelihood problem defined as the π that maximizes F on a closed interval I_π of possible π's,

$$F_\pi(p)=\pi^{y_r}(1-\pi)^{n-y_r}(\pi+\Delta_j)^{x_i}(1-\pi-\Delta_j)^{m-x_i}$$

$$I_\pi=[\max\{0,-\Delta_j\},\ \min\{1,1-\Delta_j\}] \qquad [9]$$

$z_{FM,L}$ and $z_{FM,U}$ are χ^2-distributed with 1 degree of freedom. The two treatments are equivalent if both one-sided tests are rejected at a predetermined level of α, usually $\alpha=0.05$, that is, if

$$z_{FM,L}>3.841 \text{ and } z_{FM,U}>3.841 \qquad [10]$$

which are the upper and lower $\alpha=0.05$ tails of the standard normal distribution.

These quantities can be calculated from data (see an example in Box 2 and supporting information). Calculating $\tilde{\sigma}_{FM}$ relies on finding $\tilde{\pi}$, which can be done using Solver in MS Excel or other powerful mathematical software, such as Mathematica and Matlab. Farrington & Manning (1990) provided a closed form solution for $\tilde{\pi}$ in their appendix.

Multiple comparisons to a common control treatment

In ecological risk assessment, experiments frequently have more than one test treatment compared to a single control treatment, requiring a statistical procedure for making multiple comparisons. The statistical design is a completely randomized one-way treatment structure, $\mathbf{x}=\mathbf{D}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\mathbf{x}$ is the vector of measured responses, $\mathbf{D}$ is the known design matrix, $\boldsymbol{\beta}$ is a vector of unknown fixed effects, which is estimated by the treatment means, and $\boldsymbol{\varepsilon}$ is a random error vector with $E[\boldsymbol{\varepsilon}]=\mathbf{0}$. One solution is to use equivalence tests based on confidence intervals designed for multiple comparisons (Berger & Hsu, 1996a, 1996b).

For normal data, let $\overline{X}_1, \cdots, \overline{X}_K$ be the means in K test treatments and $\overline{C}$ be the mean of the control treatment. We would like to consider K hypotheses:

$$H_0^i: \Delta_L \geqslant \frac{\overline{X}_i}{\overline{C}} \text{ or } \frac{\overline{X}_i}{\overline{C}} \geqslant \Delta_U \text{ and}$$

$$H_a^i: \Delta_L < \frac{\overline{X}_i}{\overline{C}} \text{ and } \frac{\overline{X}_i}{\overline{C}} < \Delta_U \text{ for } i=1, \cdots, K \qquad [11]$$

Each of these K hypotheses could be tested using a TOST, but it is not possible to test all K hypotheses simultaneously because an IUT does not allow for the possibility of rejecting some but not all of the hypotheses. In addition, because up to K hypotheses could be true, it is necessary to have an adjustment for testing multiple comparisons. An appropriate approach is to construct confidence intervals for each of these K hypotheses. Berger & Hsu (1996a) show that if the confidence interval, constructed from a TOST, is entirely contained in the equivalence interval $[\Delta_L, \Delta_U]$, the two means are equivalent. The confidence interval for the multiple comparisons can be constructed from [5] and [6], which give

$$\frac{\overline{X}-\Delta_L\overline{C}}{S\sqrt{1/m+\Delta_L^2/n}}>t_{\alpha,\nu} \text{ and } \frac{\overline{X}-\Delta_U\overline{C}}{S\sqrt{1/m+\Delta_U^2/n}}<-t_{\alpha,\nu}$$

where S is estimated from the variance of $\boldsymbol{\varepsilon}$. These can be rearranged to give the following confidence interval

$$\left[\left(\frac{\overline{X}_i}{\overline{C}}-\frac{t_{\alpha,\nu}S\sqrt{1/m_i+\Delta_L^2/n}}{\overline{C}}\right),\left(\frac{\overline{X}_i}{\overline{C}}+\frac{t_{\alpha,\nu}S\sqrt{1/m_i+\Delta_U^2/n}}{\overline{C}}\right)^+\right] \qquad [13]$$

To reject the null hypothesis, the confidence interval must be inside the equivalence interval. Note first that the confidence interval is constructed with $t_{\alpha,\nu}$, and not the standard $t_{\alpha/2,\nu}$. Second, because there are multiple comparisons to the same control, Dunnett's t should be used instead of Student's t with $\nu = n + \sum_i m_i - (K+1)$ (Box 3 and supporting information). This results in a strictly conservative test, and more accurate $p = 0.05$-level tests are available for balanced data (Giani & Straβburger, 1994). Bonferroni corrections are not appropriate because the treatment comparisons are correlated by using the same control.

If the experiment with multiple treatments is replicated several times, the same approach can be used with some adjustments in the values because the underlying statistical model is more complex. In this case, the statistical model is $\mathbf{x} = \mathbf{D\beta} + \mathbf{ZU} + \boldsymbol{\varepsilon}$, with everything defined as above with $\mathbf{Z}$ the design matrix for the replication of the experiments and $\mathbf{U}$ is a vector of the random effects associated with this replication. In equation [13], S is calculated from the variance of $\boldsymbol{\varepsilon}$ in the new model, $\overline{C}$ and $\overline{X}_i$ are least square means across the R replications of the experiment, $m_i = \sum_r m_{ir}$, $n = \sum_r n_r$, and $\nu = n + \sum_i m_i - R(K+1)$ (Box 4).

Equivalence standards and tests

An important issue for equivalence tests is the determination of equivalence standards. Equivalence standards, Δ_L and Δ_U (or θ_L and θ_U), are determined by a combination of regulatory, ecological and statistical considerations. The statistical considerations are related to the sample size necessary to attain an acceptable power of the equivalence test. For example, to test the equivalence of proportions (equation [9]), a sample size of 50 provides suitable power for a standard of ±20%, but this sample size is inadequate for an equivalence standard of ±10%, when a sample size >150 would be necessary.

Ecological and regulatory considerations will determine what is biologically equivalent and socially acceptable. In general, there is a large class of ecological problems that has hardly been addressed in applied ecology. When are two ecological systems ecologically equivalent? How much change could occur to an ecological system before it should be considered ecologically different? What are the ecologically essential structural and functional features of an ecological system? How much change can an ecological system tolerate before its essential ecological features are harmed? How much can components of an ecological system change without changing the ecologically essential features of the whole system in which they are embedded?

We do not presume to answer these questions, because it is likely that considerable empirical ecological research will be necessary before meaningful answers can be formulated. There exist formal theoretical conditions under which ecological systems are equivalent or nearly so (Iwasa *et al.*, 1987, 1989), but these conditions are so strict and narrow that they cannot be readily implemented, and empirical criteria are needed. One approach for setting ecological equivalence standards has been to use data from historical tests that can be related to potentially significant effects (Bertoletti *et al.*, 2007; Phillips *et al.*, 2001). For example, cladocerans have been extensively used to evaluate the toxicity of aquatic pollutants. In some cases, there are a sufficient number of laboratory toxicity bioassays that have been associated with potential ecological effects that it is possible to estimate the level of toxicity that could cause environmental harm (Bertoletti *et al.*, 2007). However, such data sets are uncommon, so this approach is of limited applicability. No such data sets exist for GMO ecological risk assessment.

Natural variability of the control treatment has been often advocated as an approach for determining equivalence standards (Barrett *et al.*, 2015; EFSA, 2010; Hong *et al.*, 2014; Kang & Vahl, 2014; Vahl & Kang, 2016; van der Voet *et al.*, 2011). The rationale is that if the control has high variability, then any test treatment must be more different, because the high variability requires a larger equivalence standard. Although this may be true in food safety research, temporal and spatial correlations in many ecological fac-

tors may result in the covariance among treatment and control responses being as important as the variance in control response. Large positive covariance would imply that the control variance overestimates the relevant natural variance, and leaves doubt as to how to estimate the relevant natural variance. More importantly, the potential irreversibility of ecological change might argue for tighter equivalence standards. Reversibility may be associated with ecosystem resilience, whereas ecological hysteresis may be indicative of irreversibility (Biggs *et al.*,2009). Thus, it may be more appropriate to define ecological equivalence standards in terms of degree of concern, namely the minimum ecological effect that is sufficient to cause harm (Perry *et al.*, 2009).

The criteria for establishing ecological equivalence standards should center on the ecological risks that society wants to avoid (Andow, 2011). Consequently, human values will be an important consideration in establishing ecological equivalence standards. For example, to construct equivalence standards to evaluate the effects of a GM crop on a generalist biological control agent, many of the most significant social values are embedded in the crop/yield loss relationship. As the central purpose of biological control of crop pests is to minimize crop yield loss, the value of biological control can be measured by the reduction in crop yield loss from pests. With a quantitative relationship between the density of a biological control agent and the suppression of the pest population, these two relationships can be combined so that a change in natural enemy density can be related to a change in crop yield loss. This combined relationship can be used to establish ecological equivalence standards related to ecological value (Andow,2011).

Inferences from equivalence tests

To illustrate how equivalence tests can support biosafety decisions, we use the examples in Box 1-4. While this discussion focuses on GMO biosafety, it should be clear that with other examples, the discussion can be generalized to many areas of environmental risk assessment. Many researchers studying GMO or *Bt* toxicity have based their conclusions of biosafety on standard hypothesis testing (e.g., Lawo *et al.*, 2009; Lundgren & Wiedenmann, 2002; Meissle & Romeis, 2009; Romeis *et al.*, 2004; von Burg *et al.*, 2010). These researchers make claims for biosafety, but in reality, they have committed the statistical error of accepting the null hypothesis, by concluding that there were no harmful effects. Here, we demonstrate how to make sound inferences of "no effect" based on equivalence tests. We have reanalyzed the data of eight examples to illustrate how equivalence tests differ from standard hypothesis tests (Table 1). The data were obtained from Paula *et al.*(2016), Paula & Andow (2016) and Guo *et al.*(2008), and are described in detail in Box 1-4. In all eight examples, the standard hypothesis test led to the conclusion that an effect of the Cry toxin was not detected (the standard null hypothesis was not rejected). The equivalence tests, using ecologically strict equivalence standards of ±20%, allowed us to conclude that in six of the eight examples, the effect of the Cry toxin was equivalent to the effect of the control. These results enable a sound conclusion of "no effect" and support for a biosafety determination of the Cry toxin for *Harmonia axyridis* development time (Cry1F), *Cycloneda sanguinea* development time (Cry1F and combined Cry1Ac and Cry1F), and *Chrysopa pallens* development time (GK12, Nu COTN 99B, and a mixture). In other words, in these respects, Cry toxins are "safe" for *H. axyridis*, *C. sanguinea* and *Ch. pallens*. For *C. sanguinea*, the sample sizes for these tests were quite modest ($n = 8$ and $n = 10$), which shows that equivalence tests do not require large sample sizes. The remaining two cases were also revealing, as they were statistically indeterminate, neither rejecting the null hypothesis for the standard test or for the equivalence test. Development time of *C. sanguinea* on Cry1Ac was not equivalent to the control, but this treatment had a small sample size ($n = 6$), and might become equivalent with higher replication. Mortality of *Brevicoryne brassicae* on Cry1Ac, had a p-value of 0.105 under the standard hypothesis test, with mortality on Cry1Ac estimated to be 31% compared to 19% on the control diet.

In this case, increased replication might result in detection of a significant effect of Cry1Ac and a determination of non-equivalence. In any event, the equivalence test returned the accurate result that mortality of *B. brassicae* on Cry1Ac was not equivalent to the control and this test does not support a biosafety decision. Equivalence tests allow sound inference about biosafety, while standard hypothesis tests do not.

Table 1　Comparison of results from standard hypothesis tests and equivalence tests for the examples in Box 1-4

Species	Parameter	Toxin	Standard hypothesis test	Equivalence test	Source
Harmonia axyridis	Development time	Cry1F	No difference detected	Equivalent	Box 1
Brevicoryne brassicae	Mortality	Cry1Ac	No difference detected	Not equivalent	Box 2
Cycloneda sanguinea	Development time	Cry1Ac	No difference detected	Not equivalent	Box 3
Cycloneda sanguinea	Development time	Cry1F	No difference detected	Equivalent	Box 3
Cycloneda sanguinea	Development time	Cry1F and Cry1Ac	No difference detected	Equivalent	Box 3
Chrysopa pallens	Development time	GK12, Cry1Ab/c fusion	No difference detected	Equivalent	Box 4
Chrysopa pallens	Development time	Nu COTN 99B, Cry1Ac	No difference detected	Equivalent	Box 4
Chrysopa pallens	Development time	Mixture	No difference detected	Equivalent	Box 4

Burden of proof

Finally, we note that equivalence statistics have direct and significant bearing on the debates about the burden of proof. Hobbs & Hilborn (2006) stated that the burden of proof has traditionally been on those who argued for regulatory intervention to stop pollution, i.e., pollution is allowed until its harms can be proven. Similarly, for invasive species risk assessment, the potential invader is assumed to be safe until proven to cause environmental harm (Simberloff, 2005). Standard hypothesis testing is well-suited for these cases, as it can only establish whether there is a difference, whether there is environmental harm. However, this approach has allowed substantial pollution and the establishment of several harmful invasive species (Simberloff, 2005), and as a general approach for environmental management, it has come under considerable criticism (e.g., Diamond *et al.*, 2012; Hanson, 2011; Kristofersson & Navrud, 2005).

In risk assessment, demonstration of biosafety is equally important as demonstration of harm. Equivalence tests are one way to establish a burden of proof of biosafety, by requiring demonstration of equivalence. However, equivalence tests are more flexible than this simple application of a "proof of safety" concept might imply. It is possible to consider the equivalence standard as a function of ecological value, and to test equivalence under different standards (ecological value). For example, a risk assessor could assess whether an environmental stressor is likely to reduce biological control of a pest thereby causing 5% more yield loss, i.e., using a 5% equivalence standard. An additional equivalence test could be performed to evaluate if the stressor is likely to be equivalent to the control at 2% or 1% yield loss levels (more stringent equivalence standards). The probability of equivalence will decline as the standard becomes smaller, so the *p*-values of the tests will increase (less likely to reject the null hypothesis that they are different). If the *p*-values of these three tests were respectively 0.02, 0.04, and 0.23, the analyst could conclude that if a 5% or 2% yield loss can be tolerated, the stressor and the control are equivalent with respect to their effects on biological control, but they are not equivalent if only 1% yield loss can be tolerated. When the magnitude of an insignificant risk can be differentiated from a significant risk, it will be possible to develop equivalence standards, and these can be used to establish a burden of proof of safety in ecological risk assessment.

Equivalence tests can support a burden of proof of safety, and this shift does not necessarily create additional assessment costs. The cost of an equivalence test will depend primarily on the sample size and error variation, which depend primarily on the planned equivalence standard. A stricter equivalence standard will require a larger sample size and will have a higher cost than a test with a more lax equivalence standard. Because ecological systems often exhibit functional re-

dundancy (Rosenfeld, 2002), and some indirect species interactions attenuate as the pathway lengthens (Abrams *et al.*, 1996), many functionally-based ecological equivalence standards may turn out to be lax. Andow (2011) suggested that an equivalence standard for a generalist biological control agent would probably be larger than the standard ±20%, and consequently, the cost of an equivalence test may be substantially lower than what is currently required under the standard hypothesis testing procedures.

Acknowledgements: The USDA regional research project NC-205 to DAA and Rockefeller Resident Scholar in Bellagio Fellowships to DAA and GLL partially supported this work.

Box 1. Equivalence test for two-sample, normal data

The data originate from Paula *et al.* (2016), who used an artificial tritrophic system to test if the toxin Cry1F, which occurs in *Bt* maize and cotton in Brazil, adversely affected an important biological control agent of agricultural pests, the coccinellid predator *Harmonia axyridis*. This experiment measured the effect of Cry1F on larval development time of the predator. Aphid prey (*Myzus persicae*) were allowed to feed for 24 h on a holidic diet in small cages with and without Cry1F at 20 μg · mL^{-1} diet before being exposed to the predator. Neonate predator larvae were transferred daily into fresh cages to consume the aphids, and development time from neonate to pupa was recorded.

Step 1. Specify equivalence standards. Values of $\Delta_L = 0.80$ and $\Delta_U = 1.25$ were specified, which correspond to ±20% similarity.

Step 2. Enter the data. Let $X_1, \cdots, X_m$ denote the untransformed development times (days) of *H. axyridis* exposed to Cry1F via *M. persicae*, $m = 39$ (test treatment). Let $C_1, \cdots, C_n$ denote development times of control *H. axyridis*, $n = 30$ (control treatment). In this example, under standard hypothesis testing, these were not significantly different.

(Control, C_i)		(Test, X_i)		
11	9	10	11	10
11	9	11	15	10
12	11	12	13	11
12	11	9	11	10
11	11	10	14	10
12	13	11	12	10
11	9	11	10	10
13	8	10	10	9
13	9	10	10	9
12	8	7	9	
11	10	11	10	
12	9	9	10	
13	8	10	9	
10	9	13	10	
11	11	13	10	

Step 3. Calculate test statistics as indicated in equation [6].

Statistic	Value
$\overline{X}$	10.5128
m	39
$\overline{C}$	10.6667
n	30
$\overline{X}-\Delta_L\overline{C}$	1.97949
$\overline{X}-\Delta_U\overline{C}$	−2.8205
S	1.53772
$\sqrt{1/m+\Delta_L^2/n}$	0.21674
$\sqrt{1/m+\Delta_U^2/n}$	0.27879
T_L	5.93944
T_U	−6.5792

We have assumed $\sigma_X^2 = \sigma_Y^2$. If $\sigma_X^2 \neq \sigma_Y^2$, Welch's *t*-test with Welch-Satterthwaite degrees of freedom should be used, although there is no need to round the calculated *df*s to an integer as sometimes recommended (USEPA, 2010).

Step 4. Conduct the TOST using equation [5]. For the example, the left tail of the *t*-distribution with $\alpha = 0.05$ and $\nu = 67$ is $t_{\alpha,\nu} = 1.66792$.

Null hypothesis	Alternative hypothesis	H_0 rejected?
$H_{01}: T_L \leqslant t_{\alpha,\nu}$	$H_{a1}: T_L > t_{\alpha,\nu}$	Rejected
$H_{02}: T_U \geqslant -t_{\alpha,\nu}$	$H_{a2}: T_U < -t_{\alpha,\nu}$	Rejected

Conclusion: Equivalence. The immature development time of the predator on the Cry1F treatment is equivalent to that in the control treatment.

Box 2. Equivalence test for two-sample, binomial data

The data originate from Paula & Andow (2016), who used an artificial holidic diet to test if the *Bt* toxin Cry1Ac adversely affected an important non-target herbivore, the aphid *Brevicoryne brassicae*. This experiment measured the effect of Cry1Ac on the survival of reproductive apterous aphids during a three-day period. Five equal-sized apterous aphids were allowed to feed continuously on a holidic diet in small cages with and without Cry1Ac at 20 μg · mL^{-1} diet. Twice daily, the number of dead aphids was counted, and the data record the total number that died during the experimental period and the number that survived.

Step 1. Specify equivalence standards. Here $\Delta_L = -0.20$ and $\Delta_U = 0.20$.

Step 2. Enter contingency table data. Number of surviving and dead aphids that fed on a diet with 20 μg · mL^{-1} Cry1Ac (test i) or control diet with no Cry1Ac. Under standard hypothesis testing, these were not significantly different ($\mathrm{LR}X^2 = 2.63$, 1 df, $p = 0.105$).

	Dead	Surviving	$\hat{\pi}$	Row totals
Control	13	54	0.1940	67
Test	22	48	0.3143	70
Column totals	35	102		137

Step 3. Find $\tilde{\pi}$ either by maximizing the likelihood in equation [10] or using formulas in Farrington & Manning (1990, Appendix). USEPA (2010) recommends using a normal approximation to the data with arcsin-sqrt transformed $\hat{\pi}$ values and Welch's t-test. The method described here does not rely on the normal approximation, and therefore is suitable for smaller sample sizes and for $\hat{\pi}$ close to 0 or 1.

	Lower	Upper
I_π	[0.2, 1]	[0, 0.8]
$\hat{\pi}_i - \hat{\pi}_c - \Delta_j$	0.3203	−0.0797
$\tilde{\pi}$	0.1915	0.3636
$\tilde{\sigma}^2_{FM}$	0.005714	0.005408
z_{FM}	17.950	1.176

Step 4. Conduct the TOST using equation [11], with a critical value = 3.841. In this case, $z_{FM,L} > 3.841$ and $z_{FM,U} < 3.841$.

Conclusion: Nonequivalence. The lower one-sided test is rejected, but the upper one-sided test is not rejected. Therefore the two treatments are not equivalent. The survival rate of the aphid feeding on Cry1Ac was not equivalent to the control.

Box 3. Equivalence test for multiple-sample, normal data

The data originate from Paula *et al.* (2016), who used an artificial tritrophic system to test if the *Bt* toxins Cry1Ac alone, Cry1F alone, or Cry1Ac/Cry1F together adversely affected an important biological control agent, the coccinellid predator *Cycloneda sanguinea*. This experiment measured the effect on larval development time of the predator from neonate to pupa, and was designed to test if the two toxins interacted with synergistic effects. Aphid prey (*Myzus persicae*) were allowed to feed for 24 hours before predator exposure on a holidic diet in small cages with and without Cry1Ac or Cry1F or both together. Neonate predator larvae were transferred daily into fresh cages to consume the aphids.

Step 1. Specify equivalence standards. Values of $\Delta_L = 0.80$ and $\Delta_U = 1.25$ were specified, which correspond to ±20% similarity.

Step 2. Enter data. Let $X_{1i}, \cdots, X_{mi}$ denote the untransformed measurements on the m_i larvae in the i^{th} test treatment and $C_1, \cdots, C_n$ denote the untransformed measurements on the n larvae in the control treatment. The data are larval development times of *C. sanguinea* reared on *M. persicae* aphids that fed on an artificial diet with 20 μg · mL^{-1} Cry1Ac (Test 1), 20 μg · mL^{-1} Cry1F (Test 2), or both 20 μg · mL^{-1} Cry1Ac and 20 μg · mL^{-1} Cry1F (Test 3). The control treatment was a control diet with no Cry toxin. There were three test treatments, so $K = 3$, with $m_1 = 6$, $m_2 = 7$, $m_3 = 12$, and $n = 19$. In this example, under standard hypothesis testing, none of the test treatments were significantly different from the control and there was no interaction of the two toxins.

Control	Test 1	Test 2	Test 3	Control	Test 1	Test 2	Test 3
11	9	9	9	10			11
9	7	9	10	10			11
9	8	10	10	8			
10	10	9	10	9			
9	8	8	9	12			
10	9	7	12	9			
10		7	9	8			
9		10	8	8			
9			10	8			
12			11				

Step 3. Calculate test statistics [13] and find the appropriate value for the one-tailed Dunnett's t, based on ν, K and α.

Statistic	Control	Test 1	Test 2	Test 3
Mean	9.474	8.50	8.625	10.00
m_i	19	6	8	12
$\overline{X}_i-\overline{C}$		−0.9737	−0.8487	0.5263
Δ_L		0.8	0.8	0.8
Δ_U		1.25	1.25	1.25
S		1.16791	1.16791	1.16791
$\sqrt{1/m_i+\Delta_L^2/n}$		0.447606	0.398352	0.342078
$\sqrt{1/m_i+\Delta_U^2/n}$		0.498902	0.455233	0.406903
CI_L		0.777912	0.804236	0.964374
CI_U		1.030205	1.031759	1.164016
H_0		not reject	reject	reject

The critical value for Dunnett's t was calculated for $\nu=41$ df, $K=3$ comparisons, and $\alpha=0.05$, using SAS (see below). The critical value is 2.16217.

We can also determine if Test 3 is equivalent to Test 1 and Test 2, which evaluates the hypothesis that there is no interaction between the two toxins (all calculations are not shown).

Comparison	CI_L	CI_U
Test 3 versus Test 1	0.73547	0.98304
Test 3 versus Test 2	0.75939	0.98585

Step 4. Compare the lower and upper confidence intervals, CI_L *and* CI_U, with the equivalence interval (0.80, 1.25). If the *CI*s are entirely within the equivalence interval (0.80, 1.25), then the test treatment mean is equivalent to the control treatment.

Conclusions: (1) Test 1 is not equivalent to the control, while Test 2 and Test 3 are equivalent to the control. The development time of the predator feeding on aphids exposed to Cry1Ac was not equivalent to the control, while for the Cry1F and the combination of both toxins, they were equivalent. (2) Both Test 1 and Test 2 are not equivalent to Test 3, which implies that the hypothesis of no interaction cannot be rejected, i.e., there might be an interaction.

Calculation of critical value, x, for one-sided Dunnett's t using SAS:

```
data;
array lambda{3}; //lambda{i}=sqrt (m_i/(m_i+n))
x = probmc ("dunnett1", ., 0.95, 41, 3, of lambda 1-lambda 3);
```

Box 4. Equivalence test for multiple-sample, normal data with replicated experiments

The data originate from Guo *et al.* (2008), who used a plant-based laboratory tritrophic system to test if larval development time (neonate to pupa) of an important biological control agent, *Chrysopa pallens*, differed when feeding on aphids from the cotton varieties Simian 3 (control), GK12 (with Cry1Ab/Ary1Ac

fusion protein), NuCOTN 99B (Cry1Ac), or alternately feeding on aphids from the three varieties. Aphid prey (*Aphis gossypii*) were collected on excised leaves from field plants and given to individual predators in petri dishes. Fresh aphids were supplied daily.

Step 1. Specify equivalence standards. Values of $\Delta_L = 0.80$ and $\Delta_U = 1.25$ were specified, which correspond to ±20% similarity.

Step 2. Enter data. Let X_{1ir}, …, X_{mir} denote the untransformed measurements on the m_{ir} larvae in the i^{th} test treatment and r^{th} experimental replicate, and C_{1r}, …, C_{nr} denote the untransformed measurements on the n_r larvae in the r^{th} experimental replicate for the control treatment. The data are larval development times (days) of *C. pallens* reared on *M. persicae* aphids that fed on cotton variety GK12 (Test 1), NCOTN 99B (Test 2), or an alternating mixture of aphids (Test 3). The control treatment was the non-*Bt* variety Simian 3. There were three test treatments ($K = 3$) and three replications of the experiment ($R = 3$) with $m_{1r} =$ 19, 17, and 18, $m_{2r} = 18$, 18, 17, $m_{3r} = 18$, 17, 18, and $n_r = 19$, 19, 18. In this example, under standard hypothesis testing, none of the test treatments were significantly different from the control, as indicated in the ANOVA table.

Source	*DF*	*SS*	*MS*	*F*	*P*
Treatment	3	1.896516	0.632172	2.017	0.1126
Experimental replicate	8	0.627450	0.078431	0.250	0.9803
Error	204	63.923117	0.313349		

Step 3. Calculate test statistics [13] and find the appropriate value for the one-tailed Dunnett's t, based on ν, K and α, as in Box 3.

Statistic	Control	Test 1	Test 2	Test 3
Mean	7.6257	7.7609	7.8322	7.8693
m_i	56	54	53	53
$\overline{X}_i - \overline{C}$		0.1351	0.2065	0.2436
Δ_L		0.8	0.8	0.8
Δ_U		1.25	1.25	1.25
S		0.559775	0.559775	0.559775
$\sqrt{1/m_i + \Delta_L^2/n}$		0.173052	0.174059	0.174059
$\sqrt{1/m_i + \Delta_U^2/n}$		0.215454	0.216263	0.216263
CI_L		0.99134	1.00054	1.00540
CI_U		1.05057	1.06005	1.06491
H_0		reject	reject	reject

The critical value for Dunnett's t was calculated for $\nu = 204$ *df*, $K = 3$ comparisons, and $\alpha = 0.05$. The critical value is 2.07698.

Step 4. Compare the lower and upper confidence intervals, CI_L and CI_U, with the equivalence interval (0.80, 1.25). If the *CI*s are entirely within the equivalence interval (0.80, 1.25), then the test treatment mean is equivalent to the control treatment.

Conclusion: Equivalence. All three Test treatments are equivalent to the control. The larval development time of the predator feeding on aphids exposed to Cry toxins in *Bt* cotton plants was equivalent to the control plant.

References

Abrams P A, Menge B A, Mittelbach G G, Spiller D A and Yodzis P, 1996. The role of indirect effects in food webs // Polis G A and Winemiller K O. *Food Webs: Integration of Patterns & Dynamics*. New York: Springer: 371-395.

Agresti A, 1996. *An Introduction to Categorical Data Analysis*. New York, NY: Wiley.

Andow D A, 2011. Assessing unintended effects of GM plants on biological species. *Journal für Verbraucherschutz und Lebensmittelsicherheit*, 6(S1): 119-124.

Barrett T J, Hille K A, Sharpe R L, Harris K M, Machtans H M and Chapman P M, 2015. Quantifying natural variability as a method to detect environmental change: definitions of the normal range for a single observation and the mean of M observations. *Environmental Toxicology and Chemistry*, 34: 1185-1195.

Berger R L, 1982. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24: 295-300.

Berger R L and Hsu J C, 1996a. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11: 283-303.

Berger R L and Hsu J C, 1996b. Rejoinder: bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11: 315-319.

Bertoletti E, Buratini S V, Prósperi V A, Araújo R P A and Werner L I, 2007. Selection of relevant effect levels for using bioequivalence hypothesis testing. *Journal of the Brazilian Society of Ecotoxicology*, 2: 139-145.

Biggs R, Carpenter S R and Brock W A, 2009. Turning back from the brink: detecting an impending regime shift in time to avert it. *Proceedings of the National Academy of Sciences of the United States of America*, 106: 826-831.

Brosi B J and Biber E G, 2009. Statistical inference, Type Ⅱ error, and decision making under the US Endangered Species Act. *Frontiers in Ecology and the Environment*, 7: 487-494.

Chen J J, Tsong Y and Kang S H, 2000. Tests for equivalence or noninferiority between two proportions. *Therapeutic Innovation & Regulatory Science*, 34: 569-578.

Diamond J, Denton D, Anderson B and Phillips B, 2012. It is time for changes in the analysis of whole effluenttoxicity data. *Integrated Environmental Assessment and Management*, 8: 351-358.

EFSA (European Food Safety Authority), 2010. Scientific opinion on statistical considerations for the safety evaluation of GMOs. *EFSA Journal*, 8(1): 1250.

Farrington CP and Manning G, 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9: 1447-1454.

Giani G and Straβburger K, 1994. Testing and selecting for equivalence with respect to a control. *Journal of the American Statistical Association*, 89: 320-329.

Guo J Y, Wan F H, Dong L, Lövei G L and Han Z J, 2008. Tri-trophic interactions between *Bt* cotton, the herbivore *Aphis gossypii* Glover (Homoptera: Aphididae), and the predator *Chrysopa pallens* (Rambur) (Neuroptera: Chrysopidae). *Environmental Entomology*, 37: 263-270.

Hanson N, 2011. Using biological data from field studies with multiple reference sites as a basis for environmental management: the risks for false positives and false negatives. *Journal of Environmental Management*, 92: 610-619.

Hobbs N T and Hilborn R, 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self-teaching. *Ecological Applications*, 16: 5-19.

Hoenig J M and Heisley D M, 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistician*, 55: 19-24.

Hong B, Fisher T L, Sult T S, Maxwell C A, Mickelson J A, Kishino H and Locke M E H, 2014. Model-based tolerance intervals derived from cumulative historical composition data: application for substantial equivalence assessment of a genetically modified crop. *Journal of Agricultural and Food Chemistry*, 62: 9916-9926.

Iwasa Y, Andreasen V and Levin S A, 1987. Aggregation in model ecosystems. I. Perfect aggregation. *Ecological Modelling*, 37(3-4): 287-302.

Iwasa Y, Levin S A and Andreasen V, 1989. Aggregation in model ecosystems II. Approximate aggregation. *Mathematical Medicine and Biology*, 6: 1-23.

Kang Q and Vahl C I, 2014. Statistical analysis in the safety evaluation of genetically-modified crops: equivalence tests. *Crop Science*, 54: 2183-2200.

Kristofersson D and Navrud S, 2005. Validity tests of benefit transfer — are we performing the wrong tests? *Environmental and Resource Economics*, 30: 279-286.

Lawo N C, Wäckers F L and Romeis J, 2009. Indian *Bt* cotton varieties do not affect the performance of cotton aphids. *PLoS ONE*, 4(3): e4804.

Liu J P and Chow S C, 1996. Comment: Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11: 306-312.

Lundgren J G and Wiedenmann R N, 2002. Coleopteran-specific Cry3Bb toxin from transgenic corn pollen does not affect the fitness of a non-target species, *Coleomegilla maculata* DeGeer (Coleoptera: Coccinellidae). *Environmental Entomology*, 31: 1213-1218.

Meissle M and Romeis J, 2009. The web-building spider *Theridion impressum* (Araneae: Theridiidae) is not adversely affected by *Bt* maize resistant to corn rootworms. *Plant Biotechnology Journal*, 7: 645-656.

Nakagawa S and Foster T M, 2004. The case against retrospective statistical power analyses with an introduction to power analysis. *Acta Ethologica*, 7: 103-108.

Patterson S D and Jones B, 2005. *Bioequivalence and Statistics in Clinical Pharmacology*. Boca Raton, USA: Chapman & Hall/CRC.

Paula D P and Andow D A, 2016. Differential Cry toxin detection and effect on *Brevicoryne brassicae* and *Myzus persicae* (Hemiptera: Aphidinae) feeding on artificial diet. *Entomolo-*

gia Experimentalis et Applicata, 159: 54-60.

Paula D P, Andow D A, Bellinati A, Timbó R V, Souza L M, Pires C S S and Sujii E R, 2016. Limitations in dose-response and surrogate species methodologies for risk assessment of Cry toxins on arthropod natural enemies. *Ecotoxicology*, 25: 601-607.

Perry J N, ter Braak C J F, Dixon P M, Duan J J, Hails R S, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F and van der Voet H, 2009. Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. *Environmental Biosafety Research*, 8: 65-78.

Phillips B M, Hunt J W, Anderson B S, Puckett H M, Fairey R, Wilson C J and Tjeerdema R, 2001. Statistical significance of sediment toxicity test results: threshold values derived by the detectable significance approach. *Environmental Toxicology and Chemistry*, 20: 371-373.

Raybould A, 2010. Reducing uncertainty in regulatory decision-making for transgenic crops: more ecological research or clearer environmental risk assessment? *GM Crops*, 1: 25-31.

Romeis J, Dutton A and Bigler F, 2004. *Bacillus thuringiensis* toxin (Cry1Ab) has no direct effect on larvae of the green lacewing *Chrysoperla carnea* (Stephens) (Neuroptera: Chrysopidae). *Journal of Insect Physiology*, 50: 175-183.

Romeis J, Hellmich R L, Candolfi M P, Carstens K, De Schrijver A, Gatehouse A M R, Herman R A, Huesing J E, McLean M A, Raybould A, Shelton A M and Waggoner A, 2011. Recommendations for the design of laboratory studies on non-target arthropods for risk assessment of genetically engineered plants. *Transgenic Research*, 20: 1-22.

Rosenfeld J S, 2002. Functional redundancy in ecology and conservation. *Oikos*, 98: 156-162.

Sasabuchi S, 1980. A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika*, 67: 429-439.

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

Simberloff D, 2005. The politics of assessing risk for biological invasions: the USA as a case study. *Trends in Ecology and Evolution*, 20: 216-222.

Suter G W I I, 2006. *Ecological Risk Assessment*. 2 ed. Boca Raton, FL: CRC Press.

USEPA, 2010. *National Pollutant Discharge Elimination System Test of Significant Toxicity Implementation Document: An Additional Whole Effluent Toxicity Statistical Approach for Analyzing Acute and Chronic Test Data. EPA* 833-*R*-10-003. Washington, DC: USEPA.

Vahl C I and Kang Q, 2016. Equivalence criteria for the safety evaluation of a genetically modified crop: a statistical perspective. *The Journal of Agricultural Science*, 154: 383-406.

van der Voet H, Perry J N, Amzal B and Paoletti C, 2011. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. *BMC Biotechnology*, 11: 15.

von Burg S, Müller C B and Romeis J, 2010. Transgenic disease-resistant wheat does not affect the clonal performance of the aphid *Metopolophium dirhodum* Walker. *Basic and Applied Ecology*, 11: 257-263.

Westlake W J, 1981. Response to T.B.L. Kirkwood: bioequivalence testing — a need to rethink. *Biometrics*, 37: 589-594.

(责任编辑:杨郁霞)