Animal Behaviour 145 (2018) 77–85

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

A new statistical method to test equivalence: an application in male and female eastern bluebird song

Evangeline M. Rose [a,*], Thomas Mathew [b], Derek A. Coss [a], Bernard Lohr [a], Kevin E. Omland [a]

[a] *Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, MD, U.S.A.*
[b] *Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, U.S.A.*

ARTICLE INFO

Article history:
Received 16 March 2018
Initial acceptance 14 May 2018
Final acceptance 27 July 2018

MS. number: A18-00204R

Keywords:
eastern bluebird
equivalence testing
female song
hypothesis testing
publication bias

Understanding complex natural systems requires approaches with minimal statistical limitations and biases. Previously, however, the fields of behaviour, ecology and evolution generally have been restricted to evaluating differences between statistical distributions. This framework can be very powerful. Unfortunately, it has also created a bias in these, and many other, fields towards publications focusing only on phenomena for which there are statistically significant differences. However, there is a wide range of questions that would benefit from reversing the existing paradigm and testing, instead, for equivalence. We have adapted the two one-sided test (TOST), also known as equivalence testing, from pharmaceutical science to be applicable to behavioural and ecological studies. We created a repeated measures analysis that allows researchers to statistically examine similarities between distributions. We compared song structure in male and female eastern bluebirds, *Sialia sialis*, as a case study for this method. We failed to find significant differences between male and female songs via a more traditional test, repeated measures ANOVA. Therefore, no definitive conclusion could be drawn about the similarities or differences in song structure. However, our repeated measures equivalence test showed that, based on five standard measures of song variation, male and female eastern bluebirds sing statistically equivalent songs. Our study highlights the presence of complex female song in a temperate songbird species. Additionally, we provide a new statistical test useful for expanding the statistical toolbox to assess new behavioural and ecological questions, and to help counteract publication bias in our fields.

Hypothesis testing is one of the foundations of scientific inquiry and a pillar upon which behavioural and ecological studies are often built. Classic hypothesis testing sets up a dichotomy in which a researcher either rejects or fails to reject the null hypothesis of equality. This arrangement allows researchers to determine, in the face of uncertainty, if two statistical distributions are significantly different. However, in the fields of behaviour, ecology and evolution, reversing this dichotomy is often useful and necessary (Table 1). There are many scenarios in which a researcher might ask whether two distributions are equal (e.g. territory sizes, niche spaces, boldness metrics). Whereas the traditional dichotomy can only assess differences, equivalence testing should be the primary tool to test hypotheses of equivalence and thus provides an opportunity to go beyond traditional null-hypothesis testing.

Equivalence testing is a modified hypothesis test that places the burden of proof on equivalence rather than difference (Limentani, Ringo, Ye, Bergquist, & McSoreley, 2005; Walker & Nowacki, 2011). Also called a two one-sided test (TOST), equivalence testing assumes a null hypothesis in which the means of two distributions are not equivalent. The alternative hypothesis is then that the two means are equivalent within an acceptable preset limit theta (Fig. 1). This theta is set by the researcher and is defined as the highest tolerated effect size difference. The use of a theta value allows researchers to look at equivalence between means within a small range as opposed to looking for equality, which implies two distributions that are identical in every way. TOST was originally designed to test bioequivalence of pharmaceutical treatments. In pharmaceutical testing, it is standard to use a very stringent value (ln(1.25)) as theta to ensure that inferior treatments do not make it to market (Chow & Liu, 2008; FDA, 2001; Hauschke, Steinijans, & Pigeot, 2007; Patterson & Jones, 2016). This value (ln(1.25)) indicates equivalence if the mean ratio of two distributions is approximately one plus or minus 2%. However, the theta used in

* Correspondence: E.M. Rose, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A.
E-mail address: eshank1@umbc.edu (E. M. Rose).

https://doi.org/10.1016/j.anbehav.2018.09.004
0003-3472/

Table 1
Difference in null and alternative hypotheses between traditional difference testing and equivalence hypothesis testing (modified from Walker & Nowacki, 2011)

Type of study	Null hypotheses	Alternative hypothesis
Difference testing	No difference between groups $\mu_1 = \mu_2$	Groups differ $\mu_1 \neq \mu_2$
Equivalence testing	Groups are not equivalent $\lvert\mu_1 - \mu_2\rvert \geq \Theta$	Groups are equivalent $\lvert\mu_1 - \mu_2\rvert < \Theta$

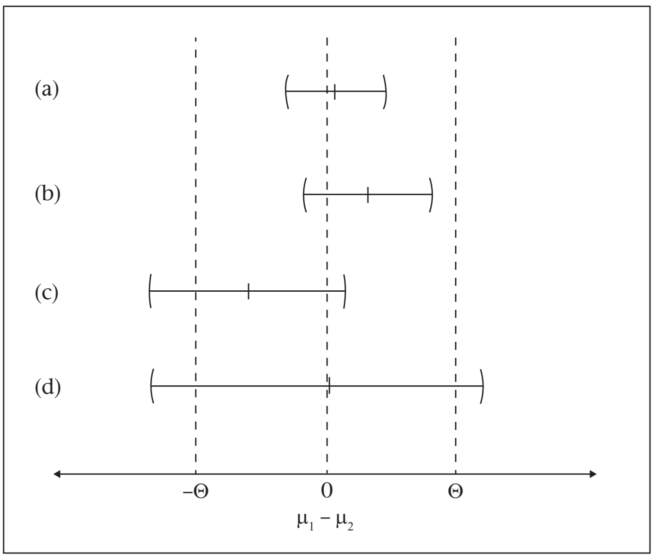

Figure 1. Four mean differences tested with the two one-sided test (TOST) are represented (modified from Limentani et al., 2005). Each line (a)–(d) represents a comparison of two distributions. The tick mark in the middle of each interval represents the difference of two means with a confidence interval around it. If theta on the *X* axis represents the largest tolerated effect size, then distributions compared in (a) and (b) are equivalent, but not the distributions in (c) and (d). This is because distributions (a) and (b) fall entirely within the preset limit theta. Additionally, the distributions in (d) are equal (not significantly different using traditional tests), because the difference in means equals zero, but they are not equivalent because the error around the difference in means is greater than theta.

TOST can be adapted to other studies in which case theta could be relaxed or tightened to include biologically relevant similarity. Limentani et al. (2005) provide guidelines for modifying theta. They conclude that theta should always be greater than s/sqrt(n), which accounts for failure due to imprecision, and less than any biologically relevant value chosen to test the data against (Limentani et al., 2005). For example, the ln(1.25) may be unnecessarily stringent if an animal's visual or auditory receptor organ is not sensitive enough to detect differences of that magnitude.

Although TOST was originally developed for pharmaceutical testing, it is increasingly being used in a wider range of fields such as psychology (Rogers, Howard, & Vessey, 1993), chemistry (Roy, 1997), environmental science (McBride, 1999), medicine (Munk, Hwang, & Brown, 2000) and software engineering (Dolado, Carmen, & Mark, 2014). In a few instances, TOST also has been used in ecological and behavioural studies. For example, Heisler, Somers, and Poulin (2016) used TOST to examine species richness, dominance and evenness in a meta-analysis of past studies. TOST has also been used to examine the effects of marking or capturing animals (Jones, Smith, Bebus, & Schoech, 2016; Knotts & Griffen, 2016). We propose to expand the use of equivalence testing in the fields of behaviour and ecology. Specifically, we modified this test to be more applicable to field studies and then used it to assess similarity between male and female song structure in eastern bluebirds, *Sialia sialis*.

Female Birdsong

The presence and elaboration of birdsong has long been an important focus of the study of sexual selection and complex signals. Birdsong has classically been considered an example of an elaborate male trait under sexual selection for territory defence and mate attraction (Catchpole & Slater, 2008). However, it is becoming increasingly apparent that song, as a complex vocal signal, is present in both sexes across many oscine passerines. In fact, Odom, Hall, Riebel, Omland, and Langmore (2014) found that female song is not only widespread but is also likely ancestral in songbirds. Previously, female song was thought to be rare or absent in temperate species (Langmore, 1998; Morton, 1996; Riebel, 2003; Slater & Mann, 2004), and less frequent and complex than male song when present (Arcese, Stoddard, & Hiebert, 1988; Beletsky, 1982; Price, Lanyon, & Omland, 2009). However, recent studies are expanding our knowledge of female song, both in temperate (e.g. Halkin, 1997; Krieg & Getty, 2015; Mahr, Evans, Thonhauser, Griggio, & Hoi, 2016) and tropical (e.g. Odom et al., 2016) regions, and the possible functions that elaborate female traits may serve. Therefore, in their recent review of female song, Odom and Benedict (2017, p. 38) defined song as 'any long, complex vocalization given by birds, especially during resource defence or mate-attraction situations'. We note here that this is not an all-inclusive definition of song due to the short and simple songs of some species. However, it is an important step towards acknowledging the prevalence of female song throughout the avian phylogeny.

Beginning a few decades ago, researchers began to document the presence and structure of female song. To help explain emerging patterns in female song, several researchers sought to classify female song function and structure in comparison to male song (Cain & Langmore, 2016; Ritchison, 1983). Many studies found that female songs, while more complex than typical calls, may be shorter (Odom et al., 2016), less variable overall (Brunton & Li, 2006; Hoelzel, 1986) or more variable (Johnson & Kermott, 1990; Pavlova, Pinxten, & Eens, 2005) and have a different (often decreased) bandwidth/frequency range (Kasumovic, 2003; Odom et al., 2016) than male song. However, other studies have shown that both males and females of some species sing (or may sing) structurally similar songs (Arcese et al., 1988; Campbell et al., 2016; Johnson & Kermott, 1990; Pavlova et al., 2005; Reichard et al., 2018). Furthermore, females may even learn their songs from both male and female tutors (Evans & Kleindorfer, 2016), and in some cases, females sing at higher rates than conspecific males (Illes, 2015; Illes & Yunes-Jimenez, 2009). In this study, we used equivalence testing to ask whether male and female eastern bluebirds have equivalent song characteristics based on five standard measures of acoustic variability.

Eastern bluebirds are a sexually dimorphic, socially monogamous thrush (Turdidae) found throughout central and eastern North America (Gowaty & Plissner, 2015). A mated pair maintains a territory and may have two to three broods per breeding season on that territory (Gowaty & Plissner, 2015). Male bluebirds are prolific singers, singing up to 20 songs per min during the breeding season (Ritchison & Huntsman, 2003). In addition to singing at a high rate, male eastern bluebirds are also capable of singing a wide variety of songs. Ritchison and Huntsman (2003) found that male bluebirds in central Kentucky, U.S.A., had repertoires containing anywhere from 40 to 81 song types. Across these various song types males generally sang within the frequency range of 2–4 kHz and produced songs with 3.02 ± 1.70 notes (mean ± SD) (Huntsman & Ritchison, 2001). Female song has also been documented in eastern