Three Statistical Methods for Evaluating Null Effects

Authors: Xu Yuepei, Lu Chunlei, Wang Jun, Song Qiongya, Jia Binbin, Hu Chuanpeng, Hu Chuanpeng

Date: 2021-04-25T00:00:00+00:00

Abstract

In psychological research, researchers may need to evaluate null effects in two situations: first, to infer the absence of an effect; second, when unexpected non-significant results occur, requiring differentiation between whether the effect truly does not exist or whether the current data fail to provide sufficient evidence. However, the commonly used Null Hypothesis Significance Test (NHST) cannot directly evaluate null effects. In recent years, three methods—equivalence testing, Bayesian estimation, and Bayes factor—have been gradually adopted to assess null effects: within the frequentist statistical framework, equivalence testing infers whether an effect is null by examining whether the effect falls within the Smallest Effect Size of Interest (SESOI) via p-values; within the Bayesian statistical framework, Bayesian estimation infers whether an effect is null by comparing the overlap between the Highest Density Interval of the posterior distribution and the practical equivalence region; while the Bayes factor assesses the relative support that the current data provide for the null hypothesis versus the alternative hypothesis, thereby inferring the relative degree of support for the null hypothesis. This article demonstrates the practical application of these three methods through the analysis of two real datasets. Each method has its own characteristics: equivalence testing is logically an extension of NHST and can be easily extended from traditional statistics; the interpretation of Bayes factor is relatively intuitive with clear logic; Bayesian estimation possesses strong flexibility and can be extended to more research questions. These three methods for evaluating null effects may assist psychological researchers in making reasonable statistical inferences and research decisions in practical studies.

Full Text

Introduction

In psychological research, researchers need to evaluate whether an effect is absent in two distinct situations: (1) when the research design or hypothesis requires demonstrating the absence of an effect, and (2) when researchers intended to reject a null effect but failed to do so (i.e., unexpectedly obtained $p > 0.05$ results), necessitating further distinction between insufficient evidence and genuine effect absence. However, conventional null hypothesis significance testing (NHST) cannot provide evidence supporting a null effect. In recent years, three methods—equivalence testing, Bayesian estimation, and Bayes factors—have been increasingly employed to evaluate null effects. This article introduces the principles of these three methods and demonstrates their practical application through two case studies. Each method has distinct characteristics: equivalence testing logically extends NHST and is easily adaptable from traditional statistics; Bayes factors offer intuitive interpretation with clear logical foundations; and Bayesian estimation provides greater flexibility, extending to a broader range of research questions. These three methods for evaluating null effects may assist psychological researchers in making appropriate statistical inferences and research decisions.

Conceptual Foundations for Evaluating Null Effects

Two primary approaches exist for evaluating null effects. One approach establishes a sufficiently small interval that can practically be considered zero, used to assess null effects (Meyners, 2012; Rogers et al., 1993). This interval is termed the "smallest effect size of interest," abbreviated as SESOI. When the target effect falls within the SESOI, researchers can consider the effect size practically zero and negligible. Two methods adopt this approach: equivalence testing under the frequentist framework and Bayesian estimation under the Bayesian framework. The alternative approach, employed by Bayes factors, sidesteps the question of whether the effect size is precisely zero by comparing the likelihood of the observed data under the null hypothesis (assuming zero effect) versus the alternative hypothesis (assuming a non-zero effect), thereby determining which hypothesis the data support.

Equivalence Testing

Equivalence testing extends traditional NHST to evaluate whether the current effect size is sufficiently small. Its logic derives from the minimal-effects test (Murphy, Myors, & Wolach, 2014). NHST compares an effect size to zero, assessing whether the probability of observing the current data under the assumption of zero effect (H_0) is sufficiently small to reject the null hypothesis (Figure 1A). If researchers instead specify H_0 as an interval, such as [-0.1, 0.1], rejecting this null hypothesis requires the effect size to be significantly greater than 0.1 or significantly less than -0.1 (Figure 1B), necessitating two one-sided

tests. This procedure is called a minimal-effects test.

Equivalence testing reverses the effect intervals for H_0 and H_1 from the minimal-effects test: H_1 falls within the interval while H_0 falls outside it (Lakens, McLatchie, Isager, Scheel, & Dienes, 2018; Lakens, Scheel, & Isager, 2018). With an SESOI of [-0.1, 0.1], the null hypothesis posits that the effect size lies outside this interval—either greater than 0.1 or less than -0.1 (Figure 1C), representing "the presence of a meaningful effect." The alternative hypothesis states that the effect size falls within [-0.1, 0.1], meaning it is too small to be considered "meaningful." If the data reject the null hypothesis, researchers can accept the alternative hypothesis of "no meaningful effect."

Beyond differing in meaning from traditional NHST hypotheses, equivalence testing imposes stricter requirements on null hypothesis specification. Whereas NHST assumes an effect size of exactly zero, equivalence testing requires researchers to specify the range of the null hypothesis—the interval outside the alternative hypothesis (SESOI). Based on existing research and practical considerations, SESOI can be established through specific approaches (see supplementary materials: osf.io/6mzr9), always requiring justification.

Figure 1. Schematic illustration of equivalence testing and Bayesian estimation principles. (A) Traditional null hypothesis significance testing; (B) Minimal-effects test; (C) Equivalence testing. ΔL represents the lower bound of SESOI, ΔU represents the upper bound; H_0: null hypothesis, H_1: alternative hypothesis. In Bayesian estimation inference, the highest density interval (HDI) and region of practical equivalence (ROPE) are combined to assess the credibility of effect size . Three possible outcomes exist: accepting the null effect (D), inability to make a definitive judgment (E), and rejecting the null effect (F).

In practice, equivalence testing requires two one-sided tests (TOST) comparing the observed data against the SESOI boundaries ΔL and ΔU. One test' s null hypothesis states that the observed effect size is less than ΔL; the other' s null hypothesis states it exceeds ΔU. The final inference combines both results: only when both p-values from the TOST are below the α level can researchers reject the null hypothesis and accept the alternative hypothesis ("no meaningful effect") following NHST logic. This constitutes statistical equivalence—the effect is sufficiently small to be considered negligible in the studied population. However, if either p-value exceeds α, researchers cannot reject the null hypothesis ("meaningful effect exists"), and equivalence cannot be supported (Lakens, Scheel, & Isager, 2018).

Notably, equivalence testing can also be implemented through parameter estimation. Under the frequentist framework, researchers estimate the effect size and its confidence interval (王珺 et al., 2019), then assess the proportion of overlap between this interval and the SESOI (Tryon, 2001).

Bayesian Estimation

Unlike frequentist equivalence testing, Bayesian estimation evaluates null effects within the Bayesian statistical framework. Bayesian and frequentist statistics differ fundamentally in their interpretation of probability. Frequentist probability represents the expected frequency across infinite repeated sampling—an outcome of long-run behavior. Bayesian probability represents the credibility of an event given available information (Kruschke, 2014; McElreath, 2018). In inferential statistics, frequentist approaches treat population parameters as fixed values, whereas Bayesian approaches treat them as random draws from probability distributions that update as data accumulate. The core of Bayesian statistics is Bayes' rule. When sampling to estimate a population parameter ( ), Bayes' rule yields:

$$P(\theta|data) = \frac{P(\theta)P(data|\theta)}{P(data)}$$

where $P(\theta|data)$ represents the posterior distribution—the probability distribution of unknown parameters given the data; $P(\theta)$ represents the prior distribution—beliefs about parameter values before observing data; $P(data|\theta)$ represents the likelihood—the probability or probability density of the current data given parameter value ; and $P(data)$ represents the total likelihood across all possible parameters.

Given a prior distribution and data likelihood, the posterior distribution represents the parameter' s probability distribution after incorporating both prior information and empirical data. In essence, Bayesian statistics continuously updates posterior distributions as data accumulate, thereby changing credibility across parameter values (Kruschke & Liddell, 2018).

When applying Bayesian estimation to evaluate null effects, statistical inference proceeds by comparing the parameter' s posterior distribution to its probability distribution under zero effect (Kirkwood & Westlake, 1981; Rouder, 2014; Westlake, 1976). The posterior distribution is represented by the highest density interval (HDI), while the zero-effect distribution is specified as the region of practical equivalence (ROPE)—a negligible effect interval including zero, analogous to SESOI in equivalence testing. After establishing ROPE, researchers examine the overlap between the 95% HDI and ROPE to assess the null effect, yielding three possible outcomes: accepting the null effect (Figure 1D), inability to make a definitive judgment (Figure 1E), or rejecting the null effect (Figure 1F). Specifically, when the 95% HDI falls entirely within ROPE, the most credible parameters are practically equivalent to zero, supporting acceptance of the null effect. When the 95% HDI partially overlaps ROPE, only some credible parameter values are equivalent to zero, precluding definitive conclusions. When the 95% HDI falls completely outside ROPE, the most credible parameters are all non-zero, warranting rejection of the null effect (Kruschke, 2011). In summary,

researchers evaluate null effects by comparing HDI against a ROPE constructed around zero.

Notably, Bayesian estimation is fundamentally a model-fitting process based on data, allowing researchers to employ different priors and models. This process requires careful consideration of prior distribution 合理性 and MCMC sampling convergence (see Depaoli & Schoot, 2017).

Bayes Factors

Although Bayes factors belong to Bayesian statistics, their approach to evaluating null effects differs from Bayesian estimation. The fundamental logic involves model comparison to obtain the relative credibility of different models given the data. It addresses which model the data better support—corresponding to H_0 or H_1 models in NHST. In the formula above, $P(data|\theta)$ represents not only the likelihood based on parameter priors but also the probability of observing current data if model H_0 or H_1 were true. The Bayes factor is defined as the ratio of these probabilities (Keysers, Gazzola, & Wagenmakers, 2020; Wagenmakers et al., 2018):

$$BF_{01} = \frac{P(data|H_0)}{P(data|H_1)}$$

The subscript in BF_{01} places 0 before 1, indicating BF_{01} as the Bayes factor for H_0 relative to H_1. Conversely, BF_{10} inverts the numerator and denominator, representing H_1 relative to H_0. A BF_{01} of 9 means the data are nine times more likely under H_0 than under H_1. After calculating Bayes factors, researchers can interpret the relative strength of evidence supporting the two models based on their magnitude. For interpreting Bayes factors, Lee and Wagenmakers (2013) provide classification statements based on Jeffreys (1961). For example, BF_{01} between 3 and 10 indicates moderate evidence supporting the null hypothesis (H_0).

As a Bayesian inference method, Bayes factors also involve prior selection. Priors are typically determined from previous research, such as using effect sizes and their distributions from meta-analyses. For original studies without relevant meta-analyses, a common practice is to use a standardized prior—for instance, in Bayesian t-tests, a Cauchy distribution for the alternative hypothesis prior (Rouder, Speckman, Sun, Morey, & Iverson, 2009), $\delta \sim \text{Cauchy}(x = 0, \gamma = 1)$. To make the alternative prior more realistic, the popular R package BayesFactor sets the default prior to Cauchy(0, 0.707).

Defining SESOI and ROPE

Both equivalence testing and Bayesian estimation employ an interval defining a sufficiently small or negligible effect. In equivalence testing, this is called the smallest effect size of interest (SESOI), while Bayesian estimation terms

it the region of practical equivalence (ROPE). Researchers in other fields use alternative names, such as the interval of clinical equivalence in clinical research (Lesaffre, 2008) or equivalence interval in pharmacology (Schuirmann, 1987). These concepts are fundamentally similar, all defining a small interval including zero effect—or a more realistic null effect. Given ROPE' s similarity to SESOI, we discuss this from the SESOI perspective.

By examining the relationship between the target effect and this interval, researchers can infer whether data support the null effect, reject it, or preclude judgment (Kruschke & Meredith, 2020; Lakens, Scheel, & Isager, 2018). For a given effect size interval from current data, a more lenient SESOI may result in the interval falling entirely within SESOI, supporting a null effect conclusion. Conversely, a narrower SESOI may lead to incomplete overlap, yielding an inconclusive result. Thus, SESOI specification directly influences null effect evaluation conclusions.

SESOI specification requires case-by-case analysis, but researchers must justify their choices regardless of method (Lakens et al., 2018). When previous research has explored the effect of interest, those results can inform SESOI. For example, Simonsohn (2015) recommends setting equivalence boundaries in replication studies at the effect size detectable with 33% power from previous research, reasoning that effects from studies with less than 33% power have more than a 66% probability of producing non-credible significant results (Simonsohn, Nelson, & Simmons, 2014). However, Simonsohn' s recommendation is not the only approach; Kordsmeyer and Penke (2017) suggest setting equivalence boundaries at the average effect size from previous studies and testing whether current data show significantly smaller effects. Yet this method cannot exclude influences of randomness and publication bias from prior research. Other perspectives propose setting equivalence boundaries at the threshold where previous studies just achieved significance (Lakens et al., 2018). A potentially more robust method uses the lower bound of effect size confidence intervals (90% or 95%) from meta-analyses as equivalence boundaries (Perugini, Gallucci, & Costantini, 2014). Finally, note that SESOI and ROPE interpretations differ between frequentist and Bayesian frameworks (Kruschke & Liddell, 2018; Kruschke & Meredith, 2020).

Figure S1. Workflow for three statistical methods evaluating null effects. Equivalence testing and Bayesian estimation require specifying an equivalence interval for the target effect beforehand, whereas Bayes factors do not. For implementation, equivalence testing, Bayesian estimation, and Bayes factors can be performed using R packages TOST, BEST, and BayesFactor, respectively; Bayes factors can also be implemented in JASP. Each method follows its specific rules for evaluating null effects, yielding conclusions supporting, rejecting, or remaining inconclusive about the null effect. When results are inconclusive, researchers may consider increasing sample size or adjusting experimental design for re-evaluation.

References

Kordsmeyer, T. L., & Penke, L. (2017). The association of three indicators of developmental instability with mating success in humans. *Evolution and Human Behavior*, *38*(6), 704–713. doi:10.1016/j.evolhumbehav.2017.08.002

Kruschke, J., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. doi:10.3758/s13423-016-1221-

Kruschke, J., & Meredith, M. (2020). BEST: Bayesian estimation supersedes the t-test. R package version 0.5.2. Retrieved from https://CRAN.R-project.org/package=BEST

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963

Lesaffre, E. (2008). Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, *66*(2),

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. doi:10.1177/1745691614528519

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*(6), 657–680. doi:10.1007/bf01068419

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. doi:10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. J. C. L. C. (2014). P-curve: A key to the file drawer. doi:10.1037/a0033242

Note: Figure translations are in progress. See original paper for figures.