

A Bayesian Alternative to the Two One-Sided Tests (TOST) Procedure: State of the Art and a Proposal for a Generalized Bayesian Equivalence Model

Antonio Monleon-Getino

amonleong@ub.edu

University of Barcelona

Method Article

Keywords: Bayesian statistics, equivalence testing, TOST, Bayes factor, ROPE, power prior, historical borrowing, biosimilars, bioequivalence, Jeffreys–Lindley paradox, Edgeworth expansion

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11307458/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A Bayesian Alternative to the Two One-Sided Tests (TOST)
Procedure:
State of the Art and a Proposal for a Generalized Bayesian
Equivalence Model
Integrating Continuous and Frequency-Based Prior Information,
with Theoretical Properties and Proofs

Antonio Monleón-Getino
BIOST3, Section of Statistics
Department of Genetics, Microbiology and Statistics
Universitat de Barcelona (UB)
amonleong@ub.edu

October 7, 2026

Abstract

Equivalence testing is a central statistical problem in bioequivalence, non-inferiority studies, biosimilar development, and reproducibility research. The classical solution, the Two One-Sided Tests (TOST) procedure, tests the composite null hypothesis that a parameter of interest lies outside a pre-specified equivalence interval $(-\delta, \delta)$. Although widely used, TOST inherits well-known limitations of frequentist hypothesis testing: it does not directly quantify the probability that the parameter lies within the equivalence region, it cannot formally incorporate external or historical evidence, and it treats equivalence and non-equivalence asymmetrically. Bayesian alternatives exist in the literature, based on posterior probability mass in a Region of Practical Equivalence (ROPE), on Bayes factors for interval null hypotheses, and on an extensive methodological literature on borrowing historical information via power priors, commensurate priors, and meta-analytic-predictive priors. However, a unified and practically implementable framework that (i) formally incorporates prior information coming from either continuous historical data or historical frequency (count) data, and (ii) controls the influence of that prior information through an explicit, tunable discounting mechanism, is still lacking. In this article we review the state of the art in Bayesian equivalence testing and historical borrowing, and we propose the *Generalized Bayesian TOST with Power-Discounted Priors* (GBT-PP), a model that unifies Normal-Normal and Beta-Binomial conjugate updating with power priors and a dual decision rule combining posterior ROPE probability and an interval-null Bayes factor. Beyond the applied proposal, we establish five original theoretical results for GBT-PP: monotonicity of the posterior mean in the discounting weight w (Proposition 1); asymptotic consistency of the decision rule, independently of w (Theorem 2); a Jeffreys–Lindley-type paradox for the interval-null Bayes factor under an increasingly diffuse reference prior (Theorem 4); a closed-form Bayesian sample-size formula quantifying the sample size saved by historical borrowing (Corollary 5); and a deterministic Edgeworth approximation to the posterior equivalence probability for the two-proportion case, avoiding reliance on Monte Carlo alone (Proposition 6). All five results are proved in full in Appendix C. We illustrate the method with two worked examples (a continuous bioequivalence scenario and a frequency-based comparison of cure rates), fully reproducible via the R implementation in Appendix B.

Keywords: Bayesian statistics, equivalence testing, TOST, Bayes factor, ROPE, power prior, historical borrowing, biosimilars, bioequivalence, Jeffreys–Lindley paradox, Edgeworth expansion

1 Introduction

Equivalence testing addresses a question that is, in a strict logical sense, the opposite of the question addressed by standard significance testing: instead of asking whether an effect differs from zero, it asks whether an effect is small enough, in absolute value, to be considered practically negligible (Lakens et al., 2020). Statistical equivalence testing was originally developed outside pharmacometrics, in the context of comparing experimental groups in the behavioral sciences (Rogers et al., 1993), before becoming the regulatory standard for bioequivalence assessment (Schuirmann, 1987; Wellek, 2010). This distinction is critical in bioequivalence trials, non-inferiority studies, biosimilar development, replication research, and, more generally, in any setting where the scientific or regulatory conclusion of interest is the *absence* of a meaningful difference rather than its presence (Meyners, 2012).

The standard frequentist solution is the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987; Wellek, 2010), in which two one-sided null hypotheses,

$$H_{01} : \theta \leq -\delta \quad \text{and} \quad H_{02} : \theta \geq \delta,$$

are tested at level α ; equivalence is declared if both are rejected, which is algebraically equivalent to requiring that a $(1 - 2\alpha) \times 100\%$ confidence interval for θ be entirely contained in $(-\delta, \delta)$. This equivalence between the intersection-union test and the confidence-interval inclusion rule was formally established by Berger and Hsu (1996).

TOST is simple, well understood, and implemented in essentially every statistical package. However, it suffers from three limitations that are particularly relevant in modern evidence-based decision making:

1. It does not produce a direct probabilistic statement such as “the probability that the true difference is practically negligible is p ”; it only allows rejection or non-rejection of a composite null hypothesis.
2. It has no native mechanism for incorporating prior or historical evidence (e.g., from prior bioequivalence studies, analytical characterization data, or historical control frequencies), even when such evidence is abundant and its formal use is increasingly encouraged by regulators and by an extensive clinical-trials literature on historical borrowing (Spiegelhalter et al., 1994; Viele et al., 2014; Schmidli et al., 2014; FDA, 2026; Niazi, 2026).
3. It inherits the general asymmetry of null-hypothesis significance testing. In an ordinary two-sided test of difference, non-significance is routinely, but incorrectly, interpreted as evidence of equivalence (Lakens et al., 2020; Dienes, 2014) – motivating equivalence-specific procedures such as TOST in the first place. TOST itself, however, retains a residual, directional asymmetry of its own: because its composite null hypothesis encodes *non*-equivalence ($H_{01} : \theta \leq -\delta$ or $H_{02} : \theta \geq \delta$), rejecting it is positive evidence *for* equivalence, but failing to reject it is, symmetrically, not evidence *for* non-equivalence either – TOST simply remains silent in that case.

Bayesian statistics offers, at least in principle, a coherent solution to all three problems: posterior probabilities provide direct probabilistic statements, priors provide a formal channel for historical

evidence, and Bayes factors provide symmetric evidence for and against competing hypotheses (Kruschke and Liddell, 2018). The purpose of this article is fourfold: (i) to review the existing Bayesian equivalence-testing literature, together with the closely related literature on historical borrowing via power priors and their extensions; (ii) to identify a specific methodological gap, namely the lack of a unified, practically usable framework that accommodates both continuous and frequency-based (count) prior information with an explicit and tunable discounting mechanism; (iii) to propose, formalize, and illustrate such a framework, which we call the *Generalized Bayesian TOST with Power-Discounted Priors* (GBT-PP); and (iv) to establish, with full proofs, a small set of original theoretical properties of GBT-PP (Section 5 and Appendix C) that were, to our knowledge, not previously available in closed form for this model class.

2 State of the Art

2.1 Classical Equivalence Testing: TOST

The TOST procedure, introduced by Schuirmann (1987) for bioequivalence assessment and later systematized by Wellek (2010), remains the regulatory standard for average bioequivalence. Its theoretical justification as an intersection-union test, and its formal equivalence with the confidence-interval inclusion rule, was established by Berger and Hsu (1996). Outside pharmacometrics, analogous ideas were proposed considerably earlier for comparing experimental groups in psychology (Rogers et al., 1993), and a broad methodological review of equivalence testing across application domains is given by Meyners (2012). The main practical strength of TOST is its direct correspondence with a simple confidence-interval rule (the $(1 - 2\alpha) \times 100\%$ CI-inclusion rule), which makes it easy to compute, communicate, and audit. Its main conceptual weakness, discussed extensively by Lakens et al. (2020), is that failing to reject H_{01} and H_{02} is not logically the same as having evidence *for* equivalence, and the procedure provides no quantitative measure of the strength of that evidence.

2.2 Bayesian Approaches to Equivalence and to Null Hypotheses

Two broad Bayesian strategies for equivalence testing have been developed in parallel.

Posterior-probability / ROPE approaches. The idea of replacing a t -test with full Bayesian posterior estimation, and examining the posterior distribution directly rather than a p -value, was popularized by Kruschke (2013) and later formalized as the “Bayesian New Statistics” by Kruschke and Liddell (2018), who argue that estimation with credible intervals and explicit posterior probabilities should replace binary hypothesis-testing decisions. Within this framework, an equivalence region (Region Of Practical Equivalence, ROPE) is defined a priori, and a decision is made by examining whether the highest density interval (HDI) of the posterior falls entirely inside, entirely outside, or partially overlaps the ROPE. This approach naturally generalizes TOST: the ROPE plays the role of $(-\delta, \delta)$, and the HDI plays the role of the confidence interval, but the underlying quantity, the posterior distribution, additionally allows computation of the exact posterior probability mass inside the ROPE. The ROPE/HDI approach is now implemented in mature, widely used software, notably the `bayestestR` package (Makowski et al., 2019), which computes ROPE probabilities and Bayes factors for a broad class of Bayesian models; our proposal can be seen as a minimal, TOST-specific complement to such general tools, focused specifically on the power-discounted historical-borrowing case. Dienes (2014) provides a closely related discussion of how Bayesian methods, more generally, allow non-significant frequentist results to be reinterpreted as

positive evidence for a null or equivalence hypothesis, directly reinforcing the motivation given in Section 1.

Bayes factor approaches for interval nulls. Morey and Rouder (2011) develop Bayes factors for testing interval null hypotheses of the form $H_0 : \theta \in (-\delta, \delta)$ against $H_1 : \theta \notin (-\delta, \delta)$, generalizing the point-null Bayes factor (Savage-Dickey ratio) to an interval, or “approximate equality,” null (see Remark 1 in Section 3.4 for a note on this hypothesis-labeling convention, which is deliberately reversed relative to the TOST hypotheses of Section 3.2). This is, to our knowledge, the most direct Bayesian analogue of the TOST null hypothesis structure, because it explicitly targets the same interval hypothesis that TOST tests, but yields a continuous measure of relative evidence (the Bayes factor) rather than a binary rejection decision. Lakens et al. (2020) provide a detailed practical comparison of Bayes factors and frequentist equivalence tests, showing that the two approaches are conceptually different (evidence for a null model versus rejection of a set of alternative models) but often converge in practice when priors and equivalence margins are chosen consistently.

Kpekenpa and Muthukumarana (2018) extend Bayesian equivalence testing to two-arm trials with *binary* outcomes, an important and comparatively underdeveloped case relative to the continuous-outcome literature. They develop the likelihood, prior, and posterior distributions for the risk difference (or related contrast) between two binomial proportions, examine a normal approximation to the posterior via a Taylor expansion, and implement full posterior inference via MCMC, applying the method to actual two-arm mortality trial data. This work is directly relevant to our proposal because it demonstrates that Bayesian equivalence testing for frequency (count) data is both feasible and already partially developed, but it does not address how *historical* frequency data (as opposed to only the current trial’s binomial counts) should be formally incorporated as prior information, nor does it unify the binary and continuous cases within a single decision framework. Our own Proposition 6 provides a closed-form alternative to their Taylor-based normal approximation for the two-proportion posterior, retaining an explicit, auditable skewness-correction term.

2.3 Historical Borrowing and Power Priors

A separate but closely related methodological literature addresses *how* historical data should be formally incorporated into a current analysis, independently of the equivalence-testing context. Ibrahim and Chen (2000) introduced the power prior, in which a historical likelihood is raised to a power $w \in [0, 1]$ before being combined with an initial prior, exactly the mechanism we adopt in Section 4. However, Neuenschwander et al. (2009) showed that this original, *unnormalized* power prior has an important technical deficiency: when w is treated as a free parameter rather than fixed by the analyst, the resulting joint prior on (θ, w) does not behave as intuitively expected, motivating the *normalized* power prior. Ibrahim et al. (2015) provide a comprehensive review of the power prior and its normalized, hierarchical, and multi-source extensions. Hobbs et al. (2011) propose *commensurate* priors, a hierarchical alternative in which the degree of borrowing is itself modeled as a function of the empirical agreement (“commensurability”) between historical and current data, rather than fixed a priori as in the simple power prior. Schmidli et al. (2014) develop *meta-analytic-predictive* (MAP) priors, together with a robust mixture extension that automatically discounts historical information when it conflicts with the current data, an approach that has become particularly influential in biosimilar development. Viele et al. (2014) provide a widely cited, regulator-facing overview of the use of historical control data in clinical trials, comparing power priors, hierarchical models, and MAP priors, and discussing the conditions under which historical borrowing is scientifically justified. Finally, Gamalo-Siebers et al. (2017) illustrate a related Bayesian extrapolation approach in the context of pediatric drug development, a use case structurally similar to biosimilar equivalence assessment in that substantial prior (adult, or originator-product) evidence

is available before the trial of primary interest is conducted.

Taken together, this literature establishes that historical borrowing is a mature and well-studied methodological problem in clinical trials generally, but, to our knowledge, none of these borrowing mechanisms has been formally integrated into a TOST-style *equivalence-testing* decision rule that treats continuous and frequency-based historical data symmetrically – the gap that motivates the present proposal (Section 2.5). Nor, to our knowledge, has this integration been accompanied by closed-form asymptotic, sample-size, or paradox-characterization results of the kind we provide in Section 5.

2.4 Regulatory Adoption of Bayesian Methods

Bayesian methods for regulatory decision-making have a long history. Spiegelhalter et al. (1994) provide an early and still widely cited synthesis of Bayesian approaches to randomized trials, including the explicit use of prior distributions representing external evidence and skepticism. More recently, the U.S. FDA’s guidance on Bayesian statistics for medical device trials (FDA, 2010) formally recognized the use of informative priors derived from historical or analogous device data as an acceptable basis for regulatory decision-making, and remains a concrete, verifiable regulatory precedent for the kind of historical-borrowing framework proposed here. The FDA released, in January 2026, a draft guidance on the use of Bayesian methodology in clinical trials of drugs and biological products (FDA, 2026), representing a significant evolution in its regulatory approach to evidence evaluation. Niazi (2026) provides a detailed review of this guidance, with particular emphasis on its implications for biosimilar development and the broader shift toward a justification-first approach to Bayesian clinical trial design. Together, these regulatory sources, combined with the historical-borrowing methodology reviewed in Section 2.3 (Viele et al., 2014; Schmidli et al., 2014; Gamalo-Siebers et al., 2017), strengthen the practical motivation for the present work: if regulators are moving toward justification-first, evidence-integrating Bayesian frameworks, there is a concrete need for equivalence-testing tools that can formally combine historical continuous evidence (e.g., analytical or pharmacokinetic data) and historical frequency evidence (e.g., historical event or cure rates) within the same statistical model.

2.5 Open Gaps

From this review we identify the following gaps, which motivate our proposal:

- Existing Bayesian equivalence frameworks (ROPE/HDI, interval-null Bayes factors) are typically presented separately, and separately for continuous data (Kruschke and Liddell, 2018; Kruschke, 2013; Morey and Rouder, 2011) or for binary/frequency data (Kpekpema and Muthukumarana, 2018).
- The historical-borrowing literature is mature and well developed for general clinical-trial endpoints (Ibrahim and Chen, 2000; Neuenschwander et al., 2009; Ibrahim et al., 2015; Hobbs et al., 2011; Schmidli et al., 2014; Viele et al., 2014), but none of these mechanisms has, to our knowledge, been integrated into a TOST-style Bayesian *equivalence* framework that treats continuous and frequency-based historical data symmetrically within a single decision rule.
- None of the reviewed equivalence-testing approaches provides an explicit, tunable mechanism for discounting historical prior information (continuous or frequency-based) as a function of its relevance or reliability, a feature that is standard in the power-prior and commensurate-prior literature (Ibrahim and Chen, 2000; Hobbs et al., 2011) but has not been incorporated into a TOST-style Bayesian equivalence framework.

- There is no widely available, minimal, reproducible implementation that lets an applied user move directly from either (a) a historical mean and standard deviation, or (b) a historical 2×2 frequency table, to a posterior probability of equivalence, complementary to general Bayesian-inference software such as `bayestestR` (Makowski et al., 2019).
- Finally, none of the reviewed frameworks provides closed-form theoretical results (monotonicity, asymptotics, paradox characterization, or sample-size formulas) for the power-discounted interval-null setting; these are provided here for the first time for GBT-PP (Section 5).

3 Mathematical Framework

3.1 Notation and the Equivalence Region

Let θ denote the parameter of scientific interest (e.g., a mean difference, a log-ratio of means, or a difference of two proportions), and let $\delta > 0$ denote a pre-specified equivalence margin, chosen on clinical, regulatory, or practical grounds, such that any $|\theta| < \delta$ is considered scientifically negligible. Define the equivalence region

$$\mathcal{E} = (-\delta, \delta).$$

3.2 Frequentist TOST

The TOST procedure tests

$$H_{01} : \theta \leq -\delta \quad \text{vs.} \quad H_{02} : \theta \geq \delta,$$

declaring equivalence if both one-sided tests reject at level α . Equivalently, letting $\hat{\theta}$ be the point estimate and $CI_{1-2\alpha}(\hat{\theta})$ the corresponding $(1 - 2\alpha) \times 100\%$ confidence interval (Berger and Hsu, 1996),

$$\text{Equivalence declared} \iff CI_{1-2\alpha}(\hat{\theta}) \subset \mathcal{E}.$$

3.3 Bayesian Equivalence via Posterior Probability Mass (ROPE)

Given a prior $p(\theta)$ and data D with likelihood $L(\theta; D)$, the posterior is

$$p(\theta \mid D) \propto L(\theta; D) p(\theta).$$

Define the posterior equivalence probability

$$P_{eq} = \Pr(\theta \in \mathcal{E} \mid D) = \int_{-\delta}^{\delta} p(\theta \mid D) d\theta. \quad (1)$$

Following Kruschke (2013); Kruschke and Liddell (2018), a natural Bayesian analogue of the TOST decision rule is

$$\text{Equivalence declared} \iff HDI_{1-2\alpha}(\theta \mid D) \subset \mathcal{E}, \quad (2)$$

where $HDI_{1-2\alpha}$ is the highest posterior density interval of mass $1 - 2\alpha$, chosen to match the credibility level of the classical $(1 - 2\alpha)$ TOST confidence interval. Rule (2) is a stricter, set-based criterion; P_{eq} in Eq. (1) is a continuous, more informative summary and is the primary quantity we use below.

3.4 Bayes Factor for Interval Null Hypotheses

Following Morey and Rouder (2011), define the complementary hypotheses

$$H_0 : \theta \in \mathcal{E}, \quad H_1 : \theta \notin \mathcal{E},$$

under a single encompassing prior $p(\theta)$. The interval-null Bayes factor is

$$BF_{01} = \frac{\Pr(H_0 | D) / \Pr(H_1 | D)}{\Pr(H_0) / \Pr(H_1)} = \frac{P_{eq} / (1 - P_{eq})}{\pi_{eq} / (1 - \pi_{eq})}, \quad (3)$$

where $\pi_{eq} = \Pr(\theta \in \mathcal{E})$ is the *prior* probability mass in $\mathcal{E}$, and P_{eq} is the posterior probability mass in $\mathcal{E}$ from Eq. (1). Equation (3) is a generalization of the Savage-Dickey density ratio to an interval (rather than point) null, and is the quantity we use, jointly with P_{eq} , in the dual decision rule proposed below.

Remark 1 (Note on hypothesis-labeling conventions). The labels $H_0 : \theta \in \mathcal{E}$ and $H_1 : \theta \notin \mathcal{E}$ above follow Morey and Rouder (2011)’s convention for interval-null Bayes factors, in which the more constrained (nested) hypothesis is labeled H_0 , consistent with the Savage–Dickey and Bayesian model-comparison tradition. This is the *opposite* convention to the frequentist TOST hypotheses of Section 3.2 (H_{01}, H_{02}), where *non-equivalence* is the hypothesis to be rejected, for regulatory burden-of-proof reasons. We retain Morey and Rouder’s original labeling here to remain consistent with that literature, but emphasize that H_0 in Eq. (3) and H_{01}, H_{02} in Section 3.2 denote different hypotheses and should not be conflated.

This same quantity is, in principle, computable with general-purpose software such as `bayestestR` (Makowski et al., 2019) for arbitrary posterior samples; our contribution is the explicit coupling of Eq. (3) with the power-discounted priors of Section 4, and the theoretical characterization of its limiting behaviour in Theorem 4.

4 Proposed Model: Generalized Bayesian TOST with Power-Discounted Priors (GBT-PP)

4.1 Rationale

We propose a framework that (i) accepts prior information in either of the two most common forms encountered in practice, continuous summary statistics (a historical mean and standard deviation) or frequency/count data (a historical 2×2 or s/f table), (ii) discounts that prior information through a single interpretable weight $w \in [0, 1]$, following the power-prior construction of Ibrahim and Chen (2000), and (iii) combines the ROPE probability P_{eq} and the interval-null Bayes factor BF_{01} into a single, pre-registered dual decision rule, addressing the fact that P_{eq} alone can be sensitive to the choice of prior variance/concentration while BF_{01} alone can behave counter-intuitively with very large samples (the same large-sample sensitivity that motivates the interval-null formulation in the first place, cf. Morey and Rouder, 2011).

We deliberately adopt the simplest, *fixed- w* version of the power prior, rather than the normalized power prior or the hierarchical commensurate-prior alternative. This is a conscious simplification, not an oversight: Neuenschwander et al. (2009) show that the unnormalized power prior has undesirable properties precisely when w is treated as a random, estimated parameter; by fixing w at a small set of pre-registered, analyst-chosen values and reporting a sensitivity curve (Section 6) rather than estimating w from the data, we avoid this specific technical pathology while retaining the interpretability advantages emphasized by Ibrahim et al. (2015). We view the fully hierarchical

treatments of borrowing developed by Hobbs et al. (2011) (commensurate priors) and Schmidli et al. (2014) (robust MAP priors) as natural and promising extensions of GBT-PP, discussed further in Section 8 and Section 9.

4.2 Prior Elicitation from Continuous Historical Data

Suppose historical data are summarized by $(\bar{x}_0, s_0, n_0)$ (mean, SD, sample size of a previous, comparable study or set of studies). Define the power-discounted prior

$$p_w(\theta \mid D_0) \propto [L(\theta; D_0)]^w p_0(\theta), \quad (4)$$

with $w \in [0, 1]$ a discount factor and $p_0(\theta)$ a (vague) initial prior. Under the usual normal approximation

$$L(\theta; D_0) \propto \exp \left[-\frac{n_0}{2s_0 < sup > 2} (\theta - \bar{x}_0) < /sup > 2 \right],$$

Eq. (4) reduces, with a flat $p_0(\theta)$, to

$$\theta \sim \mathcal{N} \left(\bar{x}_0, \frac{s_0 < sup > 2}{w n_0} \right), \quad (5)$$

i.e., the historical information contributes an *effective sample size* of $w n_0$ observations to the analysis, exactly as in the original power-prior construction of Ibrahim and Chen (2000): $w = 1$ corresponds to using the historical data with its full original weight, $w = 0$ corresponds to completely discarding it (recovering a vague, data-free prior), and intermediate values of w interpolate continuously between the two. This is the sense in which GBT-PP is a *generalized* TOST: setting $w = 0$ for all historical inputs recovers, in the large- n limit, the classical TOST decision (Corollary 3 below); $w > 0$ formally injects historical evidence, with w itself playing the role of a pre-registered, auditable “trust level” assigned to that evidence.

4.3 Prior Elicitation from Historical Frequency (Count) Data

Suppose instead that historical information is available as a 2×2 or s/f table, i.e., as s_0 successes out of n_0 historical trials (e.g., a historical cure rate, response rate, or conversion rate). Applying the same power-prior construction (4) to a Binomial likelihood with a Beta initial prior $p_0(\theta) = \text{Beta}(a_0, b_0)$ gives, by conjugacy (Appendix A.2),

$$\theta \sim \text{Beta}(a_0 + w s_0, b_0 + w f_0), \quad f_0 = n_0 - s_0. \quad (6)$$

As in the continuous case, $w s_0$ and $w f_0$ are the effective (discounted) numbers of historical successes and failures contributed to the prior; $w = 0$ recovers the initial prior $\text{Beta}(a_0, b_0)$ unaltered, and $w = 1$ uses the full historical table.

4.4 Posterior Update and the GBT-PP Decision Rule

Continuous case. Combining the power-discounted prior (5) with the current-study likelihood $\bar{x} \mid \theta \sim \mathcal{N}(\theta, s < sup > 2/n)$ yields, by standard Normal-Normal conjugacy (Appendix A.1), the posterior

$$\theta \mid D \sim \mathcal{N}(\theta_n, \tau_n < sup > 2), \quad (7)$$

where

$$\tau_n < sup > 2 = \left(\frac{w n_0}{s_0 < sup > 2} + \frac{n}{s < sup > 2} \right)^{-1}, \quad (8)$$

$$\theta_n = \tau_n < sup > 2 \left(\frac{w n_0}{s_0 < /sup > 2} \bar{x}_0 + \frac{n}{s < sup > 2} \bar{x} \right). \quad (9)$$

Frequency case. Combining (6) with current data s successes out of n trials yields, again by conjugacy,

$$\theta \mid D \sim \text{Beta}(a_0 + w s_0 + s, b_0 + w f_0 + f), \quad f = n - s. \quad (10)$$

Dual decision rule. For either case, P_{eq} is computed from the posterior via Eq. (1) (in closed form for (7)–(9); via the incomplete Beta function, or the Edgeworth approximation of Proposition 6 for two-proportion contrasts, for (10)). A central methodological point, which we make explicit here to avoid any ambiguity, is that BF_{01} in Eq. (3) is *not* computed under a single encompassing prior shared with the posterior. Specifically:

- The posterior P_{eq} is obtained from the power-discounted prior $p_w(\theta \mid D_0)$ of Eqs. (5)/(6), which depends on w and, in the continuous case, is centered at the historical mean $\bar{x}_0$.
- The prior probability $\pi_{eq} = \Pr(\theta \in \mathcal{E})$ used in the denominator of BF_{01} is computed instead under a separate, *fixed reference prior* $p_{ref}(\theta)$ – e.g., $\mathcal{N}(0, \sigma_0 < /sup > 2)$ in the continuous case – that does *not* depend on w and is not centered at $\bar{x}_0$ (Remark 3, Theorem 4). We therefore write π_{eq} without a w subscript throughout, to emphasize that it is evaluated under p_{ref} , not under $p_w(\cdot \mid D_0)$ with the current-study data removed.

This decoupling is a deliberate design choice, not an inconsistency: Theorem 4 shows that computing π_{eq} under an increasingly diffuse member of the *same* power-discounted prior family would drive BF_{01} to $+\infty$ trivially, for any data and regardless of w . Decoupling the reference prior used for π_{eq} from the power-discounted prior used for the posterior is therefore necessary for BF_{01} to remain a meaningful, auditable quantity. As a consequence, BF_{01} as used in this article is a hybrid quantity – the ratio of posterior odds under $p_w(\theta \mid D_0, D)$ to prior odds under $p_{ref}(\theta)$ – rather than a classical Bayes factor computed under one single encompassing prior throughout; we adopt this hybrid definition following the design recommendation of Remark 3, and flag it explicitly here so that it is not mistaken for the standard single-prior Savage–Dickey construction of Morey and Rouder (2011).

GBT-PP declares equivalence under the pre-registered dual rule

$$\text{Equivalence declared} \iff P_{eq} \geq \gamma \text{ and } BF_{01} \geq BF < sup > *, \quad (11)$$

for pre-specified thresholds $\gamma \in (0, 1)$ (e.g., $\gamma = 0.95$) and $BF < /sup > * > 1$ (e.g., $BF < sup > * = 3$, following conventional Bayes-factor evidence bands). Rule (11) requires *both* sufficient posterior mass in $\mathcal{E}$ *and* sufficient relative evidence for H_0 over H_1 compared to what the reference prior alone implied, which guards against the two respective failure modes of using P_{eq} or BF_{01} in isolation (Section 5.4 below; see also the discussion of Bayes-factor behaviour with large samples in Morey and Rouder, 2011).

Remark 2. All quantities in (7)–(11) are indexed by w ; we write $\theta_n(w)$, $\tau_n < /sup > 2(w)$, $P_{eq}(w)$, $BF_{01}(w)$ whenever the dependence on the discount weight needs to be made explicit, as in Section 5. π_{eq} is, by construction (see above), evaluated under the fixed reference prior $p_{ref}(\theta)$ and therefore does not carry a w index.

5 Theoretical Properties of GBT-PP

The applied literature on power priors and historical borrowing (Ibrahim and Chen, 2000; Neuen-schwander et al., 2009; Hobbs et al., 2011; Schmidli et al., 2014; Viele et al., 2014) establishes qualitative properties of discounting (more borrowing $\Rightarrow$ more influence of historical data; less data $\Rightarrow$ more sensitivity to the prior) but, to our knowledge, provides no closed-form characterization of these properties specifically for the interval-null, equivalence-testing setting of Section 4. In this section we establish five such results for GBT-PP, all proved in full in Appendix C.

5.1 Monotonicity of the Posterior Mean in the Discount Weight

Proposition 1 (Monotonicity of θ_n in w). *For the continuous-data posterior mean $\theta_n(w)$ in Eq. (9),*

$$\frac{\partial \theta_n}{\partial w} = \frac{\frac{n_0}{s_0 < \sup > 2} \cdot \frac{n}{s < \sup > 2} (\bar{x}_0 - \bar{x})}{\left(\frac{w n_0}{s_0 < \sup > 2} + \frac{n}{s < \sup > 2} \right)^2}. \quad (12)$$

Consequently, $\theta_n(w)$ is monotone increasing in w if $\bar{x}_0 > \bar{x}$, monotone decreasing if $\bar{x}_0 < \bar{x}$, and constant ($= \bar{x}$) if $\bar{x}_0 = \bar{x}$; in all cases $\theta_n(w)$ interpolates continuously between $\theta_n(0) = \bar{x}$ and $\theta_n(1)$, the latter being the fully-borrowed precision-weighted mean of $\bar{x}_0$ and $\bar{x}$.

Proof. See Appendix C.1. □

Proposition 1 makes precise, quantitatively, the qualitative statement — common in the borrowing literature but rarely derived in closed form for this specific model — that historical information “pulls” the current-study estimate toward the historical mean at a rate governed entirely by the ratio of precisions $\frac{n_0/s_0 < \sup > 2}{n/s < \sup > 2}$, independently of δ .

5.2 Asymptotic Consistency of the Decision Rule

Theorem 2 (Asymptotic consistency of GBT-PP, uniformly in w). *Fix $w \in [0, 1]$, $n_0, s_0 > 0$, $\delta > 0$. Suppose the current-study estimator satisfies $\bar{x} \xrightarrow{P} \theta_0$ and $s < \sup > 2 \xrightarrow{P} \sigma < \sup > 2 \in (0, \infty)$ as $n \rightarrow \infty$. Then*

$$\tau_n < \sup > 2(w) \xrightarrow{P} 0, \quad \theta_n(w) \xrightarrow{P} \theta_0, \quad (13)$$

and, if $\theta_0 \neq \pm\delta$,

$$P_{eq}(w) \xrightarrow{P} \mathbb{K}\{|\theta_0| < \delta\} \quad \text{for every } w \in [0, 1]. \quad (14)$$

Proof. See Appendix C.2. □

Corollary 3 (Large-sample agreement with TOST). *Under the conditions of Theorem 2, and for any fixed $w \in [0, 1]$, the GBT-PP decision rule (11) and the classical TOST decision rule of Berger and Hsu (1996) converge in probability to the same decision as $n \rightarrow \infty$: both declare equivalence with probability tending to 1 if $|\theta_0| < \delta$, and with probability tending to 0 if $|\theta_0| > \delta$.*

Proof. See Appendix C.2. The argument combines (14) with the classical consistency of the TOST confidence-interval rule (an immediate consequence of $CI_{1-2\alpha}(\hat{\theta})$ shrinking to the point θ_0 ; see Berger and Hsu, 1996, Section 2). □

Theorem 2 and Corollary 3 show that historical borrowing under GBT-PP is asymptotically *harmless* in the following precise sense: regardless of how much weight w is assigned to historical information, and regardless of whether that historical information happens to be reliable, the current-study data eventually dominate the posterior and GBT-PP converges to exactly the decision that TOST would reach with the same current-study data alone. This addresses, formally, a natural regulatory concern about power-prior discounting (cf. the general discussion in Viele et al., 2014) that is left qualitative in the literature we reviewed in Section 2.3.

5.3 A Jeffreys–Lindley-Type Paradox for the Interval-Null Bayes Factor

Theorem 4 (Jeffreys–Lindley paradox for BF_{01} under a diffuse reference prior). *Fix any discount weight $w \in [0, 1]$ (the result below concerns only the reference prior used to compute π_{eq} , not the power-discounted prior of Section 4.2 used for the posterior, and therefore holds regardless of w ; we state it for $w = 0$ purely for notational simplicity, with no loss of generality). Let the reference prior be $p_{ref}(\theta) = \mathcal{N}(0, \sigma_0 < /sup > 2)$, proper for every finite $\sigma_0 > 0$. Then*

$$\begin{aligned}\pi_{eq}(\sigma_0) &= \Phi\left(\frac{\delta}{\sigma_0}\right) - \Phi\left(\frac{-\delta}{\sigma_0}\right) \\ &= \frac{2\delta}{\sigma_0} \phi(0) + O(\sigma_0 < /sup > -3) \xrightarrow{\sigma_0 \rightarrow \infty} 0,\end{aligned}\tag{15}$$

and, for any fixed data D such that $P_{eq} \in (0, 1)$ is bounded away from 0 and 1,

$$BF_{01}(\sigma_0) = \frac{P_{eq}/(1 - P_{eq})}{\pi_{eq}(\sigma_0)/(1 - \pi_{eq}(\sigma_0))} \xrightarrow{\sigma_0 \rightarrow \infty} \infty,\tag{16}$$

independently of the observed data.

Proof. See Appendix C.3. □

Theorem 4 is a translation, into the interval-null setting of Morey and Rouder (2011), of the classical Jeffreys–Lindley paradox: as the reference prior becomes increasingly diffuse, the prior odds in favor of the (interval) null tend to zero *faster* than the corresponding posterior odds can, so the Bayes factor diverges to $+\infty$ trivially, for any data whatsoever. This has an important practical consequence for how the reference prior of the Bayes-factor component should be chosen, which we record as a design recommendation.

Remark 3 (Design recommendation). Theorem 4 formally justifies the practical recommendation, already qualitatively present in the Bayes-factor literature (Morey and Rouder, 2011; Lakens et al., 2020), that the reference prior variance $\sigma_0 < /sup > 2$ used to compute π_{eq} in the Bayes-factor component of the dual rule (11) must be *fixed and pre-registered*, and should not be allowed to become arbitrarily large (“objective” or “flat”) without a corresponding upper bound, since BF_{01} can otherwise be made arbitrarily favorable to H_0 by the analyst’s choice of σ_0 alone, with no data-dependence whatsoever.

5.4 A Closed-Form Bayesian Sample-Size Formula

Corollary 5 (Sample size required for a target posterior equivalence probability). *Assume $\theta_0 = 0$ (the design is powered under exact equivalence) and let $\gamma \in (0, 1)$ be a target posterior probability. Let $z_\gamma = \Phi < /sup > -1\left(\frac{1+\gamma}{2}\right)$. Then, for fixed w , n_0 , s_0 , δ , and $s < /sup > 2$ treated as known,*

$$P_{eq}(w) \geq \gamma \iff \tau_n(w) \leq \frac{\delta}{z_\gamma}$$

$$\iff n \geq n < \text{sup} > \star(w) := s < \text{sup} > 2 \left(\frac{z_\gamma < \text{sup} > 2}{\delta < \text{sup} > 2} - \frac{w n_0}{s_0 < \text{sup} > 2} \right). \quad (17)$$

In particular, $n < \text{sup} > \star(0) = s < \text{sup} > 2 z_\gamma < \text{sup} > 2 / \delta < \text{sup} > 2$ is exactly the classical fixed-margin sample size, and

$$\Delta n < \text{sup} > \star(w) := n < \text{sup} > \star(0) - n < \text{sup} > \star(w) = \frac{w n_0 s < \text{sup} > 2}{s_0 < \text{sup} > 2} \quad (18)$$

is the number of current-study observations “saved” by borrowing $w n_0$ effective historical observations.

Proof. See Appendix C.4. □

Eq. (18) is, to our knowledge, a new closed-form design result: it expresses the sample size saving from historical borrowing directly in units of the current study’s own variance ($s < \text{sup} > 2 / s_0 < \text{sup} > 2$ converts historical to current-study-equivalent information), for the specific case of an interval-null (equivalence) target rather than a point-alternative power calculation. It connects directly to the general borrowing literature reviewed in Section 2.3 (Viele et al., 2014; Schmidli et al., 2014), providing the explicit design-stage correlate that this literature discusses only qualitatively or via simulation for the equivalence-testing case.

Remark 4 (Practical caveat: $n < \text{sup} > \star(w) < 0$). If the discounted historical information is sufficiently strong relative to the target precision, i.e., if $w n_0 / s_0 < \text{sup} > 2 \geq z_\gamma < \text{sup} > 2 / \delta < \text{sup} > 2$, Eq. (17) yields $n < \text{sup} > \star(w) \leq 0$: formally, the target posterior probability γ is already met by the (discounted) historical information alone, before any current-study data are collected. While mathematically valid, this situation should not, in general, be interpreted as a literal recommendation to run a trial with zero or negative sample size; it instead signals that the design is historical-information-dominated, a situation that is unlikely to be considered an adequate basis for a regulatory equivalence claim without at least a minimal confirmatory current-study sample.

5.5 An Edgeworth Approximation for the Two-Proportion Case

For the frequency-based version of GBT-PP with $\theta = p_1 - p_2$ (e.g., a treatment arm with posterior (10) contrasted against an independent control-arm posterior), P_{eq} has no closed form and is currently obtained by Monte Carlo (Appendix A.2). We provide a deterministic alternative.

Proposition 6 (Edgeworth approximation to P_{eq} for a difference of two Beta posteriors). *Let $p_1 \sim \text{Beta}(a_1, b_1)$ and $p_2 \sim \text{Beta}(a_2, b_2)$ be independent, and let $\theta = p_1 - p_2$. Write, for $i = 1, 2$,*

$$\begin{aligned} \mu_i &= \frac{a_i}{a_i + b_i}, \\ \sigma_i < \text{sup} > 2 &= \frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)}, \\ \gamma_{1,i} &= \frac{2(b_i - a_i) \sqrt{a_i + b_i + 1}}{(a_i + b_i + 2) \sqrt{a_i b_i}}, \end{aligned} \quad (19)$$

and let $\mu_\theta = \mu_1 - \mu_2$, $\sigma_\theta < \text{sup} > 2 = \sigma_1 < \text{sup} > 2 + \sigma_2 < \text{sup} > 2$,

$$\gamma_{1,\theta} = \frac{\gamma_{1,1} \sigma_1 < \text{sup} > 3 - \gamma_{1,2} \sigma_2 < \text{sup} > 3}{\sigma_\theta < \text{sup} > 3}. \quad (20)$$

Then, with $z < sup > \pm = (\pm\delta - \mu_\theta)/\sigma_\theta$, the first-order Edgeworth expansion gives

$$\begin{aligned} \hat{P}_{eq} < /sup > Edge = & \left[\Phi(z < sup > +) - \frac{\gamma_{1,\theta}}{6} \phi(z < /sup > +)(z < sup > +2 - 1) \right] \\ & - \left[\Phi(z < /sup > -) - \frac{\gamma_{1,\theta}}{6} \phi(z < sup > -)(z < /sup > -2 - 1) \right] + O(\sigma_\theta < sup > 3), \end{aligned} \quad (21)$$

which reduces to the plain normal-moment-matching approximation $\Phi(z < /sup > +) - \Phi(z < sup > -)$ when $\gamma_{1,\theta} = 0$ (e.g., $a_1 = b_1$ and $a_2 = b_2$ jointly).

Proof. See Appendix C.5. □

Proposition 6 provides a deterministic, auditable alternative to the Monte Carlo estimator $\hat{P}_{eq} < /sup > MC$ used elsewhere in this paper (Appendix A.2, Section 7.2), improving on the plain normal approximation used, in a related context, by Kpekpena and Muthukumarana (2018) via a first-order skewness correction of order $O(n < sup > -1/2)$ (since $\gamma_{1,i} = O((a_i + b_i) < /sup > -1/2)$). The discrepancy $|\hat{P}_{eq}^{Edge} - \hat{P}_{eq}^{MC}|$ can be reported as a diagnostic of the adequacy of the normal approximation to the Beta posterior, and is illustrated numerically in Section 7.2.

6 Sensitivity Analysis

Because w is fixed rather than estimated (Section 4.1), GBT-PP requires reporting how the decision in Eq. (11) varies over a pre-registered grid of plausible values $w \in \{0, 0.25, 0.5, 0.75, 1\}$, rather than a single point estimate of w . Proposition 1 guarantees that $\theta_n(w)$ varies *monotonically* and *smoothly* over this grid for the continuous case, so that a sensitivity plot of $P_{eq}(w)$ and $BF_{01}(w)$ against w is guaranteed to be free of erratic, non-monotone artefacts that would otherwise be difficult to distinguish from numerical error; this is a direct practical payoff of Proposition 1 beyond its theoretical interest. We recommend reporting the full curves $P_{eq}(w)$ and $BF_{01}(w)$ for $w \in [0, 1]$ (not only the grid points) since both are available in closed form for the continuous case via Eq. (9), and via the deterministic Edgeworth approximation of Proposition 6 for the two-proportion case, so that no simulation is required to produce the full sensitivity curve.

7 Worked Examples

7.1 Continuous Outcome: Bioequivalence-Type Scenario

Consider a historical dataset with $\bar{x}_0 = 0.02$, $s_0 = 0.15$, $n_0 = 24$, and a current study with $\bar{x} = 0.03$, $s = 0.12$, $n = 30$, on a scale where the equivalence margin is set at $\delta = 0.05$. For the Bayes-factor component we fix, per Remark 3, a pre-registered reference prior standard deviation $\sigma_0 = 10 s_0 / \sqrt{n_0} \approx 0.3062$, held fixed and *independent of w* (consistent with Remark 3 and the implementation in Appendix B), evaluated *without* borrowing the historical mean shift into the location of the reference prior (i.e., π_{eq} is computed under $\mathcal{N}(0, \sigma_0 < sup > 2)$, isolating the effect of prior *precision* from prior *location*). Table 1 reports $\theta_n(w)$, $\tau_n(w)$, $P_{eq}(w)$, π_{eq} , and $BF_{01}(w)$ from Eqs. (9) and (3), computed exactly from these formulas using the function `gbtpp_continuous` of Appendix B.

Consistent with Proposition 1, $\theta_n(w)$ decreases monotonically toward $\bar{x}$'s precision-weighted combination with $\bar{x}_0$ as w increases, since $\bar{x}_0 = 0.02 < \bar{x} = 0.03$; consistent with Theorem 2, increasing w pulls the posterior toward the historical value without changing the sign of the underlying qualitative conclusion. Because the reference-prior standard deviation σ_0 used for the Bayes-factor

Table 1: Continuous case. Values computed from Eqs. (9) and (3) with $\bar{x}_0 = 0.02$, $s_0 = 0.15$, $n_0 = 24$, $\bar{x} = 0.03$, $s = 0.12$, $n = 30$, $\delta = 0.05$. Following Remark 3, the reference-prior standard deviation used to compute π_{eq} is fixed at $\sigma_0 = 10 s_0/\sqrt{n_0} \approx 0.3062$ for *all* values of w (i.e., σ_0 does not depend on w), so π_{eq} is constant across the row.

w	$\theta_n(w)$	$\tau_n(w)$	$P_{eq}(w)$	π_{eq}	$BF_{01}(w)$
0.00	0.03000	0.02191	0.8192	0.1297	30.40
0.25	0.02887	0.02063	0.8471	0.1297	37.18
0.50	0.02796	0.01955	0.8702	0.1297	44.96
0.75	0.02723	0.01863	0.8893	0.1297	53.90
1.00	0.02661	0.01783	0.9053	0.1297	64.15

component is fixed and does not depend on w (Remark 3), the prior probability π_{eq} is constant across the entire row ($\pi_{eq} \approx 0.130$), and $BF_{01}(w)$ increases monotonically with w , tracking the increase in $P_{eq}(w)$: as more historical information is borrowed, the posterior concentrates further inside $\mathcal{E}$ relative to a fixed external benchmark, so both components of the dual decision rule (Eq. 11) move in the same direction. This is the intended behavior of BF_{01} once σ_0 is correctly pre-registered and held fixed, as opposed to the spurious, w -dependent divergence that would occur if σ_0 were allowed to depend on w (the specific failure mode characterized in Theorem 4 and guarded against in Remark 3).

7.2 Frequency Outcome: Comparison of Cure Rates

Consider a historical trial with $s_0 = 45$ cures out of $n_0 = 50$ patients, and a current trial with $s = 42$ cures out of $n = 50$ patients, combined with a Laplace (Beta(1,1)) initial prior, so that, from Eq. (10), the treatment-arm posterior is Beta($a_{ref}(w)$, $b_{ref}(w)$) with

$$a_{ref}(w) = 43 + 45w, \quad b_{ref}(w) = 9 + 5w. \quad (22)$$

We contrast this treatment-arm posterior against an independent, non-borrowed control-arm posterior $p_2 \sim \text{Beta}(38, 14)$ (representing 37 cures out of 50 control patients, Laplace prior), and set the equivalence margin at $\delta = 0.10$ on the risk-difference scale. P_{eq} is computed three ways: (a) the Monte Carlo estimator of Appendix A.2 with $M = 200,000$ draws, reporting the Monte Carlo standard error $\widehat{SE}(\hat{P}_{eq}) = \sqrt{\hat{P}_{eq}(1 - \hat{P}_{eq})/M}$; (b) the Edgeworth approximation of Proposition 6; and (c) the plain normal-moment approximation ($\gamma_{1,\theta} = 0$ in Eq. 21), for comparison. The values below were obtained with the function `gbtpp_frequency` of Appendix B.

Unlike the continuous example of Section 7.1, here increasing w *decreases* P_{eq} : the historical cure rate ($s_0/n_0 = 45/50 = 0.90$) exceeds the current-study cure rate ($s/n = 42/50 = 0.84$), so borrowing historical information shifts the treatment-arm posterior mean upward (from $\mu_1 \approx 0.827$ at $w = 0$ to $\mu_1 \approx 0.863$ at $w = 1$), *widening* the gap with the control-arm posterior ($\mu_2 \approx 0.731$) rather than narrowing it, and moving $\theta = p_1 - p_2$ further away from the equivalence region $(-0.10, 0.10)$. This is a direct frequency-data analogue of Proposition 1: the direction in which borrowing moves the posterior, and hence $P_{eq}(w)$, is governed entirely by the sign of the discrepancy between the historical and current rates, here $s_0/n_0 - s/n = 0.90 - 0.84 > 0$, exactly as $\bar{x}_0 - \bar{x}$ governs the direction of movement in the continuous case. Practically, this example illustrates that historical borrowing is not automatically favorable to a finding of equivalence: it helps when the historical estimate is closer to the comparator than the current estimate, and hurts otherwise, reinforcing

Table 2: Frequency case. Treatment-arm posterior $\text{Beta}(a_{ref}(w), b_{ref}(w))$ from Eq. (22); control-arm posterior $\text{Beta}(38, 14)$; $\delta = 0.10$. Monte Carlo column uses $M = 200,000$ draws; $\widehat{SE}$ is the Monte Carlo standard error of $\hat{P}_{eq} < /sup > MC$.

w	a_{ref}	b_{ref}	$\hat{P}_{eq}^{MC} (\widehat{SE})$	$\hat{P}_{eq}^{Edge}$	$\hat{P}_{eq}^{Normal}$
0.00	43.00	9.00	0.5112 (0.0011)	0.5117	0.5120
0.25	54.25	10.25	0.4478 (0.0011)	0.4468	0.4431
0.50	65.50	11.50	0.3983 (0.0011)	0.3973	0.3915
0.75	76.75	12.75	0.3603 (0.0010)	0.3593	0.3526
1.00	88.00	14.00	0.3314 (0.0009)	0.3298	0.3228

the recommendation of Section 5 to report the full sensitivity curve $P_{eq}(w)$ rather than a single, borrowed point estimate. As before, the Edgeworth approximation of Proposition 6 tracks the Monte Carlo estimate closely across the whole range of w (maximum absolute discrepancy 0.0016), and is consistently closer to Monte Carlo than the plain normal approximation, confirming the practical value of the skewness correction even in this asymmetric-borrowing scenario.

8 Discussion

The theoretical results of Section 5 clarify, in closed form, several properties of GBT-PP that are typically discussed only qualitatively, or checked only by simulation, in the general historical-borrowing literature (Ibrahim and Chen, 2000; Neuenschwander et al., 2009; Hobbs et al., 2011; Schmidli et al., 2014; Viele et al., 2014). Proposition 1 shows that the influence of historical information on the posterior location is entirely governed by a precision ratio, independently of the equivalence margin δ ; Theorem 2 and Corollary 3 show that this influence is asymptotically self-correcting, in the specific sense that GBT-PP converges to the same decision as classical TOST regardless of w ; Theorem 4 identifies a subtle but consequential failure mode in which an ostensibly “objective” or “non-informative” reference prior for the Bayes-factor component of the dual rule can produce arbitrarily strong evidence for equivalence with no dependence on the data at all (Remark 3); and Corollary 5 translates the qualitative benefit of borrowing into an explicit, closed-form sample-size reduction, expressed in units directly comparable to a standard fixed-margin design.

We emphasize that GBT-PP does not introduce a fundamentally new *updating* mechanism: the underlying conjugate calculus (Normal-Normal, Beta-Binomial) and the power-prior discounting device are both well established (Ibrahim and Chen, 2000; Ibrahim et al., 2015). Its contribution is the specific combination of (i) a unified treatment of continuous and frequency-based historical information within a single interval-null decision rule, and (ii) the theoretical characterization, given here for the first time in closed form for this model class, of how that combination behaves asymptotically (Theorem 2), under prior mis-specification of diffuseness (Theorem 4), and at the design stage (Corollary 5).

9 Limitations and Extensions

The results of Section 5 inherit the same simplifying assumptions as the applied model of Section 4. First, w is fixed rather than estimated; extending Proposition 1 and Theorem 2 to the normalized power prior of Neuenschwander et al. (2009), in which w is itself assigned a hyperprior, is a natural next step, though Neuenschwander et al. (2009) show that the normalization constant is generally

not available in closed form, which would likely force a fully simulation-based treatment of the analogues of Theorems 2 and 4. Second, Theorem 2 assumes a single, homogeneous historical source; extending it to the multi-source, hierarchical commensurate-prior setting of Hobbs et al. (2011), or to the robust mixture MAP priors of Schmidli et al. (2014) (which are specifically designed to handle prior-data conflict, a scenario not explicitly modeled here), would substantially broaden its applicability, at the cost of losing the closed forms available in Sections 5.1–5.4. Third, Proposition 6 is a first-order asymptotic approximation; a second-order Edgeworth or saddlepoint refinement would likely be required for historical or current sample sizes below approximately 20–30 per arm, where the skewness correction itself becomes non-negligible relative to the normal term.

10 Conclusion

We have reviewed the state of the art in Bayesian equivalence testing and historical borrowing, and proposed GBT-PP, a power-discounted generalization of the Bayesian TOST framework that unifies continuous and frequency-based prior information within a single dual decision rule. Beyond the applied proposal, we have established five original theoretical results – monotonicity of the posterior mean in the discount weight (Proposition 1), asymptotic consistency of the decision rule and its large-sample agreement with classical TOST (Theorem 2, Corollary 3), a Jeffreys–Lindley-type paradox for the interval-null Bayes factor under a diffuse reference prior (Theorem 4), a closed-form Bayesian sample-size formula quantifying the benefit of historical borrowing (Corollary 5), and a deterministic Edgeworth approximation for the two-proportion equivalence probability (Proposition 6) – all proved in full in Appendix C. Together, these results place GBT-PP on a rigorous mathematical footing while preserving a practical, reproducible implementation (Appendix B).

Acknowledgments

The author is grateful to Jordi Ocaña Rebull (Section of Statistics. Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona), whose long experience in bioequivalence methodology and generous, critical reading of an earlier draft led to a clearer treatment of the asymmetry inherent to the TOST null hypothesis and of the hypothesis-labeling conventions used for the interval-null Bayes factor. His comments are gratefully acknowledged.

References

- Berger, R.L., Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4), 283–302.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781.
- U.S. Food and Drug Administration (2010). *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. Center for Devices and Radiological Health.
- U.S. Food and Drug Administration (2026). *Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products: Draft Guidance for Industry*. Docket No. FDA-2025-D-3217. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), January 2026.

- <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-bayesian-methodology-clinical-trials-drug-and-biological-products>
- Gamalo-Siebers, M., et al. (2017). Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharmaceutical Statistics*, 16(4), 232–249.
- Hobbs, B.P., Carlin, B.P., Mandrekar, S.J., Sargent, D.J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3), 1047–1056.
- Ibrahim, J.G., Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1), 46–60.
- Ibrahim, J.G., Chen, M.-H., Gwon, Y., Chen, F. (2015). The power prior: theory and applications. *Statistics in Medicine*, 34(28), 3724–3749.
- Kpekpena, C., Muthukumarana, S. (2018). Bayesian Equivalence Testing and Meta-Analysis in Two-Arm Trials with Binary Data. *Computational and Mathematical Methods in Medicine*, 2018, 8134132. <https://doi.org/10.1155/2018/8134132>. PMID: 30159005; PMCID: PMC6109545.
- Kruschke, J.K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, 142(2), 573–603.
- Kruschke, J.K., 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.
- Lakens, D., McLatchie, N., Isager, P.M., Scheel, A.M., Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, 75(1), 45–57.
- Makowski, D., Ben-Shachar, M.S., Lüdtke, D. (2019). `bayestestR`: describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40), 1541.
- Meyners, M. (2012). Equivalence tests – a review. *Food Quality and Preference*, 26(2), 231–245.
- Morey, R.D., Rouder, J.N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406–419.
- Neuenschwander, B., Branson, M., Spiegelhalter, D.J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28), 3562–3566.
- Niazi, S.K. (2026). The FDA’s adoption of Bayesian methodology: transforming clinical trial justification from biosimilars to broader drug development. *Frontiers in Medicine*, 13, 1790396. <https://doi.org/10.3389/fmed.2026.1790396>. PMID: 42058420; PMCID: PMC13122188.
- Rogers, J.L., Howard, K.I., Vessey, J.T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3), 553–565.
- Schmidli, H., Gsteiger, S., Roychoudhury, S., O’Hagan, A., Spiegelhalter, D., Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4), 1023–1032.

- Schuirman, 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.
- Spiegelhalter, D.J., Freedman, L.S., Parmar, M.K.B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A*, 157(3), 357–387.
- Viele, K., et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1), 41–54.
- Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd ed. Chapman & Hall/CRC.

A Derivations of the Posterior Distributions

A.1 Normal-Normal Case

Let $D_0 = (\bar{x}_0, s_0, n_0)$ and $D = (\bar{x}, s, n)$ with the normal-approximation likelihoods

$$\begin{aligned} L(\theta; D_0) &\propto \exp \left[-\frac{n_0}{2s_0} (\theta - \bar{x}_0)^2 \right], \\ L(\theta; D) &\propto \exp \left[-\frac{n}{2s} (\theta - \bar{x})^2 \right]. \end{aligned} \quad (23)$$

With a flat initial prior $p_0(\theta) \propto 1$, the power-discounted prior (4) is

$$\begin{aligned} p_w(\theta | D_0) &\propto \exp \left[-\frac{w n_0}{2s_0} (\theta - \bar{x}_0)^2 \right] \\ &= \mathcal{N} \left(\bar{x}_0, \frac{s_0}{w n_0} \right), \end{aligned} \quad (24)$$

giving Eq. (5). Multiplying by $L(\theta; D)$ and completing the square in θ gives a Gaussian posterior with natural (precision) parameters adding,

$$\begin{aligned} \tau_n &= \frac{w n_0}{s_0} + \frac{n}{s}, \\ \theta_n &= \frac{w n_0 \bar{x}_0 + n \bar{x}}{s_0 + s}, \end{aligned} \quad (25)$$

which is Eq. (9).

A.2 Beta-Binomial Case

With initial prior $p_0(\theta) = \text{Beta}(a_0, b_0)$ and historical likelihood $L(\theta; D_0) \propto \theta^{a_0-1} (1-\theta)^{b_0-1}$, the power-discounted prior is

$$\begin{aligned} p_w(\theta | D_0) &\propto \theta^{a_0-1} (1-\theta)^{b_0-1} \\ &= \text{Beta}(a_0, b_0), \end{aligned}$$

giving Eq. (6). Multiplying by the current-data Binomial likelihood $\theta < sup > s(1 - \theta) < /sup > f$ and using Beta-Binomial conjugacy gives Eq. (10). For a contrast $\theta = p_1 - p_2$ of two independent Beta posteriors, $P_{eq} = \Pr(-\delta < p_1 - p_2 < \delta)$ has no closed form; the Monte Carlo estimator is

$$\hat{P}_{eq}^{MC} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}\{-\delta < p_1 < sup > (m) - p_2 < /sup > (m) < \delta\},$$

$$p_i < sup > (m) \stackrel{\text{iid}}{\sim} \text{Beta}(a_i, b_i), \quad (26)$$

with standard error $\widehat{SE}(\hat{P}_{eq} < /sup > MC) = \sqrt{\hat{P}_{eq}^{MC}(1 - \hat{P}_{eq}^{MC})/M}$, converging at the usual Monte Carlo rate $O(M < sup > -1/2)$.

B R Implementation

Listing 1: R implementation of GBT-PP. The reference-prior standard deviation σ_0 used in `gbtpp_continuous` for the Bayes-factor component is a required, pre-registered argument (see Remark 3).

```

1  ## =====
2  ## GBT-PP: Generalized Bayesian TOST with Power-Discounted Priors
3  ## Script faithful to Monleon-Getino (2026), Sections 4-7 and
4  ## Appendices A-C, plus illustrative plots (Section 6 sensitivity
5  ## analysis and TOST-style posterior visualization).
6  ##
7  ## Two gaps in the Appendix B listing are completed here because the
8  ## article's text (Section 7.2) explicitly requires them to reproduce
9  ## Table 2, even though the printed code did not include them:
10 ##   (1) the simple Normal approximation to Peq (gamma_theta = 0) in the
11 ##       frequency case, alongside Monte Carlo and Edgeworth;
12 ##   (2) the dual decision rule (Eq. 11) and the sample-size formula
13 ##       (Corollary 5, Eqs. 17-18), which the text formalizes but does
14 ##       not render as executable code.
15 ## Everything else follows the Appendix B listing as written.
16 ## =====
17
18 ## -----
19 ## SECTION 4.2 / 4.4 / APPENDIX A.1
20 ## Continuous case: Normal-Normal prior with power discounting
21 ## Eqs. (4)-(9)
22 ## -----
23 gbtpp_continuous <- function(x0_bar, s0, n0, x_bar, s, n, w, delta,
24                             sigma0 = NULL, sigma0_max_mult = 10) {
25
26   if (is.null(sigma0)) {
27     sigma0 <- sigma0_max_mult * s0 / sqrt(n0)
28   }
29
30   n0_eff <- w * n0
31   prec_hist <- ifelse(n0_eff > 0, n0_eff / s0^2, 0)
32   prec_cur <- n / s^2
33

```

```

34 tau2      <- 1 / (prec_hist + prec_cur)
35 theta_n   <- tau2 * (prec_hist * x0_bar + prec_cur * x_bar)
36 tau_n     <- sqrt(tau2)
37
38 Peq <- pnorm((delta - theta_n) / tau_n) - pnorm((-delta - theta_n) / tau
   _n)
39
40 pi_eq <- pnorm(delta / sigma0) - pnorm(-delta / sigma0)
41
42 post_odds <- Peq / (1 - Peq)
43 prior_odds <- pi_eq / (1 - pi_eq)
44 BF01 <- post_odds / prior_odds
45
46 list(theta_n = theta_n, tau_n = tau_n, Peq = Peq,
47       pi_eq = pi_eq, BF01 = BF01, sigma0 = sigma0)
48 }
49
50 ## -----
51 ## SECTION 4.3 / APPENDIX A.2
52 ## Frequency case: Beta-Binomial prior with power discounting
53 ## Eqs. (6), (10), (20)-(21), (26)
54 ## -----
55 gbtpf_frequency <- function(s0, n0, s, n, w, a0 = 1, b0 = 1,
56                             delta, a2 = NULL, b2 = NULL,
57                             M = 2e5, seed = 123, return_draws = FALSE) {
58
59   f0 <- n0 - s0
60   f  <- n  - s
61
62   a_ref <- a0 + w * s0 + s
63   b_ref <- b0 + w * f0 + f
64
65   if (is.null(a2) || is.null(b2)) {
66     stop("Provide control-arm posterior parameters a2, b2 for the ",
67          "two-proportion contrast theta = p1 - p2 (Section 4.3).")
68   }
69
70   beta_moments <- function(a, b) {
71     mu    <- a / (a + b)
72     sig2  <- a * b / ((a + b)^2 * (a + b + 1))
73     gam1  <- 2 * (b - a) * sqrt(a + b + 1) / ((a + b + 2) * sqrt(a * b))
74     c(mu = mu, sig2 = sig2, gam1 = gam1)
75   }
76   m1 <- beta_moments(a_ref, b_ref)
77   m2 <- beta_moments(a2, b2)
78
79   mu_theta <- unname(m1["mu"] - m2["mu"])
80   sig_theta <- unname(sqrt(m1["sig2"] + m2["sig2"]))
81   gam_theta <- unname((m1["gam1"] * m1["sig2"]^1.5 -
82                       m2["gam1"] * m2["sig2"]^1.5) / sig_theta^3)
83
84   zplus <- (delta - mu_theta) / sig_theta
85   zminus <- (-delta - mu_theta) / sig_theta
86

```

```

87  set.seed(seed)
88  p1 <- rbeta(M, a_ref, b_ref)
89  p2 <- rbeta(M, a2, b2)
90  theta <- p1 - p2
91  Peq_MC <- mean(theta > -delta & theta < delta)
92  SE_MC <- sqrt(Peq_MC * (1 - Peq_MC) / M)
93
94  edge_cdf <- function(z) pnorm(z) - (gam_theta / 6) * dnorm(z) * (z^2 -
95    1)
96  Peq_Edge <- edge_cdf(zplus) - edge_cdf(zminus)
97
98  Peq_Normal <- pnorm(zplus) - pnorm(zminus)
99
100  out <- list(a_ref = a_ref, b_ref = b_ref,
101    mu_theta = mu_theta, sig_theta = sig_theta, gam_theta = gam_
102    theta,
103    Peq_MC = Peq_MC, SE_MC = SE_MC,
104    Peq_Edge = Peq_Edge, Peq_Normal = Peq_Normal)
105  if (return_draws) out$theta_draws <- theta
106  out
107 }
108
109 ## -----
110 ## SECTION 5 / Eq. (11): dual decision rule
111 ## -----
112 gbtp_decision <- function(Peq, BF01, gamma = 0.95, BF_star = 3) {
113   decision <- (Peq >= gamma) && (BF01 >= BF_star)
114   list(decision = decision, Peq_pass = Peq >= gamma,
115     BF_pass = BF01 >= BF_star)
116 }
117
118 ## -----
119 ## SECTION 5 / Corollary 5, Eqs. (17)-(18): approximate sample size
120 ## -----
121 gbtp_sample_size <- function(delta, sigma, w, n0, gamma = 0.95, s0 =
122   sigma) {
123   z_gamma <- qnorm((1 + gamma) / 2)
124   prec_hist <- (w * n0) / s0^2
125   prec_needed <- (z_gamma / delta)^2
126   prec_cur_needed <- prec_needed - prec_hist
127   if (prec_cur_needed <= 0) {
128     return(list(n_required = 0,
129       note = "Historical data (discounted by w) already suffice.
130       "))
131   }
132   n_required <- ceiling(prec_cur_needed * sigma^2)
133   list(n_required = n_required, prec_hist = prec_hist,
134     prec_needed = prec_needed)
135 }
136
137 ## -----
138 ## SECTION 6: sensitivity analysis over a grid of discount weights w
139 ## -----

```

```

136 gbtp_sensitivity_continuous <- function(w_grid = c(0, 0.25, 0.5, 0.75, 1)
137   , ...) {
138   res <- lapply(w_grid, function(w) {
139     out <- gbtp_continuous(w = w, ...)
140     data.frame(w = w, theta_n = out$theta_n, tau_n = out$tau_n,
141               Peq = out$Peq, pi_eq = out$pi_eq, BF01 = out$BF01)
142   })
143   do.call(rbind, res)
144 }
145 gbtp_sensitivity_frequency <- function(w_grid = c(0, 0.25, 0.5, 0.75, 1),
146   ...) {
147   res <- lapply(w_grid, function(w) {
148     out <- gbtp_frequency(w = w, ...)
149     data.frame(w = w, a_ref = out$a_ref, b_ref = out$b_ref,
150               Peq_MC = out$Peq_MC, SE_MC = out$SE_MC,
151               Peq_Edge = out$Peq_Edge, Peq_Normal = out$Peq_Normal)
152   })
153   do.call(rbind, res)
154 }
155 ## =====
156 ## SECTION 7.1 - Worked example: continuous case (bioequivalence)
157 ## Reproduces Table 1 of the article.
158 ## =====
159 sigma0_fixed <- 10 * 0.15 / sqrt(24)
160
161 table1 <- gbtp_sensitivity_continuous(
162   w_grid = c(0, 0.25, 0.5, 0.75, 1),
163   x0_bar = 0.02, s0 = 0.15, n0 = 24,
164   x_bar = 0.03, s = 0.12, n = 30,
165   delta = 0.05,
166   sigma0 = sigma0_fixed
167 )
168 cat("\n--- Table 1: continuous case (Section 7.1) ---\n")
169 print(table1, digits = 4)
170
171 ## =====
172 ## SECTION 7.2 - Worked example: frequency case (cure rates)
173 ## Reproduces Table 2 of the article.
174 ## =====
175 table2 <- gbtp_sensitivity_frequency(
176   w_grid = c(0, 0.25, 0.5, 0.75, 1),
177   s0 = 45, n0 = 50,
178   s = 42, n = 50,
179   a0 = 1, b0 = 1,
180   delta = 0.10,
181   a2 = 38, b2 = 14,
182   M = 2e5, seed = 123
183 )
184 cat("\n--- Table 2: frequency case (Section 7.2) ---\n")
185 print(table2, digits = 4)
186
187 ## =====

```

```

188 ## Dual decision rule example (w = 1, continuous case)
189 ## =====
190 res_w1 <- gbtpc_continuous(x0_bar = 0.02, s0 = 0.15, n0 = 24,
191                           x_bar = 0.03, s = 0.12, n = 30,
192                           w = 1, delta = 0.05, sigma0 = sigma0_fixed)
193 decision_w1 <- gbtpc_decision(res_w1$Peq, res_w1$BF01)
194 cat("\nDual decision at w = 1 (continuous case):\n")
195 print(decision_w1)
196
197 ## =====
198 ## Sample-size example (Corollary 5) at w = 0.5
199 ## =====
200 n_needed <- gbtpc_sample_size(delta = 0.05, sigma = 0.12, w = 0.5,
201                               n0 = 24, gamma = 0.95, s0 = 0.15)
202 cat("\nRequired current-study n at w = 0.5, gamma = 0.95:\n")
203 print(n_needed)
204
205
206 ## =====
207 ## =====
208 ## PLOTS
209 ## Four figures, each illustrating a key result of the article.
210 ## All use base R graphics only (no extra packages required).
211 ## =====
212 ## =====
213
214 ## Colorblind-friendly palette (Wong, 2011)
215 col_blue <- "#0072B2"
216 col_orange <- "#D55E00"
217 col_green <- "#009E73"
218 col_grey <- "grey50"
219
220 ## -----
221 ## FIGURE 1. Sensitivity curves for the continuous case (Section 6)
222 ##
223 ## WHAT IT SHOWS: how the posterior equivalence probability Peq and the
224 ## Bayes factor BF01 change as the discount weight w increases from 0
225 ## (fully discounted / ignore historical data) to 1 (fully pooled /
226 ## historical data treated as equally informative as current data).
227 ##
228 ## WHY IT MATTERS: this is the core robustness check proposed in
229 ## Section 6 of the article. A flat or smoothly increasing Peq curve
230 ## with w indicates that the equivalence conclusion is NOT an artifact
231 ## of over-borrowing from historical data (Remark 3). The horizontal
232 ## reference lines mark the decision thresholds gamma = 0.95 (for Peq)
233 ## and BF* = 3 (for BF01), as used in the dual decision rule, Eq. (11).
234 ## -----
235 plot_sensitivity_continuous <- function(table1, gamma = 0.95, BF_star = 3)
236 {
237   op <- par(mfrow = c(1, 2), mar = c(4.5, 4.5, 3, 1))
238
239   ## Panel A: Peq vs w
240   plot(table1$w, table1$Peq, type = "b", pch = 19, lwd = 2,
241        col = col_blue, ylim = c(0, 1),

```

```

241     xlab = "Discount weight (w)",
242     ylab = expression(P[eq]~"(posterior equivalence probability)"),
243     main = "A. Equivalence probability vs. w")
244 abline(h = gamma, col = col_grey, lty = 2, lwd = 1.5)
245 text(0.02, gamma + 0.03, paste0("gamma = ", gamma), col = col_grey,
246      adj = 0, cex = 0.85)
247 grid(col = "grey90")
248
249 ## Panel B: BF01 vs w (log scale, since BF can span orders of magnitude)
250 plot(table1$w, table1$BF01, type = "b", pch = 19, lwd = 2,
251      col = col_orange, log = "y",
252      xlab = "Discount weight (w)",
253      ylab = expression(BF[0][1]~"(Bayes factor for H0: equivalence)"),
254      main = "B. Bayes factor vs. w")
255 abline(h = BF_star, col = col_grey, lty = 2, lwd = 1.5)
256 text(0.02, BF_star * 1.15, paste0("BF* = ", BF_star), col = col_grey,
257      adj = 0, cex = 0.85)
258 grid(col = "grey90")
259
260 par(op)
261 mtext("Figure 1: Sensitivity analysis, continuous case (Section 6, Table
262      1)",
263      side = 3, line = -1.5, outer = TRUE, cex = 0.9, font = 2)
264 }
265
266 ## -----
267 ## FIGURE 2. TOST-style posterior visualization (continuous case)
268 ##
269 ## WHAT IT SHOWS: the posterior distribution of theta (the parameter
270 ## of interest, e.g. a mean difference) under a chosen w, together with
271 ## the equivalence region E = (-delta, delta). The shaded area under
272 ## the curve within E is exactly Peq.
273 ##
274 ## WHY IT MATTERS: this is the direct graphical counterpart of the
275 ## generalized Bayesian TOST (Eq. 1 and Section 3). It makes visible,
276 ## in a single picture, why Peq is high (most posterior mass inside E)
277 ## or low (posterior concentrated outside E, or straddling the
278 ## boundary), which a table of numbers alone does not convey.
279 ## -----
280 plot_posterior_continuous <- function(theta_n, tau_n, delta, Peq, w,
281                                     xlim_mult = 5) {
282   xs <- seq(theta_n - xlim_mult * tau_n, theta_n + xlim_mult * tau_n,
283            length.out = 1000)
284   ys <- dnorm(xs, theta_n, tau_n)
285
286   plot(xs, ys, type = "n",
287        xlab = expression(theta~"(parameter of interest)"),
288        ylab = "Posterior density",
289        main = paste0("Figure 2: Posterior distribution and equivalence ",
290                      "region (w = ", w, ")"))
291
292   ## Shade the equivalence region E = (-delta, delta)
293   x_in <- xs[xs > -delta & xs < delta]
294   y_in <- dnorm(x_in, theta_n, tau_n)

```

```

294 polygon(c(-delta, x_in, delta), c(0, y_in, 0),
295         col = adjustcolor(col_blue, alpha.f = 0.35), border = NA)
296
297 ## Full posterior curve
298 lines(xs, ys, col = col_blue, lwd = 2)
299
300 ## Equivalence margins and posterior mean
301 abline(v = c(-delta, delta), col = col_orange, lty = 2, lwd = 1.5)
302 abline(v = theta_n, col = col_grey, lty = 3, lwd = 1.5)
303
304 legend("topright", bty = "n", cex = 0.85,
305       legend = c(expression("Posterior of"~theta),
306                 expression("Equivalence margins"~-delta*, "*delta"),
307                 "Posterior mean",
308                 paste0("Peq = ", round(Peq, 3))),
309       col = c(col_blue, col_orange, col_grey, NA),
310       lty = c(1, 2, 3, NA), lwd = c(2, 1.5, 1.5, NA))
311 }
312
313 ## -----
314 ## FIGURE 3. Sensitivity curves for the frequency case, comparing the
315 ## three approximation methods to Peq (Section 7.2 / Proposition 6)
316 ##
317 ## WHAT IT SHOWS: Peq computed via Monte Carlo (gold standard), the
318 ## Edgeworth approximation, and the simple Normal approximation, all
319 ## plotted together across the discount-weight grid.
320 ##
321 ## WHY IT MATTERS: this directly validates Proposition 6 of the
322 ## article: the Edgeworth correction should track the Monte Carlo curve
323 ## much more closely than the simple Normal approximation, especially
324 ## when the Beta posteriors are skewed (small/large a_ref, b_ref). If
325 ## the three lines are visually indistinguishable, the Normal
326 ## approximation is adequate for this dataset; if they diverge, the
327 ## skewness correction (gamma_theta) matters.
328 ## -----
329 plot_sensitivity_frequency <- function(table2) {
330   plot(table2$w, table2$Peq_MC, type = "b", pch = 19, lwd = 2,
331       col = col_blue, ylim = range(c(table2$Peq_MC, table2$Peq_Edge,
332                                     table2$Peq_Normal)),
333       xlab = "Discount weight (w)",
334       ylab = expression(P[eq]),
335       main = "Figure 3: Peq by approximation method, frequency case")
336   lines(table2$w, table2$Peq_Edge, type = "b", pch = 17, lwd = 2,
337       col = col_orange, lty = 2)
338   lines(table2$w, table2$Peq_Normal, type = "b", pch = 15, lwd = 2,
339       col = col_green, lty = 3)
340   grid(col = "grey90")
341   legend("bottomleft", bty = "n", cex = 0.85,
342       legend = c("Monte Carlo (gold standard, Eq. 26)",
343                 "Edgeworth approximation (Prop. 6)",
344                 "Simple Normal approximation"),
345       col = c(col_blue, col_orange, col_green),
346       pch = c(19, 17, 15), lty = c(1, 2, 3), lwd = 2)
347 }

```

```

348
349 ## -----
350 ## FIGURE 4. TOST-style posterior visualization (frequency case)
351 ##
352 ## WHAT IT SHOWS: the Monte Carlo-simulated posterior distribution of
353 ##  $\theta = p_1 - p_2$  (risk difference) at a chosen  $w$ , with the
354 ## equivalence region  $E = (-\delta, \delta)$  shaded.
355 ##
356 ## WHY IT MATTERS: analogous to Figure 2 but for the Beta-Binomial
357 ## model; it is the empirical (simulation-based) counterpart used to
358 ## sanity-check the analytic Edgeworth/Normal approximations plotted
359 ## in Figure 3.
360 ## -----
361 plot_posterior_frequency <- function(theta_draws, delta, Peq_MC, w) {
362   dens <- density(theta_draws)
363
364   plot(dens, col = col_blue, lwd = 2,
365        xlab = expression(theta == p[1] - p[2]~"(risk difference)"),
366        ylab = "Posterior density (Monte Carlo)",
367        main = paste0("Figure 4: Simulated posterior of ",
368                      "theta = p1 - p2 (w = ", w, ")"))
369
370   x_in <- dens$x[dens$x > -delta & dens$x < delta]
371   y_in <- dens$y[dens$x > -delta & dens$x < delta]
372   polygon(c(-delta, x_in, delta), c(0, y_in, 0),
373          col = adjustcolor(col_blue, alpha.f = 0.35), border = NA)
374
375   abline(v = c(-delta, delta), col = col_orange, lty = 2, lwd = 1.5)
376   abline(v = mean(theta_draws), col = col_grey, lty = 3, lwd = 1.5)
377
378   legend("topright", bty = "n", cex = 0.85,
379         legend = c("Monte Carlo posterior density",
380                   expression("Equivalence margins"~-delta*, "*delta"),
381                   "Posterior mean",
382                   paste0("Peq (MC) = ", round(Peq_MC, 3))),
383         col = c(col_blue, col_orange, col_grey, NA),
384         lty = c(1, 2, 3, NA), lwd = c(2, 1.5, 1.5, NA))
385 }
386
387 ## =====
388 ## GENERATE THE FOUR FIGURES
389 ## =====
390
391 ## Figure 1: sensitivity curves, continuous case
392 plot_sensitivity_continuous(table1, gamma = 0.95, BF_star = 3)
393
394 ## Figure 2: posterior + equivalence region at  $w = 1$  (continuous case)
395 #dev.new() # opens a new plotting window; remove this line if running
396 # in RStudio / RMarkdown where each plot call is shown separately
397 plot_posterior_continuous(theta_n = res_w1$theta_n, tau_n = res_w1$tau_n,
398                          delta = 0.05, Peq = res_w1$Peq, w = 1)
399
400 ## Figure 3: sensitivity curves across approximation methods, frequency
    case

```

```

401 #dev.new()
402 plot_sensitivity_frequency(table2)
403
404 ## Figure 4: Monte Carlo posterior + equivalence region (frequency case, w
    = 1)
405 res_freq_w1 <- gbtp_frequency(s0 = 45, n0 = 50, s = 42, n = 50,
406                               w = 1, a0 = 1, b0 = 1, delta = 0.10,
407                               a2 = 38, b2 = 14, M = 2e5, seed = 123,
408                               return_draws = TRUE)
409 #dev.new()
410 plot_posterior_frequency(theta_draws = res_freq_w1$theta_draws,
411                           delta = 0.10, Peq_MC = res_freq_w1$Peq_MC, w = 1)

```

C Proofs of Section 5

C.1 Proof of Proposition 1

Write $\lambda_0 = n_0/s_0 < \sup > 2$ and $\lambda = n/s < \sup > 2$ (fixed, data-determined precisions). From Eq. (9),

$$\theta_n(w) = \frac{w\lambda_0\bar{x}_0 + \lambda\bar{x}}{w\lambda_0 + \lambda}.$$

Differentiating with respect to w using the quotient rule,

$$\frac{\partial\theta_n}{\partial w} = \frac{\lambda_0\bar{x}_0(w\lambda_0 + \lambda) - (w\lambda_0\bar{x}_0 + \lambda\bar{x})\lambda_0}{(w\lambda_0 + \lambda)^2} = \frac{\lambda_0\lambda(\bar{x}_0 - \bar{x})}{(w\lambda_0 + \lambda)^2},$$

which is Eq. (12) upon substituting back $\lambda_0 = n_0/s_0 < \sup > 2$, $\lambda = n/s < \sup > 2$. Since $\lambda_0, \lambda > 0$ and $(w\lambda_0 + \lambda)^2 > 0$ for all $w \in [0, 1]$, the sign of $\partial\theta_n/\partial w$ is exactly the sign of $\bar{x}_0 - \bar{x}$, proving monotonicity; continuity of $\theta_n(w)$ in w and the endpoint values $\theta_n(0) = \bar{x}$, $\theta_n(1) = (\lambda_0\bar{x}_0 + \lambda\bar{x})/(\lambda_0 + \lambda)$ follow by direct substitution. ■

C.2 Proof of Theorem 2 and Corollary 3

Since w, n_0, s_0 are fixed, $\lambda_0 = wn_0/s_0 < \sup > 2$ is a fixed positive constant. By hypothesis $s < \sup > 2 \xrightarrow{P} \sigma < \sup > 2 \in (0, \infty)$, so $n/s < \sup > 2 \xrightarrow{P} \infty$ as $n \rightarrow \infty$ (since n grows while $s < \sup > 2$ converges to a finite positive limit). Hence

$$\begin{aligned} \tau_n < \sup > -2(w) &= \lambda_0 + n/s < \sup > 2 \xrightarrow{P} \infty, \\ &\Rightarrow \tau_n < \sup > 2(w) \xrightarrow{P} 0, \end{aligned} \tag{27}$$

proving the first part of (13). For the posterior mean,

$$\begin{aligned} \theta_n(w) &= \frac{\lambda_0\bar{x}_0 + (n/s < \sup > 2)\bar{x}}{\lambda_0 + n/s < \sup > 2} \\ &= \bar{x} \cdot \frac{1}{1 + \lambda_0 s < \sup > 2/n} + \bar{x}_0 \cdot \frac{\lambda_0 s < \sup > 2/n}{1 + \lambda_0 s < \sup > 2/n}. \end{aligned} \tag{28}$$

Since $\lambda_0 s < \sup > 2/n \xrightarrow{P} 0$ (fixed $\lambda_0, s < \sup > 2 \xrightarrow{P} \sigma < \sup > 2$ finite, $n \rightarrow \infty$) and $\bar{x} \xrightarrow{P} \theta_0$, Slutsky's theorem gives $\theta_n(w) \xrightarrow{P} \theta_0$, proving the second part of (13).

For (14), write $P_{eq}(w) = \Phi\left(\frac{\delta - \theta_n(w)}{\tau_n(w)}\right) - \Phi\left(\frac{-\delta - \theta_n(w)}{\tau_n(w)}\right)$. If $|\theta_0| < \delta$, both $(\delta - \theta_n(w))/\tau_n(w) \xrightarrow{P} +\infty$ and $(-\delta - \theta_n(w))/\tau_n(w) \xrightarrow{P} -\infty$ (since $\tau_n(w) \xrightarrow{P} 0$ while the numerators converge in probability to the fixed nonzero constants $\delta - \theta_0 > 0$ and $-\delta - \theta_0 < 0$ respectively), so $P_{eq}(w) \xrightarrow{P} 1 - 0 = 1$. If $|\theta_0| > \delta$ (say $\theta_0 > \delta$), both numerators converge to negative constants, so both terms converge in probability to 0 and $P_{eq}(w) \xrightarrow{P} 0$; the case $\theta_0 < -\delta$ is symmetric. This holds for every fixed $w \in [0, 1]$ since $\lambda_0 = wn_0/s_0 < sup > 2$ is bounded and drops out of the limit, proving (14).

For Corollary 3: by Berger and Hsu (1996), the classical TOST/CI-inclusion rule declares equivalence iff $CI_{1-2\alpha}(\hat{\theta}) \subset (-\delta, \delta)$, and since $CI_{1-2\alpha}(\hat{\theta}) = \hat{\theta} \pm z_{1-\alpha} s/\sqrt{n}$ shrinks to the point θ_0 in probability as $n \rightarrow \infty$, this rule declares equivalence with probability $\rightarrow 1$ if $|\theta_0| < \delta$ and $\rightarrow 0$ if $|\theta_0| > \delta$ – the same limiting behavior established for $P_{eq}(w)$ above. Since the $BF_{01}(w) \geq BF < /sup > *$ component of (11) does not override a $P_{eq}(w) \rightarrow \{0, 1\}$ limit at fixed thresholds $\gamma < 1$, $BF < sup > * < \infty$, the full dual rule inherits the same limiting agreement with TOST. ■

C.3 Proof of Theorem 4

For $\theta \sim \mathcal{N}(0, \sigma_0 < /sup > 2)$,

$$\pi_{eq}(\sigma_0) = \Pr(-\delta < \theta < \delta) = \Phi(\delta/\sigma_0) - \Phi(-\delta/\sigma_0) = 2\Phi(\delta/\sigma_0) - 1.$$

A Taylor expansion of Φ about 0 gives $\Phi(u) = \frac{1}{2} + \phi(0)u + O(u < sup > 3)$ for small u , so with $u = \delta/\sigma_0 \rightarrow 0$ as $\sigma_0 \rightarrow \infty$,

$$\begin{aligned} \pi_{eq}(\sigma_0) &= 2 \left[\phi(0) \frac{\delta}{\sigma_0} + O(\sigma_0 < /sup > -3) \right] \\ &= \frac{2\delta}{\sigma_0} \phi(0) + O(\sigma_0 < sup > -3) \rightarrow 0, \end{aligned}$$

proving (15). For (16): since $P_{eq} \in (0, 1)$ is fixed (does not depend on σ_0 , as it is computed from the data-informed posterior, not from the reference prior), the posterior odds $P_{eq}/(1 - P_{eq})$ are a fixed positive finite constant. As $\sigma_0 \rightarrow \infty$, $\pi_{eq}(\sigma_0) \rightarrow 0$ by the above, so the prior odds $\pi_{eq}/(1 - \pi_{eq}) \rightarrow 0$. Hence

$$BF_{01}(\sigma_0) = \frac{P_{eq}/(1 - P_{eq})}{\pi_{eq}(\sigma_0)/(1 - \pi_{eq}(\sigma_0))} \rightarrow \frac{\text{finite positive constant}}{0 < /sup > +} = +\infty,$$

for any data D giving $P_{eq} \in (0, 1)$, proving (16). ■

C.4 Proof of Corollary 5

Under $\theta_0 = 0$, $\theta_n(w) \approx 0$ for the design calculation (the design-stage approximation standard in Bayesian sample-size arguments), so, from Eq. (1) applied to a $\mathcal{N}(0, \tau_n < sup > 2(w))$ posterior,

$$\begin{aligned} P_{eq}(w) = 2\Phi(\delta/\tau_n(w)) - 1 \geq \gamma &\iff \Phi(\delta/\tau_n(w)) \geq \frac{1+\gamma}{2} \\ &\iff \delta/\tau_n(w) \geq z_\gamma \iff \tau_n(w) \leq \delta/z_\gamma, \end{aligned} \quad (29)$$

using monotonicity of Φ . Squaring and substituting $\tau_n < /sup > 2(w) = 1/(wn_0/s_0 < sup > 2 + n/s < /sup > 2)$,

$$\tau_n(w) \leq \delta/z_\gamma \iff \frac{1}{wn_0/s_0 < sup > 2 + n/s < /sup > 2} \leq \frac{\delta < sup > 2}{z_\gamma < /sup > 2}$$

$$\begin{aligned}
&\iff wn_0/s_0 < \sup > 2 + n/s < /sup > 2 \geq \frac{z_\gamma < \sup > 2}{\delta < /sup > 2} \\
&\iff n \geq s < \sup > 2 \left(\frac{z_\gamma < /sup > 2}{\delta < \sup > 2} - \frac{wn_0}{s_0 < /sup > 2} \right) = n < \sup > \star(w), \quad (30)
\end{aligned}$$

which is Eq. (17). Setting $w = 0$ gives $n < /sup > \star(0) = s < \sup > 2z_\gamma < /sup > 2/\delta < \sup > 2$, the classical fixed-margin sample size. Finally,

$$\begin{aligned}
\Delta n < /sup > \star(w) &= n < \sup > \star(0) - n < /sup > \star(w) \\
&= s < \sup > 2 \left[\frac{z_\gamma < /sup > 2}{\delta < \sup > 2} - \left(\frac{z_\gamma < /sup > 2}{\delta < \sup > 2} - \frac{wn_0}{s_0 < /sup > 2} \right) \right] \\
&= \frac{wn_0 s < \sup > 2}{s_0 < /sup > 2},
\end{aligned}$$

which is Eq. (18). ■

C.5 Proof of Proposition 6

Let $Z_i = (p_i - \mu_i)/\sigma_i$, $i = 1, 2$, be the standardized Beta variables, with third cumulant $\kappa_{3,i} = \gamma_{1,i}\sigma_i < \sup > 3$ given by the standard Beta skewness formula (Eq. 19). The first-order Edgeworth expansion for the CDF of a standardized random variable Z with skewness γ_1 and negligible excess kurtosis contribution is the classical result

$$F_Z(z) = \Phi(z) - \frac{\gamma_1}{6}\phi(z)(z < /sup > 2 - 1) + O(n < \sup > -1). \quad (31)$$

Since p_1 and p_2 are independent, the cumulants of $\theta = p_1 - p_2$ add linearly for the mean and variance, $\mu_\theta = \mu_1 - \mu_2$, $\sigma_\theta < /sup > 2 = \sigma_1 < \sup > 2 + \sigma_2 < /sup > 2$, and for the third cumulant,

$$\kappa_{3,\theta} = \kappa_{3,1} + \kappa_{3,-p_2} = \kappa_{3,1} - \kappa_{3,2} = \gamma_{1,1}\sigma_1 < \sup > 3 - \gamma_{1,2}\sigma_2 < /sup > 3,$$

using that the third cumulant of $-p_2$ is the negative of the third cumulant of p_2 . Dividing by $\sigma_\theta < \sup > 3$ gives the standardized skewness of θ ,

$$\gamma_{1,\theta} = \frac{\kappa_{3,\theta}}{\sigma_\theta < /sup > 3} = \frac{\gamma_{1,1}\sigma_1 < \sup > 3 - \gamma_{1,2}\sigma_2 < /sup > 3}{\sigma_\theta < \sup > 3},$$

which is Eq. (20). Applying (31) to $Z = (\theta - \mu_\theta)/\sigma_\theta$ at the two evaluation points $z < /sup > + = (\delta - \mu_\theta)/\sigma_\theta$ and $z < \sup > - = (-\delta - \mu_\theta)/\sigma_\theta$, and using

$$P_{eq} = \Pr(-\delta < \theta < \delta) = F_Z(z < /sup > +) - F_Z(z < \sup > -),$$

gives directly Eq. (21). When $\gamma_{1,\theta} = 0$ the correction term vanishes and $\hat{P}_{eq} < /sup > Edge$ reduces to $\Phi(z < \sup > +) - \Phi(z < /sup > -)$, the plain normal-moment-matching approximation; a sufficient (though not necessary) condition for $\gamma_{1,\theta} = 0$ is $\gamma_{1,1}\sigma_1 < \sup > 3 = \gamma_{1,2}\sigma_2 < /sup > 3$, which holds in particular when $a_1 = b_1$ and $a_2 = b_2$ (both marginal Beta distributions symmetric), since then $\gamma_{1,1} = \gamma_{1,2} = 0$. ■

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [RscriptBayesiantost.txt](#)