Statistics and Its Interface Volume 14 (2021) 489–502

Demystify Lindley's paradox by connecting P-value and posterior probability

Guosheng Yin* and Haolun Shi

In the hypothesis testing framework, p-value is often computed to determine whether to reject the null hypothesis or not. On the other hand, Bayesian approaches typically compute the posterior probability of the null hypothesis to evaluate its plausibility. We revisit Lindley's paradox and demystify the conflicting results between Bayesian and frequentist hypothesis testing procedures by casting a two-sided hypothesis as a combination of two one-sided hypotheses along the opposite directions. This formulation can naturally circumvent the ambiguities of assigning a point mass to the null and choices of using local or non-local prior distributions. As p-value solely depends on the observed data without incorporating any prior information, we consider non-informative prior distributions for fair comparisons with p-value. The equivalence of p-value and the Bayesian posterior probability of the null hypothesis can be established to reconcile Lindley's paradox. More complicated settings, such as multivariate cases, random effects models and non-normal data, are also explored for generalization of our results to various hypothesis tests.

AMS 2000 subject classifications: Primary 62A01, 62F15; secondary 62F03.
Keywords and phrases: Bayesian posterior probability, Hypothesis testing, Interpretation of p-value, Point null hypothesis, Two-sided test.

1. INTRODUCTION

Lindley's paradox [1] refers to a case in the hypothesis testing framework where the Bayesian and frequentist approaches produce opposite conclusions for certain choices of the likelihood function or prior distribution. The paradox is of paramount importance as it highlights the major differences between the frequentist and the Bayesian approaches to hypothesis tests.

Extensive research has been conducted to reconcile the differences between Bayesian and frequentist analysis [2]. Frequentist hypothesis testing commonly relies on the computation of p-value [3], which is defined as the probability of obtaining the results at least as extreme as the observed one given the null hypothesis being true. Frequentist methods do not utilize any prior information but the observed data, and thus for a fair comparison, non-informative prior distributions should be used in the Bayesian analysis. In particular, Berger and Sellke [4], Berger and Delampady [5], and Casella and Berger [6] investigate the relationships between the p-value and Bayesian measure of evidence against the null hypothesis for hypothesis testing. Robert [7] discuss the Jeffreys–Lindley paradox by considering the role of the prior hypothesis probability. Sellke, Bayarri, and Berger [8] propose to calibrate p-values for testing precise null hypotheses. More recently, extensive discussions on modern statistical inference in a special issue of *The American Statistician* highlight several insights regarding the role of p-value and Bayesian statistics [9, 10, 11, 12, 13, 14, 15]. One important yet rarely visited issue in the reconciliation between frequentist and Bayesian approaches is the ambiguity on prior specification with the point null and composite alternative hypotheses in the Bayesian paradigm [6, 16]. Greenland and Poole [17] discuss the p-value as the probability measure of the distance, and several discussions on the p-value from the Bayesian perspectives are also provided [18, 19, 20, 21]. In particular, Shi and Yin [22] make a new interpretation of p-value as the posterior probability of the null hypothesis under both one- and two-sided hypothesis tests by slightly twisting the definition of the posterior probability under non-informative priors, which contradicts one of the statements of [23].

We revisit Lindley's paradox, exploring the connection between the frequentist p-value and Bayesian posterior probability. We emphasize that the paradox may result from certain choices of the prior distribution (e.g., the witch hat prior—a point mass at the null and flat elsewhere), or certain sampling distributions. We provide various formulations and show that the p-value and the posterior probability of the null have an asymptotic equivalence relationship under non-informative priors, leading to a reconciliation of the paradox. Moreover, we extend the results to non-normal data and multivariate tests, as well as hypothesis testing of variance components under random effects models.

The rest of the paper is organized as follows. In Section 2, we present a motivating example to demonstrate how a point null hypothesis in a two-sided test can be reformulated as a combination of two one-sided tests, which naturally reconciles Lindley's paradox. In Section 3, we analyze Lindley's paradox original example in depth and show that the

arXiv: 2002.10883
*Corresponding author.

paradox can be resolved under a negative binomial distribution rather than a binomial distribution. Section 4 considers hypothesis testing under more complicated settings and demonstrates the generalization of our result. In Section 5, we present real data examples to illustrate reconciliation of Bayesian and frequentist inferences, and Section 6 concludes with some discussion.

2. MOTIVATING EXAMPLE

In the two-sided hypothesis testing framework, it may happen that the Bayesian and frequentist approaches produce opposite conclusions. Such a conflict is primarily caused by certain choices of the prior distribution (e.g., the witch hat prior—a point mass at the null and flat elsewhere). By reformulating the prior distribution to be non-informative without using a point mass at the null hypothesis, e.g., a uniform prior, we can circumvent the paradox and establish the equivalence between Bayesian and frequentist inferences.

2.1 Illustration of Lindley's paradox

To illustrate the paradox, we start with a simple example. Suppose that 28,298 boys and 27,801 girls were born in a city last year. The observed proportion of male births in the city is $r = 28298/56099 \approx 0.5044297$. Let θ denote the true proportion of male births, and we are interested in testing

$$H_0 : \theta = 0.5 \quad \text{versus} \quad H_1 : \theta \neq 0.5. \tag{1}$$

2.1.1 P-value from an exact test

The number of male births follows a binomial distribution with mean $n\theta$ and variance $n\theta(1-\theta)$, where $n = 56,099$ is the total number of births. Let R denote the male sample proportion. Under the frequentist paradigm, the p-value based on the binomial exact test is

$$\Pr(R \geq r|H_0) = \sum_{x=28298}^{n} \binom{n}{x} 0.5^n \approx 0.01812363.$$

2.1.2 P-value using normal approximation

Because the sample size n is large and the observed male proportion r is not close to 0 or 1, we can use normal approximation to simplify the computation by assuming $R \sim N(\theta, \hat{\sigma}^2)$ with $\hat{\sigma}^2 = r(1-r)/n$. The frequentist approach calculates the p-value as the upper tail probability of as or more extreme than the observed data under the null distribution,

$$\begin{aligned} & \Pr(R \geq r|H_0) \\ = & \int_{28298/56099}^{\infty} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left\{ -\frac{(x-0.5)^2}{2\hat{\sigma}^2} \right\} dx \\ \approx & \ 0.01793329. \end{aligned} \tag{2}$$

Evidently, the exact and approximate p-values are very close. As the hypothesis test is two-sided, the final p-value is $2 \times 0.01793329 \approx 0.03586658$, and thus H_0 should be rejected at the typical significance level of 5%.

2.1.3 Posterior probability of H_0

If we proceed with a Bayesian approach, the usual practice is to first assign an equal prior probability to H_0 and H_1 without any preference, i.e., $P(H_0) = P(H_1) = 0.5$. Under H_0, θ has a point mass at 0.5. Under H_1, θ is not equal to 0.5 and we assign a uniform prior distribution to θ on $[0, 1]$. As a result, the posterior probability of H_0 is

$$\begin{aligned} & P(H_0|r) \\ = & \frac{P(r|H_0)P(H_0)}{P(r|H_0)P(H_0) + P(r|H_1)P(H_1)} \\ = & \frac{\exp\left\{ -\frac{(r-0.5)^2}{2\hat{\sigma}^2} \right\}}{\exp\left\{ -\frac{(r-0.5)^2}{2\hat{\sigma}^2} \right\} + \int_0^1 \exp\left\{ -\frac{(r-\theta)^2}{2\hat{\sigma}^2} \right\} d\theta} \\ \approx & \ 0.9543474, \end{aligned}$$

which strongly supports H_0.

Such conflict between Bayesian and frequentist hypothesis testing approaches may happen when the prior distribution is a mixture of a sharp peak at H_0 and no sharp features anywhere else, which is often known as Lindley's paradox. We explain as follows that such a conflicting result can be resolved if we view the two-sided hypothesis as a combination of two one-sided hypotheses, and further demonstrate the equivalence of p-value and the posterior probability of the null when a non-informative prior is used.

2.2 One-sided hypothesis test

For ease of exposition, we start with a one-sided hypothesis test,

$$H_0 : \theta \leq 0.5 \quad \text{versus} \quad H_1 : \theta > 0.5.$$

The p-value is still calculated in the same way as the upper tail probability of as or more extreme than the observed data under the null distribution. Under the normal approximation, following (2), we have p-value $= 0.01793329$.

2.2.1 Using Bayes' theorem

In the Bayesian approach, we assign a uniform prior distribution to θ, i.e., $\theta \sim \text{Unif}[0, 1]$, so the prior probabilities $P(H_0) = P(H_1) = 1/2$. Under normal approximation, the posterior probability of H_0 is

$$\begin{aligned} & P(H_0|r) \\ = & \frac{P(r|H_0)P(H_0)}{P(r|H_0)P(H_0) + P(r|H_1)P(H_1)} \end{aligned}$$

$$= \frac{\int_0^{0.5} \exp\left\{-\frac{(r-\theta)^2}{2\hat{\sigma}^2}\right\}d\theta}{\int_0^{0.5} \exp\left\{-\frac{(r-\theta)^2}{2\hat{\sigma}^2}\right\}d\theta + \int_{0.5}^{1} \exp\left\{-\frac{(r-\theta)^2}{2\hat{\sigma}^2}\right\}d\theta}$$
$$\approx 0.01793329,$$

which is the same as the p-value in (2).

2.2.2 Using the posterior distribution

Under the normal approximation, an alternative way is to first obtain the posterior distribution of θ, by assuming the prior distribution of θ to be flat, i.e., $p(\theta) \propto 1$. The posterior distribution of θ is then given by

$$P(\theta|r) \propto \exp\left\{-\frac{(\theta-\hat{\theta})^2}{2\hat{\sigma}^2}\right\},$$

i.e., $\theta|r \sim N(\hat{\theta}, \hat{\sigma}^2)$ where $\hat{\theta} = r$. As a result, we can compute

$$\begin{aligned} P(H_0|r) &= P(\theta \le 0.5|r) \\ &= \int_{-\infty}^{0.5} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left\{-\frac{(\theta - 28298/56099)^2}{2\hat{\sigma}^2}\right\}d\theta, \end{aligned}$$

which is exactly the same as the p-value in (2), because it is easy to show that

$$\begin{aligned} &\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-b)^2}{2\sigma^2}\right\}dx \\ = &\int_{b}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\}dx, \end{aligned}$$

for any values of a and b on the real line.

2.2.3 Bayesian exact Beta distribution

If we do not assume the asymptotic normal distribution, we can proceed with Bayesian exact computation. Under the Bayesian paradigm, if we adopt a uniform prior for θ, i.e., $\theta \sim \text{Beta}(1,1)$, the posterior distribution of θ is still Beta, i.e., $\theta|r \sim \text{Beta}(nr+1, n-nr+1)$. The posterior probability of the null can be directly calculated as

$$\begin{aligned} &\Pr(H_0|r) \\ = &\int_0^{0.5} \frac{\Gamma(n+2)}{\Gamma(nr+1)\Gamma(n-nr+1)} \theta^{nr}(1-\theta)^{n-nr}d\theta \\ \approx &\ 0.01793728, \end{aligned}$$

which is close to the p-value in (2). Note that this procedure does not rely upon the normal approximation. We further experiment other non-informative Beta prior distribution by choosing $\theta \sim \text{Beta}(\alpha, \beta)$ with $\alpha = \beta = 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001$, and the result is given in Table 1. Clearly, under non-informative prior distributions, the posterior probabilities of the null are very close to the p-value.

Table 1. Relationship between the posterior probability of the null hypothesis, $P(H_0|r)$, and the values of the hyperparameters in the Beta(α, β) prior distribution with $\alpha = \beta$ under the Bayesian exact Beta posterior distribution for the newborn motivating example

$\alpha = \beta$	$P(H_0\|r)$
1	0.01793728
0.1	0.01793580
0.01	0.01793565
0.001	0.01793564
0.0001	0.01793563
0.00001	0.01793563
0.000001	0.01793563

2.3 Two-sided hypothesis test

In a two-sided hypothesis test, the prior specification on the point null is often ambiguous by assigning a point probability mass. To circumvent the issue of the point mass at the null, we rewrite the two-sided hypothesis in (1) as a combination of two one-sided hypotheses:

$$(3) \qquad \begin{cases} H_0: & \theta \le 0.5 & \text{versus} & H_1: \theta > 0.5, \\ H_0: & \theta \ge 0.5 & \text{versus} & H_1: \theta < 0.5. \end{cases}$$

Under the frequentist paradigm, the p-value for the first one-sided hypothesis test in (3), $H_0: \theta \le 0.5$ versus $H_1: \theta > 0.5$, is given by

$$\Pr(R \ge r|H_0) = 1 - \Phi(28298/56099; 0.5, \hat{\sigma}^2) \approx 0.01793329,$$

where $\Phi(\cdot; \mu, \hat{\sigma}^2)$ denotes the cumulative distribution function (CDF) of a normal random variable with mean μ and variance $\hat{\sigma}^2$. The p-value for the second one-sided hypothesis test in (3), $H_0: \theta \ge 0.5$ versus $H_1: \theta < 0.5$, is given by

$$\Pr(R \le r|H_0) = \Phi(28298/56099; 0.5, \hat{\sigma}^2) \approx 0.9820667.$$

Therefore, the p-value under the two-sided hypothesis test in (3) is given by

$$\begin{aligned} p\text{-value}_2 &= 2 \times \min\{\Pr(R \le r|H_0), \Pr(R \ge r|H_0)\} \\ &= 2 \times 0.01793329 \\ &= 0.03586658. \end{aligned}$$

As a counterpart, we adopt the concept of the two-sided posterior probability (PoP_2) in [22], defined as

$$\begin{aligned} \text{PoP}_2 &= 2 \times \min\{\Pr(\theta \le 0.5|r), \Pr(\theta \ge 0.5|r)\} \\ &= 2 \times \min\{0.01793329, 0.9820667\} \\ &= 0.03586658. \end{aligned}$$

Therefore, it is evident that the value of PoP_2 is the same as the frequentist two-sided p-value under normal approximation.

Furthermore, a connection between the Bayes factor and p-value can be established. If an equal prior probability is assumed for H_0 and H_1, then the Bayes factor (BF) in favor of H_0 over H_1, denoted as $\text{BF}_{0,1}$, can be calculated as the odds of the p-value,

$$\text{BF}_{0,1} = \frac{p\text{-value}}{1 - p\text{-value}}.$$

Using a uniform prior on the whole interval $[0, 1]$, we can reconcile the frequentist and Bayesian inferences. The uniform prior is non-informative and, as a result, the posterior distribution is dominated by the data. On the contrary, the witch hat prior distribution under which a point mass is assigned at 0.5 with $P(H_0) = P(H_1) = 0.5$, is in fact a highly informative prior, as the null value 0.5 is far more likely than all other possible values in the parameter space. Such an informative prior leads to enormous differences between frequentist and Bayesian inference results.

3. DEMYSTIFY LINDLEY'S PARADOX

It is well-known that Bayesian methods adhere to the likelihood principle; that is, all that we know about the data or the sample is contained in the likelihood function. If the likelihood functions with respect to the parameter of interest θ under two different sampling plans or sampling distributions are proportional, Bayesian inferences on θ should be identical based on these two sampling distributions. However, frequentist approaches may result in conflicting conclusions in the hypothesis testing framework when multiple sampling distributions are feasible.

3.1 Original coin-tossing example

In the original example provided by Lindley [1], an experiment was conducted with a coin flipped for 12 times, and 9 heads and 3 tails were observed. Let θ denote the probability of observing a head, and we test the hypotheses,

$$H_0\text{: } \theta = 0.5 \quad \text{versus} \quad H_1\text{: } \theta > 0.5.$$

We consider two proposals for the likelihood function given the observed data. Let n denote the number of tosses and let Y denote the number of heads. The random variable Y may follow a binomial distribution, $Y \sim \text{Bin}(n, \theta)$, and the likelihood function is

$$L_{\text{B}}(\theta|y) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{12}{9}\theta^9(1-\theta)^3.$$

Another proposal of the likelihood function is based on the negative binomial distribution. If we let Y be the number of heads until we observe $q = 3$ tails. The random variable Y would follow a negative binomial distribution, $Y \sim \text{Neg-Bin}(q, \theta)$, and the likelihood function is

$$L_{\text{NB}}(\theta|y) = \binom{y+q-1}{y}\theta^y(1-\theta)^q = \binom{11}{9}\theta^9(1-\theta)^3.$$

It is evident that $L_{\text{B}}(\theta|y) \propto L_{\text{NB}}(\theta|y)$, and hence the posterior distributions of θ under these two likelihood functions are the same under the Bayesian paradigm. In contrast, statistical inferences under the frequentist paradigm are quite different, because the computation of p-value is contingent on the assumed sampling distribution. The p-value under the binomial sampling distribution is

$$p\text{-value}_{\text{B}} = \Pr(y \geq 9|H_0) = \sum_{y=9}^{12}\binom{12}{y}0.5^{12} \approx 0.07299805,$$

while that under a negative binomial sampling distribution is

$$\begin{aligned} p\text{-value}_{\text{NB}} &= \Pr(y \geq 9|H_0) \\ &= \sum_{y=9}^{\infty}\binom{y+2}{y}0.5^{3+y} \approx 0.03271484. \end{aligned}$$

If the significance level for frequentist hypothesis testing is set as $\alpha = 0.05$, the two hypothesis tests would lead to contradictory results.

The difference between p-values under the binomial distribution and the negative binomial distribution can be partly attributed to the difference in the sampling space. The support under the binomial distribution is constrained by the number of tosses, i.e., ranging from 0 to 12. However, the support under the negative binomial distribution ranges from 0 to infinity. Essentially, the two p-values correspond to two different perspectives on the observed data. The frequentist paradigm assumes that the parameters are fixed but the data are random, but does not specify how the data are associated with the parameters. When multiple sampling distributions can explain the data, conflicting frequentist inferences may arise.

3.2 One-sided hypothesis test

Suppose that we conduct a one-sided hypothesis test,

$$H_0 : \theta \leq 0.5 \quad \text{versus} \quad H_1 : \theta > 0.5.$$

Under the Bayesian paradigm, if we assume a symmetric Beta prior distribution for θ, i.e., $\theta \sim \text{Beta}(\alpha, \beta)$ with $\alpha = \beta$, then the posterior distribution of θ is $\text{Beta}(y+\alpha, n-y+\beta)$. The posterior probability of the null can be computed as

$$\begin{aligned} &\Pr(H_0|y) \\ = &\int_0^{0.5}\frac{\Gamma(n+\alpha+\beta)}{\Gamma(n-y+\beta)\Gamma(y+\alpha)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta. \end{aligned}$$

The top panel of Figure 1 shows the posterior probability of H_0 under different hyperparameter values $\alpha = \beta$ from 10^{-6} to 2 in the $\text{Beta}(\alpha, \beta)$ prior distribution. Under such symmetric Beta prior distributions, the implicit probability of landing on a head for a coin toss is 0.5, which is smaller than the one observed in the actual data, $9/12 = 0.75$. When

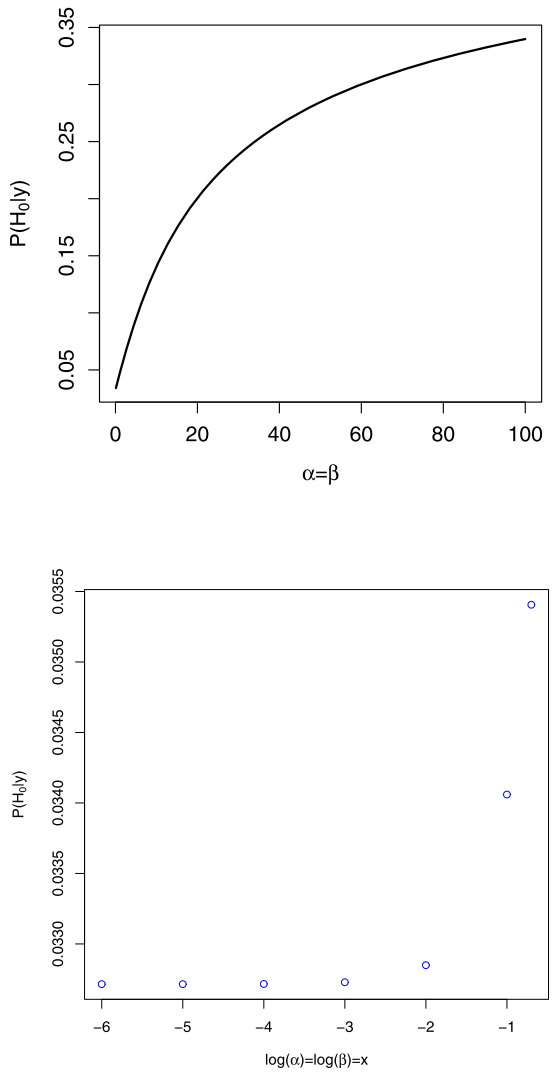

Figure 1. The posterior probability of H_0 under different Beta(α, β) prior distributions with $\alpha = \beta$ in the top panel; the zoom-in plot at the corner $(0, 0)$ by taking the log transformation of the Beta prior hyperparameters in the bottom panel.

the value of $\alpha = \beta$ increases, the prior distribution becomes more concentrated at the null value 0.5. As the information in the prior distribution strengthens, the prior plays an increasingly important role in the posterior distribution, so that the posterior probability of H_0 increases under the influence of the strengthening prior information. The bottom panel in Figure 1 shows the zoom-in plot at the corner $(0, 0)$ of the top panel by taking the log transformation of the x-axis. Table 2 shows the values of the posterior probability $P(H_0|y)$ for different values of the hyperparameters in the Beta(α, β) prior distribution with $\alpha = \beta$. The conclusion is that as the values of the hyperparameters decrease toward zero, i.e., the prior becomes more and more non-informative,

Table 2. Relationship between the posterior probability $P(H_0|r)$ and the values of the hyperparameters in the Beta(α, β) prior distribution with $\alpha = \beta$ for the original coin-tossing example in Lindley's paradox

$\alpha = \beta$	$P(H_0\|r)$
2	0.059235
1.5	0.052752
1	0.046143
0.9	0.044809
0.8	0.043471
0.7	0.042131
0.6	0.040789
0.5	0.039445
0.4	0.038099
0.3	0.036753
0.2	0.035406
0.1	0.034060
0.01	0.032849
0.001	0.032728
0.0001	0.032716
0.00001	0.032715
0.000001	0.032715

$P(H_0|y)$ approaches the p-value obtained from the negative binomial distribution but not the one from the binomial distribution.

3.3 Equivalence between negative binomial *P*-value and posterior probability

The CDF of a negative binomial distribution, $Y \sim$ Neg-Bin(q, θ), is given by

$$F_{\text{NB}}(y; q, \theta) = 1 - I_\theta(y + 1, q),$$

where $I_x(a, b)$ is the regularized incomplete Beta function defined as

$$I_x(a, b) = \frac{B(x; a, b)}{B(a, b)},$$

with

$$B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt,$$
$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt.$$

Therefore, the p-value based on the assumption $Y \sim$ Neg-Bin$(q = n - y, \theta)$ is

$$\begin{aligned} p\text{-value}_{\text{NB}} &= \Pr(Y \geq y|H_0) \\ &= 1 - F_{\text{NB}}(y - 1; q, \theta = 0.5) \\ &= I_{0.5}(y, r) = I_{0.5}(y, n - y). \end{aligned} \tag{4}$$

Under the Bayesian paradigm, if we assume a Beta(α, β) prior distribution for θ, the posterior distribution of θ is

Beta$(y+\alpha, n-y+\beta)$. The CDF of a Beta(a,b) distribution is $F_{\text{Beta}}(x;a,b) = I_x(a,b)$. Hence, the posterior probability of the null is

$$\begin{aligned} P(H_0|y) &= F_{\text{Beta}}(0.5; y+\alpha, n-y+\beta) \\ &= I_{0.5}(y+\alpha, n-y+\beta). \end{aligned} \tag{5}$$

Comparing (4) and (5), when the hyperparameters α and β are very small relative to n and y, the p-value under the negative binomial distribution is close to the posterior probability of the null. The equivalence between the negative binomial p-value and the posterior probability of the null is due to the algebraic connection between the CDF of the Beta distribution and the CDF of the negative binomial distribution, i.e., both CDFs are based upon the regularized incomplete Beta function.

3.4 Numerical study

We further conduct numerical studies to explore the relationship between the posterior probability of the null hypothesis and p-value. By mimicking the newborn male proportion example, in the first numerical experiment we set $y = 0.5044297 \times n$ while increasing n gradually. In other words, the ratio between y and n is fixed at the observed value 0.5044297, while both the values of y and n are increased to enlarge the sample size. As shown in Figure 2, the range of sample size is chosen such that p-values can cover from 0 up to around 0.5. Clearly, the p-values under the negative binomial distribution match well with the posterior probabilities of H_0, while those under the binomial distribution show some deviation, particularly for p-values near 0.5 when sample sizes are relatively small. Furthermore, we consider the case where the data are randomly generated from a binomial distribution with probability 0.5044297, and the results are shown in Figure 3. We observe that for the negative binomial distribution, the ratio between the p-value and posterior probability is maintained at 1, whereas for the binomial distribution, such a ratio is not maintained exactly but would converge to 1 as the sample size increases.

In the second numerical experiment, we follow the coin-tossing example by fixing $y/n = 9/12$, while gradually increasing n up to 120. A non-informative Beta prior, Beta$(10^{-6}, 10^{-6})$, is used. Figure 4 again shows that the p-values under the negative binomial distribution match well with the posterior probabilities of H_0, while those under the binomial distribution do not. In addition, we evaluate the results when the data are randomly generated from a binomial distribution with probability $9/12 = 0.75$, which are shown in Figure 5. We observe that the agreement between the p-value and posterior probability of H_0 is better under the negative binomial distribution than that under the binomial distribution.

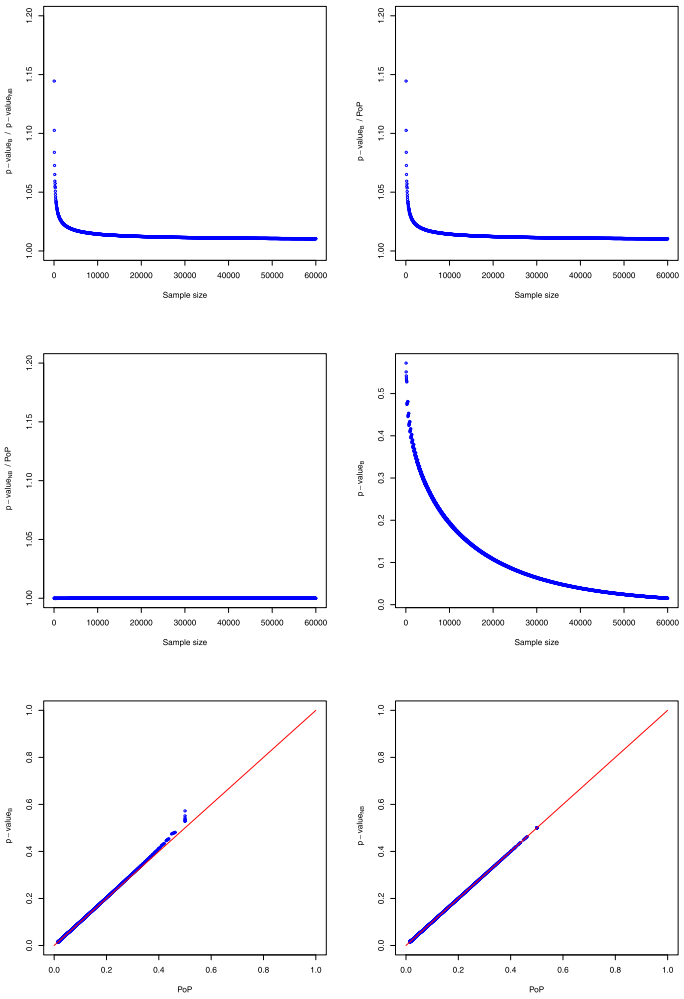

Figure 2. The ratio between p-values (p-value$_{\text{B}}$ is based on the binomial distribution, and p-value$_{\text{NB}}$ is based on the negative binomial distribution) and the posterior probability (PoP) of the null hypothesis, as sample size increases while fixing $y/n = 0.5044297$.

4. EXTENSIONS

Reconciliation between the frequentist and Bayesian inferences can be achieved not only in the case where the outcome is binary, but can also be extended to other cases where the outcomes follow a univariate or multivariate normal distribution, a non-normal distribution, or are generated from random effects models.

4.1 Hypothesis tests with normal data

We consider hypothesis tests with normal data and discuss how an equivalence relationship between the Bayesian posterior probability and frequentist p-value can be established.

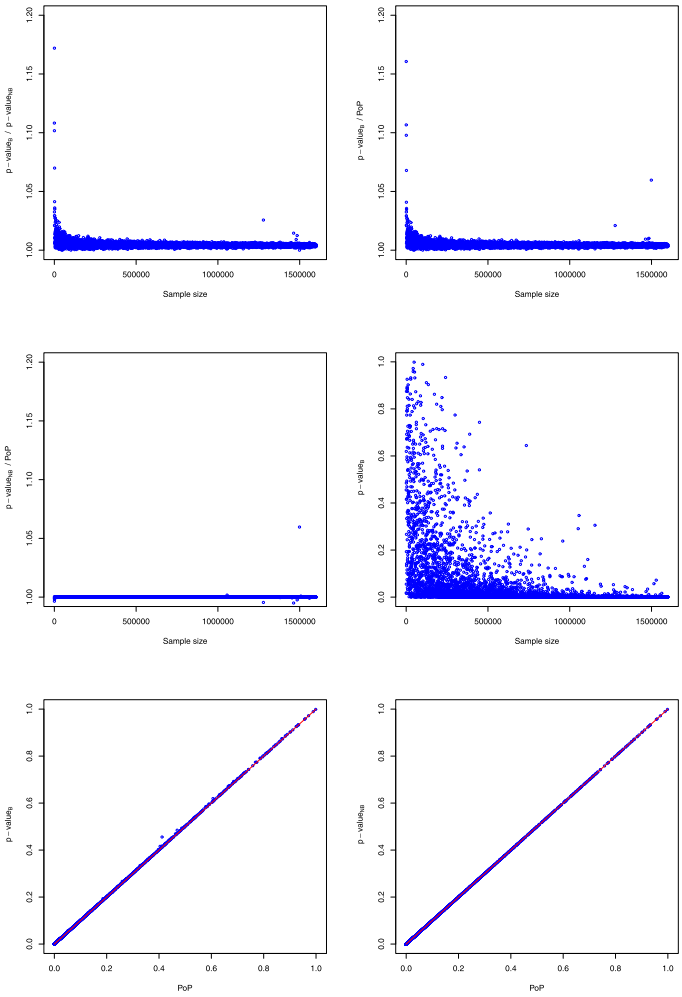

Figure 3. The ratio between p-values (p-value$_{\mathrm{B}}$ is based on the binomial distribution, and p-value$_{\mathrm{NB}}$ is based on the negative binomial distribution) and the posterior probability (PoP) of the null hypothesis, as sample size increases with outcomes generated randomly from a binomial distribution with probability equal to $y/n = 0.5044297$.

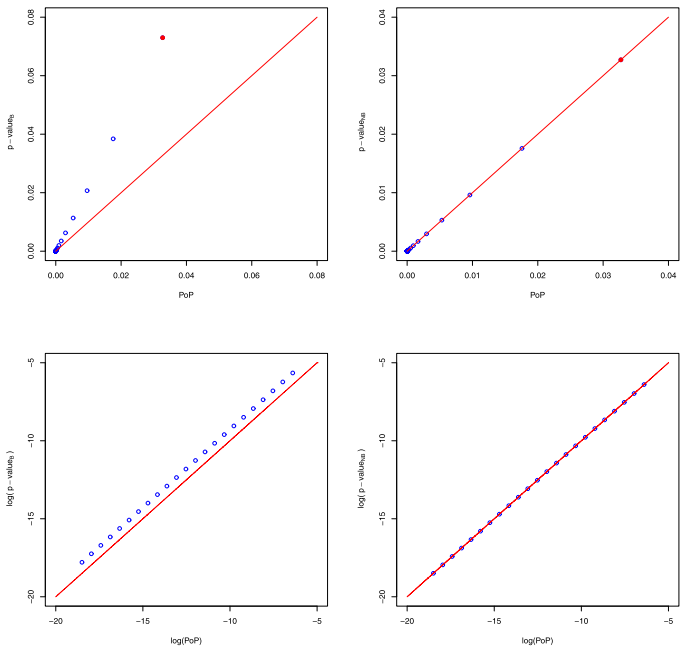

Figure 4. The relationship between p-values (p-value$_{\mathrm{B}}$ is based on the binomial distribution, and p-value$_{\mathrm{NB}}$ is based on the negative binomial distribution) and the posterior probability (PoP) of the null when y/n is fixed at 0.75. The red solid point in the first row corresponds to the original experiment with $n = 12$ and $y = 9$. The second row presents the zoom-in plot at the corner $(0, 0)$ of the first row, i.e., the logarithm of the p-value and PoP for p-values smaller than 0.002.

4.1.1 Improper flat prior

Consider a two-sample test with normal data. Let n denote the sample size for each group, and let D denote the observed data. We assume that the two groups are independent, and within each group the outcomes are independent and identically distributed as $y_{1i} \sim \mathrm{N}(\mu_1, \sigma^2)$ and $y_{2i} \sim \mathrm{N}(\mu_2, \sigma^2)$ with unknown means μ_1 and μ_2 but a known variance σ^2 for simplicity. Let $\bar{y}_1 = \sum_{i=1}^{n} y_{1i}/n$ and $\bar{y}_2 = \sum_{i=1}^{n} y_{2i}/n$ be the sample means, and $\theta = \mu_1 - \mu_2$ and $\hat{\theta} = \bar{y}_1 - \bar{y}_2$.

We are interested in the one-sided hypothesis test,

$$H_0\colon \theta \le 0 \quad \text{versus} \quad H_1\colon \theta > 0,$$

the frequentist Z-test statistic is formulated as

$$z = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{2\sigma^2/n}} = \frac{\hat{\theta}}{\sqrt{2\sigma^2/n}},$$

which follows the standard normal distribution under H_0. In a one-sided hypothesis test, the corresponding p-value is

$$\begin{aligned} p\text{-value}_1 &= \Pr(Z \ge \hat{\theta}\sqrt{n/(2\sigma^2)}|H_0) \\ &= 1 - \Phi(\hat{\theta}\sqrt{n/(2\sigma^2)}), \end{aligned} \tag{6}$$

where Z denotes the standard normal random variable and $\Phi(\cdot)$ is the corresponding CDF.

Under the Bayesian framework, if we adopt an improper flat prior distribution for θ, i.e., $p(\theta) \propto 1$, the posterior distribution of θ is

$$\theta|D \sim \mathrm{N}(\hat{\theta}, 2\sigma^2/n).$$

Hence, the posterior probability of H_0 is

$$\mathrm{PoP}_1 = \Pr(H_0|D) = \Pr(\theta \le 0|D) = 1 - \Phi(\hat{\theta}\sqrt{n/(2\sigma^2)}),$$

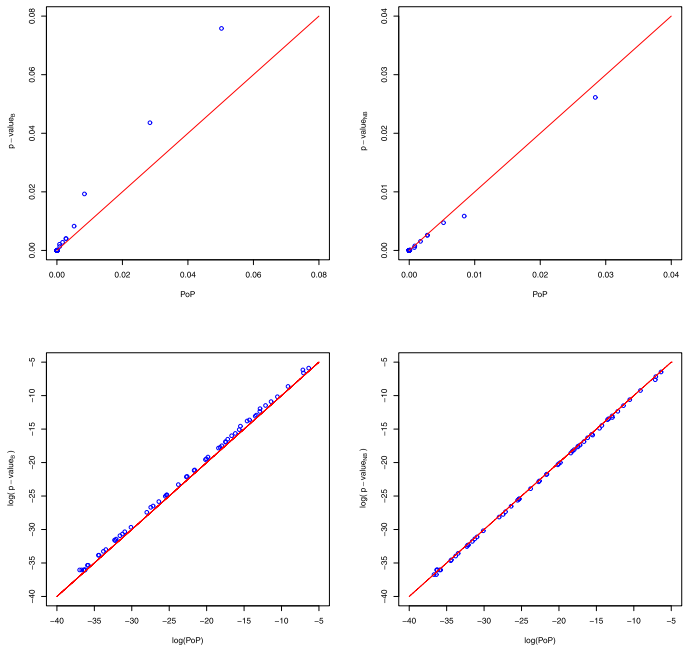

Figure 5. The relationship between p-values (p-value$_{\mathrm{B}}$ is based on the binomial distribution, and p-value$_{\mathrm{NB}}$ is based on the negative binomial distribution) and the posterior probability (PoP) of the null with outcomes generated randomly from a binomial distribution with probability equal to $y/n = 0.75$. The second row presents the zoom-in plot at the corner $(0, 0)$ of the first row, i.e., the logarithm of the p-value and PoP for p-values smaller than 0.002.

which is exactly the same as (6). We can thus establish an exact equivalence relationship between p-value and $\Pr(H_0|D)$ given an improper prior distribution of θ.

If we are interested in a two-sided hypothesis test,

$$H_0\colon \theta = 0 \quad \text{versus} \quad H_1\colon \theta \neq 0,$$

the p-value is given by

$$\begin{aligned} p\text{-value}_2 &= 2[1 - \max\{\Pr(Z \geq z|H_0), \Pr(Z \leq z|H_0)\}] \\ (7) \qquad &= 2 - 2\max\{\Phi(\hat{\theta}\sqrt{n/(2\sigma^2)}), \Phi(-\hat{\theta}\sqrt{n/(2\sigma^2)})\}. \end{aligned}$$

The two-sided test can be viewed as a combination of two one-sided tests (along the opposite directions), and thus the prior distribution can be easily specified as that in the one-sided test [22]. Otherwise, the point mass under the null hypothesis poses great challenges for Bayesian prior specifications. As a result, the two-sided posterior probability is defined as

$$\begin{aligned} \mathrm{PoP}_2 &= \Pr(H_0|D) \\ &= 2[1 - \max\{\Pr(\theta \leq 0|D), \Pr(\theta \geq 0|D)\}] \\ &= 2 - 2\max\{\Phi(\hat{\theta}\sqrt{n/(2\sigma^2)}), \Phi(-\hat{\theta}\sqrt{n/(2\sigma^2)})\}, \end{aligned}$$

which is exactly the same as the (two-sided) p-value in (7).

4.1.2 Normal prior

If the prior distribution for θ is assumed to be normal, i.e., $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$, the posterior distribution of θ is also normal, $\theta|D \sim \mathrm{N}(\tilde{\mu}, \tilde{\sigma}^2)$, where the corresponding posterior mean and the posterior variance are given by

$$\tilde{\mu} = \frac{\hat{\theta}\sigma_0^2 + \mu_0(2\sigma^2/n)}{\sigma_0^2 + 2\sigma^2/n}, \quad \tilde{\sigma}^2 = \frac{\sigma_0^2(2\sigma^2/n)}{\sigma_0^2 + 2\sigma^2/n}.$$

Under a one-sided test, the posterior probability of H_0 is

$$\begin{aligned} \mathrm{PoP}_1 &= \Pr(H_0|D) \\ &= \Pr(\theta \leq 0|D) \\ &= 1 - \Phi(\tilde{\mu}/\tilde{\sigma}) \\ &= 1 - \Phi\left(\frac{\hat{\theta}\sigma_0^2 + \mu_0(2\sigma^2/n)}{\sqrt{\sigma_0^2 + 2\sigma^2/n}} \cdot \frac{1}{\sigma_0\sqrt{2\sigma^2/n}}\right) \\ &= 1 - \Phi\left(\frac{\hat{\theta} + \mu_0(2\sigma^2/n)/\sigma_0^2}{\sqrt{1 + (2\sigma^2/n)/\sigma_0^2}} \cdot \frac{1}{\sqrt{2\sigma^2/n}}\right). \end{aligned}$$

As the prior variance increases and hence the prior distribution becomes more non-informative, then

$$\mathrm{PoP}_1 = \Pr(H_0|D) \to 1 - \Phi(\hat{\theta}\sqrt{n/(2\sigma^2)}), \quad \text{as } \sigma_0 \to \infty,$$

which equals the p-value under a one-sided hypothesis test.

For a two-sided hypothesis test, we can also assume a normal prior distribution for θ, i.e., $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$, and the asymptotic equivalence between the p-value and the posterior probability of the null can be derived similarly. In particular, we view the two-sided hypothesis test as a combination of two one-sided tests and $\Pr(\theta \leq 0|D)$ is the same as (6). For the other one-sided test, as $\sigma_0 \to \infty$,

$$\begin{aligned} \Pr(\theta \geq 0|D) &= 1 - \Phi(-\tilde{\mu}/\tilde{\sigma}) \\ &= 1 - \Phi\left(-\frac{\hat{\theta} + \mu_0(2\sigma^2/n)/\sigma_0^2}{\sqrt{1 + (2\sigma^2/n)/\sigma_0^2}} \cdot \frac{1}{\sqrt{2\sigma^2/n}}\right) \\ &\to 1 - \Phi(-\hat{\theta}\sqrt{n/(2\sigma^2)}). \end{aligned}$$

By combining the two one-sided tests, the two-sided posterior probability is given by

$$\begin{aligned} \mathrm{PoP}_2 &= \Pr(H_0|D) \\ &= 2[1 - \max\{\Pr(\theta \leq 0|D), \Pr(\theta \geq 0|D)\}] \\ &= 2 - 2\max\{\Phi(\hat{\theta}\sqrt{n/(2\sigma^2)}), \Phi(-\hat{\theta}\sqrt{n/(2\sigma^2)})\}, \end{aligned}$$

which is the same as the (two-sided) p-value in (7).

4.2 Hypothesis tests for multivariate normal data

In hypothesis testing with multivariate normal data, we consider $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where p is the dimension of the multivariate normal distribution. For ease of exposition, the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be known. Let

$D = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ denote the observed multivariate vectors, let $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ denote the sample mean vector, and thus $\bar{\mathbf{X}} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$.

Consider the one-sided hypothesis test,

$$H_0\text{: } \mathbf{c}_k^\top \boldsymbol{\mu} \le 0 \text{ for some } k = 1, \ldots, K$$
versus
$$H_1\text{: } \mathbf{c}_k^\top \boldsymbol{\mu} > 0 \text{ for all } k = 1, \ldots, K,$$

where $\mathbf{c}_1, \ldots, \mathbf{c}_K$ are K prespecified p-dimensional vectors. The likelihood ratio test statistic [24] is given by

$$Z_k = \frac{\mathbf{c}_k^\top \bar{\mathbf{X}}}{\sqrt{\mathbf{c}_k^\top \boldsymbol{\Sigma} \mathbf{c}_k / n}}, \quad k = 1, \ldots, K, \tag{8}$$

and the corresponding p-value is

$$p\text{-value}_1(k) = 1 - \Phi(Z_k).$$

The null hypothesis is rejected if all of the K p-values are smaller than α.

In the Bayesian paradigm, we assume a conjugate multivariate normal prior distribution for $\boldsymbol{\mu}$, i.e., $\boldsymbol{\mu} \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. The posterior distribution is $\boldsymbol{\mu}|D \sim \mathrm{N}_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, where

$$\begin{aligned}
\boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{\boldsymbol{\Sigma}}{n}\right)^{-1} \bar{\mathbf{X}} + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{\boldsymbol{\Sigma}}{n}\right)^{-1} \boldsymbol{\mu}_0, \\
\boldsymbol{\Sigma}_n &= \frac{1}{n}\boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{\boldsymbol{\Sigma}}{n}\right)^{-1} \boldsymbol{\Sigma}.
\end{aligned}$$

The one-sided posterior probability corresponding to $\mathbf{c}_k$ is

$$\mathrm{PoP}_1(k) = \Pr(\mathbf{c}_k^\top \boldsymbol{\mu} \le 0 | D).$$

As $\mathbf{c}_k^\top \boldsymbol{\mu}|D \sim \mathrm{N}\left(\mathbf{c}_k^\top \boldsymbol{\mu}_n, \mathbf{c}_k^\top \boldsymbol{\Sigma}_n \mathbf{c}_k\right)$, if we set $\boldsymbol{\mu}_0 = \mathbf{0}$, the one-sided posterior probability can be further derived as

$$\begin{aligned}
\mathrm{PoP}_1(k) &= 1 - \Phi\left(\frac{\mathbf{c}_k^\top \boldsymbol{\mu}_n}{\sqrt{\mathbf{c}_k^\top \boldsymbol{\Sigma}_n \mathbf{c}_k}}\right) \\
&= 1 - \Phi\left(\frac{\mathbf{c}_k^\top \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}/n\right)^{-1} \bar{\mathbf{X}}}{\sqrt{\mathbf{c}_k^\top \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}/n\right)^{-1} \boldsymbol{\Sigma} \mathbf{c}_k / n}}\right).
\end{aligned}$$

With a slight abuse of notation, let $\boldsymbol{\infty}$ denote an infinitely large positive definite matrix. When $\boldsymbol{\Sigma}_0 \to \boldsymbol{\infty}$, it is easy to verify that $\boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}/n\right)^{-1} \to \mathbf{I}_p$, where $\mathbf{I}_p$ is a p-dimensional identity matrix, and thus

$$\mathrm{PoP}_1(k) \to p\text{-value}_1(k).$$

The two-sided hypothesis test [25] can be formulated as

$$H_0\text{: } \mathbf{c}_k^\top \boldsymbol{\mu} \le 0 \text{ for some } k = 1, \ldots, K, \text{and}$$
$$\mathbf{c}_k^\top \boldsymbol{\mu} \ge 0 \text{ for some } k = 1, \ldots, K$$

versus

$$H_1\text{: } \mathbf{c}_k^\top \boldsymbol{\mu} > 0 \text{ for all } k = 1, \ldots, K, \text{ or}$$
$$\mathbf{c}_k^\top \boldsymbol{\mu} < 0 \text{ for all } k = 1, \ldots, K.$$

Based on (8), the p-value is given by

$$p\text{-value}_2(k) = 2 - 2\Phi(|Z_k|) = 2[1 - \max\{\Phi(Z_k), \Phi(-Z_k)\}].$$

The null hypothesis is rejected if all of the K p-values are smaller than α. Similar to the univariate case, we propose a definition of the two-sided posterior probability,

$$\mathrm{PoP}_2(k) = 2[1 - \max\{\Pr(\mathbf{c}_k^\top \boldsymbol{\mu} > 0|D), \Pr(\mathbf{c}_k^\top \boldsymbol{\mu} < 0|D)\}].$$

As $\boldsymbol{\Sigma}_0 \to \boldsymbol{\infty}$, the asymptotic equivalence between the posterior probability and p-value can be established along similar lines.

For illustration, we conduct a numerical study to compute the posterior probabilities of $\mathbf{c}_k^\top \boldsymbol{\mu} \le 0$ for $k = 1, \ldots, K$, and compare them with the corresponding p-values. We take $K = 2$ and $\mathbf{c}_k$ to be a unit vector with 1 on the kth element and 0 elsewhere. We assume a normal prior distribution for $\boldsymbol{\mu}$, i.e., $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \sigma_0^2 \mathbf{I}_p$. We experiment with $\sigma_0^2 = 1, 10$ and 1000 by increasing the prior variance. The relationship between the posterior probability of the null and p-value is shown in Figure 6. As the prior variance increases, the equivalence relationship becomes evident for both one-sided and two-sided multivariate tests.

4.3 Random effects model

We further consider a random effects model,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \varepsilon_{ij},$$

where y_{ij} is the outcome of observation j in cluster i, $i = 1, \ldots, n$; $j = 1, \ldots, J$, and covariates x_{ij}'s are generated from $\mathrm{Unif}(-1, 1)$. We assume the random intercept $b_i \sim \mathrm{N}(0, \tau^2)$ and the error $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$. We set the true parameter values to be $\beta_0 = 0.2$, $\beta_1 = 1$ and $\tau = \sigma = 0.5$, and the sample size $n = 100, 500$, and the cluster size $J = 2, 5$. In the Bayesian analysis, we set the prior distributions for β_0 and β_1 as $\mathrm{N}(0, 100)$, and the priors for τ^2 and σ^2 as inverse Gamma distributions, $\mathrm{IG}(0.01, 0.01)$.

We conduct hypothesis testing for both the regression coefficient and variance component; in particular, for β_1,

$$\text{Test 1: } \quad H_0 : \beta_1 \le \delta \quad \text{versus} \quad H_1 : \beta_1 > \delta,$$

and for τ,

$$\text{Test 2: } \quad H_0 : \tau^2 \le \xi \quad \text{versus} \quad H_1 : \tau^2 > \xi.$$

We vary the values of δ and ξ, and for each configuration we obtain the p-values using the Wald tests to and compare them with the posterior probabilities of the null. For the

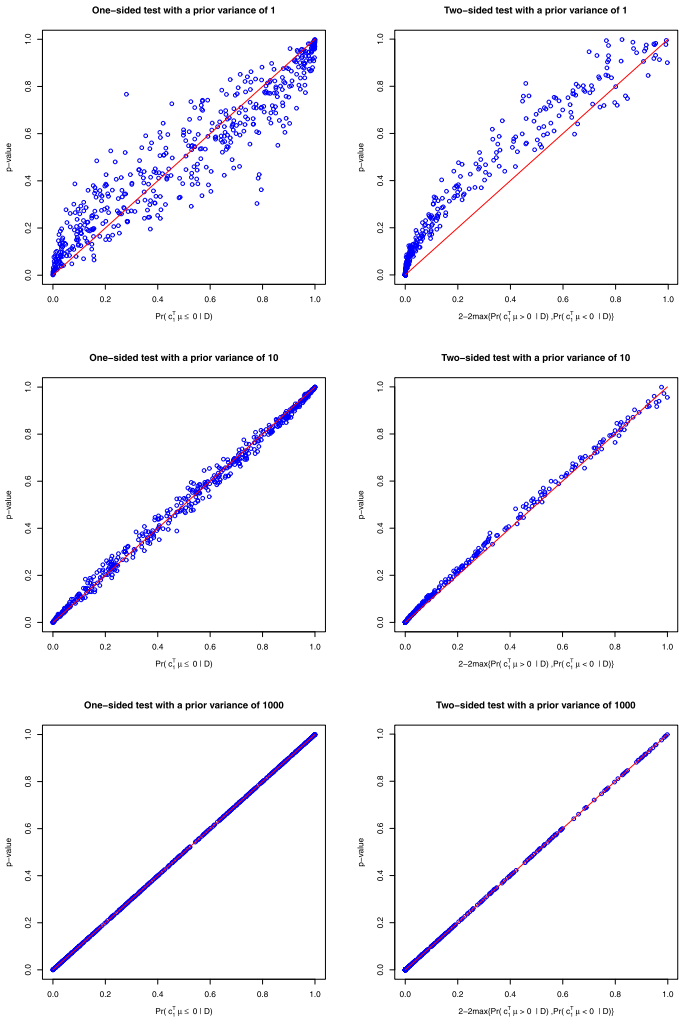

Figure 6. The relationship between the p-value and posterior probability over 1000 replications under one-sided and two-sided hypothesis tests with multivariate normal data with sample size 100, $\sigma_0^2 = 1, 10, 1000$ in the prior covariance variance matrix $\boldsymbol{\Sigma}_0 = \sigma_0^2 \mathbf{I}_p$.

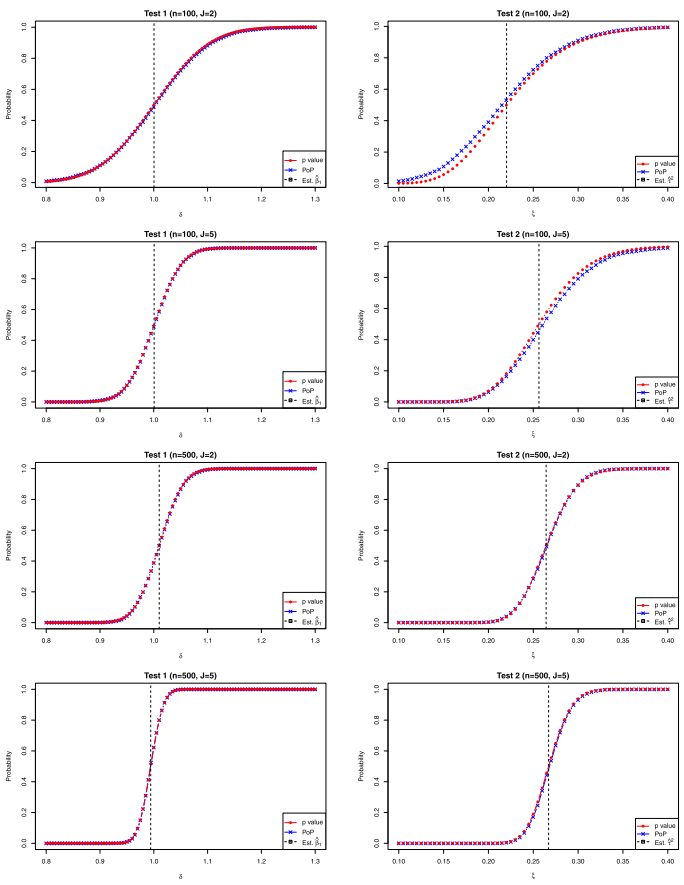

Figure 7. The relationship between the p-values and posterior probabilities for test 1 (coefficient) and test 2 (variance) with $n = \{100, 500\}$ and $J = \{2, 5\}$ under a random effects model.

frequentist test on τ^2, we use the asymptotic distribution based on the Fisher information,

$$\sqrt{N}(\hat{\tau}^2 - \tau^2) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, \frac{2\sigma^4}{J(J-1)} + \frac{2(J\tau^2+\sigma^2)^2}{J}\right),$$

and via the log transformation, we apply the delta method,

$$\sqrt{N}\{\log(\hat{\tau}^2) - \log(\tau^2)\} \xrightarrow{\mathcal{D}} \mathrm{N}\left(0, \left\{\frac{2\sigma^4}{J(J-1)} + \frac{2(J\tau^2+\sigma^2)^2}{J}\right\}\frac{1}{\tau^4}\right).$$

Figure 7 shows that the p-values and the posterior probabilities of the null hypothesis under different values of δ and ξ are very close, especially for large sample size $n = 500$. The matching pattern between the two quantities appears to be better for the tests of the regression coefficient than those of the variance component.

4.4 Hypothesis tests with non-normal data

For data that are assumed to follow a distribution other than a normal distribution, we can establish the equivalence relationship between the p-value and posterior probability using the theoretical results on asymptotical normality of the posterior probability of the half-space in [26]. As examples, we consider an exponential distribution for continuous data and a Poisson distribution for discrete data.

4.4.1 Exponential distribution

Consider a one-sample test with exponentially distributed data, $Y \sim \exp(\theta)$. Let n denote the sample size, and let D denote the observed data. Under the exponential distribution,

$$f(y) = \frac{1}{\theta}\exp(-y/\theta),$$

with mean θ and variance θ^2, we are interested in testing the hypotheses,

$$H_0 : \theta \leq \mu \quad \text{versus} \quad H_1 : \theta > \mu.$$

Let $\bar{y} = \sum_{i=1}^{n} y_i/n$ be the sample mean. Based on the Central Limit Theorem, $\bar{y} \xrightarrow{\mathcal{D}} \mathrm{N}(\theta, \theta^2/n)$, and the Wald test statistic is

$$Z = \frac{\bar{y} - \mu}{\mu/\sqrt{n}},$$

therefore the corresponding p-value is

$$p\text{-value} = 1 - \Phi(Z).$$

Under the Bayesian framework, we adopt an inverse gamma prior distribution $\theta \sim \mathrm{IG}(a, b)$, and based on the conjugate property, the posterior distribution of θ is

$$\theta | D \sim \mathrm{IG}(a + n, b + n\bar{y}).$$

The posterior probability of the null is

$$\mathrm{PoP} = \Pr(H_0|D) = \Pr(\theta \leq \mu | D) = F_{\mathrm{IG}}(\mu; a + n, b + n\bar{y}),$$

where F_{IG} represents the CDF of an inverse gamma distribution. To establish the asymptotical equivalence between the p-value and PoP, based on the theoretical results from [26], the posterior probability of H_0 converges to the standard normal CDF transformation of the likelihood ratio test statistic, $\Phi(-\sqrt{\Delta})$, where Δ is the likelihood ratio test statistic, which is asymptotically equivalent to the Wald test statistic Z^2 [27]. Therefore, we can establish the asymptotical equivalence between the p-value and posterior probability,

$$\mathrm{PoP} \to 1 - \Phi(Z) = p\text{-value}.$$

For illustration, we conduct a numerical study by simulating data from an exponential distribution with mean $\theta = 3$ and sample size 300. We are interested in testing $H_0 : \theta \leq 3$ versus $H_1 : \theta > 3$. For the prior distribution, the hyperparameters of the inverse gamma distribution are set as $a = b = 0.001$. The upper panel of Figure 8 shows the results under 1000 data replications, from which the equivalence relationship between the p-value and posterior probability is evident.

4.4.2 Poisson distribution

We consider another one-sample test where the data follow a Poisson distribution, $Y \sim \mathrm{Poi}(\lambda)$, with the probability mass function,

$$\Pr(Y = y) = \frac{\lambda^y e^{\lambda}}{y!},$$

where both the mean and variance are λ. We are interested in testing the hypotheses,

$$H_0 : \lambda \leq \mu \quad \text{versus} \quad H_1 : \lambda > \mu.$$

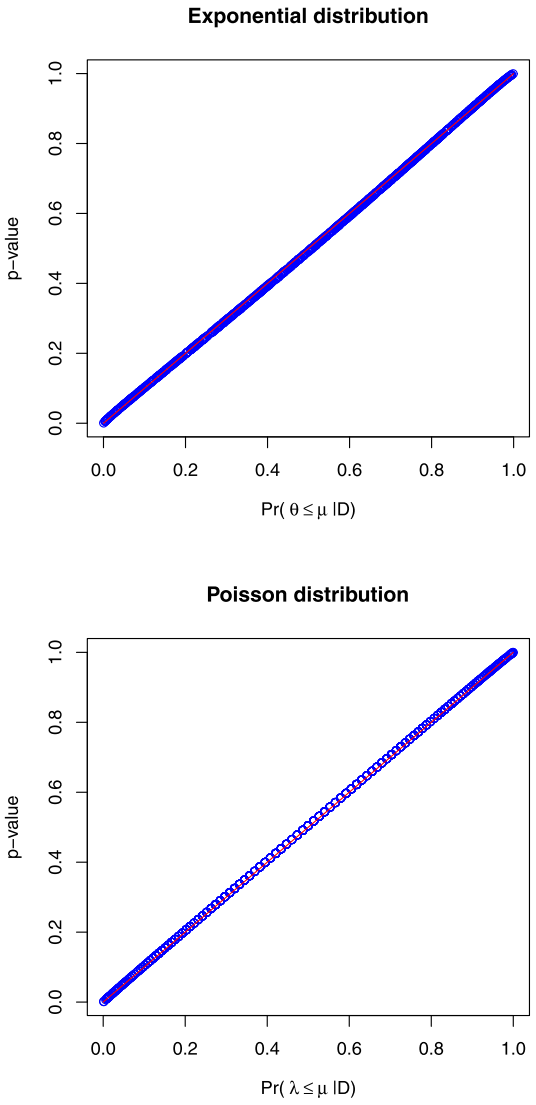

Figure 8. The relationship between the p-value and the posterior probability over 1000 replications under an exponential and a Poisson distribution with sample size 300.

Let n denote the sample size, let D denote the observed data, and $\bar{y} = \sum_{i=1}^{n} y_i/n$ is the sample mean. Based on the Central Limit Theorem, $\bar{y} \xrightarrow{\mathcal{D}} \mathrm{N}(\lambda, \lambda/n)$, and the Wald test statistic is

$$Z = \frac{\bar{y} - \lambda}{\sqrt{\lambda/n}},$$

therefore the corresponding p-value is

$$p\text{-value} = 1 - \Phi(Z).$$

For the Bayesian method, we adopt a gamma prior distribution $\lambda \sim \mathrm{Gamma}(a, b)$, and based on the conjugate

property, the posterior distribution of λ is

$$\lambda|D \sim \text{Gamma}(a + n\bar{y}, b + n).$$

The posterior probability of the null is

$$\text{PoP} = \Pr(H_0|D) = \Pr(\lambda \le \mu|D) = F_{\text{Gamma}}(\mu; a+n\bar{y}, b+n),$$

where F_{Gamma} represents the CDF of a gamma distribution. The asymptotical equivalence between the p-value and PoP follows the same derivation as in the previous section. We conduct a numerical study where data are simulated from a Poisson distribution with mean 3 and sample size 300. We are interested in testing $H_0 : \lambda \le 3$ versus $H_1 : \lambda > 3$. The hyperparameters of the gamma prior distribution are taken as $a = b = 0.001$. As shown in the lower panel of Figure 8, the p-values and the posterior probabilities are closely matched, indicating an equivalence relationship.

5. REAL DATA APPLICATION

We consider two real examples where Lindley's paradox can be constructed, leading to an unignorable influence on the interpretation of the numerical results under the frequentist and the Bayesian inferences. We consider two case studies from the Framingham Heart Study discussed in [28].

The Framingham Heart Study is a long-term and ongoing cohort study with a goal of understanding the risk factors that predispose to cardiovascular disease [29]. Initiated in 1948, the study is now on its fourth generation of participants. The demographical and medical information of the subjects is gathered every 3 to 5 years. The original cohort consisted of 5209 subjects between age 30 and 62, and an offspring cohort, which our data analysis is based upon, was added in 1971.

5.1 Cardiovascular disease among smokers/non-smokers

At the fifth examination in the Framingham Heart Study, we analyze the proportion of cardiovascular disease patients among smokers and non-smokers. Out of 744 smokers and 3055 non-smokers, 81 and 298 individuals had history of cardiovascular disease, respectively.

Let p_1 and p_2 denote the probabilities of cardiovascular disease for smokers and non-smokers, respectively, and let $\theta = p_1 - p_2$ denote the difference in the occurrence probability of cardiovascular disease. We are interested in testing the hypotheses, $H_0 : \theta = 0$ versus $H_1 : \theta \ne 0$.

The sample proportion difference is $\hat{y} = \bar{y}_1 - \bar{y}_2$, where $\bar{y}_1 = 81/744$ and $\bar{y}_2 = 298/3055$. Under normal approximation, $\hat{y} \sim \text{N}(\theta, \sigma^2)$, where $\sigma^2 = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$. Naturally, we use $\hat{\sigma}^2 = \bar{y}_1(1-\bar{y}_1)/n_1 + \bar{y}_2(1-\bar{y}_2)/n_2 = 0.000159$ to estimate σ^2. Thus, the p-value is

$$p\text{-value} = 2\{1 - \Phi(\hat{y}/\hat{\sigma})\} = 0.3692.$$

Under the Bayesian framework, Lindley's paradox would occur if we assign a prior distribution on θ with a point mass at 0, i.e., $P(H_0) = P(H_1) = 0.5$. For the case where $\theta \ne 0$, we assume a diffuse prior distribution for θ, i.e., $\theta|\theta \ne 0 \sim \text{N}(0, {\sigma_0}^2)$ with $\sigma_0 = 100$. For the case with $\theta = 0$, the marginal distribution of $\hat{y}$ given $\theta = 0$ is $\hat{y}|\theta = 0 \sim \text{N}(0, \hat{\sigma}^2)$.

Let $\phi(y; \theta, \sigma)$ denote the probability density function of a normal distribution with mean θ and standard deviation σ. Using Bayes' theorem, the posterior probability of H_0 is

$$\begin{aligned} P(H_0|\hat{y}) &= \frac{f(\hat{y}|H_0)P(H_0)}{f(\hat{y}|H_0)P(H_0) + f(\hat{y}|H_1)P(H_1)} \\ &= \frac{\phi(\hat{y}; 0, \hat{\sigma})}{\phi(\hat{y}; 0, \hat{\sigma}) + \int_{-\infty}^{\infty} \phi(\hat{y}; \theta, \hat{\sigma})\phi(\theta; 0, \sigma_0)d\theta} \\ &= 0.9998, \end{aligned}$$

which does not match with the frequentist p-value.

On the other hand, if we assume an improper prior $p(\theta) \propto 1$, the posterior distribution is given by

$$P(\theta|\hat{y}) \propto \exp\left\{ -\frac{(\theta - \hat{y})^2}{2\hat{\sigma}^2} \right\},$$

i.e., $\theta|\hat{y} \sim N(\hat{y}, \hat{\sigma}^2)$. As a result, we can compute the two-sided posterior probability,

$$\begin{aligned} \text{PoP}_2 &= 2 \times \min\{P(\theta \le 0|\hat{y}), P(\theta \ge 0|\hat{y})\} \\ &= 2 \times \min\{\Phi(-\hat{y}/\hat{\sigma}), \Phi(\hat{y}/\hat{\sigma})\} \\ &= 0.3692, \end{aligned}$$

which reconciles with the frequentist p-value.

5.2 Systolic blood pressure difference between genders

Among the 3539 subjects who participated in the seventh examination of the Framingham Heart Study, there were $n_1 = 1623$ men and $n_2 = 1911$ women. The mean systolic blood pressures for men and women were $\bar{y}_1 = 128.2$ and $\bar{y}_2 = 126.5$ respectively, and the corresponding sample standard deviations were $s_1 = 17.5$ and $s_2 = 20.1$. Let μ_1 and μ_2 denote the mean systolic blood pressures for men and women, respectively, and let $\theta = \mu_1 - \mu_2$ denote their difference. We are interested in testing whether the mean systolic blood pressures are different between men and women; that is, $H_0 : \theta = 0$ versus $H_1 : \theta \ne 0$.

The sample mean difference is $\hat{y} = \bar{y}_1 - \bar{y}_2$. Under the normal assumption, $\hat{y} \sim \text{N}(\theta, \sigma^2)$, and we approximate σ^2 using the pooled estimate $\hat{\sigma}^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)(1/n_1 + 1/n_2)/(n_1 + n_2 - 2) = 358.0862$. Thus, the p-value is

$$p\text{-value} = 2\{1 - \Phi(\hat{y}/\hat{\sigma})\} = 0.0078.$$

Under the Bayesian framework, Lindley's paradox would occur if we assign a prior distribution on θ with a point mass

at 0, i.e., $P(H_0) = P(H_1) = 0.5$. Following similar settings for θ as in the previous section, the posterior probability of H_0 is

$$\begin{aligned} P(H_0|\hat{y}) &= \frac{f(\hat{y}|H_0)P(H_0)}{f(\hat{y}|H_0)P(H_0) + f(\hat{y}|H_1)P(H_1)} \\ &= \frac{\phi(\hat{y}; 0, \hat{\sigma})}{\phi(\hat{y}; 0, \hat{\sigma}) + \int_{-\infty}^{\infty} \phi(\hat{y}; \theta, \hat{\sigma})\phi(\theta; 0, \sigma_0)d\theta} \\ &= 0.8193, \end{aligned}$$

which does not match with the frequentist p-value. Under a significance level of 0.05, the frequentist inference would reject H_0 while the Bayesian one would not. On the other hand, if we assume an improper flat prior $p(\theta) \propto 1$, the two-sided Bayesian posterior probability can be reconciled with the frequentist p-value.

6. DISCUSSION

The p-value is the most commonly used summary measure for evidence-based studies, and it has been the center of controversies and debates for decades. Recently reignited discussion over p-value has been more centered around the proposals to adjust, abandon or provide alternatives to the p-value. By definition, p-value is not the probability that the null hypothesis is true given the observed data. Contrary to the conventional notion, it does have a close correspondence to the Bayesian posterior probability of the null hypothesis under both one-sided and two-sided hypothesis tests. The necessary condition for the asymptotical equivalence relationship between the p-value and the posterior probability is primarily the assumption of a non-informative prior. Moreover, we have shown that under certain conjugate priors (e.g., normal data with a flat prior), exact equivalence can be established. Certainly, such an equivalence relationship would not hold when informative priors are used, because the p-value is computed without any prior information involved. Lindley's paradox mainly arises when a point mass is put on the parameter of interest under the null hypothesis. We circumvent the controversy by recasting a two-sided hypothesis into two one-sided hypotheses, and then the paradox can be explained: the p-value and the Bayesian posterior probability of the null hypothesis coincide.

We have established the asymptotical equivalence relationship between the p-value and posterior probability under normal approximation. Specifically, we have considered an exponential distribution for continuous data and a Poisson distribution for discrete data, and used normal approximation to establish the asymptotical equivalence relationship. Shi and Yin [22] discuss similar results for binomial data, and our results can be regarded as an extension. Moreover, we have also discussed the exact p-value for binomial data, which also leads to an evident equivalence.

ACKNOWLEDGEMENTS

We would like to thank Hengtao Zhang for his help with numerical simulations under random effects models, and three anonymous referees, Associate Editor and Editor-in-Chief (Ming-Hui Chen) for their insightful suggestions that have led to great improvements in this article. The research was supported by a grant (106200216) from the Research Grants Council of Hong Kong.

Received 11 December 2020

REFERENCES

[1] Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192. MR0087273

[2] Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? (with discussion) *Statistical Science* **18**, 1–32. MR1997064

[3] Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review* **14**, 779–804.

[4] Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association* **82**, 112–122. MR0883340

[5] Berger, J. O. and Delampady M. (1987). Testing precise hypotheses. *Statistical Science* **2**, 317–335. MR0920141

[6] Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. (with discussion) *Journal of the American Statistical Association* **82**, 106–111. MR0883339

[7] Robert, C. (1993). A note on Jeffreys–Lindley paradox. *Statistica Sinica* **3**, 601–608. MR1243404

[8] Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician* **55**, 62–71. MR1818723

[9] Benjamin, D. J. and Berger J. O. (2019). Three recommendations for improving the use of p-values. *The American Statistician* **73**, 186–191. MR3925724

[10] Betensky, R. A. (2019). The p-value requires context, not a threshold. *The American Statistician* **73**, 115–117. MR3925715

[11] Kennedy-Shaffer, L. (2019). Before $p < 0.05$ to beyond $p < 0.05$: using history to contextualize p-values and significance testing. *The American Statistician* **73**, 82–90. MR3925711

[12] Matthews, R. A. J. (2019). Moving towards the post $p < 0.05$ era via the analysis of credibility. *The American Statistician* **73**, 202–212. MR3925726

[13] McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician* **73**, 235–245. MR3925729

[14] Ruberg, S. J., Harrell Jr., F. E., Gamalo-Siebers, M., LaVange, L., Lee, J. J., Price, K. and Peck, C. (2019). Inference and decision making for 21st-century drug development and approval. *The American Statistician* **73**, 319–327. MR3925738

[15] Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician* **73**, 1–19. MR3511040

[16] Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 143–170. MR2830762

[17] Greenland, S. and Poole, C. (2013). Living with P-values: Resurrecting a Bayesian perspective. *Epidemiology* **24**, 62–68.

[18] Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: the p value fallacy. *Annals of Internal Medicine* **130**, 995–1004.

[19] Held, L. and Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application* **5**, 393–419. MR3774753
[20] Hubbard, R. and Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology* **18**: 69–88.
[21] Johnson, V. E. (2019). Evidence from marginally significant t statistics. *The American Statistician* **73**, 129–134. MR3925718
[22] Shi, H. and Yin, G. (2020). Reconnecting p-value and posterior probability under one- and two-sided tests. *The American Statistician* DOI: 10.1080/00031305.2020.1717621
[23] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician* **70**, 129–133. MR3511040
[24] Sasabuchi, S. (1980). A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika* **67**, 429–439. MR0581738
[25] Liu, H. and Berger, R. L. (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *The Annals of Statistics* **23**, 55–72. MR1331656
[26] Dudley, R. M. and Haughton, D. (2002). Asymptotic normality with small relative errors of posterior probabilities of half-spaces. *The Annals of Statistics* **30**, 1311–1344. MR1936321
[27] Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician* **36**, 153–157.
[28] Sullivan, L. M. (2007). Essentials of Biostatistics in Public Health. New York: Jones & Bartlett Learning.
[29] Mahmood, S. S., Levy, D., Vasan, R. S. and Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* **383**, 999–1008.

Guosheng Yin
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road
Hong Kong
E-mail address: gyin@hku.hk

Haolun Shi
Department of Statistics and Actuarial Science
School of Computing Science
Simon Fraser University
Burnaby, BC
Canada
E-mail address: haoluns@sfu.ca