Reconnecting p-Value and Posterior Probability Under One and Two-Sided Tests

James Park, Nakul Haridas, Dimitrios Ligas

1 Introduction-Motivation

The motivation for studying p-values in a Bayesian framework is to provide an alternative interpretation of p-values and to bridge the gap between frequentist and Bayesian approaches to statistical inference.

The p-value approach does not directly provide a measure of evidence in favor of the null or alternative hypothesis, and it relies on arbitrary thresholds for statistical significance.

In contrast, Bayesian inference provides a coherent framework for quantifying uncertainty and updating beliefs about hypotheses based on observed data. By assigning prior probabilities to hypotheses and updating them based on observed data, Bayesian inference can directly provide measures of evidence in favor of the null or alternative hypothesis.

By studying the equivalence relationship between p-values and Bayesian posterior probabilities, we can provide a Bayesian interpretation of p-values and bridge the gap between frequentist and Bayesian approaches. This can lead to a better understanding of the strengths and limitations of both approaches and provide a more comprehensive framework for statistical inference.

In this paper, Shi and Yin reassure the use of the p-value for hypothesis testing by demonstrating that the p-value is equivalent to the Bayesian posterior probability of the null hypothesis for one-sided and two-sided hypothesis tests under uninformative priors.

First, the motivating example of a two-arm clinical trial is presented. For a clinical trial to test if more participants respond to the experimental treatment compared to the standard of care practice, the data distribution can be explained as two independent binomial distributions, one for each treatment, and the success being the participants whose health outcome responded to the treatment. The null and the alternate hypothesis can be shown as:

$$H_0 : p_E \leq p_S \qquad versus \qquad H_1 : p_E \geq p_S$$

where p_E is the proportion of responders in the experimental arm, and p_S is the proportion in the standard of care arm.

In a frequentist setting, the Z-test statistic is calculated based on the number of observed responders in each arm of the trial. The critical z-score is obtained based on the α-level selected prior to observing the data, and the null hypothesis is rejected in cases the Z value exceeds the critical z-score or z_α. Under a Bayesian approach, beta prior distributions for $p_E and p_S$ are assumed along with a binomial likelihood function, which means the posterior probability will have a beta distribution given by

$$p_g|y_g \sim Beta(a_g + y_g, b_g + n - y_g),$$

where a and be are the two parameters from the beta prior, g is either E or S, and n is the number of samples in each arm. The treatment is considered when the probability of p_E greater than p_S exceeds a previously set threshold η. Authors claim that setting $\eta = 1 - \alpha$ allows controlling the Type I error rate at α.

The Type I and the Type II error rates in a frequentist setting is obtained by

$$Pr(RejectH_0|H0) = \sum_{y_E=0}^{n} \sum_{y_S=0}^{n} P(y_E|p_E = p_S)P(y_S|p_S)I(Z \geq z_\alpha)$$

$$Pr(AcceptH_0|H1) = \sum_{y_E=0}^{n} \sum_{y_S=0}^{n} P(y_E|p_E = p_S + \delta)P(y_S|p_S)I(Z < z_\alpha)$$

where δ is the difference between p_Eandp_S, and $I(\cdot)$ is the indicator function. To calculate the Bayesian Type I error rate, $\Pr(p_E > p_S|y_E, y_S) \geq 1-\alpha)$ replaces $Z \geq z_\alpha$ inside the indicator function, while $\Pr(p_E > p_S|y_E, y_S) < 1-\alpha)$ replaces $Z < z_\alpha$ for the Type II error rate.

Shown in Figure 1 are the Type I error rates and the powers calculated under the frequentist and the Bayesian frameworks, for the indicated critical z-values and the number of samples, n, which was calculated to achieve the desired power. The powers were calculated for two cases: when $p_E = p_S + 0.1$ or when $p_E = p_S + 0.15$. For most cases, the calculated Type I error rates and the powers between the two approaches overlap almost completely. The little bit of deviation in parts of the plots could be resolved by increasing the number of beta samples to be taken when calculating the Bayesian posterior probabilities. Twenty-thousand random beta samples were taken to obtain the current calculations, and any higher number would have increased the computation time significantly. Figure 1c) shows more fluctuation, and this is due to the sample size being too small (red line n=156, blue line n=298). Another thing to note is that since this is a discrete case, the calculated values on the edges of the range of p_E are prone to fluctuating more than the values in the mid-range. Overall, Figure 1 indicates that the Type I error rate and the Type II error rate for the frequentist and the Bayesian hypothesis tests of a one-sided binary case can be controlled to be approximately equal.

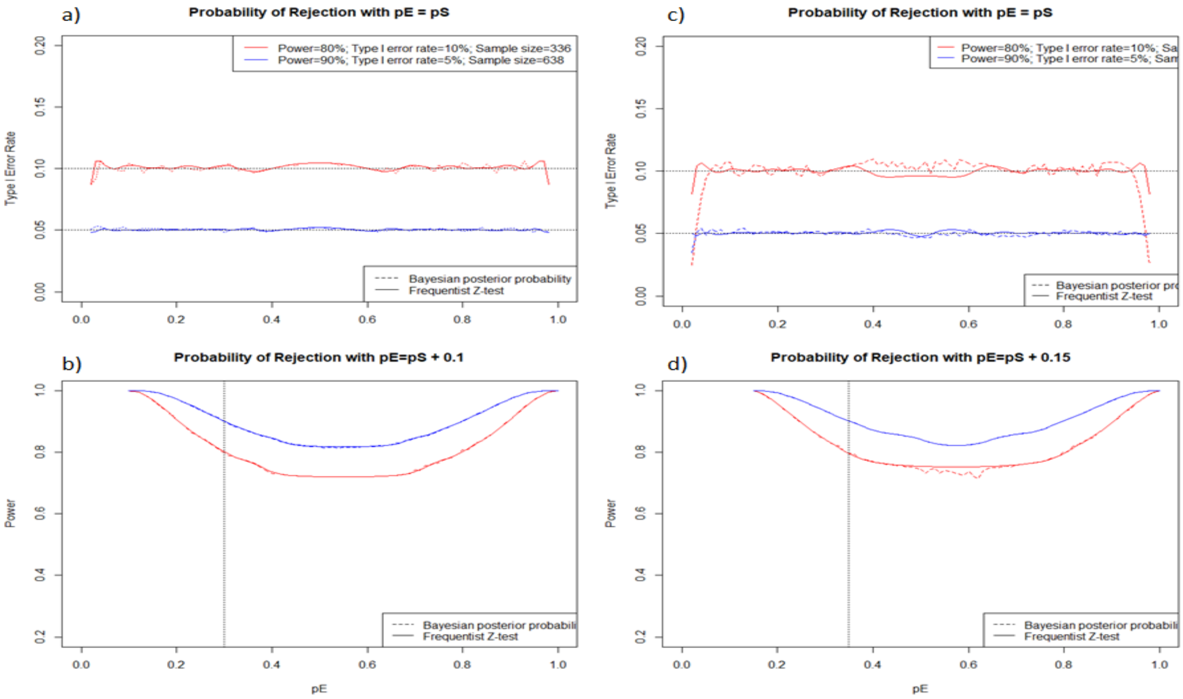

Figure 1: The Type I error rates and the powers of the tests calculated for either α=0.05 or 0.10, and the target power = 80 or 90%. For c) and d), red n=156, blue n=298

2 Hypothesis Test for Binomial Data

One-tail

Binary model has been assumed for the model and we test the following one sided Hypothesis. Initially, we test the one-sided Hypothesis as follows:

$$H_0 : p_E \leq p_S \text{ Versus } H_a : p_E \geq p_S$$

The Z Statistic is formulated as : $Z = \frac{\hat{p_e} - \hat{p_s}}{\sqrt{\frac{\hat{p_e}(1-\hat{p_e})}{n} + \frac{\hat{p_S}(1-\hat{p_S})}{n}}}$ and the p-value is: $1 - \phi(Z)$. Here $\phi(Z)$ is the CDF of the normal distribution.

Under the Bayesian approach, the Posterior probability for one tail (Pop1) is computed as follows:

$$\text{PoP}_1 = \Pr(p_e \leq p_s | y_e, y_s)$$

Two-Tail

Next, we conduct the analysis for a two-sided alternative Hypothesis, which is as follows.

$$H_0 : p_E \leq p_S \text{ Versus } H_a : p_E \neq p_S$$

The Z statistic is calculated as before, but the p-value is modified as below:

$$\text{p-value}_2 = 2 * (1 - max((\phi(Z), \phi(-Z))$$

The Posterior Probability is calculated as follows:

$$\text{PoP}_2 = 2(1 - max(\Pr(p_e \leq p_s | y_e, y_s), \Pr(p_s \leq p_e | y_e, y_s)))$$

Methodology

For testing the equivalence of the two approaches, the sample size n is taken from the set $\{20, 50, 100, 500\}$ and y_e, y_s are generated between values from 2 to n-2. The values $1, n, n-1$ are omitted as the normal approximation of the p-value is not accurate. The simulation is conducted in R, and we calculate the probabilities. Jeffrey's prior is assumed for both p_e and p_s. The same procedure is done for testing the one-sided hypothesis and two-sided hypotheses.

Results for One-Tail

The following shows the results obtained, which tells us that under uninformative priors, in the binary model, there is equivalence between p-value and one-sided posterior probability. We can see that as the sample size is increased almost all the points are on the $y = x$ line indicating their equivalence.

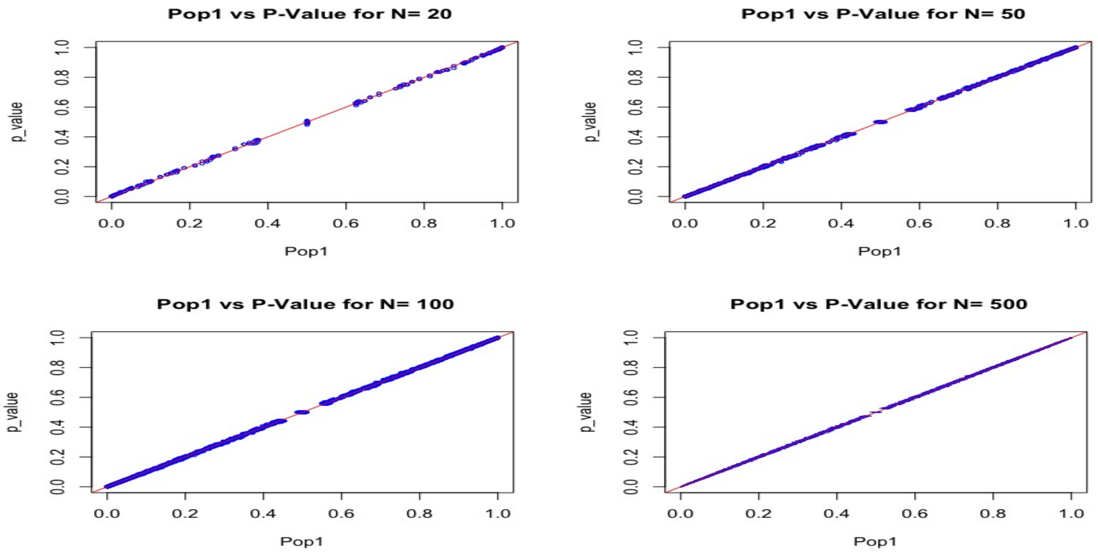

Figure 2: The relationship between p-value and posterior probability of null one-sided hypothesis testing under binary outcomes for sample sizes of 10,20,100 and 100.

Results for Two-Tail

The equivalence of the p-values and the Posterior Probability for two-sided hypothesis testing are observed under n=20,50,100 and 500. As the sample size increases, the points increasingly become clustered on the $y = x$ line.

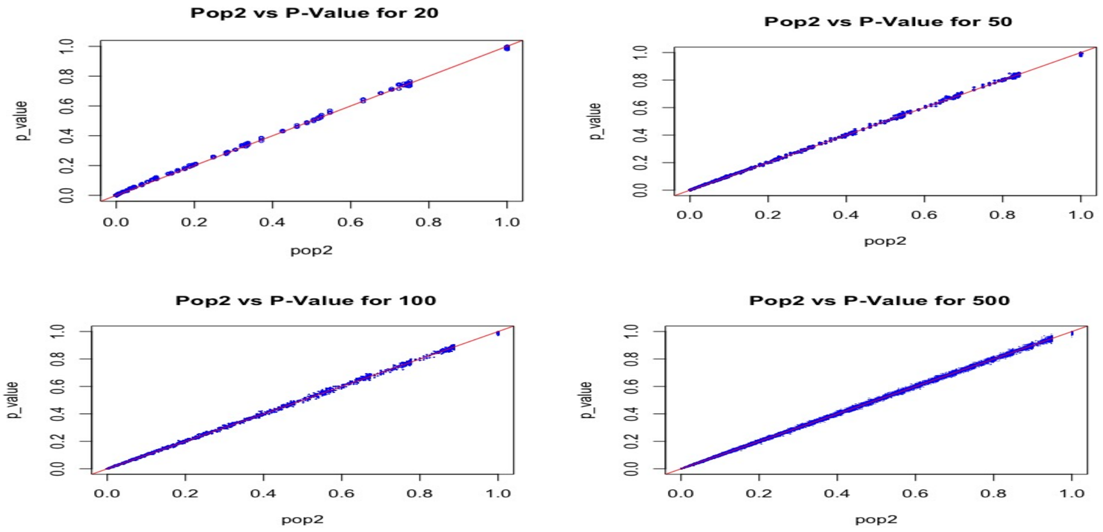

Figure 3: The relationship between p-value and posterior probability of null two-sided hypothesis testing under binary outcomes for sample sizes of 10,20,100 and 100.

3 Hypothesis Test for Normal Data

Hypothesis Test with Known Variance

Next, we examine the equivalence relationship between the p-value and the posterior probability of the null under uninformative priors for normal data. We consider a two-arm trial with normally distributed outcomes in which the objective is two compare the means between the control and the treatment groups. Let:

$$Y_{E_i}\ , Y_{S_i},\ i = 1 : n$$

denote the outcomes in the treatment and control groups respectively. The assumption is:

$$Y_{E_i} \overset{iid}{\sim} N(\mu_E, \sigma^2),\ Y_{S_i} \overset{iid}{\sim} N(\mu_S, \sigma^2)$$

The two means are unknown, and first we consider the case of known variance, which for simplicity is assumed to be $\sigma^2 = 1$. The difference of the true means is $\theta = \mu_E - \mu_S$ and comprises the parameter of interest, while its estimator is $\hat{\theta} = \bar{Y}_E - \bar{Y}_S$

I. One-sided hypothesis test-Exact Equivalence
Consider the one-sided hypothesis test:

$H_o : \theta \leq 0\ vs\ H_a : \theta > 0$. The frequentist Z-test is $Z|H_o = \frac{\hat{\theta}-0}{\sqrt{Var(\hat{\theta})}} \sim N(0,1)$ .From the independence: $Var(\hat{\theta}) = Var(\bar{Y}_E) + Var(\bar{Y}_S) = \sigma^2/n + \sigma^2/n = 2\sigma^2/n = \frac{2}{n}$ and so, $Z|H_o = \hat{\theta}\sqrt{n/2}$. Then, the p-value for the one-sided test is:

$$p\text{-value}_1 = P\left[Z > Z_o|H_o\right] = P\left[Z > \hat{\theta}\sqrt{n/2}\right] = 1 - N\left(\hat{\theta}\sqrt{n/2)}\right)$$

where N denotes the cdf of the Standard Normal Distribution.

In the Bayesian paradigm, our uncertainty about the null hypothesis given the observed data Y is quantified in the posterior of θ. The Jeffreys' prior for the likelihood parameters is $\pi(\theta, \sigma^2) \propto \frac{1}{(\sigma^2)^2}$ $(\theta = \mu_E - \mu_S)$,which for known σ^2 leads to the the improper flat prior:

$$\pi(\theta, \sigma^2) \propto 1$$

The posterior of θ is:

$$p(\theta|Y) \propto f(Y|\theta, \sigma^2)\pi(\theta, \sigma^2) \propto e^{-\frac{1}{2\sigma^2}\left[\Sigma_{i=1}^n (Y_i - m_e)^2 + \Sigma_{i=1}^n (Y_i - m_s)^2\right]}$$

$$\propto e^{-\frac{1}{2\sigma^2}\left[n\mu_E^2 + n\mu_S^2 - 2\mu_E n\bar{Y}_E - 2\mu_S n\bar{Y}_S\right]} \propto e^{-\frac{1}{2\sigma^2/n}\left[\mu_E^2 + \mu_S^2 - 2\mu_E\bar{Y}_E - 2\mu_S\bar{Y}_S\right]}$$

$$\propto e^{-\frac{1}{\sigma^2\frac{2}{n}}\left[\mu_E^2 + \mu_S^2 - 2\mu_E\bar{Y}_E - 2\mu_S\bar{Y}_S\right]}$$

which we can identify as the kernel of a normal density with mean $\bar{Y}_E - \bar{Y}_S$ and variance $\sigma^2\frac{2}{n}$. Therefore, under Jeffreys' prior with known variance, the posterior of θ, $p(\theta|Y)$ is:

$$\theta|Y \sim N\left(\bar{Y}_E - \bar{Y}_S, \sigma^2\frac{2}{n}\right)$$

The posterior probability of the null will be:

$$\text{PoP}_1 = P\left[H_o : \theta < 0|Y\right] = P\left[\frac{\theta - (Y_E - \bar{Y}_S)}{\sqrt{2/n}} \leq \frac{-(\bar{Y}_E - \bar{Y}_S)}{\sqrt{2/n}}\right] = P\left[Z \leq -\hat{\theta}\sqrt{n/2}\right] = 1 - N\left(\hat{\theta}\sqrt{n/2})\right)$$

Therefore, for a known variance and under an uninformative Jeffreys' prior, the posterior probability of the null PoP_1 and the p-value are the same, and so the equivalence relationship for the one-sided hypothesis test was established.

$$\text{PoP}_1 = P\left[H_o : \theta < 0|Y\right] = p\text{-value}_1$$

II. Two-sided hypothesis test-Exact Equivalence
For the two-sided hypothesis test: $H_o : \theta = 0 \; vs \; H_a : \theta \neq 0$ the p-value is:

$$p\text{-value}_2 = 2P\left[Z > |Z_o|\right] = 2\left(1 - P\left[Z \leq |Z_o|\right]\right) = 2\left(1 - N\left[|\sqrt{n/2}\hat{\theta}|\right]\right) = 2\left(1 - max\left\{N\left[-\sqrt{n/2}\hat{\theta}\right], N\left[\sqrt{n/2}\hat{\theta}\right]\right\}\right)$$

The authors define the posterior probability of the null for the two-sided case as:
$\text{PoP}_2 = P\left[H_o : \theta \neq 0|Y\right] = 2\left[1 - max\left\{p(\theta|Y < 0), p(\theta|Y > 0)\right\}\right] = 2\left(1 - max\left\{N\left[-\sqrt{n/2}\hat{\theta}\right], N\left[\sqrt{n/2}\hat{\theta}\right]\right\}\right)$
which is the same as the p-value.
Thus, for the two-sided hypothesis test as well, and under Jeffrey's prior with known variance, the posterior probability of the null as defined by the authors is equivalent to the p-value $\text{PoP}_2 = P\left[H_o : \theta \neq 0|Y\right] = p\text{-value}_2$

Hypothesis Test with Unknown Variance

When σ is unknown, we consider the difference $X_i = Y_{E_i} - Y_{S_i}$ which is distributed as $X_i \overset{iid}{\sim} N(\theta, 2\sigma^2)$. Then, the hypothesis test problem reduces to a 1-sample test similar to a matched pair study, with the frequentist t-test being:

$$T|H_o = \frac{\hat{\theta}}{\hat{\sigma}^2/n} \sim t_{n-1} \text{ ,with } \hat{\theta} = \bar{X} = \bar{Y}_E - \bar{Y}_S \text{ and the sample variance } \hat{\sigma}^2 = \frac{\Sigma_{i=1}^{n}(X_i - \hat{\theta})^2}{n-1}$$

The p-values for the one-sided and two-sided hypotheses are: $p\text{-value}_1 = 2P\left[t_{n-1} > T_o\right]$ and $p\text{-value}_2 = 2P\left[t_{n-1} > |T_o|\right]$.
To investigate the equivalence relationship of the p-value with the posterior probability of the null under uninformative priors, we consider Jeffreys's prior and the conjugate Normal-Gamma prior for the mean θ and the precision $\tau = \frac{1}{2\sigma^2}$
.

I. Jeffreys' Prior
The Jeffreys's prior for the likelihood parameters is $\pi(\theta, \sigma^2) \propto \frac{1}{(\sigma^2)^2}$,which is proportional to the unknown σ^2. It can be shown that the marginal posterior of θ integrating over uncertainty in σ^2 is $\theta|X \sim t_{(n)}\left[\hat{\theta}, \hat{\sigma}^2/n\right]$, i.e. a t distribution with n DoF, location $\hat{\theta}$ and variance $\hat{\sigma}^2/n$. The one-sided and two-sided posterior probabilities of the null are:

- $\text{PoP}_1 = P\left[H_o : \theta < 0|X\right]$
- $\text{PoP}_2 = P\left[H_o : \theta \neq 0|X\right] = 2\left[1 - max\left\{p(\theta|X < 0), p(\theta|Y > 0)\right\}\right]$

II. Normal-Gamma Conjugate Prior
Under the Normal-Gamma prior for the mean-precision or equivalently the Normal-Inverse-Gamma prior for the mean-variance that the authors consider, we have $(\theta, \tau) \sim N\text{-}G(\theta_0, \nu_0, a, b)$. Due to the conjugate relationship, the posterior is also a Normal-Gamma model:

$$(\theta, \tau)|X \sim N\text{-}G(M, C, A, B)$$

with posterior hyperparameters:

$$M = \frac{\nu_0\theta_0 + \Sigma_{i=1}^n X_i}{\nu_0 + n}\ , C = \nu_0 + n\ , A = a + \frac{n}{2}\ , B = b + \frac{1}{2}\Sigma_{i=1}^n (X_i - \hat{\theta})^2 + \frac{1}{2}\frac{\nu_0 n}{\nu_0 + n}(\hat{\theta} - \theta_0)^2$$

Integrating over uncertainty in τ/σ^2, it can be shown that the marginal posterior of θ is:

$$\theta|X \sim t_{(2A)}\left[M, \frac{B}{AC}\right]$$

and the posterior probabilities of the null are given as before.

Numerical Simulations

To investigate numerically the equivalence of the p-value with PoP_1 and PoP_2 under uninformative priors, we generate 1,000 data points $X_i = Y_{E_i} - Y_{S_i} \sim N(\theta, \nu)$. To acquire p-values that will span the entire (0,1) range, θ and $\nu = 2\sigma^2$ are in turn sampled as $\theta \sim N(0, 0.05)$, $\nu \sim N(1, 0.05)$, while ν is forced to assume positive values. The prior hyperparameter ν_0 controls the prior variance since the marginal prior of θ is $\theta \sim t_{(2a)}\left[\theta_0, \frac{b}{a\nu_0}\right]$. The greater the ν_0 is (or the smaller for the Normal-Inverse-Gamma version that the authors use, since it is the inverse relationship) the more informative the prior will be. Thus, to render the Normal-Gamma prior model uninformative, we set the prior hyperparameters to $\nu_0 = 0.01$ to reflect a large prior variance and $a, b = 0.01$ to reflect the absence of prior information about the precision τ. As a result of the numerical simulations, Figure 1 displays the equivalence relationship between the p-value and the posterior probability of the null under Jeffreys' and uninformative Normal-Gamma priors. To examine the sensitivity of this equivalence to the data generation mechanism, we also generate X_i from Gamma and Beta likelihoods as $X_i \sim Gamma(2, 0.5)$ and $X_i \sim Beta(0.5, 0.5)$. In this case, the mean of the respective distributions is deducted from the data to allow p-value $\in [0, 1]$. Figure 2 shows that the equivalence relationship between the p-value and the posterior probability of the null under Jeffreys' prior (for the one-sided hypothesis test) is invariant to the likelihood of the data. Lastly, we conduct a prior sensitivity analysis to investigate how informative priors and the sample size affect the results. The left panel of Figure 3 shows the relationship between the p-value and the posterior probability of the null under an informative Normal-Gamma prior with very small prior variance $\nu_0 = 1,000$ and $\theta_0 = \theta + 0.01$. Under such an informative prior, the equivalence relationship is lost, while it is gradually restored for increasing sample size. This was to be expected since for increasing sample size the likelihood dominates the prior. The right panel of Figure 3 shows that the same equivalence relationship is lost for informative priors with very small variance, and it is gradually retrieved as the prior variance increases. This result was anticipated as well since the p-value is derived exclusively from the likelihood, while the posterior probability of the null borrows information from both the prior and the likelihood. Thus, for priors that do not contribute prior knowledge, it is expected that the p-value will match the posterior probability of the null.

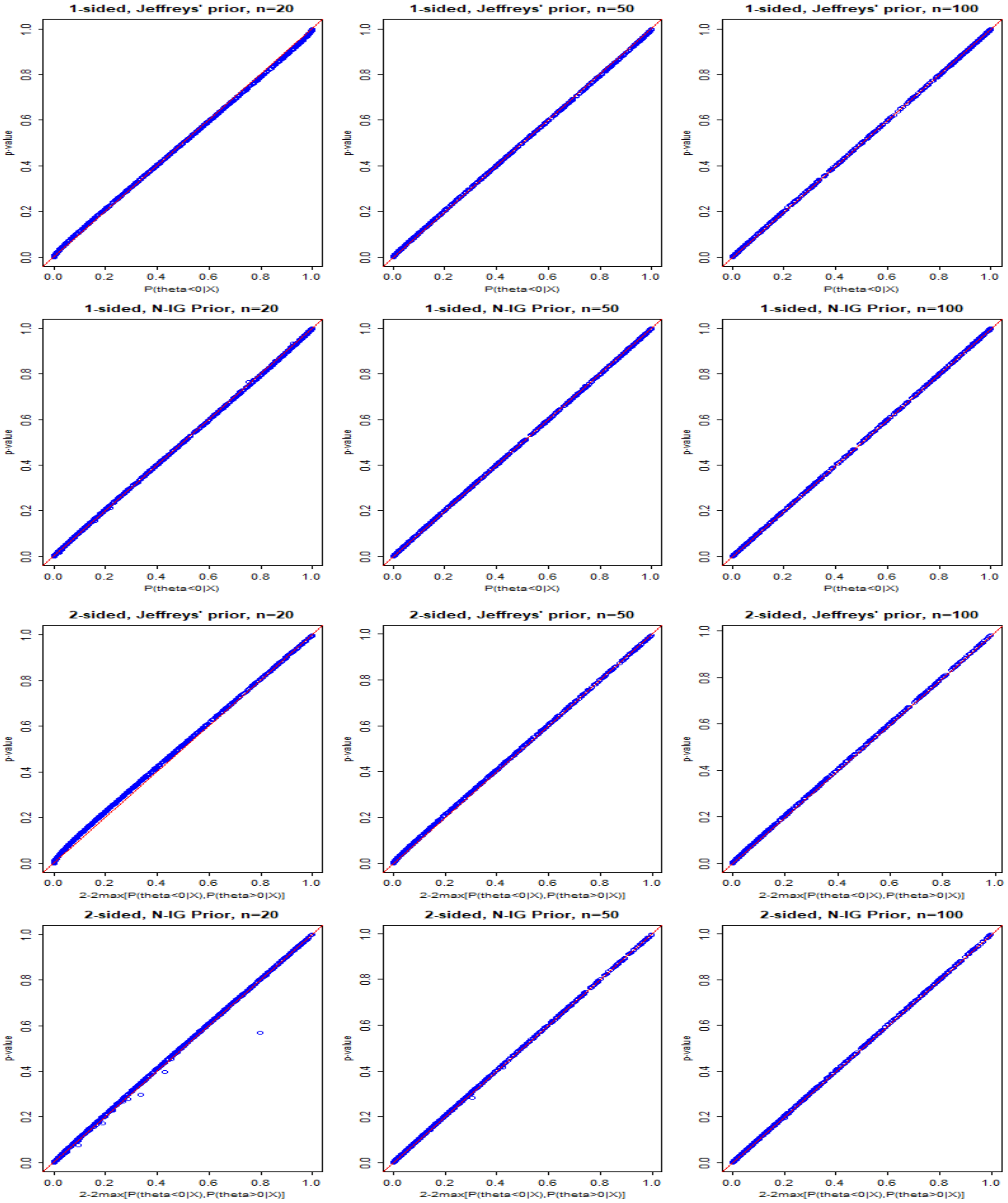

Figure 4: The relationship between the p-value and the posterior probability over 1,000 simulations for 1-sided and 2-sided hypothesis tests.

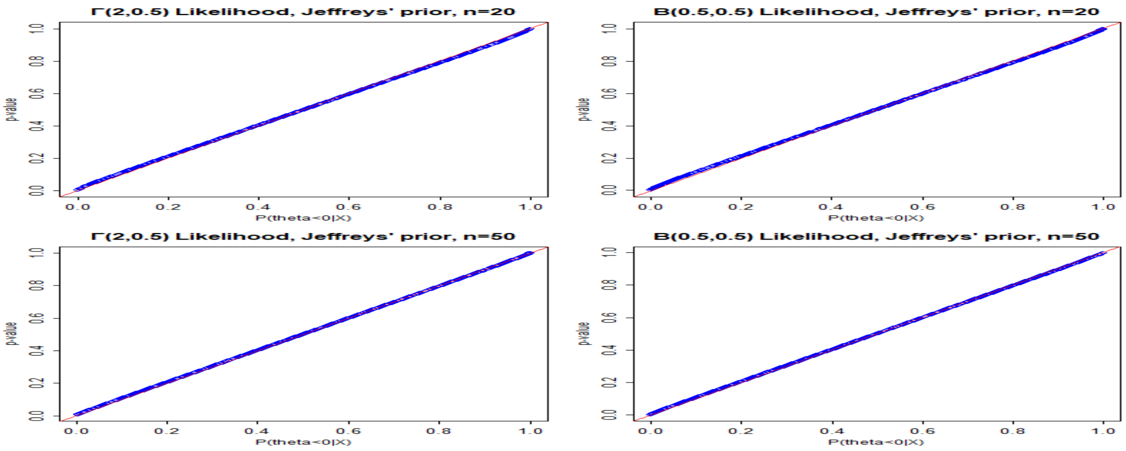

Figure 5: The relationship between the p-value and the posterior probability of the null over 1,000 simulations for 1-sided hypothesis tests with Beta and Gamma likelihoods for the outcomes.

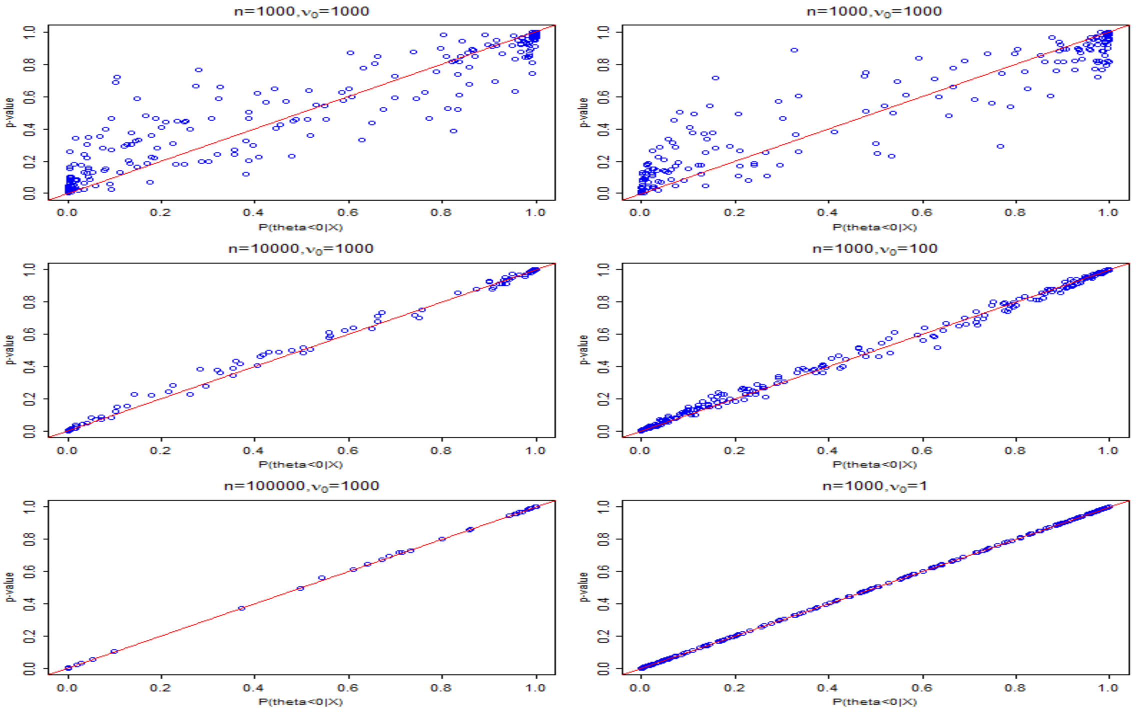

Figure 6: The relationship between the p-value and the posterior probability of the null over 1,000 simulations for 1-sided hypothesis tests with normal outcomes. Left panel: Informative normal-gamma prior under increasing sample sizes of 1,000, 10,000 and 100,000 (top to bottom). Right panel: Fixed sample size of 1,000 with increasing prior variance (top to bottom).

4 Conclusion

In conclusion, our study has provided valuable insights into the relationship between p-values and posterior probabilities in hypothesis testing for binary and normal data. We have shown that the use of Jeffreys' non-informative prior and an increasing sample size leads to a strong equivalence between these two statistics. Furthermore, we have explored the impact of an informative prior and demonstrated the importance of using a non-informative prior for fair comparison when the variance of the samples is high. As a future direction, we plan to consider the equivalence of posterior probability and p-value with various probability distributions. Overall, our study contributes to a better understanding and interpretation of these statistical measures in hypothesis testing..