Hypothesis testing of Poisson rates in COVID-19 offspring distributions

Luo, Rui

Published in:
Infectious Disease Modelling

Published: 01/12/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY-NC-ND

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/j.idm.2023.07.010

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 14/12/2023
Hypothesis testing of Poisson rates in COVID-19 offspring distributions

Rui Luo
Department of Systems Engineering, City University of Hong Kong, Kowloon Town, Hong Kong Special Administrative Region

Article info

Article history:
Received 1 August 2022
Received in revised form 14 July 2023
Accepted 18 July 2023
Available online 9 August 2023
Handling Editor: Dr Daihai He

Keywords:
Poisson distribution
Hypothesis testing
Bayes factor
Posterior predictive
p-value
COVID-19

ABSTRACT

In the present study, we undertake the task of hypothesis testing in the context of Poisson-distributed data. The primary objective of our investigation is to ascertain whether two distinct sets of discrete data share the same Poisson rate. We delve into a comprehensive review and comparative analysis of various frequentist and Bayesian methodologies specifically designed to address this problem. Among these are the conditional test, the likelihood ratio test, and the Bayes factor. Additionally, we employ the posterior predictive p-value in our analysis, coupled with its corresponding calibration procedures. As the culmination of our investigation, we apply these diverse methodologies to test both simulated datasets and real-world data. The latter consists of the offspring distributions linked to COVID-19 cases in two disparate geographies - Hong Kong and Rwanda. This allows us to provide a practical demonstration of the methodologies’ applications and their potential implications in the field of epidemiology.

© 2023 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The ubiquity of Poisson-distributed data is attributed to the strong interpretability and straightforward parameterization of the Poisson distribution. This distribution serves as a key tool for describing counting data across various contexts, including clinical trials, traffic patterns, and human communication dynamics, such as phone calls, emails, and social network connections.

The question of whether two data sets have originated from the same Poisson distribution has attracted significant interest within diverse research fields. To illustrate, a couple of guiding examples include: (1) Assessing disparities between placebo and treatment groups in clinical trials 1. (2) Evaluating whether the number of cabbage loopers is lower in fields treated with certain pesticides 2.

This paper is dedicated to the task of Poisson hypothesis testing. We aim to analyze the merits of a wide range of statistical testing approaches, both from frequentist and Bayesian perspectives. The structure of the paper is outlined as per the following segments, each addressing a key aspect of our investigation:

- Section 2: Formulation of the hypothesis testing problem for Poisson-distributed data.
- Section 3: Delving into frequentist techniques.
Section 4: Examination of the Bayes factor and the Bayesian information criterion.
Section 5: Discussion on the use of the posterior predictive \( p \)-value in the testing problem.
Section 6: A comprehensive numerical evaluation of the various approaches, showcasing their consistency in results.

In essence, this paper is driven by the aforementioned questions, and we aspire to provide a thorough examination of the methodologies involved in Poisson hypothesis testing.

2. Hypothesis testing problem formulation

In this section, we formulate the hypothesis testing problem for Poisson distribution, which is used throughout the paper. Denote \( X \) as the counting of events in unit time. Given that \( X \) follows a Poisson distribution \( \text{Poisson}(\lambda) \) with rate \( \lambda \), the probability mass function of \( X \) is given by

\[
P(X = x | \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda > 0, \quad x = 0, 1, 2, \ldots
\]  

Let \( X_1, \ldots, X_n \) and \( Y_1, \ldots, Y_m \) be independent samples, respectively, from Poisson \( \lambda_1 \) and Poisson \( \lambda_2 \). We define

\[
X = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n_1 \lambda_1),
\]  

Similarly,

\[
Y = \sum_{i=1}^{m} Y_i \sim \text{Poisson}(n_2 \lambda_2),
\]  

Let \( k_1 \) and \( k_2 \) denote the observed values of \( X \) and \( Y \), respectively. The problem of interest here is to test

\[
H_0 : \lambda_1 = \lambda_2 \quad \text{vs.} \quad H_1 : \lambda_1 \neq \lambda_2
\]  

Or equivalently,

\[
H_0 : \frac{\lambda_1}{\lambda_2} = 1 \quad \text{vs.} \quad H_1 : \frac{\lambda_1}{\lambda_2} \neq 1
\]  

3. Frequentist methods

In this section, we go through several well-known frequentist approaches which deal with the comparison of two Poisson rates (4), (5), including C-test, E-test, likelihood ratio test, and variance stabilizing transformation test. These tests are based on properties of Poisson distribution, such as.

(1) Conditional distribution of one Poisson random variable given by the sum with another independent Poisson random variable is binomial;
(2) The mean and variance of a Poisson distribution are both equal to the Poisson rate.

These frequentist tests perform consistently, as shown in Section 6. In addition, they are computationally efficient.

3.1. The conditional test

The conditional test (C-test) due to Przyborowski and Wilenski (1940) is based on the conditional distribution of \( X \) given \( X + Y = k \). Note that the conditional distribution of \( X \) given \( X + Y = k \) is binomial with the number of trials \( k \) and success probability

\[
p(\lambda_1 / \lambda_2) = \frac{(n_1/n_2)(\lambda_1/\lambda_2)}{1 + (n_1/n_2)(\lambda_1/\lambda_2)}
\]  

This C-test rejects \( H_0 \) in (5) whenever the \( p \)-value \( p^C \) satisfies the following condition:

\[
p^C = 2 \times \min\{P(X_1 \geq k_1 | k, p(1)), P(X_1 \leq k_1 | k, p(1))\} \leq \alpha
\]
where
\[
P(X_1 \geq k_1|k, p(1)) = \sum_{i=k_1}^{k} \binom{k}{i} p(1)^i (1 - p(1))^{k-i}
\]  
(8)
and \( p(1) \) is the expression in (6) with \( \lambda_1/\lambda_2 \) replaced by 1.

**Remark 1: Mid-\( p \) correction of the C-test (Gu et al., 2008).** It is well known that the C-test is an exact method for which the actual level of significance is always below the nominal level (i.e., conservative). We can use the mid-\( p \) adjusted version test suggested by Lancaster (1952, 1961) to overcome the conservativeness of the C-test. The \( p \)-value for the mid-\( p \) correction is given by
\[
p^M = \frac{p^C + p^{C^*}}{2}
\]  
(9)
where
\[
p^{C^*} = 2 \times \min\{P(X_1 \geq k_1 + 1|k, p(1)), P(X_1 \leq k_1 - 1|k, p(1))\}.
\]

3.2. Test based on estimated \( p \)-value

The test in this category due to (Krishnamoorthy & Thomson, 2004) is based on the standardized difference between \( X/n_1 \) and \( Y/n_2 \). The variance of the absolute difference \( |X/n_1 - Y/n_2| \) is given by
\[
\text{Var}(|X/n_1 - Y/n_2|) = \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}
\]  
(10)
We consider absolute value since the hypothesis (4) is two-sided. Note that \( X/n_i \) is unbiased estimate of \( \lambda_i \), \( i = 1, 2 \). This leads to an unbiased variance estimate given by
\[
\hat{V}_{XY} = \frac{X/n_1}{n_1} + \frac{Y/n_2}{n_2}
\]  
(11)
Using this unbiased variance estimate, we consider the standardized difference
\[
T_{XY} = \frac{|X/n_1 - Y/n_2|}{\sqrt{\hat{V}_{k_1,k_2}}}
\]  
(12)
as a pivot statistic for our testing problem. For a given \( (n_1, k_1, n_2, k_2) \), the observed value of the pivot statistic \( T_{XY} \) is given by
\[
T_{XY} = \frac{|X/n_1 - Y/n_2|}{\sqrt{\hat{V}_{k_1,k_2}}},
\]  
(13)
where \( \hat{V}_{k_1,k_2} \) is defined similarly as \( \hat{V}_{XY} \) in (11) with \( X \) replaced by \( k_1 \) and \( Y \) replaced by \( k_2 \). \( T_{XY} \) aids in determining how the two Poisson distributions differed from one another based on the observed \( X \) and \( Y \).

The \( p \)-value for testing (4) is
\[
P(T_{XY} \geq T_{k_1,k_2} | H_0),
\]  
(14)
which involves the unknown parameter \( \lambda_1 = \lambda_2 = \lambda \). We can estimate the \( p \)-value by replacing \( \lambda \) with its estimate \( \hat{\lambda} \)
\[
\hat{\lambda} = \frac{k_1 + k_2}{n_1 + n_2}
\]  
(15)
Using this \( \hat{\lambda} \), we estimate the corresponding \( p^E = P(T_{XY} \geq T_{k_1,k_2}) \) by
\[
\sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} \frac{e^{-n_1 \hat{\lambda}(n_1 \hat{\lambda})^{x_1}}}{x_1!} \frac{e^{-n_2 \hat{\lambda}(n_2 \hat{\lambda})^{x_2}}}{x_2!} I_{[T_{X_1,X_2} \geq T_{k_1,k_2}]}
\]  
(16)
where \( I_{[\cdot]} \) denotes the indicator function.
3.3. Likelihood ratio test

A likelihood-ratio test (LRT) uses a test statistic that compares the maximum value of the likelihood function under the null hypothesis constraint to the highest likelihood when the restriction is relaxed. Here, the likelihood function is

\[ L(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{\hat{k}_1^{k_1} \hat{\lambda}_1^{k_1} \hat{k}_2^{k_2} \hat{\lambda}_2^{k_2}}{k_1! k_2!} \]

(17)

The maximum likelihood under the parameter space subject to the "starshaped" restriction, i.e., \( \Omega^* = \{\hat{\lambda}_1, \hat{\lambda}_2|\hat{\lambda}_1 > \frac{\hat{\lambda}_2}{\hat{\lambda}_2} > 0\} \) is given by Dykstra and Robertson (1982). It is argued (Bartholomew, 1959) that such "ordered" alternative parameter space gives rise to more powerful tests.

The maximum likelihood under the whole parameter space \( \Omega = \{\hat{\lambda}_1, \hat{\lambda}_2|\hat{\lambda}_1 > 0, \hat{\lambda}_2 > 0\} \) is \( L(\hat{\lambda}_1, \hat{\lambda}_2) \), while the maximum likelihood under the \( H_0 \) restricted parameter space \( \Omega = \{\hat{\lambda}_1, \hat{\lambda}_2|\hat{\lambda}_1 = \hat{\lambda}_2\} \) is \( L(\hat{\lambda}) \).

By maximizing the log likelihood respectively we can get \( \hat{\lambda}_1 = \frac{n_1}{n_1 + n_2} \hat{\lambda}_2 = \frac{n_2}{n_1 + n_2} \hat{\lambda}_1 \). The likelihood ratio is

\[ \Lambda = \frac{L(\hat{\lambda})}{L(\hat{\lambda}_1, \hat{\lambda}_2)} = \frac{(k_1 + k_2)^{k_1+k_2}}{k_1^{k_1} k_2^{k_2}} e^{\frac{k_1}{n_1} \hat{\lambda}_1 + \frac{k_2}{n_2} \hat{\lambda}_2 + \frac{k_1 k_2}{n_1 n_2} \hat{\lambda}_1 \hat{\lambda}_2} \]  

(18)

where \( f(k,n,p) = \left( \begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \) is the probability mass function of the Binomial distribution.

An equivalent LRT statistic is \(-2 \ln \Lambda\), which has a particularly handy asymptotic distribution. If \( H_0 \) is true, then asymptotically \(-2 \ln \Lambda\) will be \( \chi^2 \) distributed with one degree of freedom.

\textbf{Remark 2: Wald test} (Davidson MacKinnon et al., 1993). In testing \( H_0: \lambda_1 = \lambda_2 \) and \( H_0: \lambda_1 \neq \lambda_2 \), another option is using the Wald test. The Wald statistic in this case is defined as

\[ Z = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\text{Var}(\hat{\lambda}_1) + \text{Var}(\hat{\lambda}_2)}} \]

(19)

\( Z \) has an asymptotic standard normal distribution. \( \lambda_i, i = 1, 2 \) can be estimated using the sample mean. As to \( \text{Var}(\hat{\lambda}_i) \), \( i = 1, 2 \), it is argued that the sample mean again is a better estimator because sample mean in a Poisson distribution is the uniformly minimum-variance unbiased estimator (UMVUE) (Rao, 1971) for the parameter. By using the sample mean instead of the sample variance, we achieve better precision.

Therefore, the Wald statistics has the form

\[ Z = \frac{\hat{\lambda}_1 - \hat{\lambda}_2}{\sqrt{\frac{n_1}{n_1} + \frac{n_2}{n_2}}} \]

(20)

3.4. Test based on variance stabilizing transformation

Anscombe (1948) derived the second order variance stabilizing transformation for a Poisson variable. If \( X \sim \text{Poisson}(\lambda) \), Anscombe showed that

\[ \text{Var}_\lambda \left( \sqrt{X + \frac{3}{8}} \right) = \frac{1}{4} + O \left( \frac{1}{\lambda} \right) \]

(21)

On this basis, we can define \( Y_i = \sqrt{X_i + \frac{3}{8}} \) and use the statistic

\[ T = 4 \sum (Y_i - \bar{Y})^2 \]

(22)

to provide a test for \( H_0 \).

From (21) we see that \( Y_i \) is approximately normal with variance \( \frac{1}{4} \) and mean \( E_i \left( \sqrt{X_i + \frac{3}{8}} \right) \). It would follow that when \( H_0 \) is true, \( T \) has approximately a \( \chi^2 \) distribution with \( n - 1 \) degree of freedom. We thus reject \( H_0 \) if \( T > \chi^2_{n-1,1-a} \).
4. Bayesian methods

Bayesian approaches are frequently employed in clinical studies to minimize sample sizes or boost power. The selection of the prior distribution is an important step in both Bayesian modeling and hypothesis testing. In this part, we will look at the Bayes factor and its several essential expansions (improvements). When there is no or little prior information available, enhanced Bayes factors avoid the problem by employing a subset of the observations as samples, the posterior of which is utilized as the prior for the remaining data.

Meng et al. (1994) and Gelman et al. (1996) are important for another prominent approach in Bayesian hypothesis testing, which focuses on posterior predictive model assessment (posterior predictive p-value). Essentially, this method examines the disparity between the model and the observed data to gather evidence for rejecting the (null) model. There is some disagreement over whether the posterior predictive p-value simply serves as a qualitative metric in hypothesis testing. However, the posterior predictive p-value has shown quantifiable efficacy with many calibration techniques presented by Hjort et al. (2006) and van Kollenburg et al. (2017). We will go over it in the next section, and concentrate on Bayes factors in this section.

4.1. Bayes factor

The Bayes factor is defined as the ratio of the posterior probability of \( H_1 \) to that of \( H_0 \), with a large Bayes factor indicating that the null hypothesis should be rejected. Specifically, the Bayes factor is defined as follows:

\[
B = \frac{P(X, Y|H_1)}{P(X, Y|H_0)}
\]  

(23)

Both \( H_0 \) and \( H_1 \) assign a prior to the corresponding model parameters, which are \( \pi_1(\lambda) \) and \( \pi_2(\lambda_1, \lambda_2) \) respectively. Marginalizing the likelihood over the prior, we have

\[
P(X, Y|H_0) = \int P(X, Y|\lambda)\pi_1(\lambda)d\lambda.
\]  

(24)

and

\[
P(X, Y|H_1) = \iint P(X|\lambda_1)P(Y|\lambda_2)\pi_2(\lambda_1, \lambda_2)d\lambda_1d\lambda_2.
\]  

(25)

The prior distribution on \( \lambda, \lambda_1, \) and \( \lambda_2 \) summarize expert opinions concerning the parameters in each of the two scenarios \( H_0 \) and \( H_1 \). This information can be collected in a variety of methods, including using previous data (Hand et al., 2011), prior elicitation of expert opinion (see notably (Garthwaite et al., 2005)), or relying on uninformative criteria (Ly et al., 2017). Following the derivation in (Sides et al., 2015), we examine conjugate priors for all three \( \lambda \)'s, such that under \( H_0, \lambda \sim \text{Gamma}(\alpha, \beta) \), and under \( H_1, \lambda \sim \text{Gamma}(a_i, b_i) \) independently. This assumption is not limiting and allows us to specify parameters of prior distributions rather than distributions themselves.

We derive the marginal likelihood of \( H_0 \) and \( H_1 \) using conjugate priors as follows

\[
P(x, y|H_0) = \int \left[ \prod_{i=1}^{n_1} \frac{x_i^{\lambda_i}e^{-\lambda_i}}{x_i!} \right] \left[ \prod_{i=1}^{n_2} \frac{y_i^{\lambda_i}e^{-\lambda_i}}{y_i!} \right] \frac{1}{\Gamma(\alpha)^{a_1}} e^{-\beta} d\lambda
\]  

\[
= \frac{\Gamma\left( \sum_{i=1}^{n_1} x_i + \alpha \right) \beta^{\alpha}}{\Gamma(\alpha)(n_1 + \alpha + \beta)^{\alpha}} \left[ \prod_{i=1}^{n_1} x_i \right] \left[ \prod_{i=1}^{n_2} y_i \right]
\]  

(26)

and
the portion of the data to be utilized in this way, and where $m_1$ under the two hypotheses in (5) is $\frac{m_1}{m_2}$. Therefore, the Bayes factor is

$$B = \frac{P(x,y|H_1)}{P(x,y|H_0)} = \frac{\Gamma(\sum_{i=1}^{n_1} x_i + \alpha_1) \Gamma\left(\frac{\sum_{i=1}^{n_2} y_i + \alpha_2}{\sum_{i=1}^{n_2} y_i + \alpha_2}\right)}{\Gamma(\alpha)(n_1 + \beta_1) \Gamma\left(\frac{\sum_{i=1}^{n_1} x_i + \alpha_1}{\sum_{i=1}^{n_1} x_i + \alpha_1}\right)}$$

Remark 3: Preference for using the likelihood odds ratio over the posterior odds for the Bayes factor. The literature often interprets the Bayes factor using the posterior odds ratio (Lavine & Schervish, 1999); however, one should note that this interpretation requires precise prior knowledge about the validity of the null or alternative hypothesis. In contrast, the likelihood odds ratio interpretation is commonly used and is considered safer in the absence of such precise prior information. This circumstance is frequently encountered when dealing with infection data from a novel type of virus where all candidate models may be deemed equally likely.

As a result, the likelihood odds ratio provides an acceptable method in the absence of such information. However, if the ratio between the two hypotheses is known, one can readily adjust the likelihood odds to the posterior odds. This adjustment can be mathematically represented as:

$$B' = \frac{P(H_1|x, y)}{P(H_0|x, y)} = \frac{P(x, y|H_1)\pi(H_1)}{P(x, y|H_0)\pi(H_0)} = B \frac{\pi(H_1)}{\pi(H_0)}$$

where $B$ and $B'$ denote the likelihood odds and the posterior odds, respectively.

Furthermore, we can utilize more informative priors, such as a non-uniform prior that represents prior knowledge of the constants. Two examples are demonstrated in Berger’s talk (Berger, 2019). Berger (Berger & Pericchi, 1996) and O’Hagan (O’Hagan, 1995) respectively proposed the intrinsic Bayes factor and the fractional Bayes factor.

The intrinsic Bayes factor solves the problem by using part of the data as a training sample. Suppose the prior of $\theta = \frac{x_i}{x_j}$ under the two hypotheses in (5) is $\pi_1(\theta_i)$, and the models each have probability density function $f_i(x|\theta_i)$, $i = 1, 2$. Let $x(l)$ signify the portion of the data to be utilized in this way, and $x(-l)$ denote the rest of the data, such that

$$0 < m_i(x(l)) < \infty, i \in \{1, 2\},$$

where $m_i(x(l))$ is the marginal density of $x(l)$ under hypothesis $H_{i-1}$, i.e.,

$$m_i(x(l)) = \int f_i(x(l)|\theta_i)\pi_i(\theta_i)d\theta_i.$$
The posterior \( \pi_i (\theta | x(l)) \) is obtained by applying Bayes’ theorem,

\[
\pi_i (\theta | x(l)) = \frac{f_i (x(l) | \theta_i) \pi_i (\theta_i)}{m_i (x(l))} \tag{32}
\]

Now, consider the Bayes factor, \( B_{10}(l) \), for the rest of the data \( x (-l) \), using \( \pi_i (\theta | x(l)) \) as the priors:

\[
B_{10}(l) = \frac{\int f_2 (x(-l) | \theta_2, f_2 (x(l)) \pi_2 (\theta_2 | x(l)) d\theta_2}{\int f_1 (x(-l) | \theta_1, f_1 (x(l)) \pi_1 (\theta_1 | x(l)) d\theta_1} = \frac{\int f_2 (x(-l) | \theta_2, f_2 (x(l)) f_2 (x(l)) \theta_2 \pi_2 (\theta_2) d\theta_2}{\int f_1 (x(-l) | \theta_1, f_1 (x(l)) \pi_1 (\theta_1) d\theta_1} \times \frac{m_1 (x(l))}{m_2 (x(l))} = B_{10} \cdot B_{01}(x(l)) \tag{33}
\]

where \( B_{10} = \frac{m_1 (x)}{m_2 (x)} \) and \( B_{01}(x(l)) = \frac{m_1 (x(l))}{m_2 (x(l))} \) represent the Bayes factors produced from the complete data \( x \) and the training samples \( x(l) \), respectively.

Note that the Bayes factor (33) depends on the specific training sample \( x(l) \). To avoid the difficulty of choosing \( x(l) \), Berger and Pericchi (1996) proposed the use of a minimal training sample to compute the Bayes factor. Then, an average over all the possible minimal training samples contained in the sample is computed. This gives the arithmetic intrinsic Bayes factor (AIBF) as

\[
B_{10}^{AI} = B_{10} \cdot \frac{1}{L} \sum_{l=1}^{L} B_{01}(x(l)), \tag{34}
\]

where \( L \) is the number of minimal training samples \( x(l) \) contained in \( x \). Also Berger and Pericchi (1998) gives the median intrinsic Bayes factor (MIBF) as

\[
B_{10}^{MI} = B_{10} \cdot ME[B_{01}(x(l))], \tag{35}
\]

where \( ME[\cdot] \) denotes the median, which would be applied to all of the training samples using the Bayes factor. The MIBF is the most dependable and extensively used intrinsic Bayes factor (Kang et al., 2006).

4.3. Fractional Bayes factor

The fractional Bayes factor (Kang et al., 2006; O’Hagan, 1995) works in the same way as the intrinsic Bayes factor, but instead of using a portion of the data to convert noninformative priors to proper priors, it takes a fraction, \( b \), of each likelihood function, \( f_i (x | \theta_i), i = 1, 2, \) and utilizes the remaining \( 1 - b \) fraction of the likelihood used for model discrimination. Then the fractional Bayes factor of \( H_1 \) vs \( H_0 \) is then calculated

\[
B_{10}^f = B_{10} \cdot \frac{\int f_1^b (x | \theta_1) \pi_1 (\theta_1) d\theta_1}{\int f_2^b (x | \theta_2) \pi_2 (\theta_2) d\theta_2} = B_{10} \cdot \frac{m_1^b (x)}{m_2^b (x)} \tag{36}
\]

where \( m_i^b (x)/m_i (x) \) denotes the correction term, which is a ratio of marginal densities computed from a fraction \( b \) of the data. We select \( b = m/n \), where \( m \) is the size of the minimal training sample size, since Berger and Pericchi (1998) and Kass and Wasserman (1995) both make compelling arguments in its favor.

**Remark 4:** **Comparison of the aforementioned Bayes factors.** The AIBF presents two problems. First, AIBF shows a high sensitivity to small perturbations in the prior hyperparameter. Second, it is relevant to the conclusions what model’s marginal is in the numerator of AIBF, i.e., \( B_{10}^M(y) \) can be different from \( 1/B_{01}^M(y) \). Berger and Pericchi (1996), using the same data and priors, show that the problems presented by the AIBF are not shown by the MIBF. The first main gain in using MIBF is the robustness with respect to small variations of the hyperparameters of the priors. The second advantage is the coherence of MIBF, since \( B_{10}^M(y) = 1/B_{01}^M(y) \). Furthermore, by definition, FBF is not sensitive at all to the number of training samples.

4.4. Implementation of the one-sided test

Following the derivation in Kang et al. (2006), we first consider the one-sided test (the original test will be considered in Remark 2):

\[
H_0 : \eta \leq \eta_0 \text{ vs. } H_0 : \eta > \eta_0 \tag{37}
\]

where \( \eta = \frac{\lambda_1}{\lambda_2} \) is the ratio of two Poisson means. The reference prior for \( \eta \) is as follows (Kang et al., 2006):
\[ \pi(\lambda, \eta) = \eta^{-2} \lambda^{-1} (n_1 + n_2 \eta)^{-2}. \]

Under \( H_0 \), the reference prior for \( \lambda \) and \( \eta \) is

\[ \pi_1(\lambda, \eta) = \eta^{-2} \lambda^{-1} (n_1 + n_2 \eta)^{-2} I[\eta \leq \eta_0], \]

and under \( H_1 \), the reference prior for \( \lambda \) and \( \eta \) is

\[ \pi_2(\lambda, \eta) = \eta^{-2} \lambda^{-1} (n_1 + n_2 \eta)^{-2} I[\eta > \eta_0], \]

where \( I[\cdot] \) is the indicator function.

### 4.4.1. Implementation of the intrinsic Bayes factor

The elements of \( B_{10} \) in intrinsic Bayes factor (33) are given by

\[
m_1(x, y) = \int_0^\infty \int_{\eta_0}^\infty f(x, y|\lambda, \eta) \pi_1(\lambda, \eta) d\lambda d\eta
\]

\[
= \frac{\Gamma(n_1 x + n_2 y + 0.5)}{\prod_{i=1}^{n_1} x_i \prod_{j=1}^{n_2} y_j!} \int_{\eta_0}^\infty \eta^{n_1 y - 0.5} (n_1 + n_2 \eta)^{-2} \eta^{n_2 \eta + 1} d\eta
\]

\[
= \frac{\Gamma(n_1 x + n_2 y + 0.5)}{\prod_{i=1}^{n_1} x_i \prod_{j=1}^{n_2} y_j!} S_1(x, y) \tag{38}
\]

and

\[
m_2(x, y) = \int_0^\infty \int_{\eta_0}^\infty f(x, y|\lambda, \eta) \pi_2(\lambda, \eta) d\lambda d\eta
\]

\[
= \frac{\Gamma(n_1 x + n_2 y + 0.5)}{\prod_{i=1}^{n_1} x_i \prod_{j=1}^{n_2} y_j!} \int_{\eta_0}^\infty \eta^{n_1 y - 0.5} (n_1 + n_2 \eta)^{-2} \eta^{n_2 \eta + 1} d\eta
\]

\[
= \frac{\Gamma(n_1 x + n_2 y + 0.5)}{\prod_{i=1}^{n_1} x_i \prod_{j=1}^{n_2} y_j!} S_2(x, y) \tag{39}
\]

where \( S_1(x, y) = \int_{\eta_0}^\infty \eta^{n_1 y - 0.5} (n_1 + n_2 \eta)^{-2} \eta^{n_2 \eta + 1} d\eta \) and \( S_2(x, y) = \int_{\eta_0}^\infty \eta^{n_1 y - 0.5} (n_1 + n_2 \eta)^{-2} \eta^{n_2 \eta + 1} d\eta \). Using minimal training sample, we only calculate the marginal densities under \( H_0 \) and \( H_1 \), respectively. The marginal density of \((x_i, y_j)\) is finite for all \( 1 \leq i \leq n_1, 1 \leq j \leq n_2 \) and hypotheses. Thus we conclude that any training sample of size two is a minimal training sample.

The marginal densities \( m_1(x, y) \) under \( H_0 \) is given by

\[
m_1(x, y) = \int_0^\infty \int_{\eta_0}^\infty f(x, y|\lambda, \eta) \pi_1(\lambda, \eta) d\lambda d\eta
\]

\[
= \frac{\Gamma(x_i + y_j + 0.5)}{x_i! y_j!} \int_{\eta_0}^\infty \eta^{x_i + y_j + 0.5} (n_1 + n_2 \eta)^{-0.5} d\eta
\]

\[
= \frac{\Gamma(x_i + y_j + 0.5)}{x_i! y_j!} T_1(x_i, y_j) \tag{40}
\]

and similarly,
\[
m_2(x, y) = \int_0^\infty \int_0^\infty f(x, y|\lambda, \eta)\pi_2(\lambda, \eta)d\lambda d\eta \\
= \frac{\Gamma(x + y + 0.5)}{x!y!} \int_0^\infty \eta^{y-0.5}(1 + \eta)^{-x+y+0.5}(n_1 + n_2\eta)^{-0.5} d\eta \\
= \frac{\Gamma(x + y + 0.5)}{x!y!} T_2(x, y)
\]

where \( T_1(x, y) = \int_0^n \eta^{y-0.5}(1 + \eta)^{-x+y+0.5}(n_1 + n_2\eta)^{-0.5} d\eta \) and \( T_2(x, y) = \int_0^\infty \eta^{y-0.5}(1 + \eta)^{-x+y+0.5}(n_1 + n_2\eta)^{-0.5} d\eta \). Therefore the AIBF of \( H_1 \) versus \( H_0 \) is given by

\[
B_{10}^{\text{AIBF}} = \frac{S_2(x, y)}{S_1(x, y)} \left[ \frac{1}{n_1n_2} \sum_{ij} T_1(x, y) \right].
\]

And the MIBF of \( H_1 \) versus \( H_0 \) is given by

\[
B_{10}^{\text{MIBF}} = \frac{S_2(x, y)}{S_1(x, y)} ME \left[ \frac{T_1(x, y)}{T_2(x, y)} \right].
\]

### 4.4.2. Implementation of the fractional Bayes factor

The element \( B_{10} \) of the fractional Bayes factor (36) is computed in the intrinsic Bayes factor in (33). Since the training sample for \( H_0 \) and \( H_1 \) consists of two sets of independent and identically distributed random variables, we formulate the correction term with the fractions \( 1/n_1 \) and \( 1/n_2 \) of each likelihood. Thus the element of correction term is given as follows

\[
m_1^b(x, y) = \int_0^\infty \int_0^\infty f_1(x|\lambda)f_2(y|\lambda, \eta)\pi_1(\lambda, \eta)d\lambda d\eta \\
= \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i! \prod_{i=1}^{n_2} y_i!} \int_0^\infty \eta^{y-0.5}(1 + \eta)^{-(x+y+0.5)}(n_1 + n_2\eta)^{-0.5} d\eta \\
= \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i! \prod_{i=1}^{n_2} y_i!} S_1(x, y, b),
\]

and

\[
m_2^b(x, y) = \int_0^\infty \int_0^\infty f_1(x|\lambda)f_2(y|\lambda, \eta)\pi_2(\lambda, \eta)d\lambda d\eta \\
= \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i! \prod_{i=1}^{n_2} y_i!} \times \int_0^\infty \eta^{y-0.5}(1 + \eta)^{-(x+y+0.5)}(n_1 + n_2\eta)^{-0.5} d\eta \\
= \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i! \prod_{i=1}^{n_2} y_i!} S_2(x, y, b).
\]

where \( S_1(x, y, b) = \int_0^n \eta^{y-0.5}(1 + \eta)^{-(x+y+0.5)}(n_1 + n_2\eta)^{-0.5} d\eta \) and \( S_2(x, y, b) = \int_0^\infty \eta^{y-0.5}(1 + \eta)^{-(x+y+0.5)}(n_1 + n_2\eta)^{-0.5} d\eta \). The label \( b \) indicates that \( b \) fraction of samples is utilized as the training data, and it is used to distinguish from \( S_1(x, y) \) and \( S_2(x, y) \) in (41), (42).

By using (47), (48) as well as (36), we have the fractional Bayes factor of \( H_1 \) versus \( H_0 \) is given by
The marginal densities

\[ p_{1} \approx \binom{n}{x, y} \frac{\Gamma(n + 1)}{\Gamma(x + 1) \Gamma(y + 1)} \]

and under \( R \), the reference prior for \( S \) is the number of samples.

\[ N_{\text{posterior}} = \binom{n}{x, y} \frac{\Gamma(n + 1)}{\Gamma(x + 1) \Gamma(y + 1)} \]

and under \( H_{1} \), the reference prior for \( \lambda \) and \( \eta \) is

\[ \pi_{2}(\lambda, \eta) = \eta^{-\lambda - \frac{1}{2}}(n_{1} + n_{2} \eta)^{-\frac{1}{2}}, \]

Then the elements of \( B_{10} \) in intrinsic Bayes factor are given by

\[ m_{1}(x, y) = \int f(x, y|\lambda) \pi_{1}(\lambda) d\lambda \]

\[ = \frac{\Gamma(n_{1} x + n_{2} y + 0.5)}{\prod_{i=1}^{n_{1}} x_{i} \prod_{i=1}^{n_{2}} y_{i}^{0.5}} \left( n_{1} + n_{2} \right)^{-\frac{1}{2}} \]

\[ \frac{\Gamma(n_{1} x + n_{2} y + 0.5)}{\prod_{i=1}^{n_{1}} x_{i} \prod_{i=1}^{n_{2}} y_{i}^{0.5}} \left( n_{1} + n_{2} \right)^{-\frac{1}{2}} \]

and

\[ m_{2}(x, y) = \int \int f(x, y|\lambda, \eta) \pi_{2}(\lambda, \eta) d\lambda d\eta \]

\[ = \frac{\Gamma(n_{1} x + n_{2} y + 0.5)}{\prod_{i=1}^{n_{1}} x_{i} \prod_{i=1}^{n_{2}} y_{i}^{0.5}} \left( n_{1} + n_{2} \right)^{-\frac{1}{2}} \]

\[ \frac{\Gamma(n_{1} x + n_{2} y + 0.5)}{\prod_{i=1}^{n_{1}} x_{i} \prod_{i=1}^{n_{2}} y_{i}^{0.5}} \left( n_{1} + n_{2} \right)^{-\frac{1}{2}} \]

where \( \tilde{S}_{1}(x, y) = (n_{1} + n_{2})^{-\frac{1}{2}} \) and \( \tilde{S}_{2}(x, y) = \int f(x, y) \eta^{0.5} - 0.5(n_{1} + n_{2} \eta)^{-\frac{1}{2}} \int f(x, y) \eta^{0.5} - 0.5(n_{1} + n_{2} \eta)^{-\frac{1}{2}} \)

The marginal densities \( m_{1}(x, y) \) under \( H_{0} \) is given by

\[ B_{10}^{F} = \frac{S_{2}(x, y) S_{1}(x, y, b)}{S_{1}(x, y) S_{2}(x, y, b)}. \]
\[ m_1(x_i, y_j) = \int_0^\infty f(x_i, y_j|\lambda) \pi_1(\lambda) d\lambda = \int_0^\infty \lambda^{x_i+y_j-1} e^{-\lambda} \frac{\Gamma(x_i+y_j+0.5)}{x_i!y_j!} \frac{\Gamma(\lambda)}{\lambda^\frac{\lambda}{2}} d\lambda, \]  
\[ m_2(x_i, y_j) = \int_0^\infty \int_0^\infty f(x_i, y_j|\lambda, \eta) \pi_2(\lambda, \eta) d\lambda d\eta = \Gamma(x_i+y_j+0.5) \frac{x_i!y_j!}{(n_1+n_2\eta)^{-0.5}} d\eta \]  
\[ = \frac{\Gamma(x_i+y_j+0.5)}{x_i!y_j!} \int_0^\infty \eta^{y_j-0.5} (1 + \eta)^{-x_i+y_j+0.5} (n_1+n_2\eta)^{-0.5} d\eta \]  
\[ = \frac{\Gamma(x_i+y_j+0.5)}{x_i!y_j!} T_2(x_i, y_j) \]  

where \( T_1(x_i, y_j) = \int_0^\infty \lambda^{x_i+y_j+0.5} \) and \( T_2(x_i, y_j) = \int_0^\infty \eta^{y_j-0.5} (1 + \eta)^{-x_i+y_j+0.5} (n_1+n_2\eta)^{-0.5} d\eta \). Therefore the AIBF and MIBF of \( H_1 \) versus \( H_0 \) can be obtained by combining into (45) and (46) respectively.

The correction terms (47), (48) in the fractional Bayes factor for the original test can be derived in a similar manner as follows

\[ m_3(x, y) = \int_0^\infty \frac{f(x|\lambda)^{\frac{\lambda}{2}}}{\prod_{i=1}^{n_1} x_i!} \frac{f(y|\lambda)^{\frac{\lambda}{2}}}{\prod_{j=1}^{n_2} y_j!} \lambda^{\frac{-\lambda}{2}} d\lambda \]  
\[ = \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i!} \frac{\Gamma(x+y+0.5)}{\prod_{j=1}^{n_2} y_j!} \left( \frac{1}{2} \right)^x \left( \frac{1}{2} \right)^y + 0.5 \]  
\[ = \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i!} \frac{\Gamma(x+y+0.5)}{\prod_{j=1}^{n_2} y_j!} S_1(x, y, b), \]  

(56)

**Table 2**

Simulated data in Section 6.2.1. \( X \) and \( Y \) are samples from Poisson (4) and Poisson (8) respectively.

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3, 3, 5, 8, 3, 2, 3, 4, 3</td>
<td>12, 15, 7, 4, 7, 3, 9, 6, 4, 10</td>
</tr>
</tbody>
</table>
and

\[
m_2^b(x,y) = \int_0^\infty \int_0^\infty f^{\text{H}}(x|\lambda)f^{\text{Y}}(y|\lambda,\eta)\pi_2(\lambda,\eta)d\lambda d\eta
\]

\[
= \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i!\prod_{i=1}^{n_2} y_i!} \int_0^{\infty} \eta^{y-0.5}(1+\eta)^{-\frac{x+y+0.5(n_1+n_2)}{2}} d\eta
\]

\[
= \frac{\Gamma(x+y+0.5)}{\prod_{i=1}^{n_1} x_i!\prod_{i=1}^{n_2} y_i!} S_2(x,y,b).
\]

\[\text{(57)}\]
where \( S_1(x, y, b) = \left( \frac{1}{2} \right)^x + y + 0.5 \) and \( S_2(x, y, b) = \int_0^\infty \eta^{0.5} (1 + \eta)^{-\frac{x}{\eta}} + y + 0.5 (n_1 + n_2 \eta)^{-0.5} d\eta \). Then the fractional Bayes factor of \( H_1 \) versus \( H_0 \) can be derived following (49).

4.5. Bayesian information criterion

Suppose we would like to determine which model best fits a given collection of data. These models are typically not all of the same dimension. Schwarz proposed the Bayesian information criterion (BIC) in (Schwarz, 1978, pp. 461–464) which provides a model selection criterion: models with lower BIC are preferred. It is Bayesian in that it uses an asymptotic result derived under the assumption that the data distribution is in the exponential family. And the Poisson distribution is a member of the exponential family.

The BIC is formally defined as

\[
\text{BIC} = -2 \ln(L) + k \ln(n)
\]

where \( \ln(L) \) is the maximized value of the likelihood function under the null hypothesis is

\[
L(\hat{\lambda}) = \frac{\hat{k}_1 \cdot \lambda^{k_1} e^{-\hat{\lambda}}}{(k_1 + k_2)!
\]

\[
(58)
\]

where \( \hat{\lambda} = \frac{k_1 + k_2}{n_1 + n_2} \) as is given in (15). The maximized likelihood under the alternative hypothesis, on the other hand, is

\[
L(\hat{\lambda}_1, \hat{\lambda}_2) = \frac{\hat{k}_1 e^{-\hat{\lambda}_1}}{k_1!} \frac{\hat{k}_2 e^{-\hat{\lambda}_2}}{k_2!}
\]

\[
(59)
\]

where \( \hat{\lambda}_1 = \frac{k_1}{n_1} \) and \( \hat{\lambda}_2 = \frac{k_2}{n_2} \).

The BIC is formally defined as
\[ \text{BIC} = k \ln(n) - 2 \ln(L), \]

\((60)\)
Table 5
This table compares different testing approaches on the COVID-19 offspring dataset. The columns correspond to those in Table 4. Due to the size of the dataset, ppp and calibrated ppp need a large number of samples to provide useful (i.e., non-zero) results. We use 10000 samples for computing the ppp and 10000 × 10000 samples for computing the calibrated ppp due to the computational limitation, and the resulting values are both 0.

<table>
<thead>
<tr>
<th>$p^C$</th>
<th>$p^D$</th>
<th>$p^{\text{lab}}$</th>
<th>ppp</th>
<th>ppp$^{\text{calib}}$</th>
<th>$B^{\text{IC}}_{10}$</th>
<th>$B^{\text{IC}}_{10}^\text{eff}$</th>
<th>$B^{\text{IC}}_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9366e-14</td>
<td>1.2866e-09</td>
<td>1.3767e-14</td>
<td>0.0000e+00</td>
<td>0.0000e+00</td>
<td>1.6359e+11</td>
<td>2.2924e+11</td>
<td>9.8001e+08</td>
</tr>
</tbody>
</table>

where $k$ is the number of parameters estimated by the model, $n$ is the sample size, and $\hat{l}$ is the model’s maximized likelihood. The BIC for the null hypothesis and the alternative hypothesis are

\[
\begin{align*}
\text{BIC}_0 & = \ln(n_1 + n_2) - 2\ln(L(\hat{\lambda})), \\
\text{BIC}_1 & = 2\ln(n_1 + n_2) - 2\ln(L(\hat{\lambda}_1, \hat{\lambda}_2)),
\end{align*}
\]

respectively. We can compute their difference as

\[
\text{BIC}_1 - \text{BIC}_0 = \ln(n_1 + n_2) - 2\ln\left(\frac{L(\hat{\lambda}_1, \hat{\lambda}_2)}{L(\hat{\lambda})}\right)
\]

Once more, it is clear that the BIC is comparable to a regularized likelihood ratio test that penalizes more complex models (such as the alternative hypothesis), and the penalization rises with sample size.

BIC has also been used in selecting inhomogeneous Poisson processes. Choiruddin et al. (2021) identify the point process as

\[
\text{BF}_1 \approx \text{BF}_2.
\]

Remark 6: Akaike information criterion (AIC) and Deviance information criterion (DIC). A model selection criterion closely related to BIC is AIC, which is defined as

\[
\text{AIC} = 2k - 2\ln(\hat{l}),
\]

where $k$ is the number of parameters and $\hat{l}$ is the maximum likelihood. Compared with (60), the AIC penalizes models less for parameters, and it does not depend directly on sample size (note the coefficient 2 versus $\ln(n)$).

The DIC generalizes the AIC for Bayesian model selections where the likelihood is evaluated at the posterior estimate. It utilizes deviance, which is a goodness-of-fit statistic defined as follows:

\[
D(\hat{\lambda}) = -2\ln(P(y|\hat{\lambda})) + C,
\]

where $\lambda$ is the model parameter. Since it cancels out when contrasting different models, $C$ is a constant that doesn’t need to be known. The DIC computes the effective number of model parameters as $p_D = D(\hat{\lambda}) - D(\bar{\lambda})$, where $\bar{\lambda}$ is the expectation of $\lambda$, e.g., the sample mean of draws from the posterior distribution. The DIC is calculated as

\[
\text{DIC} = 2p_D + D(\bar{\lambda}),
\]

which is similar to the expression of AIC, where $p_D$ penalizes complex models and $D(\bar{\lambda})$ penalizes models that fit poorly. Being simple to calculate from the samples produced by a Markov chain Monte Carlo simulation gives DIC an edge over AIC. Furthermore, Claeskens Hjort et al. (2008) show that the DIC is large-sample equivalent to the natural model-robust version of the AIC.

5. Posterior predictive $p$-value

One disadvantage of the aforementioned intrinsic Bayes factor and fractional Bayes factor is that, with large sample (either sample size is large or sample value is large, i.e., $\sum_{i=1}^{n} x_i \gg 0$), computing the integral of the correction term $S_2(x, y)$ and $S_2(x, y, b)$ in (53) and (57) will encounter numerical underflow, which limits their application dealing with large samples.

Posterior predictive $p$-value (abbreviated as “ppp”) is an appealing Bayesian tool which circumvents the troublesome integral by MCMC simulation. It is computed by calculating the fraction of predicted values that are more extreme for the test statistic than the observed value for that statistic. Posterior predictive $p$-value is essentially Bayesian analogues to "classical" $p$-value (Meng et al., 1994; Gelman et al., 1996; Gelman et al., 2013).
There have been criticisms regarding the comparatively low power of the posterior predictive \( p \)-value, i.e., it is often too conservative to reject a null hypothesis. Hjort et al. (2006) and van Kollenburg et al. (2017) proposed prior-calibrated \( p \)pp and posterior-calibrated \( p \)pp respectively to alleviate this suffering. We describe in detail the technique used to compute the posterior-calibrated \( p \)pp, and show that it outperforms the original \( p \)pp in numerical experiments (Section 6).

5.1. Brief overview of bayesian predictive density

Bayesian predictive density includes both prior predictive density and posterior predictive density. Here the "predictive" means predictive for observations. The prior predictive distribution is the distribution of (future) observation \( y \) averaged over all possible values of the parameter \( \theta \),

\[
p_x(y) = \int p(y|\theta)p(\theta)d\theta
\]  

and the posterior predictive distribution is conditional on the observed sample \( x \). More concretely, each likelihood in the parameter space is assigned a weight proportional to the posterior given \( x \),

\[
p_x(y|x) = \int p(y|\theta)p(\theta|x)d\theta
\]

Bayesian predictive densities have been widely used in many research areas, such as predicting future trends and behavior patterns, missing data analysis, and data compression and information theory. Related to the hypothesis testing problem studied in this report, they have also been used in model checking and model diagnostics (Pardoe, 2001; Gelman et al., 2013; Sinharay et al., 2006).

5.2. Posterior predictive assessment of model fitness

In this subsection, we explore model fitness analysis using posterior predictive methods. First, we describe the procedure of posterior predictive assessment using classical test statistics \( T(y) \). Then, we discuss one core concept in this field, i.e., the discrepancy variable \( D(y; \theta) \). What distinguishing \( D(y; \theta) \) from \( T(y) \) is that the discrepancy variable can as well depend on the model parameter \( \theta \), which is drawn from its posterior distribution in a Bayesian paradigm. Below is a quick overview of the posterior predictive \( p \)-value:

Overview of the posterior predictive \( p \)-value (\( p \)pp)

1) What question does it intend to answer?
   Does the model (under certain hypothesis) fit the data well?

2) What is the assumption?
   If the model fits the data well, future observation \( y_{\text{rep}} \) produced by the model should look similar to the observed data \( y \). In other words, the observed data \( y \) should look plausible under the posterior predictive distribution.

3) How does it work?
   3.1) Devise a discrepancy variable \( D(y; \theta) \) which accounts for how extreme an observation is.
   3.2) Sample \( J \) draws of \( \theta^{(i)} \) from posterior \( P(\theta|H, y) \).
   3.3) For each draw \( \theta^{(i)} \), sample a replicated data \( y_{\text{rep}}^{(i)} \) as a future observation, from the likelihood \( P(y_{\text{rep}}^{(i)}|H, \theta^{(i)}) \).
   3.4) Calculate \( p \)pp as the fraction of times that \( D(y_{\text{rep}}^{(i)}; \theta^{(i)}) \geq D(y; \theta^{(i)}) \).

4) How is it different from frequentist \( p \)-value? Frequentist \( p \)-value is defined for a test statistic \( T(y) \),

\[
p_{C} = P(T(y_{\text{rep}}^{(i)}) \geq T(y)|\theta) \quad (69)
\]

where \( \theta \) is taken as fixed, and \( p_{C} \) is a function of \( \theta \). On the contrary, the Bayesian equivalent, i.e., \( p \)pp is

\[
p_{p} = P(D(y_{\text{rep}}^{(i)}; \theta^{(i)}) \geq D(y; \theta^{(i)})) \quad (70)
\]

draws from the whole posterior distribution of \( \theta \) can be taken, and hence \( p_{p} \) is a function of both \( \theta \) and \( y \).

Please refer to the subsections below for detailed interpretations.

5.2.1. Posterior predictive assessment using classical test statistics

We use the notation \( y \) for the observed data, \( H \) for the assumed model, \( ^1 \theta \) for the unknown model parameter, and \( T \) denotes a test statistic, a function from data space to the real numbers.

To avoid confusion with the observed data, \( y \), we define \( y_{\text{rep}}^{(i)} \) as the replicated data that could have been observed with same model \( H \) and the same (unknown) parameter of \( \theta \) that produced \( y^{(i)} \). Denote the distribution of this replication by \( P=y_{\text{rep}}^{(i)}|H, \theta) \). In this notation, the classical \( p \)-value based on \( T \) is

\[ ^1 \text{The null hypothesis } H_0 \text{ denotes a model with one Poisson distribution, while the alternative hypothesis } H_1 \text{ denotes a model with two different Poisson distributions and generates two sets of observations from them respectively.}

\[ ^2 \text{Or, more accurately, as the data that would appear if the experiment that produced } y \text{ today was replicated tomorrow with the same model.} \]
\[ p_c(y, \theta) = P(T(y^{\text{rep}}) \geq T(y)|H, \theta). \]  

(71)

Hence a \( p \)-value near to 0 suggests that under the model specified by \( H \) and \( \theta \), the lack of fit in the direction of the test statistic, \( T(y) \), is unlikely. \( y \) is fixed and all randomness comes from \( y^{\text{rep}} \), as shown in (71).

In the Bayesian framework, the inference for \( \theta \) is provided by its posterior distribution, \( P(\theta|H, y) \). Correspondingly, the reference distribution of the future observation \( y^{\text{rep}} \), given \( y \), is its posterior predictive distribution. The posterior distribution of \( P(\theta|H, y) \) provides inference for \( \theta \) in the Bayesian framework. Similarly, the posterior predictive distribution of the future observation \( y^{\text{rep}} \), given \( y \), helps infer its distribution based on available information,

\[ P(y^{\text{rep}}|H, y) = \int P(y^{\text{rep}}|H, \theta)P(\theta|H, y)d\theta \]  

(72)

The observed value of \( T, T(y) \), is then displayed against the distribution of \( T(y^{\text{rep}}) \) induced by (72). The corresponding tail-area probability, or the "Bayes" \( p \)-value, analogous to (71), is

\[ p_{b}(y) = P(T(y^{\text{rep}}) \geq T(y)|H, y) = \int p_{c}(y, \theta)P(\theta|H, y)d\theta \]  

(73)

Essentially, the "Bayes" \( p \)-value is the classical \( p \)-value of (71) averaged over the posterior distribution of \( \theta \). This defines the posterior predictive \( p \)-value (also see Meng et al., 1994).

5.2.2. Posterior predictive assessment using discrepancies

The Bayesian formulation not only addresses the distracting parameters problem in classical test statistics, but it also allows for the usage of test "statistics" that rely on \( \theta \). This generalization beyond Rubin’s (1984) formulation in posterior predictive model checking (PPMC) is significant because it helps us to compare the disparity between the observed data and the postulated model directly, rather than between the data and the best fit of the model. It also makes it much easier to compute tail-area probabilities. Tsui and Weerahandi (1989) refers to such parameter-dependent test statistics as a "test variable". Meng et al. (1994) denotes it as a "discrepancy variable", which stresses that the goal here is to assess differences between a model and the data, rather than to verify whether a model is correct.

For a selected discrepancy, \( D(y; \theta) \), its reference distribution is derived from the joint posterior distribution of \( y^{\text{rep}} \) and \( \theta \),

\[ P(y^{\text{rep}}, \theta|H, y) = P(y^{\text{rep}}|H, \theta)P(\theta|H, y). \]  

(74)

Specifically, we can formally define a tail-area probability of \( D \) under its posterior reference distribution

\[ p_{b}(y) = P(D(y^{\text{rep}}, \theta) \geq D(y, \theta)|H, y). \]  

(75)

This posterior predictive \( p \)-value is well defined and calculable. The complete computation technique is described in Algorithm 1.
The posterior predictive $p$-value computed through Algorithm 1 is an estimate of $p_0$ of (75). Having obtained $\{D(y; \theta^{(j)}), D(y^{rep}(j); \theta^{(j)}), j = 1, \ldots, J\}$, Gelman et al. (1996) proposed creating a scatter-plot to perform a graphical analysis. Because graphs allow for immediate inspection of the magnitude of various values while also assisting in the detection of issues that would not be easily "visible" otherwise.

**Remark 5: Choice of the discrepancy variable $D(y; \theta)$**. There is no predefined method for selecting the discrepancy variable, and much relies on the data context and the researcher’s objectives. Discrepancies can be used to assess overall model fit using Pearson $\chi^2$-type statistics, or they can be used to assess specific parts of the model, such as appropriately capturing extreme value. Gelman et al. (1996) proposed the $\chi^2$ discrepancy as

$$X^2(y; \theta) = \sum_{i=1}^{n} \frac{(y_i - E(y_i|\theta))^2}{Var(y_i|\theta)}.$$ (77)

5.2.3. Calibration of posterior predictive $p$-value

Posterior predictive checks are simple to set up. It simply takes MCMC simulation to avoid the time-consuming or even incalculable integral in intrinsic and fractional Bayes factors. However, posterior predictive $p$-values based on these checks tend to be conservative in the sense that the distribution of posterior predictive $p$-values computed under a null model (i.e., when the data producing model and estimation model are the identical) is frequently dome shaped rather than the uniform distribution anticipated of frequentist $p$-values (Conn et al., 2018). Hjort et al. (2006) compared the $ppp$ computed under different priors. The less specific and more flat the prior is (e.g., greater variance in a normal prior), the less sensitive the $ppp$ is relative to the observed data.

The data is used twice in the $ppp$ calculation: first to update the prior to fit the data better, and secondly to estimate how unexpected the data are in comparison to the posterior parameter distribution. As a result, it is not unexpected that its distribution among plausible values of $y_{obs}$ is not uniform (Hjort et al., 2006). Indeed, as illustrated in Fig. 4, we found that the $ppp$ is more densely distributed towards 0.5. Thus the original $ppp$ often results in tests with relatively low statistical power.

Hjort et al. (2006) presented a prior-calibrated $ppp$ that takes into consideration the distribution of $y_{obs}$ as well. van Kollenburg et al. (2017) also proposed a posterior-calibrated $ppp$. In general, the calibrated $ppp$ is calculated using a double-simulation regime to re-scale the $ppp$ distribution to uniform. The calibrated $ppp$ produces better statistical power than the $ppp$. The posterior-calibrated posterior predictive $p$-value is computed following Algorithm 2, which is a modified version of the Algorithm 3 in Kollenburg (van Kollenburg et al., 2017).
Because an additional calibration step is required, the posterior-calibrated ppp may need more computational time than the standard ppp. To compute the posterior-calibrated ppp, for example, 500 ppps must be calculated, which takes up to 500 times as long as computing the normal ppp. These 500 ppps, on the other hand, may be computed in parallel, reducing calculation time significantly.

5.3. Implementation of the posterior predictive p-value in the Poisson rate test

Now consider the original test in (4). Under the null hypothesis $H_0$, $X$ and $Y$ come from the same Poisson distribution, thus $y_{obs} = \text{concatenate} (X, Y)$, the model parameter $\theta = \lambda$, which is the common Poisson rate for both $X$ and $Y$.

As to posterior sampling of $\lambda$, we use the non-informative prior Gamma (0.001, 0.001). As Gelman et al. (1996) suggested, the prior distribution for the parameters of the model need not be especially accurate. He tested different strength of the priors with both minimum and realized discrepancies and found that: As long as the prior distributions are not particularly strong, the size of the $p$-values and the conclusions reached remained essentially unchanged.

Applying the $\chi^2$ discrepancy variable (77), we define $D(y; \theta)$ as

$$D(y; \theta) = \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{\theta}.$$ (79)

Then, the ppp and posterior-calibrated ppp can be computed following Algorithm 1 and 2 respectively.

6. Numerical experiment

In this section, we use both the frequentist (C-test, E-test, likelihood ratio test) and Bayesian (intrinsic and fractional Bayes factors, ppp, and calibrated ppp) methods introduced in previous sections to analyze both simulated and real-world datasets. The code is publicly available at https://github.com/luo-lorry/Hypothesis-Test-of-Poisson-Distributions.

6.1. Comparing complexity of different methods

We analyze the time complexity and space complexity of the hypothesis testing methods discussed in previous sections. The result is shown in Table 1.
Frequentist methods are mostly efficient. Both the C-test and the E-test require a constant time (which depends on some pre-defined parameter) of evaluating the corresponding probability mass functions. The likelihood ratio test is efficient and takes $O(1)$ for both time and space.

On the other hand, running Bayesian methods—in particular, ppp and the calibrated ppp—requires more time and space. ppp takes $O(C_{\text{posterior}})$ time to generate the replicated data set (See $J$ in Algorithm 1). And the calibrated ppp even takes a multiplicity $C_{\text{calib}}$ of the time and space used by ppp (See $S$ in Algorithm 2).

Due to their tractable analytical posterior form, the basic version of the Bayes factor and the fractional Bayes factor both use constant time and space. Rather, because the data are used as training samples to create a useful prior distribution, the intrinsic Bayes factor’s time and space demands scale quadratically with the number of samples $n$.

### 6.2. Simulated dataset

In this subsection, we apply the frequentist and Bayesian methods to simulated datasets.

We first illustrate that the basic version of Bayes factor (Section 4.1) does not consistently differentiate between hypotheses when different priors are used. The reason for this is that the Bayes factor is influenced by the constants introduced by the priors. We employ the Bayes factor bound (BFB) (Benjamin & Berger, 2019; Held & Ott, 2018) to transform the frequentist $p$-value to an upper limit Bayes factor to compare the Bayes factor with the frequentist approaches. We show that, despite various degrees of power and robustness, the intrinsic Bayes factor and the fractional Bayes factor produce consistent results with the frequentist approaches.

#### 6.2.1. Basic version of Bayes factor fails to provide a consistent conclusion

We draw 10 samples (Table 2) from each of the two distributions Poisson (4) and Poisson (8), forming the X and Y used in the hypothesis test.

#### 6.2.1.1. Bayes factor bound

Bayes factor bound (BFB) is a metric for extracting information from a $p$-value. It represents the strongest case for the alternative hypothesis relative to the null hypothesis (Benjamin & Berger, 2019). For computing a Bayes factor bound from a $p$-value, we utilize the formula in Benjamin and Berger (2019):

$$BFB = \frac{1}{-e^{\log(p)}}$$

(80)

The frequentist $p$-values from C-test (Section 3.1), E-test (Section 3.2), likelihood ratio test (Section 3.3) are respectively 0.0002377, 0.0001607, 0.000076. The corresponding Bayes factor bound (80) are 193.29, 263.28, 508.17. The results are significant and indicate sufficient evidence to reject $H_0$ that the two Poisson rates are the same.

However, as we will see below, the Bayes factor reveals high posterior probability of $H_0$ under various prior selections, implying that $H_0$ is more in accord with data than $H_1$.

#### 6.2.1.2. Bayes factor and $p$-value indicate the Lindley-Bartlett paradox

Fig. 1 illustrates the Bayes factor resulting from different priors Gamma $(\alpha, \beta)$. The result vary on a drastic scale. Thus we can not draw any conclusion without a specified prior knowledge. In this sense, it is a good practice for us to use Jeffreys prior when no prior information is available.

The Lindley-Bartlett paradox (Lindley, 1957) is discovered in the Bayes factor results. We plot the Bayes factor bound corresponding to the frequentist $p$-value (C-test) as a contour line in Fig. 1. The Bayes factors with prior $(\alpha, \beta)$ chosen outside the contour line are less than the Bayes factor bound. Fig. 1 illustrates that the null hypothesis $H_0$ is preferred by the majority of Bayes factors which disagrees with the frequentist result.

The null hypothesis $H_0$ is very specific in that it does not allow for differences in the Poisson rates of the two distributions. The Bayes factor agrees with the frequentist approach unless the chosen prior aligns with the data, i.e., $\{(\alpha, \beta)|\alpha \in (1, 10), \beta \in (0.1, 1), \log(\beta) = \text{constant}\}$, as indicated in the region enclosed by the contour line in Fig. 1. When utilizing vague priors (e.g., the bottom left region in Fig. 1), Bayes factor favors the null hypothesis. Another argument is that Bayes factors automatically adjust for model complexity and prefer the simpler model.

#### 6.2.2. Frequentist and other Bayesian approaches align well with each other

To compare the frequentist and Bayesian approaches in the hypothesis testing problem, we use the simulated dataset (Table 3) from two Poisson distributions with rates $\lambda_1$ and $\lambda_2$ respectively, where $\lambda_1 = 4$ and $\lambda_2 = \lambda_1 + \Delta$, $\Delta = 0, 4, 8, 12$.

Table 4 shows the outcomes of frequentist approaches, including C-test, E-test, likelihood ratio test; and Bayesian approaches, including ppp (Section 5.2.2), calibrated ppp (Section 5.2.3), intrinsic Bayes factor (Section 4.2), and fractional Bayes factors (Section 4.3). The results are consistent across approaches. More intriguingly, we have several findings, which are as follows.

---

Note that the fractional Bayes factor may require an extra step of numerical integration.
C-test is more conservative than E-test. Therefore the mid-$p$ correction (Remark. 1) is necessary to give a more accurate result.

(2) Compared to AIBF, median intrinsic Bayes factor (MIBF) and fractional Bayes factor (FBF) are more robust, and FBF is more conservative as it uses all data as a sample (with a fraction likelihood).

(3) As expected, ppp is more conservative than other approaches. On the other hand, according to Gelman et al. (1996), supplementing the graphical assessment (scatter plots in Fig. 3) is helpful for directly inspecting the magnitude of the various quantities.

(4) The calibrated ppp is more accurate than the ppp without calibration. The histogram of ppp is more densely distributed towards 0 (Fig. 4), and thus calibrating the ppp using its reference distribution is helpful.

6.3. Offspring dataset associated with COVID-19

In this subsection, we use both frequentist and Bayesian techniques to see whether COVID-19’s spreading behavior is similar in different locations. The number of secondary infections induced by an infected case is represented by an offspring distribution, which is assumed to be Poisson. The data from Hong Kong (290 cases) (Adam et al., 2020) and Rwanda (795 cases) (Kremer et al., 2021) are publicly available. In Fig. 2, we show the histograms of the offspring in Hong Kong and Rwanda.

Based on the results shown in Table 5, the null hypothesis $H_0$ that the COVID-19 offspring distributions in Hong Kong and Rwanda have the same Poisson rate can be rejected with high confidence. The Bayes factors also indicate that the alternative hypothesis $H_1$ is preferable to $H_0$.

7. Conclusion

In this study, we employ both frequentist and Bayesian methodologies for testing hypotheses on Poisson-distributed datasets. It is crucial to note that each technique carries its distinct merits in terms of explainability, computational efficiency, and statistical power. The selection of the most appropriate approach should be a nuanced decision made by researchers, taking into account the unique context of their experimental setup and data.

Our research paves the way for future explorations in two primary areas. The first involves investigating more efficient algorithms for computing the calibrated posterior predictive $p$-value. The second anticipates the development of hybrid testing methods that synergistically incorporate the advantages of both frequentist and Bayesian techniques.

The existing literature offers valuable guidance for these future research directions. For instance, Stochastic Block Models (SBMs) have been successfully utilized for hypothesis tests on Poisson-distributed data in network science. Abbe and Sandon (2015) explored phase transition phenomena in the exact recovery of SBMs and proposed a degree-profiling algorithm based on the multivariable Poisson distribution of a node’s neighbors in each community. Yuan et al. (2022) developed a likelihood-ratio type procedure to test SBMs with bounded degrees, with the limit distribution following a power Poisson law in the growing-degree regime according to Janson’s theory (Janson, 1995). Liu et al. (2021) proposed a redefined Stochastic Block Model (RSBM) where the links generated between nodes in different blocks followed a block-node specific Poisson distribution. Beyond the area of network science, the Poisson distribution is prevalent in linguistics and is the standard distribution for describing unbounded count data (Winter & Bürkner, 2021). These diverse applications highlight the far-reaching implications of our study and the abundant opportunities for further investigation.

8. Ethical statements

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers’ bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgment

The work described in this paper was substantially supported by a grant from City University of Hong Kong (Project No. 9610639).

References


