A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-values as Simple and General p-values that Allow Double Use of the Data

Frédéric Gosselin

OPEN ACCESS Freely available online

PLoS one

A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-Values as Simple and General p-Values That Allow Double Use of the Data

Frédéric Gosselin*

Cemagref, UR EFNO, Nogent-sur-Vernisson, France

Abstract

Background: Recent approaches mixing frequentist principles with Bayesian inference propose internal goodness-of-fit (GOF) p-values that might be valuable for critical analysis of Bayesian statistical models. However, GOF p-values developed to date only have known probability distributions under restrictive conditions. As a result, no known GOF p-value has a known probability distribution for any discrepancy function.

Methodology/Principal Findings: We show mathematically that a new GOF p-value, called the sampled posterior p-value (SPP), asymptotically has a uniform probability distribution whatever the discrepancy function. In a moderate finite sample context, simulations also showed that the SPP appears stable to relatively uninformative misspecifications of the prior distribution.

Conclusions/Significance: These reasons, together with its numerical simplicity, make the SPP a better canonical GOF p-value than existing GOF p-values.

Citation: Gosselin F (2011) A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-Values as Simple and General p-Values That Allow Double Use of the Data. PLoS ONE 6(3): e14770. doi:10.1371/journal.pone.0014770

Editor: Pedro Antonio Valdes-Sosa, Cuban Neuroscience Center, Cuba

Received June 10, 2010; **Accepted** February 14, 2011; **Published** March 18, 2011

Funding: The author has no support or funding to report.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: frederic.gosselin@cemagref.fr

Introduction

Statistical model criticism, which tests a fitted statistical parametric model against observed data, is valuable for gaining more confidence in the statistical results [1-5]. Box [6] identified model criticism as one of the two main steps in statistical model development. Although many other terms have been used – model adequacy, model checking, model validation, model evaluation [3,5] –, we will use the term goodness-of-fit to refer to this confrontation between statistical model and observed data. To date, the generally preferred method has been *external* goodness-of-fit, where data used to assess the model are not those used to fit the model. The evaluation is performed either through data splitting or by comparing the model predictions against a completely different dataset [5]. External goodness-of-fit avoids using the data twice, and should result in more interpretable and less circular goodness-of-fit [7,8]. However, many researchers have proposed *internal* goodness-of-fit methods (see later), where predictions from the fitted model are compared with the observations that were used to estimate the parameters of the model. One obvious advantage of internal goodness-of-fit (GOF) is to allow fuller use of data in model checking. We will therefore focus our attention on these methods, and more precisely on GOF p-values. The GOF p-values we use are Fisherian p-values, i.e. probabilities of "seeing something [with the statistical model] as weird or weirder than you actually saw" [9]. Fisherian p-values compare the model to the data, and therefore differ from Neyman-Pearson tests which compare two models or hypotheses [9]. "Weirdness" is quantified using specific discrepancy functions, which are real-valued functions of data and of statistical model parameters. Fisherian p-values are simply calculated as the quantile of the discrepancy function calculated on the observed data in the probability distribution of discrepancy functions of data and parameters randomly generated according to some given probabilistic scheme associated to the fitted statistical model. Let us assume that, when replicating over hypothetical datasets sampled from a probabilistic model, we know these p-values have a uniform distribution on [0; 1] under assumption (A1):

(A1) the likelihood in the statistical model – or inference model, used to analyze data – is the same as the likelihood in the probabilistic model – or sampling model, used to generate data; then an extreme Fisherian p-value – i.e. a p-value very close to 0 or a p-value either very close to 0 or to 1, depending on the discrepancy function – is interpreted as contradicting (A1). The reader will find the mathematical formulation of these statements at the beginning of the Material & Methods section.

When the statistical model is fitted with Bayesian methods, these GOF p-values clearly rely on both Bayesian and frequentist ideas: they are Bayesian because the statistical parameters come either from the prior or the posterior distribution, or modifications thereof, and they are frequentist because they embed the observed data within a set of unobserved datasets sampled from a probabilistic model. This is why such methods are called calibrated Bayesian [10]. Calibrated Bayesian GOF has progres-

sively gained popularity over the last few decades, resulting in a number of more or less sophisticated techniques [6,11-18]. Calibrated Bayesian GOF differ from classical purely Bayesian methods that specify a family of alternative, more complex models and use Bayes Factors to indicate which family of models – the original or the alternative models – is the most likely [6,19]. Even though this purely Bayesian method does have some interesting features (e.g. discussion in [13]), it cannot deal with the Fisherian view of model checking, i.e. testing whether the data are consistent with a given model, without the need for an alternative hypothesis [9,10,20]. What if both the original and the alternative models were inconsistent with the data? Huber [19] qualifies these purely Bayesian procedures as 'tentative overfitting', commenting that these Bayesian methods "are based on the unwarranted presumption that by throwing in a few additional parameters one can obtain a perfectly fitting model. But how and where to insert those additional parameters often is far from obvious (...). Remember that Kepler rejected the epicyclic Ptolemaic/Copernican models because he could not obtain an adequate fit within that class." In turn, we note that emerging Bayesian GOF methods involve nonparametric alternatives [21-23], thus enriching the Bayesian GOF toolbox.

Given that frequentist statistics are believed to be more powerful than Bayesian statistics for model criticism [6,12], Little [10] viewed calibrated Bayesian p-values as an improvement over purely Bayesian p-values – and in this article we will indeed focus on calibrated Bayesian techniques. The Material and Methods section begins by proposing a brief overview of what is known on frequentist and calibrated Bayesian GOF p-values under assumption (A1) according to three criteria:

- C1: asymptotically with respect to sample size, the probability distribution of the p-value when replicating over observed datasets should be known for a variety of discrepancy functions and priors;
- C2: under reasonable finite sample sizes, the probability distribution of the p-value when replicating over observed datasets should be close to a known reference distribution for a variety of discrepancy functions and priors;
- C3: the p-values should be numerically inexpensive and relatively easy to implement based on a Monte Carlo Markov Chain or frequentist model fit [3,16].

Conditions (C1) and (C2) are required in order to use candidate GOF p-values as described above in the Fisherian perspective. Having p-values that work for very different probability distributions and any discrepancy function has an obvious advantage: it provides users with assurance that they can use the method for different kinds of statistical models, and that they have sufficient flexibility to check the model [4,15,20,24]. Condition (C3) is motivated by time constraints in the application of such methods.

As will be seen in the Material and Methods section – to which point we defer a precise definition of the p-values – some calibrated Bayesian and classical frequentist GOF p-values share the difficulty that their probability distribution is generally unknown, even asymptotically; this contradicts (C1), which makes it difficult to interpret the surprise resulting from a given p-value [14-17]. For this reason, posterior predictive p-values (p_{pop}) [4,13], which are possibly the most widely used in modern applied Bayesian settings, have come under challenge from the statistical literature [14,15,17]. Other calibrated Bayesian GOF p-values prove very computer-intensive – thus contradicting (C3). Finally, most of them do not apply to general discrepancy functions – thus contradicting (C1) and (C2). Three of the reviewed p-values – the prior predictive p-value (p_{prp}; [6]), the plug-in half-sample ML p-value ($p_{ML_{hs}}$; [25]) and the normalized sampled posterior p-value (p_{nsp}), developed in [16,18] – meet these three criteria, provided we have the same prior and likelihood in the data analysis as we had when generating data – for p_{prp} and p_{nsp}, and provided that the discrepancy function depends solely on normalized data – for p_{nsp}, on uniformized data for $p_{ML_{hs}}$– or on data – for p_{prp}. Normalized data are simple transformations of the observed data that:

(i) calculate uniformized data in [0; 1], which are the values of the empirical cumulative distribution at observed values – based on the probability distribution used in the statistical likelihood and on a suitable parameter value;

(ii) calculate the inverse cumulative function of the standard normal distribution on these uniformized data (cf. legend of Table 1 for a mathematical formulation).

The mathematical results we know for $p_{ML_{hs}}$ are limited to uniformized data. Also, we know that, in general, p_{prp} strongly depends on the prior chosen in data analysis, which is not the case for $p_{ML_{hs}}$. But is this also the case for p_{nsp}? Indeed, what happens to p_{nsp} when the prior used in data analysis does not correspond to the prior used in data generation? Also, what happens when discrepancy functions are more general, i.e. dependent on statistical parameters or on unnormalized data – which leads to

Table 1. Discrepancy functions $d(\mathbf{X},\theta,\psi)$ considered in the simulations of this paper.

Description	*General shape of the discrepancy function*
Test statistic function	$t(\mathbf{X})$
Test statistic function on normalized data	$t(\mathbf{Y})$
Other kinds of discrepancy functions	Centered mean (denoted meanc), variance (denoted varc), log-likelihood (LL)

NOTE: ψ denotes a vector of length n, composed of random numbers from the uniform distribution that are independent from each other and from all the other random variables considered. F_0 denotes the cumulative distribution function of the standard normal distribution, and $F(.|\theta,\psi)$ denotes the cumulative distribution function $F(.|\theta)$ of the density $f(.|\theta)$ of the model – or a randomized version of it when X is discrete:

$$F(\mathbf{X}|\theta,\psi)=F(\mathbf{X}-\eta|\theta)+\psi*[F(\mathbf{X}|\theta)-F(\mathbf{X}-\eta|\theta)],$$

where η is a small positive number so that $X_i-\eta$ remains bigger than the closest smaller discrete value to X_i. Normalized data are defined as $\mathbf{Y}=F_0^{-1}[F(\mathbf{X}|\theta,\psi)]$.

We considered the following t functions: mean, variance, and only in the case of unnormalized data, $p_0(\mathbf{X})=\sum_{i=1}^{n} 1_{X_i\le 0}$ and maximum (only for comparing p_{sp} with p_{pop} under the Poisson model), and only in the case of normalized data, skewness, kurtosis, and

$$Z_a(\mathbf{Y}) = -\sum_{i=1}^{n}\left[\frac{\log F_0(Y_i^*)}{n-i+.5}+\frac{\log(1-F_0(Y_i^*))}{i-.5}\right],$$

where (Y_i^*) denotes the ascending ordered version of $\mathbf{Y}$. Z_a is obtained as $Z=\int_{-\infty}^{\infty} Z_t\,dw(t)$, with the likelihood ratio statistic as Z_t and an adequate weight function $w(t)$ [37]. Centered mean and variance are the empirical mean and variance minus the mean and variance expected with θ.

doi:10.1371/journal.pone.0014770.t001

the more general sampled posterior p-value (p_{sp})? And does p_{sp} or p_{nsp} apply to discrete data? Finally, are p_{sp} or p_{nsp} more powerful than p_{pop} for detecting discrepancies between the data and the statistical model in situations when the likelihood in the statistical model is not the same as the likelihood in the probabilistic model? And how do p_{sp} and $p_{ML_{hs}}$ compare in such situations? In the second part of the paper, we study the promising p-values p_{sp} or p_{nsp} both mathematically and through simulations. Our main results are that:

(i) p_{sp} meets criterion (C1);

(ii) provided the prior distribution in the statistical analysis is equally or less informative than the prior in the probabilistic model, simulations on simple models indicate that p_{sp} has an approximately uniform distribution and fulfils criterion (C2) with sample size from several dozens to several hundreds; and

(iii) based on a specific example, p_{sp} and p_{nsp} are shown to be more powerful p-values than p_{pop} and as powerful as $p_{ML_{hs}}$.

This yields an easier way of calculating GOF p-values than the methods proposed in [7,14,17,26]. In the last part of the paper, we discuss the benefits and drawbacks of this new p-value. Leading out of this discussion, p_{sp}, p_{nsp} and $p_{ML_{hs}}$ appear to be preferable to p_{pop} and other p-values.

Materials and Methods

Review of published results

For the sake of simplicity, this section will concentrate only on the mathematical setting for continuous observations. The case of discrete valued observations will be dealt with in the next section.

Suppose that we have observed a realization $\mathbf{x_{obs}}$ of a random variable $\mathbf{X}$, $\mathbf{X}\in R^n$. We propose a parametric probability family model, $f(\mathbf{X}|\theta)$, $\theta\in\Theta\subset R^p$, for the density of $\mathbf{X}$ given θ, and a prior probability distribution $\pi(\theta)$ for θ. Although some of the results in this paper might also extend to cases where the prior is improper and the posterior is proper, we will assume (A2) throughout, i.e.:

(A2) the prior distribution is proper.

This paper will walk us through an investigation of the fit of the above statistical model with the observed data $\mathbf{x_{obs}}$. We do so by comparing the distribution of a given discrepancy function $d(\mathbf{X},\theta)$ – where $\mathbf{X}$ and θ are simulated in some way from the statistical model – with the value involving observed data, $d(\mathbf{x_{obs}},\theta)$, using the Fisherian p-value:

$$p_{m,d}(\mathbf{x_{obs}}) \equiv \mathrm{P}^{m(.)}[d(\mathbf{X},\theta) > d(\mathbf{x_{obs}},\theta)]$$

as a measure of compatibility, where $m(.)\equiv m(\mathbf{X},\theta)$ is a reference probability density for $(\mathbf{X},\theta)$ that depends on the statistical model. Each GOF p-value is defined by a reference density m and a discrepancy function d [15,20]. When the discrepancy function d does not depend on θ, Robins et al. [15] propose to shift terms and call d a test statistic function.

Our setting has so far been purely Bayesian. The frequentist part of the setting is defined by a probabilistic model for the random sampling of data $\mathbf{x}\sim m_0$ according to a given density

$$m_0(\mathbf{x}) = \int_{\Theta} f_0(\mathbf{x}|\theta)\ \pi_0(\theta)\ \mathrm{d}\theta$$

based on the parametric probability family model, $f_0(.|\theta)$, and on a prior probability distribution $\pi_0(\theta)$ – which can be a Dirac or point mass distribution. Following many authors [14-17,27], we require that, under (A1) (i.e. $f=f_0$), the probability distribution of $[p_{m,d}(\mathbf{x_{obs}})]_{\mathbf{x_{obs}}\sim m_0}$ be known at least asymptotically – i.e. when the size n of $\mathbf{x_{obs}}$ tends to infinity – and, more precisely, that this distribution be the uniform distribution on $[0;1]$, i.e.

$$\lim_{n\to\infty} P_{m_0}[p_{m,d}(\mathbf{x_{obs}}) \le s] \equiv P_{\mathbf{x_{obs}}\sim m_0}[p_{m,d}(\mathbf{x_{obs}}) \le s] = s, \forall s \in [0;1].$$

Such GOF p-values will hereafter be called *asymptotically uniform*.

The classical p-values proposed in the literature meet criterion (C3). They correspond to the following reference densities:

- the *plug-in ML density*: $m_{ML}(\mathbf{X},\theta|\mathbf{x_{obs}}) = f(\mathbf{X}|\theta)\ \delta_{\hat{\theta}}(.)$, where $\delta_{\hat{\theta}}(.)$ is the Dirac function at $\hat{\theta}$, which is the Maximum Likelihood Estimator (MLE) of θ – given $\mathbf{x_{obs}}$ and the likelihood f. Even though other values than the MLE can be used for θ in a plug-in p-value (cf. [16]), this is a reference density that is used at least implicitly in many frequentist diagnostic tools (cf. graphical tools in [1,2]);
- the *prior predictive density*: $m_{prp}(\mathbf{X},\theta) = \mathrm{f}(\mathbf{X}|\theta)\ \pi(\theta)$ [6];
- the *posterior predictive density*: $m_{pop}(\mathbf{X},\theta|\mathbf{x_{obs}}) = f(\mathbf{X}|\theta)\ P_{pop}(\theta|\mathbf{x_{obs}})$, where $P_{pop}(\theta|\mathbf{x_{obs}}) = f(\mathbf{x_{obs}}|\theta)\ \pi(\theta)/v(\mathbf{x_{obs}})$ is the posterior density of θ, given $\mathbf{x_{obs}}$, and $v(\mathbf{x_{obs}}) = \int_{\Theta} f(\mathbf{x_{obs}}|\theta)\ \pi(\theta)\ \mathrm{d}\theta$ is the marginal density of $\mathbf{x_{obs}}$ [12,13].

This paper will not go further in investigating the prior predictive p-value – dubbed p_{prp} – because of its strong dependence on the statistical prior π, in contradiction with (C2) [10,14] (also see Text S6).

With $\pi_0 = \delta_{\theta_0}$ for some fixed θ_0, and under the general assumption that the function d is a function of $\mathbf{X}$ alone that has a normal limiting distribution, Robins et al. [15] showed that the plug-in ML and posterior predictive p-values – respectively dubbed p_{ML} and p_{pop} – are asymptotically uniform when the asymptotic mean of $d(\mathbf{X})$ does not depend on θ. If the asymptotic mean of $d(\mathbf{X})$ depends on θ, then as shown by Robins et al. [15], p_{ML} and p_{pop} are generally not asymptotically uniform: more precisely, they are conservative p-values, which means the probability of extreme values is lower than the nominal probabilities from the uniform distribution. These p-values therefore only fulfill criterion (C1) if we greatly restrict the discrepancy functions considered.

This has led to the development of other p-values associated with less classical densities m, among which:

- the *post-processing* method of the posterior predictive p-value to render it a uniform p-value [17];
- the *partial posterior predictive density*: $m_{ppop}(\mathbf{X},\theta|\mathbf{x_{obs}}) = f(\mathbf{X}|\theta)\ P_{ppop}(\theta|\mathbf{x_{obs}})$, where $P_{ppop}(\theta|\mathbf{x_{obs}})$ is the partial posterior density of θ, proportional to $f(\mathbf{x_{obs}}|\mathrm{d_{obs}},\theta)\ \pi(\theta)$ where $f(\mathbf{X}|\mathrm{d_{obs}},\theta)$ is the density function of $\mathbf{X}$ conditional on the value of θ and on $d(\mathbf{X}) = \mathrm{d_{obs}} = d(\mathbf{x_{obs}})$ [14];
- the *conditional predictive density*: $m_{cp}(\mathbf{X},\theta|\mathbf{x_{obs}}) = f\left(\mathbf{X}|\hat{\theta}_{cML,obs},\theta\right) P_{cp}(\theta|\mathbf{x_{obs}})$, where $P_{cp}(\theta|\mathbf{x_{obs}})$ is the density that is proportional to $f\left(\hat{\theta}_{cML,obs},\theta\right)\ \pi(\theta)$, where $\hat{\theta}_{cML,obs}$ is the maximizer of the likelihood $f(\mathbf{x_{obs}}|\mathrm{d_{obs}},\theta)$ and where $f\left(\hat{\theta}_{cML,obs},\theta\right)$ is the marginal density of the random variable $\hat{\theta}_{cML,obs}$ evaluated at its observed value [15];
- the *plug-in half-sample ML density*: $m_{ML_{hs}}(\mathbf{X},\theta|\mathbf{x_{obs}}) = f(\mathbf{X}|\theta)\ \delta_{\hat{\theta}_{hs}}(.)$, where $\delta_{\hat{\theta}_{hs}}(.)$ is the Dirac function at $\hat{\theta}_{hs}$, which is the MLE of θ given a *half random sample of* $\mathbf{x_{obs}}$ and likelihood f [25];

– what we hereby term the *sampled posterior p-value* (p_{sp}) developed in [16,18], based on $m_{sp}(\mathbf{X},\theta|\mathbf{x_{obs}})=f(\mathbf{X}|\theta)\delta_{\tilde{\theta}}(.)$, where $\tilde{\theta}$ is a unique value of θ, which is a random sample of the posterior distribution $P_{pop}(.|\mathbf{x_{obs}})$.

With $\pi_0=\delta_{\theta_0}$ for some fixed θ_0, it has been proved mathematically, under certain assumptions, that the partial posterior predictive and conditional predictive p-values are asymptotically uniform p-values whatever the test statistic function and the prior distribution [15], thus fulfilling criterion (C1), with restrictions on discrepancy functions. However, due to criterion (C3), we will consider neither the partial posterior predictive density, nor the conditional predictive p-value [15,16] nor the post-processing method in [17] in this paper.

Durbin [28] showed that the plug-in half-sample ML p-value $p_{ML_{hs}}$ was asymptotically uniform provided it was used on uniformized data and with specific test statistic functions. This p-value has seldom been adopted, although Stephens [25] stressed its usefulness.

Johnson [16] proved that for a specific discrepancy measure, the sampled posterior p-value is also asymptotically uniform. More recently, Johnson [18] showed that if:

– the statistical model – including the prior π – is the same as the probabilistic model – including the prior π_0 – from which the data were sampled; and
– $G(s)\equiv\int_{R^n} I_{A(\theta,\mathbf{X})}(\mathbf{X})f(\mathbf{X}|\theta)\mathbf{dX}$, where $\mathrm{A}(\theta,\mathbf{X})=\{(\theta,\mathbf{X}):\mathrm{d}(\mathbf{X},\theta)\leq s\}$, depends solely on s, whatever the value of θ – i.e. in short, if $d(\mathbf{X},\theta)$ is pivotal;

then p_{sp} is not only asymptotically uniform, but is also uniform whatever the sample size. Normalized sampled posterior p-values (p_{nsp}) that use test statistics on normalized transformations of $\mathbf{X}$ possess this property. These p-values thus fulfill criteria (C1) and (C2) but with restrictions on discrepancy functions and on the prior distribution, as π must be equal to π_0.

Simulation setting

What do we know about p_{sp}, with more general discrepancy functions? We will show in the Results section that, for any discrepancy function, p_{sp} is uniform for $\pi=\pi_0$ and asymptotically uniform for $\pi\neq\pi_0$, including for discrete-valued discrepancy functions. We also wanted to include discrete-valued discrepancy functions, due to the discrete nature of either the random variables $\mathbf{X}$ or the discrepancy function. We will therefore consider the following modified p-value:

$$p_{m,d}(\mathbf{x_{obs}},\varepsilon)=\mathbf{P}^{m(.)}[d(\mathbf{X},\theta)>d(\mathbf{x_{obs}},\theta)]+\varepsilon\mathbf{P}^{m(.)}[d(\mathbf{X},\theta)=d(\mathbf{x_{obs}},\theta)],$$

where ε is drawn from a uniform distribution, independently of the other random variables.

Based on the mathematical results to come, p_{sp} appears a promising p-value that applies widely in terms of discrepancy functions, and – asymptotically – in terms of prior distributions. However, these results no longer hold when the land of asymptotia is obviously not reached, as can be the case in hierarchical models or in models that fit parameters with a limited number of observations (see, for instance, the last model in the Poisson example in [16]). Furthermore, when sample size is moderate and the statistical prior does not correspond to the data generation prior, we have no clear information on how close p_{sp} is to being uniform. We therefore used simulations to study how p_{sp} behaves in a finite sample context under four scenarios.

Objectives and scenarios. Our first scenario was performed to illustrate the uniformity results in the Results section when $f=f_0$ and $\pi=\pi_0$, while the three other scenarios were conceived to study in a finite sample context the distance to uniformity of the empirical distribution of p_{sp}, $\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon)\right]_{\mathbf{x_{obs}}\sim m_0,\varepsilon\sim U(0;1)}$, for different kinds of discrepancies between the probabilistic and statistical prior distributions:

Scenario 1: *Perfect fit between the probabilistic and statistical models*. Here, the model that generated the data and the model used to fit the data were exactly the same – including for the prior distribution.

Scenario 2: *The statistical and probabilistic models differ only by the dispersion of their priors*.

Scenario 3: *The statistical and probabilistic models differ only by the centering and dispersion of their priors*.

Scenario 4: *The statistical and probabilistic models differ only by their priors, the probabilistic prior π_0 being a Dirac distribution*. This setting is the same as in Scenario 1, except that data were generated from fixed parameters chosen at the mean of their statistical prior under Scenario 1.

Finally, we compared p_{sp} with p_{pop} and $p_{ML_{hs}}$ under Scenario 4 and a modification of Scenario 4 in which $f\neq f_0$ to illustrate the conservativeness of p_{pop} and the potentially good properties of $p_{ML_{hs}}$ under Scenario 4, and to study the difference of power between the three p-values.

Models and methods. We dealt with these issues on the following parametric models, both for data generation and data analysis, which involved conjugate priors [4] (also see Table 2):

– Poisson model: $f(x_i,\lambda)=Poisson(x_i|\lambda)$, $1\leq i\leq n$ with a Gamma prior for λ: $\lambda\sim\mathrm{Gamma}(\alpha_0,\beta_0)$;
– Normal model: $f(x_i,\theta,\sigma)=Normal(x_i|\theta,\sigma^2)$, $1\leq i\leq n$ with the priors: $1/\sigma^2\sim\mathrm{Gamma}(\alpha_0,\beta_0)$ and $\theta\sim\mathrm{Normal}(\theta_0,\sigma^2/\sigma_0^2)$;
– Bernoulli model: $f(x_i,\theta)=Bern(x_i|\theta)$, $1\leq i\leq n$ with a Beta prior for θ: $\theta\sim\mathrm{Beta}(\alpha_0,\beta_0)$.

For each dataset, α_0, β_0, θ_0 and σ_0 were held fixed in data generation and data analysis but were allowed to differ between the two phases. As conjugate priors were used, the explicit formula for the posterior distribution was known [4] and thus used under R 2.2.1 software [29] to fit the Bayesian models to the data.

Under Scenario 1, the priors were as above, with some parameters held fixed and some parameters that were capable of varying between datasets:

– for the Poisson model, constant mean and random index of dispersion of the Gamma prior: $\alpha_0/\beta_0=\theta_0=\exp(1)$ and $(\alpha_0/\beta_0^2)/(\alpha_0/\beta_0)=\rho_0\sim\mathrm{Uniform}(0;2)+.05$;
– for the Normal model, constant mean and random variance of the prior for $1/\sigma^2$: $\alpha_0/\beta_0=1$ and $\alpha_0/\beta_0^2=\rho_0\sim 10^{(\mathrm{Uniform}(0;1)-1)}$, and constant $\theta_0=0$, $\sigma_0=1$ in the prior for θ; and

Table 2. Summary of the models considered in simulations.

Poisson model	$f(x_i,\lambda)=Poisson(x_i\|\lambda)$, $1\leq i\leq n$ and $\lambda\sim\mathrm{Gamma}(\alpha_0,\beta_0)$
Normal model	$f(x_i,\theta,\sigma)=Normal(x_i\|\theta,\sigma^2)$, $1\leq i\leq n$, $1/\sigma^2\sim\mathrm{Gamma}(\alpha_0,\beta_0)$ and $\theta\sim\{\mathrm{Normal}(\theta_0,\sigma^2/\sigma_0^2)$
Bernoulli model	$f(x_i,\theta)=\mathrm{Bern}(x_i\|\theta)$, $1\leq i\leq n$ and $\theta\sim\mathrm{Beta}(\alpha_0,\beta_0)$

NOTE: For each dataset, α_0, β_0, θ_0 and σ_0 were held fixed in the data generation and data analysis steps.
doi:10.1371/journal.pone.0014770.t002

– for the Bernoulli model, $\alpha_0 = 2\,\theta_0\,\rho_0$ and $\beta_0 = 2(1-\theta_0)\,\rho_0$, with $\rho_0 \sim 10^{\text{Uniform}(0;1)}$ and $\theta_0 = .5$.

The setting of Scenario 2 is the same as in Scenario 1, except that ρ_0 is replaced in the statistical model by $\mu_{\sigma_0}\rho_0$ in the Poisson and normal cases and by $(\rho_0-1)/\mu_{\sigma_0}+1$ in the Bernoulli case, where $\log\left(\mu_{\sigma_0}\right) \sim \text{Normal}(0,1.4^2)$. Scenario 3 differs from Scenario 2 by θ_0 values in the statistical model that are no longer fixed but drawn at random according to $\theta_0 \sim \exp[\text{Normal}(1,.4^2)]$ in the Poisson case, $\theta_0 \sim \text{Normal}(0,3^2)$ in the Normal case and $\theta_0 \sim \text{Uniform}(.25,.75)$ in the Bernoulli case. The distributions for parameters μ_{σ_0} and θ_0 in Scenarios 2 and 3 were chosen to vary the levels of informativeness and off-centering of the statistical prior with respect to the probabilistic prior. Finally, in Scenario 4, data were generated from fixed parameters, chosen at the mean of their statistical prior under Scenario 1, i.e. $\lambda = \exp(1)$ in the Poisson case, $\sigma = 1$ and $\theta = 0$ in the Gaussian case, and $\theta = .5$ in the Bernoulli case.

We considered three kinds of discrepancy function, $d(\mathbf{X},\theta)$, i.e. test statistics, test statistics on normalized data and other discrepancy functions (cf. Table 1). Test statistics on normalized data were introduced because they define pivotal quantities used by [18] to find results under the condition $\pi = \pi_0$.

The number of observations in each dataset, n, was a random figure between 20 and 1,000: $n \sim \exp[3.45 * U(0;1)+3]$ with probability 0.7 and $n \sim U(250;\,1{,}000)$ with probability 0.3. n was rounded to the nearest ten or – if the value was above 200 – to the nearest hundred. We used 5,000 sampled values of $\mathbf{X}$ to calculate p-values. The programs were run either on a DELL Latitude D830 Intel Centrino T7250 or on a server with two dual-core Opteron 2.2 GHz processors and 3 Gb of RAM. One hundred thousand replicated datasets were studied under Scenarios 2 to 4 and 10,000 under Scenario 1. To illustrate the dependence of p_{prp} on the statistical prior distribution, we also calculated the p_{prp} based on 3,000 datasets under Scenario 2.

The p-value associated to each dataset and each chosen discrepancy function differed from the classical calculation for predictive p-values. Let us denote:

$$\alpha = \sum_j 1_{d\left(\mathbf{X}_j,\tilde{\theta}\right)\,>\,d\left(\mathbf{x_{obs}},\tilde{\theta}\right)} + \varepsilon \sum_j 1_{d\left(\mathbf{X}_j,\tilde{\theta}\right)\,==\,d\left(\mathbf{x_{obs}},\tilde{\theta}\right)}$$

and

$$\beta = \sum_j 1_{d\left(\mathbf{X}_j,\tilde{\theta}\right)\,<\,d\left(\mathbf{x_{obs}},\tilde{\theta}\right)} + (1-\varepsilon)\,\sum_j 1_{d\left(\mathbf{X}_j,\tilde{\theta}\right)\,==\,d\left(\mathbf{x_{obs}},\tilde{\theta}\right)},$$

where ε is a random value from the uniform distribution. Instead of the classical formula $\alpha/(\alpha+\beta)$ [4], the p-value was drawn at random from the beta distribution with the respective shape parameters $\alpha+1$ and $\beta+1$. Indeed, it can be shown that this distribution is the posterior distribution of the underlying p-value

$$p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon) = \int_{R^n} \left[1_{\{\mathbf{X};d(\mathbf{X},\theta)\,>\,d\left(\mathbf{x_{obs}},\theta\right)\}} + \varepsilon\, 1_{\{\mathbf{X};d(\mathbf{X},\theta)\,=\,d\left(\mathbf{x_{obs}},\theta\right)\}}\right] f\left(\mathbf{X}|\tilde{\theta}\right)\, d\mathbf{X},$$

once we have observed or sampled $\left(\mathbf{X}_j\right)_j$, $\tilde{\theta}$, ε and $\mathbf{x_{obs}}$, provided the prior of the p-value is uninformative [4] (p.40). In contrast, the use of $\alpha/(\alpha+\beta)$ can result in significant departures from the uniform distribution, which would be due to the calculation method and not to the underlying p-value; this would especially occur with a low number of replicated data $\left(\mathbf{X}_j\right)_j$ or to estimate the tails of the uniform distribution (see Text S9).

The resulting p-values were considered as sampled from the distribution $\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon)\right]_{\mathbf{x_{obs}} \sim m_0, \varepsilon \sim U(0;1)}$. They were numerically compared with the uniform distribution, through Kolmogorov-Smirnov tests, which are adequate and easy to calculate for such continuous valued distributions, as well as through binomial two-sided tests for the proportion of p-values that were in the 5% or 1% extremities of the $[0;1]$ interval. As stated above, we used a uniform random number ε' to ventilate between the "less extreme" and "more extreme" categories, the probability of the event when the proportion simulated from the binomial distribution was equal to the observed proportion. This guaranteed a uniform distribution of the associated p-value. For the proportion of p-values in the 5% or 1% extremities of the $[0;1]$ interval, we also calculated the posterior density of the estimated proportion from the observed number, using a beta distribution as above. We then analyzed where the posterior estimates were positioned relative to intervals around the target probabilities of 5% or 1%. For example, we distinguished cases where 95% of the estimates of the underlying proportion of p-values fell in the interval $[0;.04[$ (proportion of p-values is estimated to be non-negligibly less than 5%), from cases where 95% of the estimates fell in the interval $[.04;.06]$ (proportion of p-values is estimated to be negligibly different from 5%), and from cases where 95% of the estimates fell in the interval $]0.06;1]$ (proportion of p-values is estimated to be non-negligibly greater than 5%) (see Text S1).

Comparing p_{sp} with p_{pop} and $p_{ML_{hs}}$ under the Poisson model. Finally, for the Poisson model, we compared p_{sp} with p_{pop} and $p_{ML_{hs}}$ under Scenario 4 and a modification of Scenario 4 in which $f \neq f_0$. We used the same test statistics as above, plus the maximum function. Forty-thousand datasets were generated as in Scenario 4 or from a Polya distribution [30] with a maximum value $n_{\max}$ drawn at random from the values 4 and 5, and a mean and variance equal to those of the aforementioned Poisson distribution. The sample size was drawn at random from between 20 and 50, except for Figure 1 where it was sampled from the set (20,30,40,50,60,70,80).

The R commands to run and analyze the simulations described above can be found in Text S8.

Results

The sampled posterior p-value: mathematical results

p_{sp} is uniform when $f = f_0$ and $\pi = \pi_0$. The following lemma extends Johnson's [18] results on test statistics applied on normalized data to general discrepancy functions, including discrete-valued discrepancy functions:

Lemma: Assume that $\pi = \pi_0$ is proper, and $f = f_0$ – so that assumptions (A1) and (A2) are met. Then, for every discrepancy function d, the probability distribution of $\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon)\right]_{\mathbf{x_{obs}} \sim m_0, \varepsilon \sim U(0;1)}$ is uniform, i.e.

$$P_{m_0,U}\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon) \le s\right] \equiv \int_{R^n} \int_0^1 1_{\left\{p_{m_{sp},d}\left(\mathbf{x_{obs}},\varepsilon\right) \le s\right\}}\, m_0(\mathbf{x_{obs}})\, d\varepsilon\, d\mathbf{x_{obs}} = s, \forall s \in [0;1].$$

Proof. The proof of this Lemma follows the same line as the proof of the Lemma in [18]. For the sake of clarity, let us denote

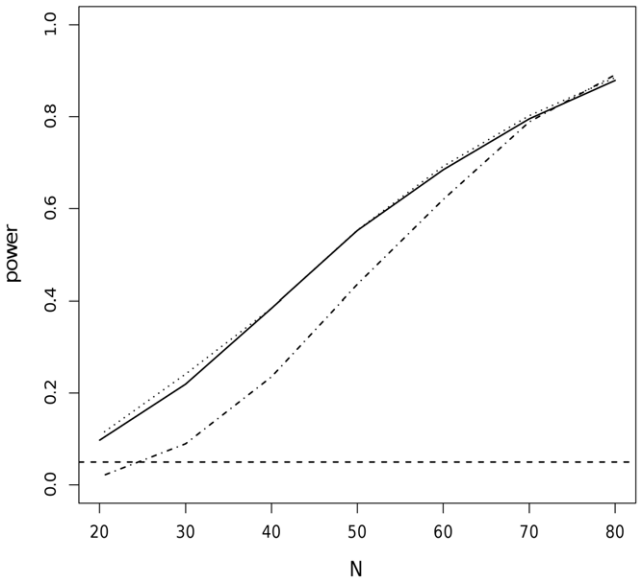

Figure 1. Power of the sampled posterior (p_{sp}; solid line), half-sample ML ($p_{ML_{hs}}$; dotted line) and posterior predictive (p_{pop}; dotted-dashed line) p-values. Power of the p-values p_{sp} (solid line), $p_{ML_{hs}}$ (dotted line) and p_{pop} (dotted-dashed line) used with the maximum test statistic to detect departures from the Poisson distribution at the level of p = 0.05 when data are distributed according to a Polya distribution with $n_{\max} = 5$. Power is plotted as a function of sample size N varying between 20 and 80. p_{sp} and $p_{ML_{hs}}$ were equivalent in terms of power, and both were more powerful than p_{pop} except at the highest sample sizes. The dotted baseline level corresponds to p = 0.05.
doi:10.1371/journal.pone.0014770.g001

$g(\mathbf{X}, \mathbf{x_{obs}}, \theta, \varepsilon) = 1_{\{\mathbf{X};\, d(\mathbf{X},\theta) > d(\mathbf{x_{obs}},\theta)\}} + \varepsilon 1_{\{\mathbf{X};\, d(\mathbf{X},\theta) = d(\mathbf{x_{obs}},\theta)\}}$, so that $p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon) = \int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right) d\mathbf{X}$. Then, by simply substituting the place where the marginal density $v(\mathbf{x_{obs}})$ occurs in the integrals,

$$P_{m_0,U}\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon) \le s\right] = .../...$$

$$= \int_{\Theta}\int_{R^n}\int_{\Theta}\int_{0}^{1} 1_{\left[\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right) d\mathbf{X}\right] \le s}$$
$$P_{pop}\left(\tilde{\theta}|\mathbf{x_{obs}}\right) f(\mathbf{x_{obs}}|\theta_0)\ \pi(\theta_0)\, d\varepsilon d\tilde{\theta}\, d\,\mathbf{x_{obs}} d\,\theta_0$$

$$= \int_{\Theta}\int_{R^n}\int_{\Theta}\int_{0}^{1} 1_{\left[\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right) d\mathbf{X}\right] \le s}$$
$$\frac{f\left(\mathbf{x_{obs}}|\tilde{\theta}\right)\pi(\tilde{\theta})}{v(\mathbf{x_{obs}})} f(\mathbf{x_{obs}}|\theta_0)\ \pi(\theta_0)\, d\varepsilon d\tilde{\theta}\, d\,\mathbf{x_{obs}} d\,\theta_0$$

$$= \int_{\Theta}\int_{R^n}\int_{0}^{1} 1_{\left[\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right) d\mathbf{X}\right] \le s}$$
$$\left[\int_{\Theta} \frac{f(\mathbf{x_{obs}}|\theta_0)\ \pi(\theta_0)}{v(\mathbf{x_{obs}})}\, d\,\theta_0\right] f\left(\mathbf{x_{obs}}|\tilde{\theta}\right)\ \pi(\tilde{\theta})\, d\,\varepsilon\, d\,\mathbf{x_{obs}} d\tilde{\theta}$$

$$= \int_{\Theta}\int_{R^n}\int_{0}^{1} 1_{\left[\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right) d\mathbf{X}\right] \le s}$$
$$f\left(\mathbf{x_{obs}}|\tilde{\theta}\right)\, d\,\varepsilon\, d\,\mathbf{x_{obs}}\, \pi(\tilde{\theta})\, d\tilde{\theta}$$

However, in this last equation, conditional on $\tilde{\theta}$, $d\left(\mathbf{X},\tilde{\theta}\right)$ and $d\left(\mathbf{x_{obs}},\tilde{\theta}\right)$ in function g have the same probability distribution and are independent. Then, still conditional on $\tilde{\theta}$, due to the very definition of g, $\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right) d\mathbf{X}$ has a uniform distribution between 0 and 1 when $\mathbf{x_{obs}}$ then ε are sampled as specified in the integral. For this reason, the above formula can be rewritten as:

$$P_{m_0,U}\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon) \le s\right] = \int_{\Theta}\int_{0}^{1} 1_{p\,\le\, s}\ \pi(\tilde{\theta})\, dp\ d\tilde{\theta} = s,$$

which yields our result.

p_{sp} is asymptotically uniform when $f = f_\theta$ and $\pi \neq \pi_0$. The above result shows that p_{sp} is uniform provided (A1), (A2) and the statistical prior π – which generates the posterior distribution $P_{pop}(\theta|\mathbf{x_{obs}})$ – is the same as the probabilistic prior π_0. We can extend this result when both priors differ by showing that that under conditions:

- on the likelihood – including the identifiability of the model, and the independence of observations;
- on the priors – including that for every θ such that $\pi_0(\theta) > 0$, we must have $\pi(\theta) > 0$;
- on the discrepancy function – its continuity relative to θ;
- on the parameter space Θ – its compactness;

then, p_{sp} is asymptotically uniform under (A1) and (A2).

Sketch of proof. If we assume that the parameter space Θ is compact, that the model is identifiable and that the random variables are independent and identically distributed, i.e. $f(\mathbf{X}|\theta) = \prod_i f(x_i|\theta)$, then when the size n of the sample $\mathbf{x_{obs}}$ drawn from $f(.|\theta_0)$ for a given θ_0 tends to infinity, whatever the neighborhood A of θ_0, $\lim_{n\to\infty} \int_{R^n} P_{pop}(\theta \in A|\mathbf{x_{obs}})\, f(\mathbf{x_{obs}}|\theta_0)\, d\,\mathbf{x_{obs}} = 1$, i.e. $\lim_{n\to\infty} P_{pop}(\theta \in A|\mathbf{x_{obs}}) = 1, f(.|\theta_0)$-almost surely [4] (p.587 in Appendix B). From the continuity of $d(\mathbf{X},\theta)$ relative to θ condi-

tions, we deduce that $\int_{R^n}\int_{\Theta}\int_{0}^{1} 1\left[\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\tilde{\theta},\varepsilon)\, f\left(\mathbf{X}|\tilde{\theta}\right)\, \mathrm{d}\,\mathbf{X}\right] \leq s$ $P_{pop}\left(\tilde{\theta}|\mathbf{x_{obs}}\right) f(\mathbf{x_{obs}}|\theta_0)\, \mathrm{d}\,\varepsilon \mathrm{d}\tilde{\theta}\, \mathrm{d}\,\mathbf{x_{obs}}$ is asymptotically equal to $\int_{R^n}\int_{\Theta}\int_{0}^{1} 1\left[\int_{R^n} g(\mathbf{X},\mathbf{x_{obs}},\theta_0,\varepsilon) f(\mathbf{X}|\theta_0)\, \mathrm{d}\,\mathbf{X}\right] \leq s$ $f(\mathbf{x_{obs}}|\theta_0)\, \mathrm{d}\,\varepsilon\, \mathrm{d}\,\mathbf{x_{obs}}$, which as in the proof of the above Lemma is equal to s. Since these quantities are bounded by 1, we get through an integration over θ_0 according to the prior π_0, that $\lim_{n\to\infty} P_{m_0,U}\left[p_{m_{sp},d}(\mathbf{x_{obs}},\varepsilon) \leq s\right] = s$.

We speculate that the proof in Gelman *et al.* [4] (p.587 in Appendix B) can be extended to the case where the random variables are independent but not identically distributed – i.e. $f(\mathbf{X}|\theta) = \prod_i f_i(x_i|\theta)$– provided the f_i distributions are sampled from a common probability law, making it possible to use Kolmogorov's strong law of large numbers instead of the usual law of large numbers employed in [4].

Discussion. In these conditions, p_{sp} is asymptotically uniform even when $\pi \neq \pi_0$. These results also hold when π_0 is a Dirac distribution δ_{θ_0}. Under more stringent conditions on the likelihood and the prior, these results can be made sharper – and inform on the speed of convergence – by using the convergence of the posterior distribution to normality [4,31-33].

The sampled posterior p-value: simulation results. Our above results are mathematical and mostly asymptotic. We now study the finite sample behavior of the sampled posterior p-value based on our simulations. Overall, our results for Scenario 1 – corresponding to a perfect matching of the statistical and probabilistic models – were in accordance with our expectations: p_{sp} and p_{nsp} then had behaviors compatible with uniform p-values (Text S1).

When the statistical prior had the same mode but was sharper than the probabilistic prior in Scenario 2, p_{sp} and p_{nsp} yielded poor results for the studied sample sizes (Table 3 and Text S2), in contrast with their asymptotic good behavior (previous section). Conversely, when the statistical prior was less informative than the probabilistic prior, both p-values were much closer to being uniform (Text S2), in sharp contrast with p_{prp} (Text S6).

Except in one case, p_{sp} and p_{nsp} were also not far from being asymptotically uniform in Scenario 4 when the true parameter value was equal to the mode of the statistical prior (cf. Text S4). An exception was observed for the Bernoulli model with p_{sp} and $t = var$ or $d = varc$: in this case, p_{sp} did not approach uniformity, even with relatively high sample sizes. On the whole, however, p_{sp} and p_{nsp} were further from being uniform for small sample sizes in Scenario 4 than in Scenario 2 with uninformative statistical priors.

De-centering of the statistical prior (Scenario 3) yielded p_{sp} and p_{nsp} values that were further from the uniform distribution (Table 4 and Text S3). However, p_{sp} and p_{nsp} remained relatively close to being uniform when the statistical prior was less informative than the probabilistic prior and when de-centering was not too strong.

Comparing p_{sp} against p_{pop} and $p_{ML_{hs}}$ for the Poisson model under Scenario 4 with $t = max$ showed thatp_{pop} was conservative, as expected by the mathematical results in [15] while p_{sp} and $p_{ML_{hs}}$ were closer to being uniform for sample sizes 20 and 50 (Text S5). When the true distribution was a Polya distribution instead of a Poisson distribution, p_{sp} and $p_{ML_{hs}}$ were of similar and greater power, except for the highest sample sizes where p_{pop} tended to be slightly more powerful (Figure 1). A difference in power of 10 to 20% in favor of p_{sp}, p_{nsp} or $p_{ML_{hs}}$ was not uncommon and was observed with various discrepancy functions (Figure 1 and Table 5).

Table 3. Behavior of p_{nsp} relative to the uniform distribution under Scenario 2, depending on the interval housing the statistical prior sharpness parameter μ_{σ_0}.

Interval to which $\mu_{\sigma 0}$ belongs	[.0;.40[	[.40;1.01[	[1.01;2.59[	[2.59;∞[
D	$.080^{***}$	.005	$.01^{*}$	$.009^{*}$
$P_{5\%}$	$.091^{***,+,++}$	$.049^{00}$	$.050^{00}$	$.052^{00}$
$P_{1\%}$	$.035^{***,+,++}$	$.011^{0}$	$.009^{0}$	$.011^{0}$

Kolomogorov-Smirnov distance (D) between the simulated p_{nsp} and the uniform distribution and frequency ($P_{5\%}$ and $P_{1\%}$) of p_{nsp} found at the 5% and 1% extremities of the unit interval for the Poisson model with $t = Z_a$ in Scenario 2 based on 100,000 different datasets. In this Scenario, the statistical prior $\lambda \sim Gamma\left(\sqrt{1/\left(\theta_0\mu_{\sigma_0}\rho_0\right)}, \sqrt{\theta_0/\left(\mu_{\sigma_0}\rho_0\right)}\right)$, has a different sharpness to the probabilistic prior $\lambda \sim \mathrm{Gamma}\left(\sqrt{1/(\theta_0\rho_0)}, \sqrt{\theta_0/\rho_0}\right)$. The statistics for the overall sample were $D = .019^{**}$, $P_{5\%} = .061^{***,+}$ and $P_{1\%} = .016^{***,++}$. These results illustrate that for p_{nsp} to be approximately uniform when the statistical prior is not the same as the probabilistic prior, it is preferable for the statistical prior to be less informative rather than more informative compared with the probabilistic prior. Similar results were found for other test statistics and other probability distributions (cf. Text S2).

NOTE: The notation for the significance of the tests is as follows: (*) means that the test is significant at a level between .05 and .1; * between .01 and .05; ** between .0001 and .01; *** less than .0001. The notation system for the study of the negligibility of departures from expected values is as follows, for $P_{5\%}$: 00 (respectively, 0) means 95% of the estimated values of the underlying p-value are in the interval [.045; .055] (resp. [.04; .06]); ++ (respectively, +) means 95% of the estimated values are in the interval].06; 1] (resp.].055; 1]); – (respectively, -) means 95% of the estimated values are in the interval [0; .04[(resp. [0; .045[). For $P_{1\%}$, the notations are the same but with cutoff points divided by 5.

doi:10.1371/journal.pone.0014770.t003

Discussion

Synthesis of results

In this paper, we first recap on various calibrated Bayesian methods for goodness-of-fit (GOF) p-values and extend the results found in [18] for normalized sampled posterior p-values (p_{nsp}) in different directions. We show in particular that similar results apply for the more general p_{sp} when the data are not normalized and for discrepancy functions that can be discrete-valued rather than only for continuous-valued test statistic functions. We also show that this p-value is asymptotically uniform when the statistical prior differs from the probabilistic prior ($\pi \neq \pi_0$). Through simulations, we empirically tested this p-value under $\pi \neq \pi_0$ in a finite sample context. The results show that p_{sp} has a relatively correct behavior provided that the statistical prior is "not too informative and not too uninformative", and not too far off-centered, relative to the probabilistic prior. An exception to this statement occurred in Scenario 4 with the Bernoulli model and $t = var$ or $d = varc$, for which p_{sp} was far from being uniform even for relatively large sample sizes. We think this is because the fixed parameter $\theta = .5$ used to sample $\mathbf{x_{obs}}$ was precisely the parameter value for which the variance was the largest over the full parameter space. This might correspond to a very slow convergence in this specific case or to a restriction of our asymptotic mathematical results, somewhat similar to the convergence at the edge of parameter space in [4] (Section 4.3). A simulation with $\theta = .7$ yielded a p_{sp} that was much closer to being uniform (Text S4).

Table 4. Behavior of p_{nsp} relative to the uniform distribution under Scenario 3, based on the frequency of p_{nsp}values found at the 5% extremities of the unit interval, depending on the interval of the statistical prior sharpness parameter μ_{σ_0} (in rows) and off-centering parameter $|\log(\theta_0)-1|$ (in columns).

Interval to which $\mu_{\sigma 0}$ (row) and $\|\log(\theta_0)$ -1| (column) belong	[.0;.14[	[.14;.28[	[.28;.47[	[.47; 1.73[
[.0;.40[	$.093^{***,++}$	$.106^{***,++}$	$.157^{***,++}$	$.241^{***,++}$
[.40;1.01[	$.053^{o}$	$.053^{o}$	$.054^{o}$	$.081^{***,++}$
[1.01;2.59[	$.050^{o}$	$.054^{o}$	$.046^{o}$	$.056^{*}$
[2.59;∞[	$.050^{o}$	$.053^{o}$	$.051^{o}$	$.050^{o}$

Frequency ($P_{5\%}$) of p_{nsp} that are at the 5% extremities of the unit interval, for the Poisson model with $t=Z_a$ under Scenario 3, according to the values of μ_{σ_0} and $|\log(\theta_0)-1|$, for 100,000 different simulated datasets. Similar results were found for other t functions and for the Poisson distribution, with more significant results for certain other t functions when $\mu_{\sigma_0}\in[.01;1.0[$ and $|\log(\theta_0)-1|\in[.0;2.03[$ (see Text S3).
NOTE: The notation system for the significance of the tests and the negligibility of departures from expected values are as in Table 3. Qualitatively similar results were found for $t=$mean and $t=$variance. For $t=$kurtosis and $t=$skewness, results were much less strongly and much less frequently significant.
doi:10.1371/journal.pone.0014770.t004

Based on these new results and on the review of published results (Material and Methods Section), we shortlisted three alternative methods as simple candidates of asymptotically uniform GOF p-values:

Method 1: p_{sp}, with a variety of discrepancy functions d and with not too inadequate statistical priors;

Method 2: p_{pop} or p_{ML} with a test statistic function t such that asymptotically the mean of $t(\mathbf{X})$ is not dependent on θ [15]. Examples of such functions include skewness or kurtosis for the normal distribution, skewness for the t distribution, or the ratio between the mean of the sample and its variance for a Poisson distribution;

Method 3: $p_{ML_{hs}}$, with specific test statistic functions used on uniformized data.

We will also discuss two other, more elaborate sets of methods:

Method 4: partial posterior predictive p-values (p_{ppop}) or conditional predictive p-values (p_{cp}) used only with test statistic functions, as developed and proposed by [14,15,26];

Method 5: calibrated posterior predictive p-values (p_{cpp}) as proposed in [17].

One last method could have been to use p_{pop} or p_{ML} with test statistic functions, knowing that they are conservative [15]. However, our results for p_{pop} show that we then lose a significant amount of power compared with p_{sp} and p_{nsp} (Figure 1 and Table 5). This strategy will therefore not be considered further here.

Table 5. Difference in power between p_{sp} or $p_{nsp}$and p_{pop} according to sample size (in columns) and discrepancy function (in rows).

Sample size *N*	20	50
Skewness on normalized data	.050	.103
Kurtosis on normalized data	.119	.243
Za on normalized data	.031	.135
Maximum on normalized data	.068	.204
Maximum on raw data	.083	.113

Difference in power between p_{sp} or $p_{nsp}$and p_{pop} for detecting departures from the Poisson distribution at the 5% level and when the true distribution is a Polya distribution with maximum value $n_{\max}$ equal to 5, with the sample size *N* equal to 20 or 50. Various discrepancy functions and test statistic functions are considered. The difference in power is positive, indicating more power for p_{sp} or p_{nsp}. The magnitude of the difference can be quite substantial, ranging from 0.1 to 0.25. Similar results were obtained between $p_{ML_{hs}}$ and p_{pop}, with slightly greater power differences than between p_{sp} or $p_{nsp}$and p_{pop}.
doi:10.1371/journal.pone.0014770.t005

The relative merits of candidate p-values

We hereafter discuss the merits and limits of our preferred method – Method 1 or p_{sp} – in comparison with the other candidate methods. With respect to Method 2, Method 1 has the advantage of allowing the use of various discrepancy functions whereas Method 2 requires very specific test statistic functions; this means that different aspects of the probabilistic model can be studied with Method 1 rather than only the t functions that characterize the hypothesized probabilistic distribution. We agree with [4,20,24] on the necessary adaptation of discrepancy functions to each particular situation where we might want to test departures of data from the model on case-specific features. This makes it possible to include problems involving detection of outliers ($t=min$ or $t=max$) and dependence between observations [24] in model checking. It also means that p_{sp} appears more flexible and better applicable to very different probability distributions than Method 2: for more complicated hypothesized distributions, it might be difficult to build t functions such that asymptotically the mean of $t(\mathbf{X})$ does not depend on θ.

On a more theoretical grounding, while p_{pop} and p_{ML} provided default and intuitive responses to question (b) in [34], i.e. "what replications should we compare the data to?" – p_{sp} gives a different and less intuitive answer, based on mathematical results: replications should all be sampled from the likelihood based on a unique parameters value, itself sampled from the posterior distribution, and not from multiple parameters values sampled from the same distribution (p_{pop}) or from the Maximum Likelihood parameters (p_{ML}).

In comparison with p_{cpp} (Method 5), the main advantage of p_{sp} is its much weaker computational cost inside MCMC computations, including for complicated models. By contrast, p_{cpp} entails multiplying the MCMC computational burden by the number of "repetitions" of the model on which post-processing is based. This would take from at least a hundred to a thousand times longer than p_{sp}. From our point of view, this is a major problem, especially in cases such as hierarchical models on large datasets. Therefore, the choice between Methods 1 and 5 may primarily depend on the length of time required to fit the model.

Regarding Method 4, the apparent weakness of p_{sp} compared with the results in [26] for p_{cp} is that we have no information on when the asymptotic behavior is reached – except when the whole

statistical model is the same as the probabilistic model used to sample the data. Nevertheless, our simulation results do show that provided the priors are not too informative or too uninformative, and not too far off-centered, p_{sp} is not very far from being uniform. An advantage of p_{sp} over p_{ppop} or p_{cp} is its simplicity: we do not need to calculate the calibrated likelihood of the model with respect to the test statistic, as we do for p_{ppop} or p_{cp}. Moreover, if one wishes to calculate N different p-values based on different test statistics, it can be done inside the same numerical fitting in the case of p_{sp} but must be done N times on N different calibrated likelihoods for p_{ppop} or p_{cp}. A final advantage of p_{sp} over p_{ppop} or p_{cp} is that we have mathematical results for discrepancy functions in general, rather than just for test statistic functions as is the case for p_{ppop} and p_{cp}.

Methods 1 and 3 appear very close in terms of applicability and, in the example studied, in terms of power. Their respective powers could be studied in more detail in the future. A common feature of both methods is that they give random results, in the sense that we can randomly reach different p-values for the same observed data $\mathbf{x_{obs}}$. A small advantage in favor of p_{sp} in Method 1 is that it does not require a separate fit on the half-sample, which contrasts with Method 3. A stronger advantage for Method 1 is that its asymptotic validity is proved for general discrepancy functions, whereas the mathematical results we have for $p_{ML_{hs}}$ in Method 3 only apply to specific test statistic functions of uniformized data [28]. This in particular implies that we have no mathematical result on the asymptotic uniformity of $p_{ML_{hs}}$ in Figure 1.

We therefore propose using p_{sp} and p_{nsp} as a good GOF strategy, which is unrestricted with respect to distributions and d functions and which has a reasonable numerical and coding cost. To our knowledge, these are the only p-values that have a known asymptotic probability distribution whatever the discrepancy function.

Notes on how to use the p_{sp}

This section discusses two points related to the strategy of using p_{sp}: the choice of prior distribution, and the choice of the parameter value(s) used to sample "new data" and normalize it.

First, our results indicate that we should generally prefer priors that are moderately less informative in data analysis than in data sampling (Table 4 and Appendices 2 and 3). This statement somewhat echoes similar considerations in [17] (Section 9.3). If this result were to be generalizable, it would mean that when p_{sp} indicates a significant departure from the uniform distribution, depending on whether the prior is judged as too informative (or respectively, too uninformative), the same model should be tested with less informative (or respectively more informative) priors. An alternative might be to use p_{sp} in a frequentist setting, provided the asymptotic assumption of normality of the estimators is assumed correct (cf. next section). If significant departures from a uniform distribution are still found, the probability distribution used in the likelihood should be reconsidered in data analysis.

Second, p_{sp} involves a single sampled value $\tilde{\theta}$ value of the model parameter θ, which means that the p_{sp} method might give different random results on the same dataset with the same model [18]. An alternative solution would be to use the probabilistic bounds method proposed in [18] (Section 2.3). A further potential alternative we propose, with the formalism of p_{sp} (see Table 1), could be –:

1–for each dataset $\mathbf{x_{obs}}$ and function d, draw at random $\alpha \sim U(0,1)$;

2–after MCMC, calculate the sampled posterior p-values $\left\{p_{m_{sp}(.|\theta_i),d}(\mathbf{x_{obs}},\varepsilon_i)\right\}_{(\theta_i,\varepsilon_i)}$ associated with the (θ_i) s sampled from the posterior distribution associated with $\mathbf{x_{obs}}$ and (ε_i) sampled from the uniform distribution;

3–consider the empirical α-quantile of the latter distribution.

Provided analysts use the same value for α drawn at random at the beginning of the first analysis for the same dataset, this would guarantee a better comparability of the analysis of the same dataset by different analysts.

Final global remarks

In contrast to the likelihood principle, calibrated Bayesian techniques involve the use of artificial data – i.e. data that were not observed. This makes pure Bayesians reluctant to use these techniques [35]. Indeed, internal calibrated Bayesian goodness-of-fit is sometimes considered to be a hopeless cause, where proponents want to have the cake – i.e. estimate model parameters based on all the data available – and eat it too – by confronting the fitted model to the same data that were used to fit it. Calibrated internal goodness-of-fit consequently attracts criticism for using the data twice [8]. Strikingly, p_{sp} seems to provide a nearly uniform p-value, although it uses the data $\mathbf{x_{obs}}$ twice: once to estimate the posterior distribution – from which $\tilde{\theta}$ is sampled – and once again to calculate $d\left(\mathbf{x_{obs}},\tilde{\theta}\right)$. It therefore appears to warrant the same criticisms as p_{ML} or p_{pop}, which were supposed to justify their lack of asymptotical uniformity. Johnson [16] explains it in these terms, in the context of chi-square statistics: "Heuristically, the idea [...] is that the degrees of freedom lost by substituting the grouped MLE for θ in Pearson's χ^2 statistic are exactly recovered by replacing the MLE with a sampled value from the posterior [distribution]". The proof of Lemma 1 in the Results section reveals another explanation: as we are working on *sampled* data to fit statistical models, we should also agree to work on *sampled* parameters to criticize the model. Indeed, this double sampling allowed us to make the roles of data and parameters symmetrical, enabling us to prove our mathematical results. Therefore, the problem lies less in that a GOF p-value uses data twice, but more in *how* it uses the data twice – see [36] on the need to more precisely define what we mean by "using the data twice".

We have applied p_{sp} and p_{nsp} in a Bayesian context. However, as stressed in [16], these p-values might also be used with frequentist methods when the asymptotic assumption of normality of the estimators is correct. Indeed, we applied p_{nsp} on the Poisson case by drawing a value of θ at random on the log scale from a normal distribution with the estimated mean as mean and with the estimated standard error as standard error fitted with a Poisson generalized linear model (glm). The results indicate as good a behavior as p_{nsp} used in Bayesian models under Scenarios 1 and 4 (Text S7).

Little [10] once wrote that Bayesian statistics were relatively weak for model assessment compared to frequentist statistics. Although the underused $p_{ML_{hs}}$ might be a good frequentist GOF p-value if its properties are known for more general discrepancy functions, our results highlight an even more attractive solution that mixes frequentist reasoning with a completely Bayesian modeling formulation, by using the sampled posterior p-values (p_{sp}) in a calibrated Bayesian framework. The transposition of p_{sp} into a frequentist setting has been shown to be correct in the above example, and could therefore represent another potential "frequentist" solution. However, we believe that for the not-so-infrequent cases where the normal approximation of the estimate distribution is not accurate – as can be found for binomial or Poisson regression with a high proportion of zero values – a Bayesian framework is more adequate than a frequentist setting for sampling a value of θ.

Supporting Information

Text S1 Results of Scenario 1.
Found at: doi:10.1371/journal.pone.0014770.s001 (0.23 MB DOC)

Text S2 Results of Scenario 2.
Found at: doi:10.1371/journal.pone.0014770.s002 (0.37 MB DOC)

Text S3 Results of Scenario 3.
Found at: doi:10.1371/journal.pone.0014770.s003 (0.40 MB DOC)

Text S4 Results of Scenario 4.
Found at: doi:10.1371/journal.pone.0014770.s004 (0.40 MB DOC)

Text S5 Results of Scenario 4 for the sampled posterior and the posterior predictive p-values.
Found at: doi:10.1371/journal.pone.0014770.s005 (0.06 MB DOC)

Text S6 A simple illustration of the strong dependence of prior predictive p-values on the prior distribution.
Found at: doi:10.1371/journal.pone.0014770.s006 (0.06 MB DOC)

Text S7 Results of the normalized sampled posterior p-value with frequentist Poisson generalized linear models.
Found at: doi:10.1371/journal.pone.0014770.s007 (0.07 MB DOC)

Text S8 R commands to run and analyze the simulations.
Found at: doi:10.1371/journal.pone.0014770.s008 (0.93 MB DOC)

Text S9 R commands to illustrate the discrepancy between the rough deterministic and the stochastic methods to transform α and β to p-values.
Found at: doi:10.1371/journal.pone.0014770.s009 (0.05 MB DOC)

Acknowledgments

The author thanks Elfie Perdereau for discussion on the posterior predictive checks, Vicki Moore for correcting the English, and Claude Millier and Marion Gosselin for re-reading a previous version of the paper. The author also thanks three anonymous referees for their useful remarks.

Author Contributions

Conceived and designed the experiments: FG. Performed the experiments: FG. Analyzed the data: FG. Contributed reagents/materials/analysis tools: FG. Wrote the paper: FG.

References

1. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS. New York: Springer. 528 p.
2. Harrell FE (2001) Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, USA: Springer. xxiii + 568 p.
3. O'Hagan A (2003) HSSS model criticism. In: Green PJ, Hjort NL, Richardson ST, eds. Highly Structured Stochastic Systems Oxford University Press. pp 423–444.
4. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis. Boca Raton: Chapman & Hall. 668 p.
5. Mentré F, Escolano S (2006) Prediction discrepancies for the evaluation of nonlinear mixed-effects models. J Pharmacok & Phramacod 33: 345–67.
6. Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness. J R Stat Soc Ser A 143: 383–430.
7. Evans M (1997) Bayesian inference procedures derived via the concept of relative surprise. Commun Stat - Theory Methods 26: 1125–43.
8. Evans M (2000) Comments on Asymptotic distribution of P values in composite null models by J. M. Robins, A. van der Vaart and V. Ventura. J Am Stat Assoc 95: 1160–3.
9. Christensen R (2005) Testing Fisher, Neyman, Pearson, and Bayes. Am Stat 59: 121–6.
10. Little RJ (2006) Calibrated Bayes: A Bayes/frequentist roadmap. Am Stat 60: 213–23.
11. Guttman I (1967) The use of the concept of a future observation in goodness-of-fit problems. J R Stat Soc Ser B-Stat Methodol 29: 83–100.
12. Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 12: 1151–72.
13. Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Stat Sinica 6: 733–60.
14. Bayarri MJ, Berger JO (2000) P-values for composite null models. J Am Stat Assoc 95: 1127–42.
15. Robins JM, van der Vaart A, Ventura V (2000) Asymptotic distribution of P values in composite null models. J Am Stat Assoc 95: 1143–56.
16. Johnson VE (2004) A Bayesian chi(2) test for goodness-of-fit. Ann Stat 32: 2361–84.
17. Hjort NL, Dahl FA, Hognadottir G (2006) Post-processing posterior predictive p values. J Am Stat Assoc 101: 1157–74.
18. Johnson VE (2007) Bayesian Model Assessment Using Pivotal Quantities. Bayesian Anal 2: 719–34.
19. Huber PJ (2002) Approximate models. In: Huber-Carol C, Balakrishnan N, Nikulin MS, Mesbah M, eds. Goodness-of-fit tests and model validity. Boston: Birkhäuser. pp 25–41.
20. Anscombe FJ (1963) Tests of goodness of fit. J R Stat Soc Ser B-Stat Methodol 25: 81–94.
21. Verdinelli I, Wasserman L (1998) Bayesian goodness-of-fit testing using infinite-dimensional exponential families. Ann Stat 26: 1215–41.
22. Robert CP, Rousseau J (2002) A Mixture Approach to Bayesian Goodness of Fit. Technical Report 02009, Cahier du CEREMADE, Université Paris Dauphine.
23. McVinish R, Rousseau J, Mengersen K (2009) Bayesian goodness of fit testing with mixtures of triangular distributions. Scand J Stat 36: 337–54.
24. Cox DR (2002) Karl Pearson and the Chi-squared test. In: Huber-Carol C, Balakrishnan N, Nikulin MS, Mesbah M, eds. Goodness-of-fit tests and model validity. Boston: Birkhäuser. pp 3–8.
25. Stephens MA (1978) On the half-sample method for goodness-of-fit. J R Stat Soc Ser B-Stat Methodol 40: 64–70.
26. Fraser DA, Rousseau J (2008) Studentization and deriving accurate p-values. Biometrika 95: 1–16.
27. Bayarri MJ, Berger JO (2004) The interplay of Bayesian and frequentist analysis. Stat Sci 19: 58–80.
28. Durbin J (1973) Distribution theory for tests based on the sample distribution function. Philadelphia: SIAM Publications n°9. 64 p.
29. R Development Core Team (2005) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
30. Patil GP, Boswell MT, Joshi SW, Ratnaparkhi MV (1984) Dictionary and classified bibliography of statistical distributions in scientific work. Volume 1: Discrete models. Mairland, Maryland, USA: International Co-operative Publishing House. 458 p.
31. Walker AM (1969) On the Asymptotic Behaviour of Posterior Distributions. J R Stat Soc Ser B-Stat Methodol 31: 80–8.
32. Johnson RA (1970) Asymptotic Expansions Associated with Posterior Distributions. Ann Math Stat 41: 851–64.
33. Shen X, Wasserman L (2001) Rates of convergence of posterior distributions. Ann Stat 29: 687–714.
34. Gelman A (2007) Comment: Bayesian checking of the second levels of hierarchical models. Stat Sci 22: 349–52.
35. Piccinato L (2000) Comments on Asymptotic distribution of P values in composite null models by J. M. Robins, A. van der Vaart and V. Ventura. J Am Stat Assoc 95: 1166–7.
36. Evans M (2007) Comment: Bayesian checking of the second levels of hierarchical models. Stat Sci 22: 344–8.
37. Zhang J (2002) Powerful goodness-of-fit tests based on the likelihood ratio. J R Stat Soc Ser B-Stat Methodol 64: 281–94.