Posterior predictive p-values and the convex order

Patrick Rubin-Delanchy* and Daniel John Lawson**

*Heilbronn Institute for Mathematical Research, University of Bristol, Bristol, UK
**School of Social and Community Medicine and Department of Statistics, University of Bristol, Bristol, UK

January 5, 2015

Abstract

Posterior predictive p-values are a common approach to Bayesian model-checking. This article analyses their frequency behaviour, that is, their distribution when the parameters and the data are drawn from the prior and the model respectively. We show that the family of possible distributions is exactly described as the distributions that are less variable than uniform on [0, 1], in the convex order. In general, p-values with such a property are not conservative, and we illustrate how the theoretical worst-case error rate for false rejection can occur in practice. We describe how to correct the p-values to recover conservatism in several common scenarios, for example, when interpreting a single p-value or when combining multiple p-values into an overall score of significance. We also handle the case where the p-value is estimated from posterior samples obtained from techniques such as Markov Chain or Sequential Monte Carlo. Our results place posterior predictive p-values in a much clearer theoretical framework, allowing them to be used with more assurance.

1 Introduction

In important papers on Bayesian model-checking, Meng (1994) and Gelman et al. (1996) proposed to test the fit of a model by analysing the following posterior quantity. Let $f$ be some function measuring the discrepancy between the model and the data. The question asked is: if a new dataset were generated from the same model and parameters, what is the probability that the new discrepancy would be as large? In mathematical notation this probability is written (Meng, 1994; Gelman et al., 1996, Eq. 2.8, Eq. 7)

$$P = P\{f(D^*, \theta) \geq f(D, \theta) \mid D\},$$

where $\theta$ represents the model parameters, $D$ is the observed dataset, $D^*$ is a hypothetical dataset generated from the model with parameters $\theta$, and $P(\cdot \mid D)$ is the joint posterior distribution of $(\theta, D^*)$ given $D$. A variable $P$ in the general form above is referred to as a posterior predictive p-value.

Since their introduction, which can be credited to Rubin (1984), Tsui and Weerahandi (1989), Meng (1994) or Gelman et al. (1996), depending on exact definitions, posterior predictive p-values have received a number of criticisms. First, $P$ is a p-value, and as such its interpretation is full of pitfalls. For example, it is certainly not the probability that the model is right. Second, the dependence of $f$ on the unknown $\theta$ may seem odd. Third, because the full posterior is used, rather than the prior or some cleverly chosen partial posterior, see e.g. Bayarri and Berger (2000), there is something self-fulfilling about this check; heuristically, one would expect $P$ to concentrate around $1/2$.

This last issue is really part of a bigger problem. In reality there does not seem to be a clear mathematical description of the probabilistic behaviour of $P$, except for a few insights given in the last pages of Meng (1994). Over the last two decades, statements have appeared in the literature generally suggesting that the problem is ‘hard’. For example Hjort et al. (2006) say “the
interpretation and comparison of posterior predictive p-values [is] a difficult and risky matter. Bayarri and Berger (2000) have commented that “its main weakness is that there is an apparent “double use” of the data...This double use of the data can induce unnatural behavior”. In a discussion of Gelman et al. (1996), Rubin alluded to some “conservative operating characteristics” (Rubin 1996).

If there is one important message in this article, it is that the frequency behaviour of posterior predictive p-values is precisely described as being less variable, in the convex order, than a uniform random variable on $[0, 1]$. Although the property had already been discovered by Meng (1994, Theorem 1), our main contribution is that any probability measure of this sort is the distribution of some posterior predictive p-value (Theorem 3). By some straightforward applications of the convex order we will show that these p-values are not conservative in general, for example, the $2\alpha$ bound given in Meng (1994) is achievable (Section 2.2 and Figure 2). However, when many posterior predictive p-values are combined into an overall score, the result is sometimes conservative.

A posterior predictive p-value is a meaningful quantity: it is the probability of the discrepancy being as large tomorrow as it is today. Given a sample from the posterior distribution, this probability can typically be estimated very quickly and with no hassle at all (e.g. in about one line of code). As a result, the use of this model-checking technique and its variants seems to be widespread in applied statistics. These are good reasons to seek to understand the behaviour of posterior predictive p-values in repeated samples. We do not take any position on the philosophical validity of the approach.

In fact, understanding the behaviour of posterior predictive p-values has a more general application, that is not necessarily Bayesian. Suppose we have two random objects, $X$ and $Y$, with a known joint distribution, and only $Y$ is observed. Many common statistical models have this structure. For example, $X$ might be the underlying state in a state-space model and $Y$ the observation; or $Y$ might be a point process and $X$ an underlying random intensity, as in the Cox process. In such models we often want to test something about $Y$ based on $X$. For instance, in the standard Kalman filter model, we could be interested in testing the distance, $f(Y, X)$, between the state and the observation. Ideally we would be able to observe the true p-value, $Q = P\{f(Y^*, X) \geq f(Y, X) | X, Y\}$, where $Y^*$ is a replicate of $Y$ conditional on the true $X$. However, this is impossible because $X$ is not observable. Instead, suppose we analyse $P = E(Q | Y)$. Then, under the hypothesis that the model holds, how should $P$ be distributed? The definition of $P$ above is equivalent to (1), if we replace $Y$ with $D$ and $X$ with $\theta$.

A practical issue is the interpretation of $P$ when it is estimated by simulation. This is a very common case since $P$ is usually constructed from posterior samples of $\theta$. Interestingly, a number of our results hold without any alteration in the second, but not the first, of two estimation schemes given by Gelman et al. (1996).

The rest of this article is organised as follows. Section 2 treats the case of a single posterior predictive p-value. First, we prove our main result, that there is a posterior predictive p-value for any distribution that is less variable than uniform in the convex order, in the process also deriving an extension of a famous theorem by Strassen (1965). Second, we describe this family of distributions, re-proving the $2\alpha$ bound found by Meng (1994). Third, we construct an example of a posterior predictive p-value that achieves the bound. In Section 3 we consider the problem of testing the fit of a model using multiple posterior predictive p-values. We provide tail probability bounds and some asymptotic results that can be used to make a test of the overall significance of the p-values conservative. In Section 4 we compare two schemes for evaluating the p-value from posterior samples.

2 The convex order

We start with a joint distribution over two random elements, $\theta$ and $D$. In Bayesian statistics this would normally be decomposed as a marginal distribution on $\theta$, called the prior, and a conditional distribution on $D | \theta$, called the model. For a given dataset $D$, a typical calculation of the
posterior predictive p-value would proceed as follows (Gelman et al., 1996, Section 2.3). First, simulate \( \theta_1, \ldots, \theta_M \) from the posterior distribution of \( \theta \) given \( D \), for a large \( M \). Second, for each \( \theta_i \), simulate a replicated dataset \( D_i^* \). Finally, estimate

\[
\hat{P}_M = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\{f(D_i^*, \theta_i) \geq f(D, \theta_i)\},
\]

(2)

where \( \mathbb{I} \) is the indicator function. \( P \) is the limit of \( \hat{P}_M \) as \( M \to \infty \), assuming the \( \theta_i \) are independent. We will revisit the properties of the estimate under dependence and finite \( M \) in Section 4. For now, assume that \( P \) is effectively observable for a given dataset \( D \), e.g. by making \( M \) large enough or through some analytical solution.

Our analysis focuses on the frequency behaviour of \( P \), meaning its behaviour given a specified joint distribution of \( \theta \) and \( D \). Now, because \( D \) is random, \( P \) is a random variable. It could be simulated as follows. To obtain a single realisation, we would draw \( \theta \) from the prior, and \( D \) from the model of \( D \mid \theta \). Then we would discard \( \theta \) and compute \( P \) in (1) conditional on \( D \), e.g. via (2), as if we had never seen \( \theta \). To obtain multiple independent replicates of \( P \), we would repeat this cycle, each time constructing a new \( \theta \) and \( D \).

From this point on, we always assume that \( f(D, \theta) \) is an absolutely continuous random variable. Meng (1994) makes use of the identity

\[
P = P\{f(D^*, \theta) \geq f(D, \theta) \mid D\}
= \mathbb{E}[P\{f(D^*, \theta) \geq f(D, \theta) \mid \theta, D\} \mid D],
\]

to make the following observation. For any convex function \( h \) we have \( E\{h(P)\} \leq E\{h(U)\} \), where \( U \) is a uniform random variable on \([0, 1]\). The proof of the last statement uses the fact that the quantity \( P\{f(D^*, \theta) \geq f(D, \theta) \mid \theta, D\} \) is a random variable distributed as \( U \), marginally over \( \theta \) and \( D \), and then applies Jensen’s inequality. Meng (1994) then goes on to find an upper bound \( P[P \leq \alpha] \leq 2\alpha \) for \( \alpha \in [0, 1] \).

In fact, the property being alluded to is an important stochastic order. Let \( X \) and \( Y \) be two random variables with probability measures \( \mu \) and \( \nu \) respectively. We say that \( \mu \) (respectively, \( X \)) is less variable than \( \nu \) (respectively, \( Y \)) in the convex order, denoted \( \mu \leq_{cx} \nu \) (respectively, \( X \leq_{cx} Y \)) if, for any convex function \( h \),

\[
E\{h(X)\} \leq E\{h(Y)\}.
\]

The convex order is a statement about variability: convex functions generally put more weight on the extremes. In fact, it has direct implications in terms of the first two moments of \( X \) and \( Y \). Using \( h(x) = x \) and then \( h(x) = -x \), two convex functions, we can immediately deduce that \( E(X) = E(Y) \). Then, since \( (x - E(X))^2 \leq (x - E(Y))^2 \) is a convex function in \( x \), the variance of \( X \) must be smaller than the variance of \( Y \). In this article we will say that a probability measure \( P \), and a random variable distributed as \( P \), is sub-uniform if \( P \leq_{cx} U \), where \( U \) is a uniform distribution on \([0, 1]\). Posterior predictive p-values have a sub-uniform distribution.

At first glance, Meng’s findings could seem quite conservative. They would suggest that, to be sure not to exceed a false positive rate of \( \alpha \) when the model on \((\theta, D)\) holds, we would have to multiply our posterior predictive p-value by two. Yet from practical experience, the variance result above, as well as a loose inspection of (1), we could have the impression that these p-values are already quite conservative — even the raw p-value looks too large. This raises the question of whether the bound can be improved. More generally, it would be useful to know whether the frequency behaviour of posterior predictive p-values is well described as being sub-uniform, in other words, whether the space of distributions cannot somehow be reduced. The rest of this section addresses these questions by making the following points:

1. It is possible to construct a posterior predictive p-value with any sub-uniform distribution (Theorem 3).
2. Some sub-uniform distributions achieve the \( 2\alpha \) bound (Corollary 1).
3. Therefore, some posterior predictive p-values achieve the \( 2\alpha \) bound. In fact, we can construct a simple example where this happens (Section 2.3).
2.1 A posterior predictive p-value for every sub-uniform distribution

A famous theorem by Strassen [1965] provides a fundamental interpretation of the convex order, coupling $X$ and $Y$ via a martingale:

**Theorem 1** (Strassen’s theorem). For two probability measures $\mu$ and $\nu$ on the real line the following conditions are equivalent:

1. $\mu \leq_{cx} \nu$;

2. there are random variables $X$ and $Y$ with marginal distributions $\mu$ and $\nu$ respectively such that $E(Y \mid X) = X$.

This (much more approachable) version of the theorem is due to Müller and Rüschendorf [2001]. The original version holds for more general probability measures.

Strassen’s theorem is central to our main result. Given a sub-uniform probability measure $P$, it is possible to construct a coupling, $(P, U)$, where $P$ is distributed as $P$, $U$ is uniform on $[0, 1]$, and $E(U \mid P) = P$. However, to make progress, certain awkward couplings need to be forbidden, namely, those for which the conditional random variable $U \mid P$ has some discrete components. The following theorem ensures we only need to deal with ‘nice’ couplings.

**Theorem 2** (Conditionally continuous coupling). Let $\mu$ and $\nu$ be two probability measures on the real line where $\nu$ is absolutely continuous. The following conditions are equivalent:

1. $\mu \leq_{cx} \nu$;

2. there exist random variables $X$ and $Y$ with marginal distributions $\mu$ and $\nu$ respectively such that $E(Y \mid X) = X$ and the random variable $Y \mid X$ is either singular, i.e. $Y = X$, or absolutely continuous with $\mu$-probability one.

The proof is relegated to the appendix because it is quite technical. (It may be advantageous to first consult Section 2.2 on the integrated distribution function.) On the other hand, the basic idea is quite simple. A small amount of zero-mean, continuously distributed noise is added to $X$, constructing an intermediary variable $\tilde{X}$ with distribution $\tilde{\mu}$. The noise depends on $X$ in such a way that $\mu \leq_{cx} \tilde{\mu} \leq_{cx} \nu$. Then Strassen’s theorem is used to couple $Y$ and $\tilde{X}$. Clearly $E(Y \mid X) = X$ and the details of the construction ensure that, no matter how $Y$ and $\tilde{X}$ are coupled, $Y \mid X$ is either continuous or singular.

From this we are able to construct a probabilistic structure that has more resemblance to a Bayesian model-checking setup.

**Lemma 1.** Let $\mu$ and $\nu$ be two probability measures on the real line satisfying $\mu \leq_{cx} \nu$ where $\nu$ is absolutely continuous. Then there exist real random variables $X, S, \theta$ and a collection of random variables $Y_{t \in \mathbb{R}}$ such that

$$X = E(Y_{\theta} \mid S),$$

where $X$ has marginal distribution $\mu$ and $Y_t$ has marginal distribution $\nu$ for any $t \in \mathbb{R}$.

**Proof.** By Theorem 2 there exists a coupling of real random variables, $(X, S)$, such that $S \mid X$ is continuous or singular with probability one. Let $G$ be a continuous distribution function that is positive on $\mathbb{R}$. If $S \mid X$ is singular, let $Y_t = S$ for all $t \in \mathbb{R}$. Otherwise, $S \mid X$ has a continuous distribution function, denoted $F_{S \mid X}$. If we define $Y_t$ via

$$Y_t = F_{S \mid X}^{-1}([F_{S \mid X}(S) + G(t) \mod 1], \quad t \in \mathbb{R},$$

then $Y_t \mid X$ has the same distribution as $F_{S \mid X}^{-1}(U)$, where $U$ is uniformly distributed on $[0, 1]$, therefore $Y_t \mid X$ is distributed as $S \mid X$ for any $X$. Hence, $Y_t$ has measure $\nu$ marginally.
Let \( \theta \) be a random variable with distribution function \( G \) that is independent of all previously defined random variables. If \( S \mid X \) is singular then clearly \( X = E(Y_\theta \mid S) \). Otherwise,

\[
E(Y_\theta \mid S) = E\left(F_{S|X}^{-1}\{F_{S|X}(S) + G(\theta)\} \mod 1\right)
= E\{F_{S|X}^{-1}(U)\} = E\{S \mid X\} = X.
\]

The proof is heavily reliant on the continuous coupling idea, making the step \( \Theta \) possible, and essentially allowing us to choose any distribution for \( \theta \). It would be interesting to know if this can somehow be relaxed. We are now in a position to state our main result.

**Theorem 3** (Posterior predictive p-values and the convex order). Let \( P \) be a sub-uniform probability measure. Then there exist random variables \( P, D, \theta \) and a discrepancy function \( f \) such that

\[
P = P\{f(D^*, \theta) \geq f(D, \theta) \mid D\},
\]

where \( P \) has measure \( P \), \( D^* \) is a replicate of \( D \) conditional on \( \theta \) and \( P(\cdot \mid D) \) is the joint posterior distribution of \( (\theta, D^*) \) given \( D \).

**Proof.** This time we construct a coupling \( E(S \mid P) = P \) such that \( S \mid P \) is continuous or discrete with probability one, \( P \) has marginal distribution \( P \) and \( S \) is marginally uniform on \([0, 1] \). As in the proof of Lemma \( \Theta \) we arrive at a setup

\[
U_t = F_{S|P}^{-1}\{F_{S|P}(S) + G(t)\} \mod 1, \quad t \in \mathbb{R},
\]

if \( S \mid P \) is continuous, and \( U_t = S \) otherwise, where \( G \) is some positive continuous distribution function on \( \mathbb{R} \).

Let \( D \) be a random variable that implies \( S \), i.e., there exists a function \( g \) such that \( S = g(D) \) with probability one, but that is otherwise independent of the other variables. Given the values of \( D \) and \( t \) the value of \( U_t \) is known. Therefore, we can construct a discrepancy function \( f \) such that \( f(D, t) = F^{-1}(U_t) \) with probability one, where \( F \) is a continuous survival function. Then, if \( \theta \) has distribution \( G \),

\[
P\{f(D^*, \theta) \geq f(D, \theta) \mid D\} = E(P\{f(D^*, \theta) \geq f(D, \theta) \mid \theta, D\} \mid D)
= E(U_\theta \mid D) = E(U_\theta \mid S) = P,
\]

where the last equality follows the same argument given at the end of the proof of Theorem \( \Theta \). □

It is quite telling that the proof needs a parameter-dependent discrepancy function. It seems possible that not all sub-uniform distributions are attainable if \( f \) can depend only on \( D \). In fact, in his highly influential paper on Bayesian model-checking, Rubin (1984, 1994) only considered p-values of this type,

\[
P\{f(D^*) \geq f(D) \mid D\}.
\]

It would be interesting to know if this can somehow be relaxed. We are now in a position to state our main result.

## 2.2 Characterising sub-uniformity

To help visualise the space of sub-uniform distributions, it will be useful to introduce the *integrated* distribution function of a random variable \( X \) with distribution function \( F_X \),

\[
\phi_X(x) = \int_{-\infty}^{x} F_X(t)dt,
\]

which is defined for \( x \in \mathbb{R} \). Müller and Rüschendorf (2001) analysed and made extensive use of this function and its counterpart, formed from the survival function, where \((1 - F_X)\) replaces \( F_X \) in the above. Some of their results are restated here:
1. $\phi_X$ is non-decreasing and convex;
2. Its right derivative $\phi_X^+(x)$ exists and $0 \leq \phi_X^+(x) \leq 1$;
3. $\lim_{x \to -\infty} \phi_X(x) = 0$ and $\lim_{x \to \infty} \{x - \phi_X(x)\} = E(X)$.

Furthermore, for any function $\phi$ satisfying these properties, there is a random variable $X$ such that $\phi$ is the integrated distribution function of $X$. The right derivative of $\phi$ is the distribution function of $X$, $F_X(x) = \phi^+(x)$.

Let $Y$ be another random variable with integrated distribution function $\phi_Y$. Then $X \leq_{cx} Y$ if and only if $\phi_X(x) \leq \phi_Y(x)$ for $x \in \mathbb{R}$ and $\lim_{x \to \infty} \{\phi_Y(x) - \phi_X(x)\} = 0$.

The integrated distribution function of a uniform random variable is $\phi_U(x) = x^2/2$. Figure 1 shows this function alongside some others, corresponding to sub-uniform probability measures.

From the points above, it is clear that all these functions must be non-decreasing, convex, with a right derivative between 0 and 1, always below $\phi_U(x)$, equal to 0 at 0 and 1/2 at 1. In fact, the space of sub-uniform probability measures is characterised by functions satisfying these criteria.

The dashed line in Figure 1 is of particular interest. It corresponds to a distribution, hereafter denoted $P_{2\alpha}$, which is a mixture of a point mass at $\alpha$, of probability $2\alpha$, and a uniform distribution over $[2\alpha, 1]$, of probability $(1 - 2\alpha)$.

$P_{2\alpha}$ is sub-uniform, as can be established by (analytically) comparing its integrated distribution function to $\phi_U$, and achieves the $2\alpha$ bound: if $P$ is a random variable from $P_{2\alpha}$ then $P(P \leq \alpha) = 2\alpha$.

Our next result gives a general bound on the distribution function of $X$ given the probability measure of $Y$.

**Lemma 2.** Let $X$ and $Y$ be two random variables satisfying $X \leq_{cx} Y$, with distribution functions $F_X$ and $F_Y$ respectively. For a given $\alpha \in \mathbb{R}$, let

$$h = \min \left[ 1, \max \left\{ w : w(x - \alpha) \leq \int_{-\infty}^{x} F_Y(t) dt, \ x \in \mathbb{R} \right\} \right].$$

Then $F_X(\alpha) \leq h$. Furthermore, there exists a random variable $\tilde{X}$, with distribution function $F_{\tilde{X}}$, such that $\tilde{X} \leq_{cx} Y$ and $F_{\tilde{X}}(\alpha) = h$.

A formal proof of this lemma is given in the appendix, but the basic idea is illustrated in Figure 1 with $P_{2\alpha}$: we find an integrated distribution function $\phi$ which has a maximal derivative at $\alpha$ with $\phi(x) \leq \phi_Y(x)$ for $x \in \mathbb{R}$ and $\lim_{x \to \infty} \{\phi_Y(x) - \phi(x)\} = 0$. For the case of a sub-uniform probability measure we find:
Corollary 1. Let $P$ be a sub-uniform random variable with distribution function $F_P$. Then $F_P(\alpha) \leq 2\alpha$, for $\alpha \in [0, 1/2]$. Furthermore, for any such $\alpha$, there exists a sub-uniform random variable $\tilde{P}$, with distribution function $F_{\tilde{P}}$, satisfying $F_{\tilde{P}}(\alpha) = 2\alpha$.

This corollary is only included for completeness, since everything it says is already known. The existence part of the statement is evident from $P_{2\alpha}$, and Meng (1994, Eq. 5.6) had already proved the bound.

2.3 A constructive example

Taking stock of our results so far, there exist non-conservative posterior predictive $p$-values. In fact, there must exist a posterior predictive $p$-value that is $\alpha$ with probability $2\alpha$. We now present an example.

The proof of Theorem 3 finds a continuum of random variables $U_t, t \in \mathbb{R}$ that are marginally uniform and average to $P$, i.e. $E(U_t \mid D) = P$. In fact, for the special probability measure $P_{2\alpha}$, we do not need a continuum. Let $U_0$ be a uniform random variable on $[0, 1]$ and let $U_1 = U_0 \cdot I(U_0 \geq 2\alpha) + (2\alpha - U_0) \cdot I(U_0 < 2\alpha)$. Then $(U_0 + U_1)/2$ has distribution $P_{2\alpha}$. Dahl (2006) used this decomposition to construct a (quite theoretical) posterior predictive $p$-value with distribution $P_{2\alpha}$.

Figure 2: Construction of the $2\alpha$ case. Consider a particle moving on the geometry shown, starting at $X(0) = 0$, on the far left, and travelling either clockwise or anti-clockwise. Location is measured 'clockwise' around the loop to a position $X(1) = x$. The discrepancy function is the distance travelled, measured along either of the paths $\theta = 0$ or $\theta = 1$. This can yield a posterior predictive $p$-value with distribution $P_{2\alpha}$, see details in main text.

Our example is based on a more physically interpretable Bayesian model. Let $X(t), t \in [0, \infty)$ denote the position of a particle in the geometry shown in Figure 2 as it travels from the left ($X(0) = 0$), towards the loop, and then around it, either clockwise ($\theta = 1$) or anti-clockwise ($\theta = 0$), stopping before it has gone all the way around ($X(\infty) < 1$). The two senses of rotation are equally probable a priori, $P(\theta = 0) = P(\theta = 1) = 1/2$, and the dynamics of the particle are such that the distance travelled after one unit of time is continuously distributed, with survival function $G$, density $g$ and support on $[0, 1]$.

After one unit of time, the position of the particle is observed, $X(1) = x$, recorded going clockwise around the loop. The distance travelled along the path indexed by $\theta$ is

$$d(x, \theta) = \begin{cases} x & : x \leq 1 - 2\alpha, \\
 x & : x > 1 - 2\alpha \text{ and } \theta = 1, \\
 2 - 2\alpha - x & : x > 1 - 2\alpha \text{ and } \theta = 0. \end{cases}$$

The posterior probability of $\theta$ given $x$ is

$$p(\theta \mid x) \propto g(d(x, \theta))p(\theta),$$

for $\theta = 0, 1$. We will use the distance travelled, $d$, as a discrepancy function. Let $X^*(t), t \in [0, \infty)$ be a second, hypothetical, particle in the same conditions, observed at $X^*(1) = x^*$. Given $x$ and $\theta$, the probability that the second particle would travel at least as far is $G\{d(x, \theta)\}$. Therefore, the
posterior predictive p-value is

\[ P = \mathbb{P}\{d(x^*, \theta) \geq d(x, \theta) | x\} \]

\[ = \sum_{\theta=0,1} p(\theta | x) \mathbb{P}\{d(x^*, \theta) \geq d(x, \theta) | \theta, x\} \]

\[ = \sum_{\theta=0,1} p(\theta | x) G\{d(x, \theta)\}. \]

If \( G(t) = 1 - t, \ g = 1 \), then we cannot distinguish which direction the particle took, i.e. \( p(\theta | x) = 1/2 \) for \( \theta = 0, 1 \). Then

\[ P = \begin{cases} 1 - x & x \leq 1 - 2\alpha, \\ (1 - x)/2 + (x + 2\alpha - 1)/2 = \alpha & x > 1 - 2\alpha. \end{cases} \]

Now consider how \( P \) would behave in repeated experiments. The observation \( x \), above, is now a random variable. When \( G(t) = 1 - t \), it is uniformly distributed on \([0, 1)\), so that \( P \) is distributed as \( \mathcal{P}_{2\alpha} \).

As we vary \( G \), we can construct a range of other sub-uniform distributions with \( \mathbb{P}(P \leq \alpha) > \alpha \). This example gives us a heuristic understanding of how non-conservative behaviour can occur in practice. The risk comes from a) having p-values that have a conflicting view of what is ‘extreme’ and then b) the posterior not allowing a single one to dominate.

3 Multiple testing

Suppose we have multiple discrepancy functions \( f_1, \ldots, f_m \), giving posterior predictive p-values \( P_1, \ldots, P_m \) respectively, and we want to combine these into one, overall, anomaly score. A conservative solution would be to multiply every p-value by two before any analysis. This section investigates situations where it is possible to do better.

The \( P_i \) could occur from testing a few specific hypotheses, or from more generic bulk testing of the data, in which case we might obtain, for example, a p-value for every observation. These different scenarios affect whether the p-values can be treated as independent and/or identically distributed and, also, what order of magnitude we might expect for \( m \). In the analysis below, the \( P_i \) are always assumed to be at least independent.

Fisher’s method (Mosteller and Fisher [1948]) is a popular way of combining p-values. Suppose we have classical p-values, \( U_1, \ldots, U_m \), which are independent uniform random variables on \([0, 1] \) under the null hypothesis. Then the statistic \( -2 \sum \log(U_i) \), called Fisher’s score, has a \( \chi^2 \) distribution with \( 2m \) degrees of freedom. The null hypothesis is rejected when this statistic is large.

Replacing the \( U_i \) with \( P_i \) in this procedure has an interesting asymptotic effect:

**Lemma 3** (Fisher’s method is asymptotically conservative). Let \( P_1, \ldots, P_m \) and \( U_1, \ldots, U_m \) each be sequences of independent and identically distributed sub-uniform and uniform random variables on \([0, 1] \) respectively. For \( \alpha \in (0, 1] \), let \( t_{\alpha,m} \) be the critical value defined by

\[ \mathbb{P}\left(-2 \sum_{i=1}^{m} \log(P_i) \geq t_{\alpha,m}\right) = \alpha. \]

Then there exists \( n \in \mathbb{N} \) such that

\[ \mathbb{P}\left(-2 \sum_{i=1}^{m} \log(P_i) \geq t_{\alpha,m}\right) \leq \alpha, \]

for any \( m \geq n \).

Hence, we can dispense with the conservative correction if \( m \) is large enough and the \( P_i \) are identically distributed. A formal proof is given in the appendix. Since \( \mathbb{E}\{-\log(P_i)\} \leq \mathbb{E}\{-\log(U_i)\} \), from the definition of the convex order, a direct application of the law of large numbers gets us
most of way, except for the possibility \( E \{- \log(P_i)\} = E \{- \log(U_i)\} \). In fact this exception is no problem because, surprisingly, it implies that the \( P_i \) are uniform, see [Shaked and Shanthikumar 2007, Theorem 3.A.43] or Lemma 7 in the appendix.

In the finite, non-identically distributed case we are able to derive three probability bounds. None beats the other two uniformly for all \( m \) and all significance levels (see Figure 3), but of course in practice we can use the minimum.

**Lemma 4.** Let \( P_1, \ldots, P_m \) be a sequence of independent sub-uniform random variables. Then for \( x \geq 2m \),

\[
P \left( -2 \sum_{i=1}^{m} \log(P_i) \geq x \right) \leq \min \left[ S_{2m} \left( x - 2m \log 2 \right), \right.
\]

\[
\left. m / \left[ m + \left\{ \left( x - 2m \right) / 2 \right\} \right], \exp \left\{ m - x / 2 - m \log(2m/x) \right\} \right],
\]

where \( S_k \) is the survival function of a \( \chi^2 \) variable with \( k \) degrees of freedom.

The first uses the \( 2\alpha \) bound directly (Corollary 1). The second uses bounds on the mean and variance of \( - \log(P_i) \) (given in Lemma 7 in the appendix) and then applies the Chebyshev-Cantelli inequality. The third is based on a bound on the moment generating function of \( - \log(P_i) \). The derivation details are in the appendix.

Figure 3 presents the behaviour of the different bounds under different conditions. For a given \( m \) (20 on the left and a billion on the right) and \( \alpha \), we compute the critical value, \( t_{\alpha,m} \). The curves show the bound given by each formula, i.e. inputting \( x = t_{\alpha,m} \) in Lemma 4 as \( \alpha \) ranges from \( 10^{-5} \) to 0.1. For low \( \alpha \), the bound based on the moment generating function, marked MGF, is by far the best. For large \( m \), the bound based on multiplying every p-value by two, essentially the method we are trying to beat, performs very poorly.

The most significant p-value, \( \min(P_i) \), is another useful statistic. Here, it turns out that we cannot improve over doubling the value (which would be the result of doubling every \( P_i \)). Interestingly, this results in a simple asymptotic rule for how to correct the corresponding overall p-value.

**Lemma 5.** Let \( P_1, \ldots, P_m \) and \( U_1, \ldots, U_m \) each be sequences of independent sub-uniform and uniform random variables on \([0, 1] \), respectively. For \( x \in [0, 1] \), let \( q = 1 - (1 - x)^m = P(\min(U_i) \leq x) \). For
\[ \Pr\{\min(P_i) \leq x\} \leq 1 - (1 - 2x)^m, \]
\[ = 1 - \{2(1 - q)^{1/m} - 1\}^m, \]
which is no larger than \(2q\) and tends to \(2q - q^2\) as \(m \to \infty\). Furthermore, this bound is achievable if the \(P_i\) are independent and identically distributed.

**Proof.** Let \(F_i\) denote the distribution function of \(P_i\). Then
\[ \Pr\{\min(P_i) \leq x\} = 1 - \prod_i \{1 - F_i(x)\} \]
\[ \leq 1 - (1 - 2x)^m \]
\[ = 1 - \{2(1 - q)^{1/m} - 1\}^m \]
\[ = 1 - (1 - q)(2 - (1 - q)^{-1/m})^m \]
\[ \to 1 - (1 - q)^2 = 2q - q^2 \]
using Corollary 1 in the second line (and the fact that the bound is achievable), and the formulae
\[ (1 + a/m)^m \to \exp(a) \text{ and } m(a^{1/m} - 1) \to \log(a) \]
in the fifth line. The expression \(\{2(1 - q)^{1/m} - 1\}^m\)
is an increasing function of \((1 - q)^{1/m}\), which is itself increasing in \(m\), therefore the composition is increasing. Hence, \(1 - \{2(1 - q)^{1/m} - 1\}^m\) attains its maximum at \(m = 1\), where it is \(2q\). \(\square\)

## 4 Estimation schemes

In practice the posterior predictive p-value will often be estimated by simulation. We now characterise the distribution of the estimate. Assume that, for any \(D\), we can sample a sequence \(\theta_1, \ldots, \theta_M\), that may or may not be dependent, from the posterior distribution of \(\theta\) given \(D\). Furthermore, for any \(\theta_i\), we can simulate a replicate dataset \(D_i^\ast\) independently. These are fairly usual conditions. A typical reason for the \(\theta_i\) to be dependent is if they are consecutive states from a Markov chain Monte Carlo algorithm.

Suppose \(M = 1\) in \(2\). \(P\) is estimated from one indicator, \(\hat{P}_1 = I\{f(D_1^\ast, \theta_1) \geq f(D, \theta_1)\}\). Since \(f(D^\ast, \theta_1)\) and \(f(D, \theta_1)\) are identically distributed, marginally, \(\hat{P}_1\) is a Bernoulli random variable with success probability 1/2 (remember \(f(D, \theta)\) is absolutely continuous). This not a sub-uniform random variable; in fact, with respect to the convex order, \(\hat{P}_1\) is the maximal random variable that has mean 1/2 and support on [0, 1] \cite{ShakedandShanthikumar2007} (Theorem 3.A.24). Although the point is somewhat pedantic, for any fixed and finite \(M\) the calculation \(2\) will usually return identically zero or one with some positive probability, so that estimate is rarely strictly sub-uniform.

Instead, suppose it is possible to compute \(\Pr\{f(D^\ast, \theta_i) \geq f(D, \theta_i)|\theta_i, D\}\) for any \(\theta_i\) and \(D\), and consider the alternative estimate, also mentioned in \cite{Gelmanetal1996} (Section 2.3),
\[ Q_M = \frac{1}{M} \sum_{i=1}^M \Pr\{f(D_i^\ast, \theta_i) \geq f(D, \theta_i)|\theta_i, D\}. \]

Intuitively, this estimate should do better because it is as if an infinite number of draws of \(D_i^\ast\) were made for every \(\theta_i\). Again, suppose \(M = 1\). Viewed over the joint distribution of \(\theta\) and \(D\), the variable \(Q_1\) is a uniform random variable over [0, 1] (compare to \(\hat{P}_1\) which was Bernoulli). To see this, first note that the random variable \(\Pr\{f(D^\ast, \theta_i) \geq f(D, \theta_i)|\theta_i, D\}\) is distributionally identical to \(\Pr\{f(D^\ast, \theta) \geq f(D, \theta) | \theta, D\} = W\), say. Then the conditional random variable \(W | \theta\) is uniform (for the same reason any classical p-value is uniform). Therefore \(W\) is also uniform marginally.

The estimate \(Q_M\) is an average of uniform random variables which, regardless of any independence, must be sub-uniform \cite{ShakedandShanthikumar2007} (Theorem 3.A.36). Therefore, remarkably, much of the stochastic behaviour of \(Q_M\) can also be understood by the methods of this article. We have shown:

**Lemma 6.** Let \(f\) be a function of \(D\) and \(\theta\), which in turn have a joint distribution such that \(f(D, \theta)\) is an absolutely continuous random variable. For a fixed \(M\), let \(\theta_1, \ldots, \theta_M\) be replicates of
θ given D, with arbitrary dependence, and let \( D^*_i \) be an independent replicate of D given \( \theta_i \), for \( i = 1, \ldots, M \). Then the estimate \( Q_M \), defined in (3), is sub-uniform. In particular, \( \Pr(Q_M \leq \alpha) \leq 2\alpha \), for \( \alpha \in [0, 1/2] \).

5 Conclusion

We have shown that the family of distributions that are less variable than uniform on \([0,1]\), in the convex order, fully characterise the frequency behaviour of posterior predictive p-values. From the properties of this order we established various probability bounds that can be used for conservative testing. Most of the resulting recommendations are really quite simple, e.g., multiply the p-value by two or, Fisher’s method is asymptotically conservative.

There are other approaches to Bayesian model-checking, such as partial (Bayarri and Berger 2000) or recalibrated (Hjort et al. 2006) predictive p-values, which circumvent any need for bounds by creating a perfectly uniform statistic. Of course these methods have their own problems (e.g. a computational burden) but they do address an issue that remains largely unsolved in this article, which is that for everyday models and data, posterior predictive p-values seem to be very conservative. One of key objectives in the future has to be to find simply identifiable subclasses of models and tests for which our bounds can be reduced. For example, we conjecture that the p-values of Rubin (1984), Equation (4), which do not allow the test to depend on the parameter, can be bounded differently.

Acknowledgements

PRD is funded by the Wellcome Trust and Royal Society on Grant Number WT104125AIA. DJL is funded by the Heilbronn Institute for Mathematical Research.

Appendix

Proof of Theorem 2. It is straightforward to prove (and already known) that the existence of the martingale representation implies the convex order, by Jensen’s inequality. Here we focus on the converse statement. We will rely on the properties of integrated distribution functions, given at the beginning of Section 2.2.

Let \( \phi_X \) and \( \phi_Y \) be the integrated distribution functions of \( \mu \) and \( \nu \) respectively, so that \( \phi_X(x) \leq \phi_Y(x) \) for \( x \in \mathbb{R} \). If \( \phi_X(x) = \phi_Y(x) \) for all \( x \) then let \( Y \mid X = X \) and the proof is finished. Otherwise, because both functions are continuous the set \( \{ x \in \mathbb{R} : \phi_X(x) < \phi_Y(x) \} \) can be partitioned into a countable set of open intervals \( C_i, i \in \mathbb{N} \). Consider one such interval, \( C = (a,b) \) (allowing \( a = -\infty \) and \( b = \infty \)). First we show that it is possible to construct a linear interpolation of \( \phi_X \) over \( C \), denoted \( \phi_X^* \), at a set of points of \( \mu \)-measure 0 such that \( \phi_X^*(x) \leq \phi_Y(x) \) for \( x \in C \). Choose a point \( x_0 \in C \) of \( \mu \)-measure 0 and fix some \( \beta \in (0,1) \). We construct the interpolating points \( x_j, j \in \mathbb{Z} \) recursively from \( x_0 \). We show how to construct \( x_1 \) from \( x_0 \), then \( x_2 \) from \( x_1 \) and so on. The interpolating points \( x_{-1}, x_2, \ldots \) are created similarly. For \( j \in \mathbb{N} \) let

\[
x'_j + 1 = \sup \{ x \in [x_j, b) : \forall \alpha \in [0,1] : \alpha \phi_X(x_j) + (1 - \alpha)\phi_X(x) \leq \phi_Y(\alpha x_j + (1 - \alpha)x) \}.
\]

If \( x'_j + 1 = \infty \), which is only possible if \( b = \infty \), let \( x_{j+1} = x'_j + 1 = b = \infty \). Otherwise choose \( x_{j+1} \) to be a point in \( [x'_j + 1 - \beta(x'_j + 1 - x_j), x'_j + 1] \) such that \( \mu(\{x_{j+1}\}) = 0 \). Stop the procedure if \( x_{j+1} = b \). We claim that for any \( x \in [x_0, b) \), sup \( \{ j \in \mathbb{N}_0 : x_j \leq x \} < \infty \). Otherwise, for any \( \epsilon > 0 \) there would exist \( j \in \mathbb{N} \) such that \( x'_j + 1 - x_j \leq \epsilon \) and a solution for \( \alpha \) to \( \alpha \phi_X(x_j) + (1 - \alpha)\phi_X(y) \geq \phi_Y(\alpha x_j + (1 - \alpha)y), \) where \( y = \min(x_j + 2\epsilon, x) \). Then \( \phi_Y(x_j) \leq \phi_Y(y) \leq \phi_X(x_j) + (y - x_j) \), first using the fact that both \( \phi_X \) and \( \phi_Y \) are non-decreasing and then using \( \phi_X^* \leq 1 \). This implies \( \phi_Y(x_j) \leq \phi_X(x_j) \leq 2\epsilon \). Therefore the functions \( \phi_X \) and \( \phi_Y \) would come arbitrarily close to each other over the closed interval \([x_0, x]\). Since both are continuous, by the extreme value theorem we would have \( \phi_X(z) = \phi_Y(z) \) for some \( z \in [x_0, x] \), which is impossible since \( z \in C \).
By a similar construction we form \( x_{-1}, x_{-2}, \ldots \) The set of all intervals \( (x_j, x_{j+1}) \) constructed for every \( C_i \) is countable. Denote these by \( I_n = (l_n, u_n), \) \( n \in \mathbb{N} \), let \( S = \mathbb{R} \setminus (\cup I_n) \) and finally define the Markov kernel from \( \mathbb{R} \) onto \( \mathbb{R} \),

\[
K(x, dy) = \begin{cases} 
\delta_x & x \in S, \\
k_n(x, dy) & x \in I_n,
\end{cases}
\]

where \( \delta_x \) denotes the point mass at \( x \), and \( k_n(x, dy) \) is a Markov kernel with the following properties. For every \( x \in I_n \), \( k_n(x, dy) \) is absolutely continuous, \( \int_{I_n} k_n(x, dy) = 1 \) and \( \int_{I_n} yk_n(x, dy) = x \). Furthermore, for any measurable set \( A \subseteq I_n \) such that

\[
\int_{I_n} k_n(x, A) \mu(dx) = 0,
\]

there is no \( p \) in the support of \( \mu \) such that \( k_n(p, A) > 0 \).

An example of an admissible choice for \( k_n \) would be for \( k_n(x, dy) \) to be a uniform distribution over the interval centered at \( x \) with length \( 2 \min(x-l_n, u_n-x) \). To see this, suppose that \( k_n(p, A) = v > 0 \) for \( p \) in the support of \( \mu \). It is clear from our choice of \( k_n \) that there is an open neighbourhood \( N \) of \( p \) for which \( \sup_Y |k_n(p, Y) - k_n(x, Y)| \leq v/2 \) for any \( x \in N \), the supremum taken over sets \( Y \) in the \( \sigma \)-algebra of \( \mu \). Therefore,

\[
\int_{I_n} k_n(x, A) \mu(dx) \geq \int_N k_n(x, A) \mu(dx) \geq v/2 \int_N \mu(dx) > 0.
\]

The last inequality comes from \( N \) being an open neighbourhood of a supported point.

Let \( \hat{X} \) be the random variable that results from applying \( K \) to \( X \). We now show \( \hat{\mu} \leq_{cx} \nu \) where \( \hat{\mu} \) is the probability measure of \( \hat{X} \). For any \( x \in S \) the kernel does not allow movement from the right to the left or the left to the right of \( x \). Therefore, for \( x \in S \),

\[
\phi\hat{X}(x) = E\{(x - \hat{X})_+\} = P(\hat{X} \leq x)E(x - \hat{X} \mid \hat{X} \leq x) = E\{(x - X)_+\} = \phi_X(x),
\]

using the fact that \( K \) is mean-preserving. For \( x \in I_n \) the convexity of \( \phi_X \) implies

\[
\phi\hat{X}(x) \leq \frac{(x - l_n)}{(u_n - l_n)} \phi\hat{X}(l_n) + \frac{(u_n - x)}{(u_n - l_n)} \phi\hat{X}(u_n) = \frac{(x - l_n)}{(u_n - l_n)} \phi_X(l_n) + \frac{(u_n - x)}{(u_n - l_n)} \phi_X(u_n) \leq \phi_Y(x),
\]

using \( l_n, u_n \in S \) for the equality and the construction of \( l_n \) and \( u_n \) for the second inequality. Since \( \hat{\mu} \) has the same expectation as \( \mu \) and therefore \( \nu \), we conclude \( \hat{\mu} \leq_{cx} \nu \) (Shaked and Shanthikumar 2007, p. 110).

Finally, by Strassen’s theorem there exists a random variable \( Y \) with probability measure \( \nu \) such that \( E(Y \mid \hat{X}) = \bar{X} \). This random variable satisfies \( E(Y \mid X) = X \). For \( p \in S \), the random variable \( Y \mid X = p \) is singular except potentially at the set of interpolating points \( \{x_n, n \in \mathbb{N} \} \) which was constructed to have \( \mu \)-measure 0. Suppose there exists a supported point \( p \in I_n \) and \( q \in \mathbb{R} \) such that \( E(\delta_q \mid X = p) > 0 \). Then since

\[
E(\delta_q \mid X = p) = \int E(\delta_q \mid \hat{X} = y) k_n(p, dy),
\]

there must exist a set \( A \subseteq I_n \) such that \( k_n(p, A) > 0 \) and \( E(\delta_q \mid \hat{X} \in A) > 0 \). Since \( Y \) is absolutely continuous we also have \( \hat{\mu}(A) = 0 = \int_{I_n} k_n(x, A) \mu(dx) \) violating the construction of \( k_n \). Hence there are no supported points in \( \cup I_n \) only potentially a set of \( \mu \)-measure 0 in \( S \) such that \( Y \mid X \) is neither singular nor absolutely continuous. \( \square \)
Proof of Lemma 2. Let $\phi_X$ and $\phi_Y$ denote the integrated distribution functions of $X$ and $Y$ respectively. The function $\phi_X$ is non-negative, continuous and convex, therefore the set \( \{ w : w(x - p) \leq \phi_Y(x), x \in \mathbb{R} \} \) is non-empty (it contains 0) and closed. Hence, the maximum in \( \ref{eq:prooflem2} \) is well-defined. For $x \in \mathbb{R}$ we have

\[
F_X(\alpha)(x - \alpha) = \phi_X^+(\alpha)(x - \alpha) \\
\leq \phi_X(\alpha) + (x - \alpha)\phi_X^+(\alpha) \\
\leq \phi_X(x) \leq \phi_Y(x),
\]

using the non-negativity and convexity of $\phi_X$. Hence, $F_X(\alpha) \leq h$. If $h = 1$ then the singular random variable $\tilde{X} = E(Y)$ satisfies $\tilde{X} \leq_{cx} Y$ and $F_{\tilde{X}}(\alpha) = h$. Otherwise, the set \( \{ x \in \mathbb{R} : h(x - \alpha) \leq \phi_Y(x) \} \) is closed and non-empty, again containing 0. Therefore $\beta = \max\{ x : h(x - \alpha) \leq \phi_Y(x) \}$ is well-defined and satisfies $h(\beta - \alpha) = \phi_Y(\beta)$. If $h \leq 1$, consider

\[
\phi(x) = \begin{cases} 
0 & x \leq \alpha, \\
\frac{h(x - \alpha)}{\alpha} & \alpha \leq x \leq \beta, \\
\phi_Y(x) & \alpha \geq \beta.
\end{cases}
\]

This is a valid integrated distribution function, in particular, it is convex because $\phi_Y^+(\beta) \geq h$ (otherwise $\phi_Y$ and $h(x - \alpha)$ would cross). Moreover, $\phi(x) \leq \phi_Y(x)$ and $\lim_{x \to \infty} \{ \phi(x) - \phi_Y(x) \} = 0$. Let $\tilde{X}$ be a random variable with integrated distribution function $\phi$. Then $X \leq_{cx} Y$, and $F_X(\alpha) = \phi^+(\alpha) = h$.

The proofs of lemmas 3 and 1 both need the following result.

Lemma 7. Let $P$ be a sub-uniform probability measure. Then either i) $P$ is uniform on $[0,1]$ or ii) $E\{-\log(P)\} < E\{-\log(U)\} = 1; \quad \text{var}\{-\log(P)\} < \text{var}\{-\log(U)\} = 1,$

where $U$ is a uniform random variable on $[0,1]$.

Proof. Shaked and Shanthikumar (2007, Theorem 3.A.43) provide the following theorem. If $X \leq_{cx} Y$ and for some strictly convex function $h$ we have $E\{h(X)\} = E\{h(Y)\}$ then $X$ is distributed as $Y$. The function $-\log(x)$ is strictly convex, therefore either $P$ is uniform or $E\{-\log(P)\} < E\{-\log(U)\}$. If the latter is true, then

\[
\text{var}\{-\log(P)\} = E[-\log(P) - E\{-\log(P)\}]^2 \\
< E[-\log(P) - E\{-\log(U)\}]^2, \\
\leq E\{\log(U) + 1\}^2 \\
= \text{var}\{-\log(U)\}
\]

In the second line, the fact that the expected square distance from the mean is smaller than from any other point is used, and in the fourth we used the fact that $(\log(x) + 1)^2$ is convex.

Proof of Lemma 3. Let $X_i = -2\log(P_i)$, $\mu_X = E(X_i)$, $Y_i = -2\log(U_i)$, and $E(Y_i) = \mu_Y$. If $\mu_X = \mu_Y$ then by Lemma 7 the $P_i$ are uniform on $[0,1]$ and we are done. The statement is also true if $\alpha = 1$. Therefore assume $\mu_X < \mu_Y$, $\alpha \in (0,1)$ and let $t \in (\mu_X, \mu_Y)$. By the weak law of large numbers there exists an $n' \in \mathbb{N}$ such that, for $m \geq n'$,

\[
P\left(\sum_{i=1}^{m} Y_i \geq mt\right) \geq \alpha,
\]

so that $t_{\alpha,m} \geq mt$. Therefore, for $m \geq n'$,

\[
P\left(\sum_{i=1}^{m} X_i \geq t_{\alpha,m}\right) \leq P\left(\sum_{i=1}^{m} X_i \geq mt\right).
\]

Again by the law of large numbers, the right-hand side tends to zero. Hence there exists an $n \geq n'$ such that it is bounded by $\alpha$ for $m \geq n$.  

\[13\]
Proof of Lemma 4. Let $R_m = -2 \sum \log(P_i)$. From Corollary 1 we have $U_i/2 \leq_{st} P_i$, for $i = 1, \ldots, m$, where $U_1, \ldots, U_m$ are independent uniform random variables on $[0, 1]$ and $\leq_{st}$ denotes the usual stochastic order [Shaked and Shanthikumar, 2007, Chapter A.1]. This implies $-\log(P_i) \leq_{st} -\log(U_i/2)$. Because the usual stochastic order is closed under convolution (Shaked and Shanthikumar, 2007, Theorem 1.A.3), we have $R_m \leq_{st} -2 \sum \log(U_i) + 2m \log 2$. The sum $-2 \sum \log(U_i)$ has a $\chi^2$ distribution with $2m$ degrees of freedom, proving the first bound. Lemma 7 implies $E(R) \leq 2m$ and $\text{var}(R) \leq 4m$. Therefore, using Cantelli’s inequality,

$$
P[R_m \geq x] \leq \frac{\text{var}(R)}{\text{var}(R) + \{x - E(R)\}^2}
$$

for $x \geq 2m$. This proves the second bound. Finally, the moment generating function of $R_m$ is $E(\exp(tR_m)) = \prod E(P_i^{-2t})$ for $t \geq 0$. For $t \in [0, 1/2)$ each $E(P_i^{-2t}) \leq E(U^{-2t}) = (1 - 2t)^{-1}$ since $x^{-2t}$ is a convex function in $x$ for $x \in [0, 1]$. Using Markov’s inequality,

$$
P(R_m \geq x) = P(\exp(tR_m) \geq \exp(tx))
\leq \exp(-tx) E(\exp(tR_m))
\leq \exp(-tx - m \log(1 - 2t)),
$$

for $t \in [0, 1/2)$. The minimum of this function is at $t = 1/2 - m/x$, giving the third bound. 

References

Bayarri, M. and Berger, J. O. (2000). P values for composite null models. Journal of the American Statistical Association, 95(452):1127–1142.

Dahl, F. A. (2006). On the conservativeness of posterior predictive p-values. Statistics & probability letters, 76(11):1170–1174.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4):733–760.

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). Post-processing posterior predictive p values. Journal of the American Statistical Association, 101(475):1157–1174.

Meng, X.-L. (1994). Posterior predictive p-values. The Annals of Statistics, 22(3):1142–1160.

Mosteller, F. and Fisher, R. A. (1948). Questions and answers. The American Statistician, 2(5):pp. 30–31.

Müller, A. and Rüschendorf, L. (2001). On the optimal stopping values induced by general dependence structures. Journal of applied probability, 38(3):672–684.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics, 12(4):1151–1172.

Rubin, D. B. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4):733–760. Discussion of Gelman, Meng and Stern.

Shaked, M. and Shanthikumar, J. G. (2007). Stochastic orders. Springer.

Strassen, V. (1965). The existence of probability measures with given marginals. The Annals of Mathematical Statistics, 36(2):423–439.

Tsui, K.-W. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. Journal of the American Statistical Association, 84(406):602–607.