Invariant P-values for Model Checking and Checking for Prior-data Conflict

by

Michael Evans
Department of Statistics
University of Toronto

and

Gun Ho Jang
Department of Statistics
University of Toronto

Technical Report No. 0803, June 24, 2008

TECHNICAL REPORT SERIES
UNIVERSITY OF TORONTO
DEPARTMENT OF STATISTICS

Invariant P-values for Model Checking and Checking for Prior-data Conflict

Michael Evans and Gun Ho Jang
Department of Statistics
University of Toronto

Abstract: P-values have been the focus of considerable criticism based on various considerations. Still the P-value represents one of the most commonly used statistical tools. When we are assessing the suitability of a single hypothesized distribution, it is not clear that there is a better choice of a measure of surprise. This paper is concerned with the definition of appropriate model-based P-values for model checking and checking for prior-data conflict.

Key words and phrases: P-values, invariance under transformations, discrepancy measures for model checking, checking for prior-data conflict.

1 Introduction

The use of P-values is common in statistical practice. Despite this it is reasonable to say that the logical foundations for the P-value are somewhat weak. This has lead to a variety of criticisms of P-values and even to doubts as to their correctness. The purpose of this paper is to examine the foundations of the P-value concept and attempt to provide a version of the P-value that addresses at least some of the issues raised concerning their validity.

The following situation arises in many statistical contexts and could be considered almost the archetypal statistical problem. Suppose we observe a value $x_0 \in \mathcal{X}$ and this value was presumed to have been generated via a prescribed probability measure P on $\mathcal{X}$. The question of interest is then: given the evidence presented by x_0, is P a reasonable choice? In certain situations we could answer this with a categorical no, e.g., suppose that P concentrates on C and $x_0 \notin C$. While this can arise, it is typical in applications that x_0 is a possible value from P but, if x_0 is in a region where P assigns relatively little probability, then we feel we have evidence against P. Note that x_0 not occurring in such a region is not evidence in favor of P, as there are many probability distributions with this property and we are not selecting among them. In general, we are only looking here for evidence that suggests that a specific P is inappropriate.

As an example of this, consider model checking where P corresponds to the conditional distribution of the data given a minimal sufficient statistic, or where P corresponds to the distribution of an ancillary statistic. Then evidence against P is evidence against assumptions we have made as part of a statistical analysis. Model checking is an important and necessary part of statistical analyses. In Bayesian analyses P could correspond to the prior predictive distribution of a minimal sufficient statistic given an ancillary and, as discussed in Evans and Moshonov (2006, 2007), we want to assess whether or not the observed value of the minimal sufficient statistic is a reasonable value from this distribution. This is a check as to whether or not the prior is in conflict with the data.

In general then, we are looking for a measure of how surprising the observed value x_0 is as a possible value from P. A common approach to this is to say that we need to prescribe a real-valued discrepancy statistic $T : \mathcal{X} \to R^1$, so that in some sense $T(x)$ measures how divergent the value x is, and then compute the P-value

$$P(T(x) \geq T(x_0)). \tag{1}$$

If this P-value is small, then we interpret this as evidence that x_0 is a surprising value and so we have evidence against P. In general, no guidance is provided as to how the statistic T is to be chosen with respect to P. Further, it can be noted that some restrictions on T are necessary if (1) is to have an appropriate interpretation. In particular, the right tail of the distribution of T should be the only region that has relatively low probability. Otherwise, we could have an extreme value of $T(x_0)$ in the left tail, or a value that occurs near a shallow anti-mode, that lead to a reasonable value of (1) and yet $T(x_0)$ could still be considered as surprising and so evidence against P.

As an example of this consider the situation where $\#(\mathcal{X}) = k < \infty$, x_0 corresponds to a sample of n and we use the Chi-squared statistic T as the discrepancy statistic. Then, for n large, (1) corresponds to the right-tail of a Chi-squared$(k-1)$ distribution. If k is also large, however, a value of T in the left-tail does not provide evidence against P via (1), even though we know that it is very unlikely for P to have produced such a value. We will see in Example 1, however, that this apparent ambiguity can be explained when we are more careful about defining P-values.

The choice of the discrepancy statistic T also poses some problems. It seems clear that the choice of such a statistic should be made prior to seeing the data. Further, if we think of choosing T to check for some particular discrepancy, then there seems to be no reason why we should be restricted to a specific choice but may well have $T = (T_1, \ldots, T_m)'$ where the T_i are different discrepancy statistics. Clearly, it is more appropriate then to compare $T(x_0)$ to P_T rather than compute (1) for each T_i, i.e., the dependencies among the T_i should be taken into account when making the assessment, as $T(x_0)$ may be a surprising value even when each $T_i(x_0)$ isn't. We note, however, that (1) does not tell us how to proceed when we have multiple discrepancy statistics.

We will subsequently argue that there are contexts where we do not want the statistician to be free to choose the discrepancy measures and their associated

P-values, but want them to be determined by the model. This arises in checking for prior-data conflict and we will discuss this issue in section 5. In section 2 we discuss a particular context where the computation of an appropriate P-value seems uncontroversial. We argue that this is the basis for the general development of P-values for the problems considered in this paper. In sections 3 and 4 we develop a general approach for the computation of P-values based on an observed measured response. The central idea is that volume distortions must not affect the computation of P-values. We will refer to such a P-value as a *model based P-value*. In section 6 we discuss some computational issues.

Many other criticisms of P-values are often cited. In particular, a common complaint, is that in reality the data is never distributed exactly as P, but P may be adequate for the application at hand in the sense that it provides a good approximation. If we observe enough data, however, any reasonable P-value will detect the discrepancy and lead to evidence against P. Of course, this cannot be viewed as a criticism of the P-value in question, as it is doing the right thing. Rather it suggests that in such problems we really do have to say what size of discrepancies are meaningful and then assess whether or not the discrepancy detected is to be taken seriously. So the P-value is not the end of the story in model assessment and cannot cover-up modelling inadequacies, namely, situations where we really can't say what discrepancies are meaningful. We do not view the necessity of taking into account practical significance, as opposed to statistical significance, as a criticism of the P-value.

In Schervish (1996) the use of certain frequentist P-values was examined as a measure of support for a hypothesis. The analysis there demonstrated convincingly that there is little logical support for this. As mentioned, we are using P-values as measures of surprise, not support, and restricting our discussion to the situation where we have a single P. In Berger and Delampady(1987) and Berger and Selke(1987) comparisons are made between frequentist P-values and Bayesian measures of evidence in the context of hypothesis testing, i.e., assessing the evidence in favor of a point null hypothesis $H_0 = \{\theta_0\} \subset \Theta$. Arguments are presented there in favor of the Bayesian measures. For our discussion here, however, we are restricting to a single P and agree that we might proceed very differently in situations where alternatives to P are prescribed, i.e., the computation of the P-values we discuss here may well not be appropriate in such situations, as we have more information available.

2 Model-based P-values with Discrete P

Suppose that P is discrete with probability function given by $p(x) = P(\{x\})$. Then an obvious model-based P-value for checking whether or not x_0 is a surprising value from P is given by

$$P(p(x) \leq p(x_0)). \tag{2}$$

We see that (2) is the probability of observing a value whose probability of occurrence is no greater than the probability of occurrence of the observed x_0.

If (2) is small then it seems clear that x_0 is surprising and we have evidence against P.

Note that the appropriate inequality in (2) is less than or equal to, as we want no evidence against x_0 when P is uniform. Further, we see that (2) handles values in either tail, values that lie between modes and also multidimensional x. It seems reasonable to refer to (2) as a pure *model-based* P-value. Values that are surprising are identified by the model and not by the statistician's choices.

We may, however, have a discrepancy statistic T of interest. Then P_T is discrete with probability function p_T and the model-based P-value induced by T is given by

$$P_T(p_T(t) \leq p_T(T(x_0))). \tag{3}$$

For example, suppose that $C \subset \mathcal{X}$ is such that $P(C)$ is very small. Then, with $T = I_C$, (3) equals $P(C)$ and we have evidence against P when $x_0 \in C$.

We see that (2) determines whether or not x_0 is a surprising value based solely on the smallness of its probability of occurrence when compared to the probabilities of occurrence of other values. If $p(x_0)$ is very small compared to the other possibilities, then (2) seems like an appropriate measure of surprise. Consider a situation, however, where $\mathcal{X}$ has a large finite cardinality k and $p(x_0) = (1 \pm \epsilon)/k$, $p(x) = 1/k \mp \epsilon/k(k-1)$ otherwise and ϵ is very small. In the first case the P-value is $1-(1+\epsilon)/k$ and we have no evidence against P, while in the second case it is $(1-\epsilon)/k$ and we have evidence against P. So even though the probability distributions are very similar, the P-values are quite different. This points to the need generally to consider discrepancy statistics as checks on P rather than relying solely on (2).

We now consider an important example.

Example 1. *Multinomial*$(1, \theta_1, \ldots, \theta_k)$

Suppose we observe a sample $x_{10}, \ldots, x_{n0}$ that is supposed to have come from the Multinomial$(1, \theta_1, \ldots, \theta_k)$ distribution where $\theta_1, \ldots, \theta_k$ are known values that are all positive. Then, denoting the cell counts by $T(x_{10}, \ldots, x_{n0}) = (t_{10}, \ldots, t_{k0})$, the P-value (2) is given by

$$P_T(\theta_1^{t_1} \cdots \theta_k^{t_k} \leq \theta_1^{t_{10}} \cdots \theta_k^{t_{k0}}) \tag{4}$$

where P_T is the Multinomial$(n, \theta_1, \ldots, \theta_k)$ distribution. We can write (4) as

$$P_T\left(\sum_{i=1}^{k} (\ln \theta_i)\sqrt{\theta_i(1-\theta_i)}\left(\frac{t_i - n\theta_i}{\sqrt{n\theta_i(1-\theta_i)}}\right) \leq \sqrt{n}\sum_{i=1}^{k}(\ln \theta_i)\left(t_{i0}/n - \theta_i\right)\right).$$

Putting $\sigma^2 = \sum_{i=1}^{k}(\ln \theta_i)^2\theta_i(1-\theta_i) - 2\sum_{i<j}(\theta_i \ln \theta_i)(\theta_j \ln \theta_j)$, then

$$\sum_{i=1}^{k} (\ln \theta_i)\sqrt{\theta_i(1-\theta_i)}\left(\frac{t_i - n\theta_i}{\sqrt{n\theta_i(1-\theta_i)}}\right) \xrightarrow{D} N\left(0, \sigma^2\right)$$

so (4) converges in probability to $\Phi(\sqrt{n}\sum_{i=1}^{k}(\ln \theta_i)\left(t_{i0}/n - \theta_i\right)/\sigma)$ and the joint asymptotic normality of the t_{i0}/n implies that (4) is asymptotically uniform, when the probabilities $\theta_1, \ldots, \theta_k$ are correct.

Now observe that generally $\sum_{i=1}^{k}(\ln\theta_i)\,(t_{i0}/n-\theta_i) \stackrel{a.s.}{\to} \sum_{i=1}^{k}(\ln\theta_i)\,(p_i-\theta_i)$ for some p_i, and so we will find evidence against the probabilities $\theta_1,\ldots,\theta_k$, for large enough n, whenever $\sum_{i=1}^{k}(\ln\theta_i)\,(p_i-\theta_i) < 0$. Note that this holds whenever the expected value of $-\ln\theta_i$ under the true distribution is greater than the entropy of the assumed distribution. If we take $E(-\ln\theta_i)$ as a measure of diffuseness of a Multinomial$(1,p_1,\ldots,p_k)$ distribution, then this says we will inevitably find evidence against P whenever the true distribution is more diffuse but not otherwise. For example, when $k=2$ and $\theta\neq 1/2$, this is equivalent to $p\ln(\theta/(1-\theta)) < \theta\ln(\theta/(1-\theta))$ which occurs when $\theta>1/2$ and $p<\theta$ or when $\theta<1/2$ and $p>\theta$. When each $\theta_i=1/k$, then we will never find evidence against the uniform distribution and this makes sense as the uniform distribution is the most diffuse distribution.

The check based on (4) will only detect certain discrepancies and this is true of most discrepancy statistics. Of course, we can also consider other discrepancy statistics and perhaps it is natural to consider T itself. So in this case we need to evaluate

$$P_T\left(p_n(t)\le p_n(t_0)\right) \tag{5}$$

where $p_n(t) = \binom{n}{t_1\ldots t_k}\theta_1^{t_1}\cdots\theta_k^{t_k}$. In the Appendix we provide a proof of the following result.

Theorem 1. Suppose that $\theta_i\neq 0$ for $i=1,\ldots,k$ and we have a sample of n from a Multinomial$(1,p_1,\ldots,p_k)$ distribution. Then, as $n\to\infty$,
(i) when $\theta_i=p_i$ for all i,

$$-\ln p_n(t)-\frac{1}{2}\sum_{i=1}^{k}\ln\theta_i+\frac{k-1}{2}\ln 2\pi n \stackrel{P}{\to}\sum_{i=1}^{k}\frac{(t_i-n\theta_i)^2}{2n\theta_i}$$

and so has a limiting Chi-squared$(k-1)$ distribution,
(ii) when $\theta_i=p_i$ for all i, the P-value (5) converges in probability to
$P\left(X\ge\sum_{i=1}^{k}(t_{i0}-n\theta_i)^2/n\theta_i\right)$ where $X\sim$ Chi-squared$(k-1)$,
(iii) when $\theta_i\neq p_i$ for some i, the P-value (5) converges in probability to 0.

We note that (iii) says that the model-based P-value based on T will always detect when the assumed distribution is wrong, provided n is large enough. Also, we see that the Pearson Chi-squared test statistic arises directly as an approximation when computing the model-based P-value (5) and this adds support for the use of this statistic. In the typical development of the Chi-squared test, the statistic is developed via intuition and then its asymptotic distribution is derived using the delta theorem while, in Pearson (1900), the statistic is developed from the quadratic form in the exponential of a multivariate normal density approximating the multinomial distribution. Finally, we see that computing the right-tail probability for the Chi-squared test is the correct approximation to (5). So, for large n, if the Chi-squared statistic is small, then (5) is large. Bayesian uses of the Chi-squared test statistic for model checking are discussed in Johnson (2004).

3 Model-based P-values with General P

Additional considerations arise when P is a continuous measure. We restrict our attention to the absolutely continuous case so that P has density f with respect to to a support measure μ. The natural analog of (2) is then

$$P(f(x) \leq f(x_0)) \tag{6}$$

and this is the probability of observing a value whose density is no greater than the density of the observed x_0.

The P-value (6), however, has a disturbing feature. Suppose we change the support measure from μ to ν where $\nu(A) = \int_A g(x)\,\mu(dx)$ for some integrable, nonnegative g. Then the density of P with respect to ν is f/g and (6) becomes $P(f(x)/g(x) \leq f(x_0)/g(x_0))$ which will generally be quite different than (6).

Another manifestation of the nonuniqueness of (6) arises when we consider 1-1 transformations of x. Suppose that $\mathcal{X}$ is an open subset of R^k, μ_k is volume measure, and $W : \mathcal{X} \to \mathcal{X}$ is 1-1 and sufficiently smooth. Then the density of $w = W(x)$ with respect to μ_k is given by $f_W(w) = f(W^{-1}(w))J_W\left(W^{-1}(w)\right)$ where $J_W(x)$ is the reciprocal of the Jacobian determinant of W at x. We see that (6) applied to w becomes

$$P_W(f_W(w) \leq f_W(w_0)) = P(f(x)J_W\,(x) \leq f(x_0)J_W\,(x_0))$$

where $w_0 = W(x_0)$ and again this will typically be different than (6) unless J_W is constant, e.g., when $\mathcal{X} = R^k$ and W is an affine transformation.

So it is clear that we cannot just write down a density and compute (6) as a model-based P-value. Still, the fact that we can do this in a satisfactory way in the discrete case, leads us to believe that there must be an appropriate resolution of this problem in more general contexts.

In measure-theoretic terms a density f, with respect to a support measure μ, is seen simply as a device to compute probabilities. In statistical contexts, however, a density plays a somewhat greater role. For example, if $f(x_1) > f(x_2)$, then we want to say that the probability of x_1 occurring is greater than the probability of x_2 occurring. For this to hold we can't allow f to be defined in an arbitrary fashion. In effect we need to have that $P(A)/\mu(A) \to f(x)$, as the set A converges to $\{x\}$, as then $P(A) \approx f(x)\mu(A)$ when A is close to $\{x\}$. Further, to compare the probabilities of two points x_1 and x_2 we need $A_i \to \{x_i\}$ with $\mu_k(A_1) = \mu_k(A_2)$ and then, for example, we can say that the probability of x_1 occurring is greater than the probability of x_2 occurring when $f(x_1) > f(x_2)$. The mathematics of making this precise is discussed, for example, in Rudin (1974), under the topic of differentiating one measure with respect to another. To use this here we suppose that $\mathcal{X}$ is an open subset of R^k and P is absolutely continuous with respect to volume measure μ_k.

It then seems natural to choose $\mu = \mu_k$ as it weights sample points equally and so $f(x)$ expresses the essence of how the probability measure is behaving at x. This is analogous to using counting measure as the support measure in the discrete case as then $f(x)$ has a direct interpretation as the probability of

x. In fact any measure that weights points equally will be a positive multiple of volume measure and (6) is invariant under these choices for μ. More generally $\mathcal{X}$ could be a manifold with locally Euclidean structure and with μ being geometric measure—the analog of volume measure on such a space—see Tjur (1974) for more details.

Given that we have settled on a specific support measure, the issue is then how to deal with the noninvariance of (6) under smooth, 1-1 transformations $W : \mathcal{X} \to \mathcal{X}$. It might seem that the only way to obtain invariance in general is to add an ingredient to the problem. We argue, however, that such an ingredient is actually implicitly part of any statistical problem where we are using a continuous distribution to model a measured response x. When we take this into account we can derive a version of (6) that is invariant under smooth transformations and that serves as a sensible definition of a model-based P-value. For note that, when we use a continuous probability distribution to model a variable that is being measured as part of some observational process, we are in fact thinking of the distribution as an approximation to an underlying discrete reality. For example, we measure variables to some fixed accuracy and so there is an underlying discreteness to the sample space.

To develop an invariant P-value we first show that (6) arises as an approximation to a P-value based on an appropriate discrete response. Suppose then that the underlying discreteness translates into a value x lying in a set $B_n(x)$ such that $\{B_n(x) : x \in \mathcal{X}\}$ forms a partition of $\mathcal{X}$ with $\mu_k(B_n(x))$ finite and constant in x, and such that $B_n(x)$ converges 'nicely' (see Rudin (1974)) to x as $n \to \infty$. We then have that $P(B_n(x))/\mu_k(B_n(x)) \to f(x)$ as $n \to \infty$ as long as f is continuous at x. So for n large, $P(B_n(x)) \approx f(x)\mu_k(B_n(x))$ and $f(x)$ serves as surrogate for the probability of x, at least when we are comparing the probabilities of different values of x occurring. Note that the constancy of $\mu_k(B_n(x))$ in x is necessary for this interpretation of $f(x)$. As a particular example of this, suppose that $\mathcal{X} = R^1$ and we partition R^1 using $\{((i-1)/n, i/n] : i \in Z\}$ and $B_n(x)$ is the set $((i-1)/n, i/n]$ that contains x.

Rather than observing x, the essential discreteness of the problem means that we will observe some $x_n(x) \in B_n(x)$ and the probability of observing $x_n(x)$ is $P(B_n(x))$. Note that implicitly x_0 is one of the values assumed by x_n. Then for the discrete response variable x_n, the appropriate P-value (2) is given by

$$\sum_{\{x_n(x):P(B_n(x))\leq P(B_n(x_0))\}} P(B_n(x)). \tag{7}$$

We then want to show that (6) serves as an approximation to (7).

While such a result seems intuitively plausible, a general proof is not straightforward. We require some regularity conditions as we cannot expect such an approximation to hold if we allow f and the partition $\{B_n(x) : x \in \mathcal{X}\}$ to be too general. For this we use the theory of contented sets and functions as discussed in Loomis and Sternberg (1968) where it is used to develop the Riemann integral. Essentially, a bounded set A is *contented* if its μ_k-measure can be approximated arbitrarily closely by the μ_k-measure of a finite union of disjoint rectangles contained in A and also by the μ_k-measure of a finite union of disjoint rectangles

containing A. A bounded function f with compact support is *contented* if it can be approximated arbitrarily closely by step functions. Further, we say that a function f is locally constant at x if we can find an open set containing x on which f is constant. For $x_0 \in \mathcal{X}$ let $LC(x_0) = \{x : f(x) = f(x_0), f \text{ is locally constant at } x\}$. In the Appendix we prove the following result.

Theorem 2. Suppose that
(i) $\mathcal{X}$ is a contented subset of R^k with positive content,
(ii) $B_n(x)$ is a rectangle containing x with $\mu_k(B_n(x))$ finite and constant in x, and such that $B_n(x)$ converges nicely to x as $n \to \infty$,
(iii) $\{B_n(x) : x \in R^k\}$ forms a partition of R^k with $\{B_{n+1}(x) : x \in R^k\}$ a subpartition of $\{B_n(x) : x \in R^k\}$ and $\sup_{x \in R^k} \text{diam}(B_n(x)) \to 0$ as $n \to \infty$,
(iv) f is a continuous density function on $\mathcal{X}$ with $f^{-1}A$ contented for any interval A and such that $LC(x_0)$ is contented with $\mu_k(f^{-1}f(x_0) \cap LC(x_0)^c) = 0$, then (7) converges to (6) as $n \to 0$

Theorem 2 establishes that the appropriate discrete P-value, in the sense that we will always be measuring x to some finite accuracy, is indeed approximated by the continuous version given by (6), provided n is large enough.

Although the result will hold under weaker conditions, the conditions specified seem to apply in typical applications. Condition (iv) controls the behavior of f and in particular prevents it from being too 'wiggly' so that points in contours of f are either points where the function is locally constant or part of a null set. For example, the condition holds for all piecewise smooth f. This condition can be substantially weakened if $P(f(x) = c) = 0$ for every c. In essence the distribution of $f(x)$ can have a discrete component but our conditions imply that this can really only arise by f being constant on sets where it is locally constant. The conditions on the partitions $\{B_n(x) : x \in R^k\}$ are stronger than needed. In particular, we could allow for more general sets than rectangles. Further, it is implicit in Theorem 2 that the accuracy of the discretization is effectively the same across the sample space. We could allow for this accuracy to vary across the sample space and this would determine a different approximation to (7). While this is reasonable, we do not pursue this further here but note that the situation we have considered is very common.

We may, however, base the P-value on a statistic T, such as a discrepancy statistic, and use the observed value $T(x_0)$. The question then is: given the initial discretization on $\mathcal{X}$ as determined by the measurement process, how should we take this into account? For, even if T is 1-1, it will give rise to volume distortions and we do not want these volume distortions to affect our P-value. This is the heart of the invariance issue and we discuss this in the next section.

4 Invariant P-values for General Statistics

Suppose $T : \mathcal{X} \to \mathcal{T}$ is a general statistic, and we want to compare $t_0 = T(x_0)$ to P_T to assess whether or not we have evidence against P. When P_T is discrete, it would seem that the relevant P-value is as discussed in Section 2 (see Example

4). In the continuous case, however, additional complexities arise. This is because T may distort volume and we need to ensure that the P-value we use does not depend on these distortions.

Suppose first that $\mathcal{X}$ and $\mathcal{T}$ are open subsets of R^k and that T is 1-1 and smooth. Then a partition element $B_n(x) \subset \mathcal{X}$, with measure $\mu_k(B_n(x))$, is transformed into $TB_n(x)$ with measure $\mu_k(TB_n(x)) = \mu_k(B_n(x))J_T^{-1}(x')$ for some $x' \in B_n(x)$, while the density of the transformed response with respect to μ_k (Euclidean measure on R^k) is $f_T(t) = f(T^{-1}(t))J_T(T^{-1}(t))$. Accordingly, we cannot use the P-value $P_T\left(f_T(t) \leq f_T(t_0)\right)$ to assess whether or not t_0, or equivalently $x_0 = T^{-1}(t_0)$ is surprising, since the density $f_T(t)$ depends on volume distortions and the sets $TB_n(x)$ are no longer necessarily of equal volume. There is clearly an easy fix for this, however, as we simply correct for this volume distortion and compute the P-value

$$P_T\left(f_T(t)/J_T(T^{-1}(t)) \leq f_T(t_0)/J_T(T^{-1}(t))\right) = P(f(x) \leq f(x_0)). \tag{8}$$

With this refinement the P-value introduced in section 3 becomes invariant under 1-1, smooth transformations of the response, i.e., we retain as part of the problem prescription how the continuous probability model is approximating an essentially discrete response.

In general, however, T will not be 1-1. Suppose then, that $\mathcal{X}$ is an open subset of R^k and $\mathcal{T}$ is an open subset of R^l where $l \leq k$. Let f_T denote the density of P_T with respect to μ_l and suppose this is continuous. Suppose that T is sufficiently smooth so that for each $t \in \mathcal{T}$ the set $T^{-1}\{t\}$ is a Riemann manifold with geometric measure on $T^{-1}\{t\}$ denoted by ν_t. For example, when T is 1-1, then $T^{-1}\{t\}$ is a 0-dimensional Riemann manifold and ν_t is counting measure. Results in Tjur (1974) show that, in general,

$$f_T(t) = \int_{T^{-1}\{t\}} f(x)\left|\det(dT(x) \circ dT'(x))\right|^{-1/2} \nu_t(dx) \tag{9}$$

where dT is the differential of T. Formula (9) shows directly how f_T is affected by volume distortions. For, at $x \in T^{-1}\{t\}$ the contribution to the density value $f_T(t)$ is distorted by the factor $J_T(x) = \left|\det(dT(x) \circ dT'(x))\right|^{-1/2}$. Accordingly, just as we do in the 1-1 case, we adjust the integrand in (9) by dividing by the factor $J_T(x)$ to obtain

$$f_T^*(t) = \int_{T^{-1}\{t\}} f(x)\,\nu_t(dx)$$

as the appropriate density to use. Note that f_T^* is the density of P_T with respect to the measure $(f_T(t)/f_T^*(t))\mu_l(dt)$ and the ratio $f_T(t)/f_T^*(t)$ measures the effect of the volume distortion induced by T on the density f_T. We then compute the P-value

$$P_T(f_T^*(t) \leq f_T^*(t_0)) = P(f_T^*(T(x)) \leq f_T^*(T(x_0))) \tag{10}$$

to assess whether or not $t_0 = T(x_0)$ is a surprising value from P_T. We see that (10) depends only on the density assignment f on the original response

space, which is determined by how we are approximating an essentially discrete response, and the preimage sets of T.

We have the following simple but significant result for (10).

Theorem 3. When $\mathcal{X}$ is an open subset of R^k, $T : \mathcal{X} \to \mathcal{T}$ is onto with $\mathcal{T} \subset R^l$ open, and T is sufficiently smooth, then the P-value given by (10) is invariant under 1-1 smooth transformations of $\mathcal{T}$.

Proof: Suppose W is a 1-1, smooth transformation defined on $\mathcal{T}$ and $w = W(t)$. Then, $(W \circ T)^{-1}\{w\} = T^{-1}\{t\}$ and $f^*_{W \circ T}(w) = \int_{T^{-1}\{t\}} f(x)\, \nu_t(dx) = f^*_T(t)$.

We now consider some applications and note that these support (10) as the appropriate definition of an invariant P-value.

Example 2. *T a smooth 1-1 transformation.*

Suppose that $T : \mathcal{X} \to \mathcal{T}$ is 1-1. Then $T^{-1}\{t\}$ is a singleton set. Any discrete set of points is a 0-dimensional Riemann manifold and geometric measure is counting measure. Therefore, $f^*_T(t) = \int_{T^{-1}\{t\}} f(x)\, \nu_t(dx) = f(T^{-1}\{t\})$ and (10) equals (8).

Example 3. *T a smooth k-1 transformation.*

Suppose that $T^{-1}\{t\} = \{x_1(t), \ldots, x_k(t)\}$ for each t. Then we have that $f_T(t) = \sum_{i=1}^k f(x_i(t)) J_T(x_i(t))$ and note that the volume distortion $J_T(x_i(t))$ could vary with i. In this case, we have that ν_t is counting measure and the corrected density is $f^*_T(t) = \sum_{i=1}^k f(x_i(t))$. As in Example 2, it seems clear here how we need to correct for volume distortions and, as such, it provides strong support for (10) as the relevant P-value.

The following example shows that (10) gives the correct answer in the discrete case as well.

Example 4. *P_T is discrete.*

First suppose that P is discrete so we can consider $\mathcal{X}$ as a 0-dimensional Riemann manifold with geometric measure equal to counting measure and similarly for $\mathcal{T}$. In this case ν_t is counting measure on $T^{-1}\{t\}$, $J_T(x) \equiv 1$ and so $f^*_T(t) = \int_{T^{-1}\{t\}} f(x)\, \nu_t(dx) = \sum_{x \in T^{-1}\{t\}} P(X = x) = p_T(t)$ is the probability function of T. Note dT is just the identity so there is no volume distortion.

When P is continuous then, for those t with $p_T(t) > 0$, we have that ν_t is μ_k restricted to $T^{-1}\{t\}$. Accordingly, we have that $p_T(t) = \int_{T^{-1}\{t\}} f(x)\, \nu_t(dx) = f^*_T(t)$. So, in general, we obtain the P-value discussed in section 2.

Example 5. *$J_T(x)$ is constant.*

Note that $J_T(x)$ is constant for all x whenever $T(x)$ is an affine transformation. So we could have $T(x) = a + Bx$ for some $a \in R^l$ and $B \in R^{l \times k}$ when $\mathcal{X} \subset R^k$. Also notice that when $x \in R^n$ and $T(x)$ is the order statistic then $J_T(x)$ is constant for all x. It is then clear from (9) that, in this case, we can compute (10) as $P_T(f_T(t) \leq f_T(t_0))$.

For example, when $T(x) = \bar{x}$ we simply use the density of $\bar{x}$ to compute the P-value. When P is the $N_1(0, 1)$ distribution, then (10) is $2(1 - \Phi(\bar{x}))$.

As another example, suppose that T is projection on the i-th coordinate, so $J_T(x) \equiv 1$. Then $T^{-1}\{t\}$ is the set of points in $\mathcal{X}$ with i-th coordinate equal to

t, ν_t is Euclidean volume on this set, and $f_T^*(t)$ is the marginal density of the i-th coordinate. Of course, this generalizes to arbitrary coordinate projections.

Example 6. *$J_T(x)$ is constant for $x \in T^{-1}\{t\}$.*

In some ways Example 5 is the simplest situation as the volume distortion induced by T is constant on $\mathcal{X}$. We now allow for the possibility that the volume distortion is constant in $T^{-1}\{t\}$ but may vary with t.

Put $J_T^*(t) = J_T(x)$ for $x \in T^{-1}\{t\}$. From (9) we have that $f_T(t) = f_T^*(t)J_T^*(t)$ and so (10) can be computed as $P_T(f_T(t)/J_T^*(t) \leq f_T(t_0)/J_T^*(t_0))$. This permits us to avoid the integration involved in calculating $f_T^*(t)$ when we know the distribution of T and can compute $J_T(x)$ easily.

As an example, suppose that $T(x) = x'x$. Then $T^{-1}\{t\}$ is a $(k-1)$-dimensional sphere in R^k. Now $dT(x) = 2(x_1 \ldots x_k)$ so $dT(x) \circ dT'(x) = 4x'x = 4t$ and $J_T(x) = 1/2t^{1/2}$ is constant for $x \in T^{-1}\{t\}$ for every t. Note that the adjustment factor applied to $f_T(t)$ is to multiply by $2t^{1/2}$ and this is precisely the distortion caused by the "quadratic" part of the transformation. The appropriate P-value is $P_T(f_T(t)t^{1/2} \leq f_T(t_0)t_0^{1/2})$. We see that in this case we must modify the usual density that we work with.

As a particular case, suppose that $x \sim N_k(0, I)$. Then $T(x) \sim$ Chi-squared(k) with density $f_T(t) = \Gamma^{-1}(k)2^{-k/2}t^{(k/2)-1}e^{-t/2}$. Therefore, the invariant P-value is given by $P_T(t^{(k-1)/2}e^{-t/2} \leq t_0^{(k-1)/2}e^{-t_0/2})$ and only when $k = 1$ is this equivalent to $P_T(t \geq t_0)$.

Notice that when we directly observe $T \sim$ Chi-squared(k), in the sense that it is a measured variable, and we discretize using equal length intervals, then the relevant P-value is $P_T(t^{(k/2)-1}e^{-t/2} \leq t_0^{(k/2)-1}e^{-t_0/2})$. As just shown, when we take into account that T arises as a transformation of a measured variable, the P-value changes. Further, both of these P-values are two-sided when $k > 1$. In contrast, the approximate P-value that arises in Example 1, for a multinomial with $k+1$ categories, is a right-tail only P-value for the Chi-squared(k) distribution, and this follows directly from our theory.

In section 5 we discuss some further examples and, in particular, some examples where $J_T(x)$ varies with $x \in T^{-1}\{t\}$. There are also computational issues that need to be addressed in such contexts and we discuss these in section 6.

In section 5 we also discuss another use of a transformation W to assess surprise. This involves comparing the observed x_0 with the conditional distribution of x given that $W(x) = W(x_0) = w_0$. In this case the conditional density of x, with respect to geometric measure on $W^{-1}\{w_0\}$, is given by $f(x)J_W(x)/f_W(w_0)$ and it is clear that the volume distortion at x, induced by the conditioning, is given by $J_W(x)$. Accordingly the relevant P-value, based on the full data, is given by $P(f(x)/f_W(w_0) \leq f(x_0)/f_W(w_0) \,|\, W(x) = w_0) = P(f(x) \leq f(x_0) \,|\, W(x) = w_0)$. If we have a transformation T of x, then the relevant P-value is given in the following result.

Lemma 4. Suppose that $\mathcal{X}, \mathcal{W}$ and $\mathcal{T}$ are manifolds, with geometric measures $\mu_{\mathcal{X}}, \mu_{\mathcal{W}}$ and $\mu_{\mathcal{T}}$ respectively, and $W : \mathcal{X} \to \mathcal{W}, T : \mathcal{X} \to \mathcal{T}$ are onto smooth mappings. Let $\nu_{T,W,t,w}$ denote geometric measure on $T^{-1}\{t\} \cap W^{-1}\{w\}$. Then

the relevant conditional P-value based on T, given $W(x) = w_0$, is

$$P_T\left(f^*_{T,W}(t \,|\, w_0) \leq f^*_{T,W}(t_0 \,|\, w_0) \,|\, W(x) = w_0\right) \tag{11}$$

where $t_0 = T(x_0)$, and $f^*_{T,W}(t \,|\, w_0) = \int_{T^{-1}\{t\}\cap W^{-1}\{w_0\}} f(x)\, \nu_{T,W,t,w_0}(dx)$.
Proof: The conditional density of T given $W = w$ is given by $f_{T,W}(t \,|\, w) = \int_{T^{-1}\{t\}\cap W^{-1}\{w\}} (f(x)/f_W(w)) J_{(T,W)}(x)\, \nu_{T,W,t,w}(dx)$. Therefore, volume distortion induced by the transformations is $J_{(T,W)}(x)$ and the result follows.

We will need the following result concerning the composition of mappings.
Lemma 5. Suppose that $\mathcal{X}$,$\mathcal{U}$ and $\mathcal{T}$ are manifolds, with geometric measures $\mu_{\mathcal{X}}, \mu_{\mathcal{U}}$ and $\mu_{\mathcal{T}}$ respectively, and $U : \mathcal{X} \to \mathcal{U}, T : \mathcal{U} \to \mathcal{T}$ are onto smooth mappings, then

$$f^*_{T\circ U}(t) = \int_{T^{-1}\{t\}} J_T(u) \int_{U^{-1}\{u\}} f(x) J^{-1}_{T\circ U}(x) J_U(x)\, \upsilon_{U,u}(dx)\, \upsilon_{T,t}(du),$$

where $\upsilon_{U,u}$ and $\upsilon_{T,t}$ are the geometric measures on $U^{-1}\{u\}$ and $T^{-1}\{t\}$.
Proof: Suppose that $g : \mathcal{X} \to R^1$ is nonnegative, $\int_A g(x)\, \mu_{\mathcal{X}}(dx)$ is finite for compact A and let $B \subset \mathcal{T}$ be open. By the measure decomposition theorem (see Tjur (1974), Theorem 15.1) applied to $g(x)\, \mu_{\mathcal{X}}(dx)$ and $T \circ U$, we have that $\int_{\mathcal{X}} I_B(T(U(x))) g(x)\, \mu_{\mathcal{X}}(dx) = \int_B \int_{(T\circ U)^{-1}\{t\}} g(x) J_{T\circ U}(x)\, \upsilon_{T\circ U,t}(dx)\, \mu_{\mathcal{T}}(dt)$. Apply the measure decomposition theorem first to $g(x)\, \mu_{\mathcal{X}}(dx)$ and U and then to $\int_{U^{-1}\{u\}} I_B(T(U(x))) g(x) J_U(x)\, \upsilon_{U,u}(dx)\, \mu_{\mathcal{U}}(du)$ and T to obtain

$$\begin{aligned}
&\int_{\mathcal{X}} I_B(T(U(x))) g(x)\, \mu_{\mathcal{X}}(dx) \\
&= \int_{\mathcal{U}} \int_{U^{-1}\{u\}} I_B(T(U(x))) g(x) J_U(x)\, \upsilon_{U,u}(dx)\, \mu_{\mathcal{U}}(du) \\
&= \int_B \int_{T^{-1}\{t\}} \int_{U^{-1}\{u\}} g(x) J_U(x)\, \upsilon_{U,u}(dx) J_T(u)\, \upsilon_{T,t}(du)\, \mu_{\mathcal{T}}(dt).
\end{aligned}$$

From this we conclude that $\int_{(T\circ U)^{-1}\{t\}} g(x) J_{T\circ U}(x)\, \upsilon_{T\circ U,t}(dx)\, \mu_{\mathcal{T}}(dt) = \int_{T^{-1}\{t\}} \int_{U^{-1}\{u\}} g(x) J_U(x)\, \upsilon_{U,u}(dx) J_T(u)\, \upsilon_{T,t}(du)\, \mu_{\mathcal{T}}(dt)$. Putting $g(x) = f(x) J^{-1}_{T\circ U}(x)$ establishes the result.

5 Applications

Suppose that we have a statistical model $\{P_\theta : \theta \in \Theta\}$ where P_θ is a probability measure on $\mathcal{X}$ with density f_θ with respect to support measure μ_k, and Π is a prior probability measure on Θ. Let $M(A) = \int_\Theta P_\theta(A)\, \Pi(d\theta)$ denote the prior predictive on $\mathcal{X}$ with density $m(x) = \int_\Theta f_\theta(A)\, \Pi(d\theta)$ with respect to μ_k. We will investigate here the relevant P-values for assessing the model and checking for prior-data conflict in light of an observed x_0. As we will see, these P-values are invariant and depend only on the densities f_θ. In particular, the P-values do not depend on any choice of density for the prior. This makes sense because we do not directly measure the variable θ, only the response variable x.

5.1 Model Checking

If $W : \mathcal{X} \to \mathcal{W}$ is a minimal sufficient statistic, then the conditional distribution of the data given W is independent of θ and is denoted $P(\cdot \,|\, W(x) = w_0)$. Suppose we wish to check if the model makes sense in light of the observed x_0. By the converse of the factorization theorem we have that $f_\theta(x) = g_\theta(W(x))h(x)$. Lemma 4 and (11) give an invariant P-value that assesses f_θ for each θ. We have the following result.

Theorem 6. The P-value (11) is given by

$$P_T\left(h_{T,W}(t \,|\, w_0) \le h_{T,W}(t_0 \,|\, w_0) \,|\, W(x) = w_0\right), \tag{12}$$

where $h_{T,W}(t \,|\, w_0) = \int_{T^{-1}\{t\} \cap W^{-1}\{w_0\}} h(x)\, \nu_t(dx)$, i.e., it is independent of θ, and (12) is independent of the choice of h.

Proof: In the continuous case we assume that each density is continuous at any observed x_0 and restrict attention to those x_0 for which $f_\theta(x_0) > 0$. When $f_\theta(x_0) > 0$, then $g_\theta(W(x_0)) > 0$ and $g_\theta(W(x)) = g_\theta(W(x_0))$ for the event $W(x) = t_0 = W(x_0)$. We have that (11) equals

$$\begin{aligned} &P_T\left(\begin{array}{c} \int_{T^{-1}\{t\} \cap W^{-1}\{w_0\}} f_\theta(x)\, \nu_t(dx) \\ \le \int_{T^{-1}\{t_0\} \cap W^{-1}\{w_0\}} f_\theta(x)\, \nu_t(dx) \end{array} \;\middle|\; W(x) = w_0 \right) \\ &= P\left(h_{T,W}(t \,|\, w_0) \le h_{T,W}(t_0 \,|\, w_0) \,|\, W(x) = w_0\right). \end{aligned}$$

Further, if $g_\theta(T(x))h(x) = g'_\theta(T(x))h'(x)$, then

$$\begin{aligned} &P_T\left(h_{T,W}(t \,|\, w_0) \le h_{T,W}(t \,|\, w_0) \,|\, W(x) = w_0\right) \\ &= P_T\left(\begin{array}{c} \int_{T^{-1}\{t\} \cap W^{-1}\{w_0\}} g_\theta(W(x))h(x)\, \nu_t(dx) \\ \le \int_{T^{-1}\{t_0\} \cap W^{-1}\{w_0\}} g_\theta(W(x_0))h(x)\, \nu_t(dx) \end{array} \;\middle|\; W(x) = w_0 \right) \\ &= P_T\left(\begin{array}{c} \int_{T^{-1}\{t\} \cap W^{-1}\{w_0\}} g'_\theta(W(x))h'(x)\, \nu_t(dx) \\ \le \int_{T^{-1}\{t_0\} \cap W^{-1}\{w_0\}} g'_\theta(W(x_0))h'(x)\, \nu_t(dx) \end{array} \;\middle|\; W(x) = w_0 \right) \\ &= P_T\left(h'_{T,W}(t \,|\, w_0) \le h'_{T,W}(t_0 \,|\, w_0) \,|\, W(x) = w_0\right) \end{aligned}$$

and we are done.

We now consider an application of this result.

Example 6. Model checking for the *location-scale normal model.*

Suppose that $x = (x_1, \ldots, x_n)$ is a sample of n from the $N(\mu, \sigma^2)$ distribution with $\mu \in R^1$ and $\sigma^2 > 0$ unknown. Then $W(x) = (\bar{x}, ||x - \bar{x}1_n||)$ is minimal sufficient. Putting $d(x) = (x - \bar{x}1_n)/||x - \bar{x}1_n||$, we can write $x = \bar{x} + ||x - \bar{x}1_n||d$ and note that $\bar{x}, ||x - \bar{x}1_n||$ and d are statistically independent with d uniformly distributed on $S^{n-1} \cap L^\perp\{1_n\}$. In this case h is constant (so we can take it to be 1) and $W^{-1}\{(\bar{x}_0, ||x_0 - \bar{x}_0 1_n||)\}$ is the $(n-2)$-dimensional sphere $\bar{x}_0 1_n + ||x_0 - \bar{x}_0 1_n||(S^{n-1} \cap L^\perp\{1_n\})$.

It is natural here to consider functions of d as discrepancy statistics for checking the model. For example, the family $T_p \circ d = \sum_{i=1}^d d_i^p$ is of some interest as this gives effectively the skewness and kurtosis statistics when $p =$

3 and 4, respectively. In this case, $h_{T,W}(t \,|\, w_0)$ is the volume of the $(n-3)$-dimensional submanifold of $\bar{x}_0 1_n + ||x_0 - \bar{x}_0 1_n|| (S^{n-1} \cap L^{\perp}\{1_n\})$ given by $(T_p \circ d)^{-1}\{t\} \cap W^{-1}\{(\bar{x}_0, ||x_0 - \bar{x}_0 1_n||)\}$. Alternatively, from the proof of Theorem 6, we can compute the invariant P-value by assuming $(\mu, \sigma) = (0, 1)$, letting f denote the density of a sample of n from the $N(0,1)$ distribution and computing $P_{(0,1)}(f^*_{T_p \circ d}(T_p(d(x))) \leq f^*_{T_p \circ d}(T_p(d(x_0))) \,|\, (\bar{x}_0, ||x_0 - \bar{x}_0 1_n||)) = P_d(f^*_{T_p \circ d}(T_p(d)) \leq f^*_{T_p \circ d}(T_p(d_0)))$ where $f^*_{T_p \circ d}(t) = \int_{(T_p \circ d)^{-1}\{t\}} f(x)\, \upsilon_t(dx)$ and d is uniformly distributed on $S^{n-1} \cap L^{\perp}\{1_n\}$. The volume distortion induced by $T_p(d)$ can be computed explicitly as $J_{T_p \circ d}(x) = p(T_{2p-2}(d(x)) - T^2_{p-1}(d(x))/n - T^2_p(d(x)))^{-1/2}/||x - \bar{x}1_n||$. We see that this is not a function of $T_p \circ d$ and also $J_{T_p \circ d}(x) = J_{T_p \circ d}(-x)$. Since $f(x) = f(-x)$ and $(T_p \circ d)^{-1}\{-t\} = (-1)^p (T_p \circ d)^{-1}\{t\}$ when p is a nonnegative integer, we have that $f^*_{T_p \circ d}(t)$ and the density $f_{T_p \circ d}(t) = \int_{(T_p \circ d)^{-1}\{t\}} f(x) J_{T_p \circ d}(x)\, \upsilon_t(dx)$ are symmetric about 0 when p is odd. If both $f^*_{T_p \circ d}$ and $f_{T_p \circ d}$ are unimodal, then this implies that the P-values based on the densities, tail probabilities of $|T_p \circ d|$ and the invariant P-values are the same when p is odd. Although we do not have a proof, it would appear that in general, the differences among these P-values disappear with increasing n, so the need to correct for volume distortion vanishes with large sample sizes in this case.

In Figure 1 we have plotted the densities and invariant P-values for tests of skewness for several sample sizes n. The P-values are two-sided. The invariant P-values are the same as those based on the density of $T_3(d)$ and tail probabilities of $|T_3(d)|$, for all cases except $n = 3$. We note that the asymptotic approximation to the exact P-value can be poor but is quite good for $n = 100$. When $n = 3$ the invariant P-value is identically equal to 1, i.e., we never find evidence against the model. In this case, we can show that $J^{-2}_{T_3}(x) = 1/6 - T^2_3(x)$ and the density of T_3 at t is proportional to $(1 - 6t^2)^{-1/2}$ for $-1/\sqrt{6} < t < 1/\sqrt{6}$, i.e., all the density is due to the volume distortion caused by T_3. Notice too that the density is U-shaped with infinite singularities at the end-points. Accordingly, if we were to use the density for the P-value we would reject the model for values of T_3 near 0 and this doesn't make sense. It wouldn't seem to make sense to reject for large values of $|T_3|$ either, at least based on the shape of the distribution. The invariant P-value is telling us that there is no test for skewness based on T_3 when $n = 3$. Intuitively this seems reasonable as we need two degrees of freedom for location and scale and to check for skewness, we need at least two more to see if there is asymmetry about the center.

In Figure 2 we have plotted the densities and invariant P-values for tests of kurtosis for several sample sizes n. The densities are quite irregular for small sample sizes and skewed. The invariant P-values, those based on the density and tail probabilities are all different in this case. The P-values based on the densities and asymptotics are quite similar for $n = 100$ while this is not the case for the invariant P-values. This indicates that the volume distortion is still having an appreciable effect when $n = 100$.

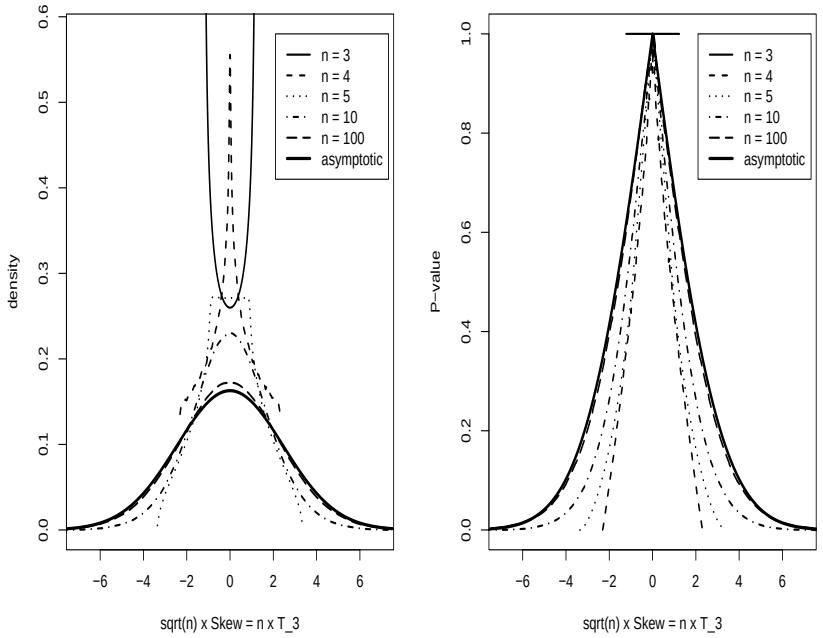

Figure 1: Densities and invariant P-values for test of skewness for various sample sizes $n$when sampling from normal.

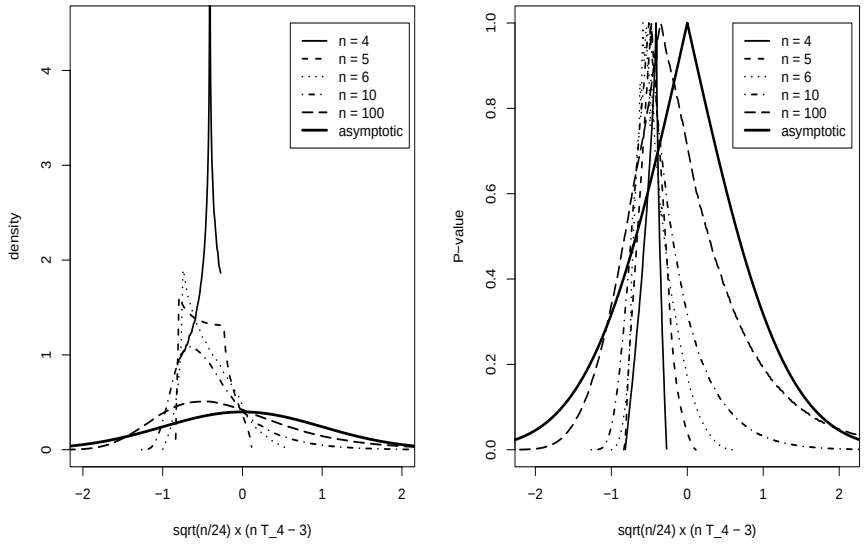

Figure 2: Densities and invariant P-values for test of kurtosis for various sample sizes n when sampling from normal.

In Figure 3 we have plotted the densities and invariant P-values based on the Jarque-Bera test statistic $n(nT_3^2/6 + (nT_4 - 3)^2/24)$ for several sample sizes n. This is clearly an attempt to assess both skewness and kurtosis simultaneously. The densities are quite irregular for small sample sizes and skewed. The P-

values based on the densities and asymptotics are quite different for $n = 100$ while this is not the case for the invariant P-values. Again this indicates that the volume distortion is still appreciable when $n = 100$.

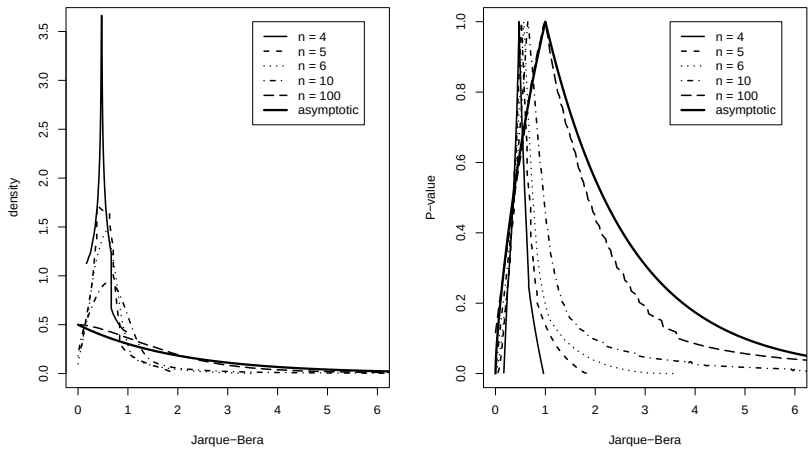

Figure 3: Densities and invariant P-values for Jarque-Bera test for various sample sizes n when sampling from normal.

When U is an ancillary statistic, then a function of U can be used to assess the model. If we consider the transformation $T \circ U$, then we must evaluate $\int_{(T\circ U)^{-1}\{t\}} f_\theta(x)\, \nu_t(dx)$ and, in general, there is no reason to suppose that this is independent of θ. When the distribution of $T \circ U$ is discrete, however, then $\int_{(T\circ U)^{-1}\{t\}} f_\theta(x)\, \nu_t(dx)$ is the probability function of $T \circ U$ and as such is independent of θ. Theorem 7 will show that the P-value based on $\int_{(T\circ U)^{-1}\{t\}} f_\theta(x)\, \nu_t(dx)$ is independent of θ for a very broad class of ancillaries.

Consider the following example which will serve as an archetype for a common situation where ancillaries arise.

Example 7. *Location-scale models.*

Suppose we have $x \in R^n$ and the model is $x = \mu 1_n + \sigma z$ where z is distributed with density f with respect to volume measure on R^n, and $\mu \in R^1, \sigma > 0$ are unknown. Then x has density $f_{\mu,\sigma}(x) = \sigma^{-n} f((x - \mu 1_n)/\sigma)$. We take the parameter space to be $\Theta = \{(\mu, \sigma) : \mu \in R^1, \sigma > 0\}$ and note that we have a group product defined on Θ via $(\mu_1, \sigma_1)(\mu_2, \sigma_2) = (\mu_1 + \sigma_1\mu_2, \sigma_1\sigma_2)$. This group acts on R^n via $(\mu, \sigma)x = \mu 1_n + \sigma x$. A maximal invariant is then given by $U(x) = (x - \bar{x}1_n)/||x - \bar{x}1_n||$ and this is ancillary. Note that $U^{-1}\{u\} = \{x : x = a1_n + cu, \text{ for some } (a, c) \in \Theta\} = \Theta u$, i.e., $U^{-1}\{u\}$ is an orbit of the group action. Clearly, this orbit is half of a 2-dimensional plane in R^n and so geometric measure ν_u is just area.

If we wish to base our checking on U itself, then we must evaluate

$$f^*_{\mu,\sigma,U}(d_0) = \int_{U^{-1}\{u\}} f_{\mu,\sigma}(x)\,\nu_u(dx) = \int_0^\infty \int_{-\infty}^\infty f_{\mu,\sigma}(a1_n + cu)\,\sqrt{n}da\,dc$$
$$= \int_0^\infty \int_{-\infty}^\infty \sigma^{-n} f\left(\frac{a-\mu}{\sigma}1_n + \frac{c}{\sigma}u\right)\sqrt{n}da\,dc$$
$$= \sigma^{-(n-2)} \int_0^\infty \int_{-\infty}^\infty f\,(a1_n + cu)\;\sqrt{n}da\,dc.$$

Accordingly the P-value for model checking is given by

$$P_U(f^*_{\mu,\sigma,U}(u) \le f^*_{\mu,\sigma,U}(u_0))$$
$$= P_U\left(\int_0^\infty \int_{-\infty}^\infty f\,(a1_n + cu)\;da\,dc \le \int_0^\infty \int_{-\infty}^\infty f\,(a1_n + cu_0)\;da\,dc\right)$$

and this is independent of the model parameter and we have a valid P-value for checking the model. If instead we use a function $T(U)$ then, an application of Lemma 5 shows that (10) is independent of (μ, σ) by the same argument, as the Jacobian factors do not depend on the parameter. Note that when f is the $N(0,1)$ density, then basing model checking on the ancillary d or the conditional distribution of the data given a minimal sufficient statistic produces the same results.

More generally suppose we have a group model $\{f_g : g \in G\}$ where G is a group, with a smooth product, acting freely and smoothly on $\mathcal{X}$ and $f_g(x) = f(g^{-1}x)J_g(g^{-1}x)$ for some fixed density f. Now suppose that $[\cdot] : \mathcal{X} \to G$ is smooth and satisfies $[gx] = g[x]$ so $U(x) = [x]^{-1}x$ is a maximal invariant and is thus ancillary. So $u = U(x) \in \mathcal{X}, x = [x]U(x)$ and $U^{-1}\{u\}$ is the orbit $\{gu : g \in G\}$. Now if ν^*_G denotes geometric measure on G we have that $\nu_u = K(u)\nu^*_G$ for some positive function K. Let $z = g^{-1}x$ so that $[z] = g^{-1}[x]$ and let $J^*_g([z])$ denote the Jacobian of the transformation $[z] \to [x]$. Then we have that

$$f^*_{gU}(u) = \int_{U^{-1}\{u\}} f_g(x)\,\nu_u(dx) = \int_{\{gu:g\in G\}} f_g([x]u)\,\nu_u(dx)$$
$$= K(u)\int_G f(g^{-1}[x]u)J_g(g^{-1}[x]u)\,\nu^*_G(d[x])$$
$$= K(u)\int_G f([z]u)J_g(u)J^*_g([z])\,\nu^*_G(d[z]).$$

Now if we can write $J_g(u)J^*_g(z) = L(u)m(g)$ for some positive functions L and m, then we have that the invariant P-value $P_U(f^*_{gU}(u) \le f^*_{gU}(u_0))$ is indeed independent of g. Further, by Lemma 5 this will also hold for $T \circ U$ as well. For example, in Example 6 $J_{(\mu,\sigma)}(u) = \sigma^{-n}$ and, with $[x] = [\bar{x}, ||x - \bar{x}1_n||]$, then $J^*_g([z]) = \sigma^2$ and this condition is satisfied. More generally, this condition is satisfied in a wide range of group models, such as those discussed in Fraser (1979). Accordingly the following result is broadly applicable.

Theorem 7. Suppose that $\{f_g : g \in G\}$ is a family of densities with respect to geometric measure $\mu_{\mathcal{X}}$ on $\mathcal{X}$, where G is a group with a smooth product, a smooth action defined on $\mathcal{X}$ and $f_g(x) = f(g^{-1}x)J_g(g^{-1}x)$. Further suppose that there exists smooth $[\cdot] : \mathcal{X} \to G$ satisfying $[gx] = g[x]$ and let $J_g^*([z])$ denote the Jacobian of the transformation $[z] \to [x]$ where $x = gz$. If there exist positive functions L and m so that $J_g(u)J_g^*(z) = L(u)m(g)$, then we have that the P-value (10) based on the ancillary $T \circ U$, with $U(x) = [x]^{-1}x$ and T smooth, is independent of the model parameter and is thus a valid check on the model.

Although we have not been able to find an example, there may exist cases where the invariant P-value based on an ancillary $T \circ U$ is not independent of the parameter. In such a case it is perhaps difficult to accept $T \circ U$ as a true ancillary, because its ancillarity is dependent on the way the transformation is distorting volumes in some essential way.

5.2 Checking for Prior-Data Conflict

In Evans and Moshonov (2006, 2007) methodology was developed for investigating the existence of a conflict existing between the prior probability assignments made for the model parameter θ, and the values of θ deemed relevant by the likelihood. If T is a minimal sufficient statistic and $U(T)$ is a maximal ancillary, then the assessment is made based upon comparing the observed value $t_0 = T(x_0)$ with $M_T(\cdot \,|\, u_0)$, the conditional prior predictive distribution of T given $U(t_0) = u_0$. The comparison was based upon the P-value $M_T(m_T(t \,|\, u_0) \leq m_T(t_0 \,|\, u_0) \,|\, u_0)$ where $m_T(\cdot \,|\, u_0)$ is the prior predictive density of T given $U(T) = u_0$, based on either counting measure or volume measure depending on whether $M_T(\cdot \,|\, u_0)$ was discrete or continuous. This choice of P-value was made primarily because there was no theory that dictated how such an assessment was to be made, but concern was expressed about the lack of invariance in the continuous case. We can now use the approach developed here to derive an appropriate invariant P-value.

The prior predictive density of x is given by $m(x) = \int_\Theta f_\theta(x)\,\Pi(d\theta)$ and note that this is just an average of the density values $f_\theta(x)$ with respect to the prior. There is no volume distortion involved in this, for if $f_\theta(x)\mu_k(B_n(x))$ is the probability of observing the discretized response $x_n(x)$ when θ is true, then $m(x)\mu_k(B_n(x))$ is the probability of observing $x_n(x)$ when $\theta \sim \Pi$ and $x \sim P_\theta$. Furthermore, $m_T^*(t) = \int_{T^{-1}\{t\}} m(x)\,\upsilon_t(dx) = \int_{T^{-1}\{t\}} \int_\Theta f_\theta(x)\,\Pi(d\theta)\,\upsilon_t(dx) = \int_\Theta \int_{T^{-1}\{t\}} f_\theta(x)\,\upsilon_t(dx)\,\Pi(d\theta) = \int_\Theta f_{\theta T}^*(x)\,\Pi(d\theta)$ and so m_T^* is obtained by averaging the densities appropriate to checking each P_θ measure individually based on any statistic T. If T is minimal sufficient with $f_\theta(x) = g_\theta(T(x))h(x)$, then $m_T^*(t) = \int_\Theta g_\theta(t)\,\Pi(d\theta) \int_{T^{-1}\{t\}} h(x)\,\upsilon_t(dx)$.

If T is a complete minimal sufficient statistic, then we can ignore ancillaries and the relevant P-value is $M_T(m_T^*(t) \leq m_T^*(t_0))$. The following result shows that we need a slight modification for the general situation.

Theorem 8. Suppose that $\{f_\theta : \theta \in \Theta\}$ is a family of densities with respect to geometric measure $\mu_{\mathcal{X}}$ on $\mathcal{X}, \Pi$ is a prior probability measure on $\Theta, T : \mathcal{X} \to \mathcal{T}$ is

a smooth mapping onto manifold $\mathcal{T}$ with geometric measure $\mu_{\mathcal{T}}$ and $U : \mathcal{T} \to \mathcal{U}$ is a smooth mapping onto manifold $\mathcal{U}$ with geometric measure $\mu_{\mathcal{U}}$. Further suppose that T is minimal sufficient and $U \circ T$ is ancillary. Then the invariant P-value based on the conditional prior predictive of T given U is

$$M_T(m_T^*(t) \leq m_T^*(t_0) \,|\, U = u_0) \tag{13}$$

where $m_T^*(t) = \int_\Theta f_{\theta T}^*(x)\, \Pi(d\theta)$.

Proof: For $t \in U^{-1}\{u_0\}$ we have that $m_T(t \,|\, u_0) = \int_\Theta f_{\theta T}(t \,|\, u_0)\, \Pi(d\theta)$. Note that $T^{-1}\{t\} \cap T^{-1}U^{-1}\{u_0\} = T^{-1}\{t\}$ when $t \in U^{-1}\{u_0\}$ and is the empty set otherwise. If $t \in U^{-1}\{u_0\}$, then

$$f_{\theta T}(t \,|\, u_0) = f_{\theta T}(t) J_U(t)/f_U(u_0) = (J_U(t)/f_U(u_0)) \int_{T^{-1}\{t\}} f_\theta(x) J_T(x)\, \nu_t(dx).$$

Therefore, removing the volume distortions due to T and U, we have that $m_T^*(t \,|\, u_0) = \int_\Theta \int_{T^{-1}\{t\}} (f_\theta(x)/f_U(u_0))\, \nu_t(dx)\, \Pi(d\theta) = m_T^*(t)/f_U(u_0)$ and the result follows.

So (13) is obtained by averaging, with respect to the prior, the relevant functions for checking each P_θ measure and then, comparing the observed value of this function with its distribution under the prior predictive given the ancillary $U(T)$. As argued in Evans and Moshonov (2006, 2007), conditioning on the ancillary is appropriate when assessing prior-data conflict, as this removes variation from the assessment that has nothing to do with the prior.

For the examples included in Evans and Moshonov (2006, 2007) the only P-values that will change, when we use these invariant P-values, are those recorded for the location-scale models. In all the other examples the volume distortions are constant, either because of discreteness or because the model was a location model. The change in the P-values for location-scale models is illustrated by the following example, and we see that this is very small.

Example 8. *Prior-data conflict for the location-scale normal model.*

For a sample x of size n from the $N(\mu, \sigma^2)$ model, $T(x) = (\bar{x}, (n-1)^{-1}||x - \bar{x}1_n||^2) = (\bar{x}, s^2)$ is minimal sufficient. Then for a prior π on (μ, σ^2)

$$m_T(\bar{x}, s^2) = \int_0^\infty \int_{-\infty}^\infty \int_{T^{-1}\{(\bar{x}, s^2)\}} f(x \,|\, \mu, \sigma^2) J_T(x)\, \upsilon_{(\bar{x}, s^2)}(dx)\, \pi(\mu, \sigma^2)\, d\mu\, d\sigma^2$$

where $f(\cdot \,|\, \mu, \sigma^2)$ is the joint density of the sample, $\upsilon_{(\bar{x}, s^2)}$ is surface area measure on the $(n-2)$-dimensional sphere $\bar{x}1_n + ||x - \bar{x}1_n||(S^{n-1} \cap L^\perp\{1_n\})$, and $J_T(x) = n^{1/2}(n-1)^{1/2}/2s$. So $m_T^*(\bar{x}, s^2) = 2s m_T(\bar{x}, s^2)/n^{1/2}(n-1)^{1/2}$ and the P-values based on m_T and m_T^* differ by very little. In fact this difference disappears as n grows. The arbitrariness of the P-value based on the density is demonstrated by the fact that, if we had instead chosen the minimal sufficient statistic to be $T(x) = (n\bar{x}, ||x - \bar{x}1_n||)$, then the P-value based on m_T equals the invariant P-value.

It is not clear that the general use of discrepancy statistics is appropriate when checking for prior-data conflict. For there is no sense in which we think

of the prior as being wrong, as it represents some individual's (or individuals') beliefs about what the true distribution is. Rather we simply look to see if there is a conflict between what the data says about θ, as expressed by the likelihood function, and the prior. Evans and Moshonov (2006, 2007) assessed this by determining whether or not the observed likelihood function is a surprising value from its prior predictive distribution given an ancillary or, equivalently, whether or not the observed value of a minimal sufficient statistic is a surprising value from its prior predictive distribution given an ancillary. With the use of the invariant P-values developed here, this assessment becomes independent of the choice of a particular form chosen for the minimal sufficient statistic.

The assessment for prior-data conflict becomes more involved when a prior is specified hierarchically. For example, suppose the prior is specified componentwise as $\pi_2(\theta_2 \,|\, \theta_1)\pi_1(\theta_1)$, where the model parameter equals (θ_1, θ_2), or where θ_2 is the model parameter and θ_1 corresponds to hyperparameters. In such a case choices are made for the π_i and, as discussed in Evans and Moshonov (2006, 2007), we wonder if perhaps only part of this specification leads to a prior-data conflict. Of course, the P-values developed there for such cases should also be modified to use invariant P-values.

We have restricted our discussion to determining appropriate P-values for checks on the sampling model, based on the conditional distribution given a minimal sufficient statistic or based on an ancillary statistic, and separate checks for prior-data conflict, based on the conditional prior predictive of a minimal sufficient statistic given an ancillary. Other authors, such as Box (1980), Meng (1994), Gelman, Meng and Stern (1996), Berger and Bayarri (2000), and Bayarri and Castellanos (2007), have recommended P-values for Bayesian model checking that combine elements of the prior and the model. We feel that our developments are also relevant to the checks recommended by these authors.

6 Computations

Implementation of invariant P-values will sometimes require the numerical evaluation of f_T^*. In Example 6 we used simulation based upon the following result.

Lemma 9. Suppose that $\mathcal{X}$ and $\mathcal{T}$ are manifolds, with geometric measures $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{T}}$ respectively, and $T : \mathcal{U} \to \mathcal{T}$ is an onto, smooth mapping. If $E(J_T^{-1}(X)) < \infty$, then $f_T^*(t) = f_T(t)E(J_T^{-1}(X) \,|\, T = t)$.

Proof: For $B \subset \mathcal{T}$, by the measure decomposition theorem, we have that

$$E(J_T^{-1}(X)I_B(T(X))) = \int_{\mathcal{X}} J_T^{-1}(x)I_B(T(X))f(x)\,\mu_{\mathcal{X}}(dx)$$
$$= \int_B \int_{T^{-1}\{t\}} f(x)\,\nu_t(dx)\,\mu_{\mathcal{T}}(dt) = \int_B f_T^*(t)\mu_{\mathcal{T}}(dt)$$

and of course $E(J_T^{-1}(X)I_B(T(X))) = \int_B E(J_T^{-1}(X) \,|\, T = t)\, f_T(t)\mu_{\mathcal{T}}(dt)$.

Therefore we generate a sample $x_1, \ldots, x_n$ from f and use the kernel density estimator $\hat{f}_T(t) = n^{-1}\sum_{i=1}^n K_h(T(x_i) - t)$ to approximate $f_T(t)$ with $K_h(t) =$

$I_{(-1,1)}(t/h)/2h$. For small $h > 0$ and n large, $2hf_T(t)E(J_T^{-1}(X)\,|\,T = t) \approx n^{-1}\sum_{i=1}^n J_T^{-1}(x_i)I_{(t-h,t+h)}(T(x_i))$. Then we approximate $E(J_T^{-1}(X)\,|\,T = t)$ by $\hat{E}(J_T^{-1}(X)\,|\,T = t) = \sum_{i=1}^n J_T^{-1}(x_i)K_h(T(x_i) - t)/\sum_{i=1}^n K_h(T(x_i) - t)$, the Nadaraya-Watson estimator. The approximation is carried out at some of the $t_i = T(x_i)$ values. Further details on this estimator and kernel regression can be found in Wand and Jones (1995).

7 Conclusions

The use of P-values is somewhat of a controversial topic in statistics. In many ways, it seems to us, however, that the P-value represents the best way of assessing whether or not an observed value x_0 from a distribution P is surprising. Perhaps the first concern about P-values arises here, as it is not clear exactly how such a P-value should be computed. Some may argue that we must make use of a real-valued discrepancy statistic $T(x)$ and compute $P_T(T \geq T(x_0))$. While there is intuitive support for this, it only seems justified when the region of relatively low probability for T is just the right-tail. Further, it doesn't really help at all when T is multivariate.

We have argued that there is a logical basis for the development of appropriate P-values for discrete models. Further, we can carry this development over to suitably regular continuous models provided that we acknowledge that our continuous models are approximations to a discrete reality and, that we make sure that volume distortions induced by transformations do not affect our P-values. This leads to results that are the same or very similar, in many examples, to the way we currently compute P-values based simply on intuition. This is satisfying, as a radical change in such a fundamental tool would make us wary. Perhaps the most important consequence is that we feel that we have resolved the issue of the noninvariance of the P-value in the continuous case, and this makes us more confident that these are appropriate measures of surprise for the problems discussed.

8 Appendix

Proof of Theorem 1
(i) We need to consider the distribution of

$$\ln p_n(t) = \ln \binom{n}{t_1 \ldots t_k} \theta_1^{t_1} \cdots \theta_k^{t_k} = \sum_{i=1}^k (t_i \ln \theta_i - \ln t_i!) + \ln n!$$

when $(t_1, \ldots, t_k) \sim \text{Multinomial}(n, \theta_1, \ldots, \theta_k)$. For $M > 0$, let

$$C_M = \{(t_1, \ldots, t_k) : \max_{i=1,\ldots,k} |t_i - n\theta_i|/n^{1/2} \leq M \text{ and no } t_i = 0\}.$$

Since $(z_1, \ldots, z_k) = (t_1 - n\theta_1, \ldots, t_k - n\theta_k)/n^{1/2} \xrightarrow{D} N_k(0, \Sigma)$ as $n \to \infty$, where

$\sigma_{ii} = \theta_i(1-\theta_i)$ and $\sigma_{ij} = -\theta_i\theta_j$ when $i \neq j$, and since no $\theta_i = 0$, there exists M_ϵ and n_{M_ϵ}, such that for all $n \geq n_{M_\epsilon}$ then $P_T(C_{M_\epsilon}) > 1-\epsilon$.

Suppose that $(t_1, \ldots, t_k) \in C_{M_\epsilon}$. By Stirling's formula $\ln t_i! = (1/2)\ln 2\pi t_i + t_i \ln t_i - t_i + \lambda(t_i)$ where $|\lambda(t_i)| < 1/12t_i$ and, using $\sum_{i=1}^k t_i = n$,

$$\begin{aligned}
&\sum_{i=1}^k (t_i \ln \theta_i - \ln t_i!) \\
&= -\sum_{i=1}^k t_i \ln \frac{t_i}{n\theta_i} - \frac{1}{2}\sum_{i=1}^k \ln \frac{t_i}{n} + \sum_{i=1}^k \lambda(t_i) - n\ln n + n - \frac{k}{2}\ln 2\pi n. \qquad \text{(A1)}
\end{aligned}$$

Now $t_i \ln(t_i/n\theta_i) = (n\theta_i + \sqrt{n}z_i)\ln(1 + z_i/\sqrt{n}\theta_i)$ and since $|z_i| \leq M$, we can choose n larger than n_{M_ϵ} so that $|z_i|/\sqrt{n}\theta_i \leq 1/2$ for all i. When $|z| \leq 1/2$,

$$\left|\ln(1+z) - z + z^2/2\right| = \left|\sum_{i=1}^\infty (-1)^{i+1}\frac{z^i}{i} - z + z^2/2\right| \leq \sum_{i=3}^\infty \frac{|z|^i}{i} \leq \frac{2}{3}|z|^3,$$

so $\ln(1+z) = z - z^2/2 + O(|z|^3)$. Applying this to the first term in (A1), gives

$$\begin{aligned}
\sum_{i=1}^k t_i \ln \frac{t_i}{n\theta_i} &= \sum_{i=1}^k (n\theta_i + \sqrt{n}z_i)\ln(1 + z_i/\sqrt{n}\theta_i) \\
&= \sum_{i=1}^k (n\theta_i + \sqrt{n}z_i)\left\{\frac{z_i}{\sqrt{n}\theta_i} - \frac{z_i^2}{2n\theta_i^2} + O\left(\frac{|z_i|^3}{n^{3/2}}\right)\right\} \\
&= \sum_{i=1}^k \frac{z_i^2}{2\theta_i} + \sum_{i=1}^k O\left(\frac{|z_i|^3}{n^{1/2}}\right) + \sum_{i=1}^k O\left(\frac{|z_i|^4}{n}\right).
\end{aligned}$$

When $|z| \leq 1/2$, then $|\ln(1+z)| \leq \sum_{i=1}^\infty |z|^i/i = |z|\sum_{i=0}^\infty |z|^i/(i+1) \leq 2|z|$ and, applying this to the second term in (A1), gives

$$\frac{1}{2}\sum_{i=1}^k \ln \frac{t_i}{n} = \frac{1}{2}\sum_{i=1}^k \ln \theta_i + \frac{1}{2}\sum_{i=1}^k \ln\left(1 + \frac{z_i}{\sqrt{n}\theta_i}\right) = \frac{1}{2}\sum_{i=1}^k \ln \theta_i + \sum_{i=1}^k O\left(\frac{|z_i|}{\sqrt{n}}\right).$$

The third term in (A1) satisfies

$$\left|\sum_{i=1}^k \lambda(t_i)\right| \leq \sum_{i=1}^k \frac{1}{12t_i} = \frac{1}{12}\sum_{i=1}^k \frac{1}{n\theta_i + \sqrt{n}z_i} = O\left(\frac{1}{n}\right).$$

Combining all this and $|\ln n! - n\ln n + n - (1/2)\ln 2\pi n| \leq 1/12n$ gives

$$-\ln p_n(t) = \sum_{i=1}^k \frac{(t_i - n\theta_i)^2}{2n\theta_i} + \frac{1}{2}\sum_{i=1}^k \ln \theta_i + \frac{k-1}{2}\ln 2\pi n + r_n(t)$$

where

$$r_n(t) = \sum_{i=1}^{k} O\left(\frac{|z_i|}{n^{1/2}}\right) + \sum_{i=1}^{k} O\left(\frac{|z_i|^3}{n^{1/2}}\right) + \sum_{i=1}^{k} O\left(\frac{|z_i|^4}{n}\right) + O\left(\frac{1}{n}\right) = O\left(\frac{1}{n^{1/2}}\right).$$

when $(t_1, \ldots, t_k) \in C_{M_\epsilon}$ since $|z_i| \leq M_\epsilon$. Then, for $\eta > 0$, we have that

$$P_T(|r_n(t)| > \eta) \leq P_T(|r_n(t)| > \eta \,|\, C_{M_\epsilon}) P_T(C_{M_\epsilon}) + P_T(C^c_{M_\epsilon}) \leq \epsilon$$

since $P_T(|r_n(t)| > \eta \,|\, C_{M_\epsilon}) = 0$ for all n large enough and so $r_n(t) \xrightarrow{P} 0$. Since $\sum_{i=1}^{k} (t_i - n\theta_i)^2 / n\theta_i \xrightarrow{D} X$ where $X \sim$ Chi-squared$(k-1)$ we have proved (i).
(ii) We can write

$$\begin{aligned} &P_T\left(\ln p_n(t) \leq \ln p_n(t_0)\right) \\ &= P_T\left(\sum_{i=1}^{k} \frac{(t_i - n\theta_i)^2}{2n\theta_i} + r_n(t) \geq \sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i} + r_n(t_0)\right). \end{aligned}$$

Then, letting G_{k-1} denote the Chi-squared$(k-1)$ distribution function,

$$\begin{aligned} &P_{T_0}\left(\left|P_T\left(\ln p_n(t) \leq \ln p_n(t_0)\right) - \left\{1 - G_{k-1}\left(\sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i}\right)\right\}\right| > \eta\right) \\ &\leq P_{T_0}\left(\left| P_T\left(\begin{array}{c} \sum_{i=1}^{k} \frac{(t_i - n\theta_i)^2}{2n\theta_i} + r_n(t) \geq \\ \sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i} + r_n(t_0) \end{array}\right) - \right.\right. \\ &\qquad \left.\left. \left\{1 - G_{k-1}\left(\sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i} + r_n(t_0)\right)\right\} \right| > \eta/2\right) + \\ &P_{T_0}\left(\left| \begin{array}{c} \left\{1 - G_{k-1}\left(\sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i} + r_n(t_0)\right)\right\} \\ - \left\{1 - G_{k-1}\left(\sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i}\right)\right\} \end{array} \right| > \eta/2\right) \end{aligned} \tag{A2}$$

Now let $\epsilon > 0$ satisfy $\epsilon < \eta/2$. When X has a continuous distribution and $X_n \xrightarrow{D} X$ then, for any $\epsilon > 0$, we have that $\sup_x |P(X \leq x) - P(X_n \leq x)| \leq \epsilon$ for all n large enough. So, since $r_n(t) \xrightarrow{P} 0$, the first term on the right in (A2) equals 0 for all n large enough. Since $r_n(t_0) \xrightarrow{P} 0$, then

$$G_{k-1}\left(\sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i} + r_n(t_0)\right) \xrightarrow{P} G_{k-1}\left(\sum_{i=1}^{k} \frac{(t_{i0} - n\theta_i)^2}{n\theta_i}\right).$$

Combining all this we have proved (ii).
(iii) First we note that if $p_i = 0$, then $P_T(t_i = 0) = 1$ and so $p_i \ln(\theta_i/p_i) = t_i \ln(\theta_i/p_i) = 0$. Now $(z_1, \ldots, z_k) = (t_1 - np_1, \ldots, t_k - np_k)/n^{1/2} \xrightarrow{D} N_k(0, \Sigma)$ as $n \to \infty$, where $\sigma_{ii} = p_i(1 - p_i)$ and $\sigma_{ij} = -p_i p_j$ when $i \neq j$. We have that

$$\ln p_n(t_0) = \sum_{i=1}^{k} t_{i0}(\ln \theta_i - \ln p_i) + \sum_{i=1}^{k} (t_{i0} \ln p_i - \ln t_{i0}!) + \ln n!. \tag{A3}$$

We can apply the same analysis to $\sum_{i=1}^{k}(t_{i0}\ln p_i - \ln t_{i0}!)$ as we did in (i) but now we must take into account that whenever $p_i = 0$, then corresponding terms are dropped. Doing this we obtain (always interpreting $0 \cdot \infty = 0$)

$$\begin{aligned} s_n(t_0) &= -\ln p_n(t_0) - \frac{1}{2}\sum_{i=1}^{k}\ln\theta_i - \left(\ln n! - n\ln n + n - \frac{k}{2}\ln 2\pi n\right) \\ &= -n\sum_{i=1}^{k}\frac{t_{i0}}{n}\ln\frac{\theta_i}{p_i} + \sum_{i=1}^{k}\frac{(t_{i0}-np_i)^2}{2np_i} - \frac{1}{2}\sum_{i=1}^{k}\ln\theta_i + \frac{1}{2}\sum_{i=1,p_i\neq 0}^{k}\ln p_i \\ &+ \frac{l}{2}\ln 2\pi n + r_n^*(t_0) \end{aligned}$$

where l is the number of $p_i = 0$ and $r_n^*(t_0) \xrightarrow{P} 0$ when sampling from the Multinomial$(1, p_1, \ldots, p_k)$ distribution. Arguing as in (ii), we have that the P-value given by (5) converges in probability to $1 - G_{k-1}(s_n(t_0))$. Under sampling from the Multinomial$(1, p_1, \ldots, p_k)$ distribution $\sum_{i=1}^{k}(t_{i0}/n)\ln(\theta_i/p_i) \xrightarrow{a.s} \sum_{i=1}^{k} p_i \ln(\theta_i/p_i)$. This equals minus the Kullback Leibler distance between the p_i and θ_i distributions and so is negative when $p_i \neq \theta_i$ for some i. It is then immediate that $s_n(t_0) \xrightarrow{P} \infty$ and this completes the proof of (iii).

Proof of Theorem 2

Let B be a bounded set formed from a union of elements of $\{B_1(x) : x \in R^k\}$, such that $P(B^c) < \epsilon$ and $x_0 \in B$. Since $P(B_n(x) \,|\, B) = P(B_n(x))/P(B)$ when $B_n(x) \subset B$ and $P(B_n(x) \,|\, B) = 0$ otherwise, we have that

$$\left|\sum_{\{x_n(x):P(B_n(x))\leq P(B_n(x_0))\}} P(B_n(x)) - P(f(x)\leq f(x_0))\right| \leq 2\epsilon +$$

$$\left|\sum_{\{x_n(x):P(B_n(x)\,|\,B)\leq P(B_n(x_0)\,|\,B)\}} P(B_n(x)\,|\,B) - P(f(x)\leq f(x_0)\,|\,B)\right| P(B).$$

So, if we prove that

$$\sum_{\{x_n(x):P(B_n(x)\,|\,B)\leq P(B_n(x_0)\,|\,B)\}} P(B_n(x)\,|\,B) \to P(f(x)\leq f(x_0)\,|\,B),$$

as $n \to \infty$, then the result will be established. Accordingly, we hereafter assume that $\mathcal{X}$ is contained in a bounded set B with $\{B_n(x) : x \in R^k\}$ a finite partition of B.

Now suppose that f is unbounded on $\mathcal{X}$ and let $\epsilon > 0$. Let $M > 0$ and $\mathcal{X}_M = \{x : f(x) < M\}$. Since $P(\mathcal{X}_M^c) \to 0$ as $M \to \infty$, we can find M such that $P(\mathcal{X}_M^c) < \epsilon$. Since $\cup_{\{x_n(x):B_n(x)\cap\mathcal{X}_M^c\neq\phi)\}} B_n(x)$ is monotonically decreasing to $\mathcal{X}_M^c$, there exists n_0 such that for all $n \geq n_0$, then

$$\left|\sum_{\{x_n(x):B_n(x)\cap\mathcal{X}_M^c\neq\phi)\}} P(B_n(x)) - P(\mathcal{X}_M^c)\right| < \epsilon.$$

Therefore, taking $B' = B \setminus \cup_{\{x: B_{n_0}(x) \cap \mathcal{X}_M^c \neq \phi)\}} B_{n_0}(x)$, and reasoning as in the preceding paragraph with B' replacing B, we see that we need only prove the result when f is bounded. We assume f is bounded hereafter.

Since f is continuous on $\mathcal{X}$, we have that $P(B_n(x))/\mu(B_n(x)) \to f(x)$ for all $x \in \mathcal{X}$. For each $x \in \mathcal{X}$ there exists $x'_n(x) \in B_n(x)$ such that $P(B_n(x)) = f(x'_n(x))\mu(B_n(x))$ and so, since $\mu(B_n(x))$ is finite and constant, (7) equals

$$\sum_{\{x_n(x): f(x'_n(x)) \leq f(x'_n(x_0))\}} f(x'_n(x))\mu(B_n(x)). \tag{A4}$$

Since $\mathcal{X}$ is contained in the union of finitely many of the $B_n(x)$, the sum in (A4) is a finite sum. Now

$$\begin{aligned}
&\sum_{\{x_n(x): f(x'_n(x)) \leq f(x'_n(x_0))\}} f(x'_n(x))\mu(B_n(x)) \\
&\quad = \sum_{\{x_n(x): f(x'_n(x)) < f(x_0)\}} f(x'_n(x))\mu(B_n(x)) + \\
&\qquad \sum_{\{x_n(x): f(x_0) \leq f(x'_n(x)) \leq f(x'_n(x_0))\}} f(x'_n(x))\mu(B_n(x)) - \\
&\qquad\quad \sum_{\{x_n(x): f(x'_n(x_0)) \leq f(x'_n(x)) < f(x_0)\}} f(x'_\delta(x))\mu(B_\delta(x)).
\end{aligned}$$

We have that

$$\sum_{\{x_n(x): f(x'_n(x)) < f(x_0))\}} f(x'_n(x))\mu(B_n(x)) \to P(f(x) < f(x_0))$$

as $n \to 0$ as the left side is an approximating Riemann sum to the right side. Further, $f(x'_n(x_0)) \to f(x_0)$ as $n \to \infty$ and so, for $\epsilon > 0$ we can find n_ϵ such that for all $n \geq n_\epsilon$, then $|f(x_0) - f(x'_n(x_0))| < \epsilon$. Accordingly,

$$\begin{aligned}
&\sum_{\{x_n(x): f(x'_n(x_0)) \leq f(x'_n(x)) < f(x_0)\}} f(x'_n(x))\mu(B_n(x)) \\
&\quad \leq \sum_{\{x_n(x): f(x_0) - \epsilon < f(x'_n(x)) < f(x_0)\}} f(x'_n(x))\mu(B_n(x)) \\
&\qquad\qquad \to P(f(x_0) - \epsilon < f(x) < f(x_0))
\end{aligned}$$

as $n \to \infty$ and this upper bound converges to 0 as $\epsilon \to 0$.

Now we have that $1 = I_{LC(x_0)} + I_{f^{-1}f(x_0) \cap LC(x_0)^c} + I_{(f^{-1}f(x_0))^c}$,

$$\begin{aligned}
&\sum_{\{x_n(x): f(x_0) \leq f(x'_n(x)) \leq f(x'_n(x_0))\}} I_{f^{-1}f(x_0) \cap LC(x_0)^c}(x'_n(x)) f(x'_n(x))\mu(B_n(x)) \\
&= f(x_0) \sum_{\{x_n(x): f(x_0) \leq f(x'_n(x)) \leq f(x'_n(x_0))\}} I_{f^{-1}f(x_0) \cap LC(x_0)^c}(x'_n(x))\mu(B_n(x))
\end{aligned}$$

$$\leq f(x_0) \sum_{x_n(x)} I_{f^{-1}f(x_0)\cap LC(x_0)^c}(x'_n(x))\mu(B_n(x)) \to f(x_0)\mu(f^{-1}f(x_0)\cap LC(x_0)^c)$$

where $\mu(f^{-1}f(x_0)\cap LC(x_0)^c) = 0$ and

$$\begin{aligned}
&\sum_{\{x_n(x):f(x_0)\leq f(x'_n(x))\leq f(x'_n(x_0))\}} I_{(f^{-1}f(x_0))^c}(x'_n(x))f(x'_n(x))\mu(B_n(x)) \\
&\leq \sum_{\{x_n(x):f(x_0)\leq f(x'_n(x))\leq f(x_0)+\epsilon\}} I_{(f^{-1}f(x_0))^c}(x'_n(x))f(x'_n(x))\mu(B_n(x)) \\
&\to P(\{x : f(x_0)\leq f(x)\leq f(x_0)+\epsilon\}\cap \left(f^{-1}f(x_0)\right)^c)
\end{aligned}$$

and this converges to $P(f^{-1}f(x_0)\cap \left(f^{-1}f(x_0)\right)^c) = 0$ as $\epsilon \to 0$.

Finally, we consider

$$\sum_{\{x_n(x):f(x_0)\leq f(x'_n(x))\leq f(x'_n(x_0))\}} I_{LC(x_0)}(x'_n(x))f(x'_n(x))\mu(B_n(x)).$$

Now $LC(x_0)$ is covered by finitely many of the $B_n(x)$. Let $LC^\epsilon(x_0)$ be the set of points in $LC(x_0)$ that lie a distance greater than ϵ from $\partial LC(x_0)$. Since $LC(x_0)$ is an open set, $LC^\epsilon(x_0) \uparrow LC(x_0)$ as $\epsilon \to 0$ and so $\mu(LC^\epsilon(x_0)) \uparrow \mu(LC(x_0))$. We can choose n_ϵ so that when $n \geq n_\epsilon$, then $x_n(x) \in LC^{2\epsilon}(x_0)$ then $x'_\delta(x) \in LC^\epsilon(x_0)$ and so

$$\begin{aligned}
&\sum_{\{x_n(x):f(x_0)\leq f(x'_n(x))\leq f(x'_n(x_0))\}} I_{LC^{2\epsilon}(x_0)}(x'_n(x))f(x'_n(x))\mu(B_n(x)) \\
&= f(x_0)\sum_{x_n(x)} I_{LC^{2\epsilon}(x_0)}(x'_n(x))\mu(B_n(x)) \\
&\leq \sum_{\{x_n(x):f(x_0)\leq f(x'_n(x))\leq f(x'_n(x_0))\}} I_{LC(x_0)}(x'_n(x))f(x'_n(x))\mu(B_n(x)) \\
&\leq \sum_{\{x_n(x):f(x_0)\leq f(x'_n(x))\leq f(x_0)+\epsilon\}} I_{LC(x_0)}(x'_n(x))(f(x_0)+\epsilon)\mu(B_\delta(x)). \qquad \text{(A5)}
\end{aligned}$$

Now the left-hand side of (A5) converges to $f(x_0)\mu(LC^{2\epsilon}(x_0))$ which converges to $f(x_0)\mu(LC(x_0)) = P(f(x) = f(x_0))$ as $\epsilon \to 0$. The right-hand side of (A5) converges to

$$\begin{aligned}
&P(LC(x_0)\cap\{x : f(x_0)\leq f(x)\leq f(x_0)+\epsilon\})+ \\
&\epsilon\mu(LC(x_0)\cap\{x : f(x_0)\leq f(x)\leq f(x_0)+\epsilon\})
\end{aligned}$$

which converges to $P(f(x) = f(x_0))$ as $\epsilon \to 0$ and this establishes the result.

Example Where Theorem 2 Fails

Let $q_1, q_2, \ldots$ be a listing of all rational numbers in $\mathcal{X} = (0,1)$. Fix a $\delta \in (0, 1/8)$ sufficiently small and let $A_0 = \mathcal{X}\cap\cup_{i=1}^\infty(q_i - \delta 2^{-i}, q_i + \delta 2^{-i})$. For $x \in A_0$, there is an interval $(a,b) \subset \mathcal{X}$ such that $x \in (a,b) \subset A_0$. For $x \in A_0$ define

$a(x) = \inf\{a \in [0,x] \,|\, (a,x) \subset A_0\}$ and $b(x) = \sup\{b \in [x,1] \,|\, (x,b) \subset A_0\}$. Then, $(a(x), b(x)) \subset A_0$ for all $x \in A_0$. Since A_0 is a countable union of intervals, there are countably many x_i's in A_0 such that $A_0 = \cup_{i=1}^{\infty}(a(x_i), b(x_i))$ and such that these intervals are disjoint.

The intervals $(a(x_i), b(x_i))$'s can be ordered to obtain the class of intervals $\{(a_i, b_i) : i = 1, 2, \ldots\}$ where the ordering is such that (a_i, b_i) satisfies $i < j$ whenever $b_i - a_i > b_j - a_j$, or $a_i < a_j$ when $b_i - a_i = b_j - a_j$. Now let $A = \mathcal{X} \cap A_0^c$ and define probability density $f(x)$ on $\mathcal{X}$ as

$$f(x) = \begin{cases} 1 & \text{if } x \in A, \\ 1 + \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(x - a_i)}{b_i - a_i} & \text{if } x \in (a_i, b_i) \text{ for some } i. \end{cases}$$

We see immediately that f is continuous on $\mathcal{X}$ and so, for every $x \in \mathcal{X}$, $\lim_{n\to\infty} P(B_n(x))/\mu_k(B_n(x)) = f(x)$ for any $B_n(x)$ shrinking nicely to x.

We need the following results.

Lemma 10. (i) volume$(A) = 1 - 2\alpha \geq 1 - 2\delta$ where $\alpha = \sum_{i=1}^{\infty}(b_i - a_i)/2$.
(ii) For any $x_0 \in A$, $P(f(x) \leq f(x_0)) = 1 - \beta - \alpha$ where $\beta = \sum_{i=1}^{\infty}(b_i - a_i)^3/\pi^2$.
Proof: (i) We have that volume$(A_0) = A_0 = \cup_{i=1}^{\infty}(a_i, b_i) = \sum_{i=1}^{\infty}(b_i - a_i) = 2\alpha$ and volume$(A) = 1 -$ volume(A_0). Further,

$$\begin{aligned} \text{volume}(A) &= \text{volume}((0,1)\backslash A_0) = \text{volume}((0,1) \cap \cap_{i=1}^{\infty}(q_i - \delta 2^{-1}, q_i + \delta 2^{-i})^c) \\ &\geq 1 - \sum_{i=1}^{\infty} \text{volume}((q_i - \delta 2^{-1}, q_i + \delta 2^{-i})) = 1 - \sum_{i=1}^{\infty} 2\delta 2^{-i} = 1 - 2\delta. \end{aligned}$$

(ii) For any $x_0 \in A$, $f(x_0) = 1$ and

$$\begin{aligned} &P(f(x) \leq f(x_0)) = P(f(x) \leq 1) = 1 - P(f(x) > 1) \\ &= 1 - P(\cup_{i=1}^{\infty}(a_i, (a_i + b_i)/2)) \\ &= 1 - \sum_{i=1}^{\infty} \int_{a_i}^{\frac{a_i + b_i}{2}} f(x)dx \\ &= 1 - \sum_{i=1}^{\infty} \left[\frac{b_i - a_i}{2} + \int_{a_i}^{\frac{a_i + b_i}{2}} \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(x - a_i)}{b_i - a_i} dx \right] \\ &= 1 - \alpha + \sum_{i=1}^{\infty} \frac{(b_i - a_i)^3}{2\pi^2} \cos \frac{2\pi(x - a_i)}{b_i - a_i} \Big|_{x=a_i}^{x=(a_i+b_i)/2} = 1 - \alpha - \sum_{i=1}^{\infty} \frac{(b_i - a_i)^3}{\pi^2} \\ &= 1 - \alpha - \beta. \end{aligned}$$

Note that $\delta \in (0, 1/8)$ and Lemma 10(i) implies that the volume of A is bigger than $3/4$.

Now let $B_n(x) = \mathcal{X} \cap ((k-1)2^{-n}, k2^{-n}]$ where $k = \lceil 2^n x \rceil$ for $x \in \mathcal{X}$. Then, $B_n(x) = (l_n(x), u_n(x)]$ for $k < 2^n$ and $B_n(x) = (l_n(x), u_n(x))$ if $k = 2^n$.

Lemma 11. The set $N(x) = \{n \in \mathbb{N} : x - l_n(x) > u_n(x) - x\}$ has infinitely many elements for $x \in \mathcal{X}$.

Proof: The binary expansion of x is given by $x = [0.x_1x_2x_3\cdots]_2 = \sum_{i=1}^{\infty} x_i 2^{-i}$. Then, $2^n x = [x_1\cdots x_n.x_{n+1}x_{n+2}\cdots]_2$ and $u_n(x) = 2^{-n} + [0.x_1\cdots x_n]_2$ if $x_{n+k} = 1$ for some $k > 0$ and otherwise, i.e., $x_{n+k} = 0$ for all $k \in \mathbb{N}$ or equivalently $x = m2^{-n}$ for some $m \in \mathbb{N}$, $u_n(x) = [0.x_1\cdots x_n]_2 = x$.

For the second case, i.e., $u_n(x) = x$. Then, $u_k(x) = x$ for all $k \geq n$. Note $l_k(x) = u_k(x) - 2^{-k}$. Thus, $x - l_k(x) = u_k(x) - l_k(x) = 2^{-k} > 0 = u_k(x) - x$ for all $k \geq n$. Hence, $N(x)$ contains infinitely many elements.

For the first case, i.e., there are infinitely many 1's in $x_1, x_2, \ldots$, assume that $N(x)$ is finite. Then there is a number $M \in \mathbb{N}$ such that $x - l_n(x) \leq u_n(x) - x$ for all $n \geq M$. Now $x - l_n(x) \leq u_n(x) - x$ implies that $x_{n+1} = 0$ since $x - l_n(x) < u_n(x) - x$ implies this and if $x - l_n(x) = u_n(x) - x$, then we would be in the second case as $u_{n+1}(x) = x$. Hence, we get $x_n = 0$ for all $n \geq 1 + M$. In other words $x = [0.x_1\cdots x_M]_2 = m2^{-M}$ for some $m \in \mathbb{N}$ and this is a contradiction. Hence $N(x)$ must contain infinitely many elements.

Let $\mathcal{A} = \{a_i : i = 1, 2, \ldots\}$ and $\mathcal{B} = \{b_i : i = 1, 2, \ldots\}$. We have the following result.

Lemma 12. Each $x \in A$ is irrational and $\mathcal{A} \cup \mathcal{B} \subset A$.
Proof: We must have that $x \in A$ is irrational since A_0 contains all the rationals in $(0, 1)$ and $A = (0, 1)\backslash A_0$. For any $x \in \mathcal{A} \cup \mathcal{B}$, we have $x \in A$ because $A = (0, 1)\backslash \cup_{i=1}^{\infty} (a_i, b_i)$.

Now if $x \in A$, then $l_n(x) < x < u_n(x)$ since $l_n(x)$ and $u_n(x)$ are rational and x is irrational by Lemma 12. Since $l_n(x) \in A_0$, there exists i such that $l_n(x) \in (a_i, b_i)$ and $b_i \leq x$ since $x \notin A_0$. Therefore, $a_i < l_n(x) < b_i \leq x$ implying $b_i = \min(\mathcal{B} \cap B_n(x))$ and we define $a_{l,n}(x) = a_i, b_{l,n}(x) = b_i$. Similarly, since $u_n(x) \in A_0$, there exists j such that $u_n(x) \in (a_j, b_j)$ and $x \leq a_j$ since $x \notin A_0$. Therefore, $x \leq a_j < u_n(x) < b_j$ implying $a_j = \max(\mathcal{A} \cap B_n(x))$ and we define $a_{u,n}(x) = a_j, b_{u,n}(x) = b_j$. Note that

$$a_{l,n}(x) < l_n(x) < b_{l,n}(x) < a_{u,n}(x) < u_n(x) < b_{u,n}(x) \tag{A6}$$

whenever $x \in A$. Note that we must have that $b_{l,n}(b) = b$ for any $b \in \mathcal{B}$ and $a_{u,n}(a) = a$ for any $a \in \mathcal{A}$.

We need the following trigonometric inequalities.

Lemma 13. (i) $|\sin x| \leq |x|$ for all $x \in \mathbb{R}$,
(ii) $\cos x \leq 1 - x^2/2 + x^4/24 = 1 - (x^2/2)(1 - x^2/12)$ for all $|x| \leq \sqrt{56}$.
Proof: (i) The result is well-known.
(ii) The trigonometric function expansion of $\cos x$ is given by

$$\cos x = \sum_{i=0}^{\infty} (-1)^i \frac{x^{2i}}{(2i)!} = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \sum_{j=2}^{\infty} \frac{x^{4j-2}}{(4j-2)!} \left(1 - \frac{x^2}{(4j-1)(4j)}\right).$$

For $|x| \leq \sqrt{56}$ and $j \geq 2$, we have $1 - x^2/[(4j-1)4j] \geq 0$ as well as

$$\cos x \leq \sum_{i=0}^{\infty} (-1)^i \frac{x^{2i}}{(2i)!} = 1 - \frac{x^2}{2} + \frac{x^4}{24}.$$

Hence, the lemma follows.

We now establish some inequalities for $P(B_n(x))$.

Lemma 14. For $x \in A$ we have that $P(B_n(x))$ satisfies

$$P(B_n(x)) \geq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))(b_{l,n}(x) - l_n(x))^2 \tag{A7}$$

and

$$\begin{aligned} &P(B_n(x)) \\ &\leq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))(b_{l,n}(x) - l_n(x))^2 \left(1 - \frac{\pi^2 (b_{l,n}(x) - l_n(x))^2}{3(b_{l,n}(x) - a_{l,n}(x))^2}\right) + \\ &(b_{u,n}(x) - a_{u,n}(x))(u_n(x) - a_{u,n}(x))^2. \end{aligned} \tag{A8}$$

Proof: Suppose $x \in A$. Let $M_n(x) = \{i \in \mathbb{N} : (a_i, b_i) \cap B_n(x) \neq \emptyset\}$.

$$\begin{aligned} P(B_n(x)) &= \int_{l_n(x)}^{u_n(x)} f(v)dv \\ &= \int_{l_n(x)}^{u_n(x)} 1 + \sum_{i \in M_n(x)} I_{(a_i,b_i)}(v) \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(v - a_i)}{b_i - a_i} dv \\ &= 2^{-n} + \int_{l_n(x)}^{b_{l,n}(x)} \frac{(b_{l,n}(x) - a_{l,n}(x))^2}{\pi} \sin \frac{2\pi(v - a_{l,n}(x))}{b_{l,n}(x) - a_{l,n}(x)} dv \\ &+ \int_{a_{u,n}(x)}^{u_n(x)} \frac{(b_{u,n}(x) - a_{u,n}(x))^2}{\pi} \sin \frac{2\pi(v - a_{u,n}(x))}{b_{u,n}(x) - a_{u,n}(x)} dv \\ &= 2^{-n} + \frac{(b_{l,n}(x) - a_{l,n}(x))^3}{2\pi^2} (\cos \frac{2\pi(b_{l,n}(x) - l_n(x))}{b_{l,n}(x) - a_{l,n}(x)} - 1) \qquad \text{(A9)} \\ &+ \frac{(b_{u,n}(x) - a_{u,n}(x))^3}{2\pi^2} \left(1 - \cos \frac{2\pi(u_n(x) - a_{u,n}(x))}{b_{u,n}(x) - a_{u,n}(x)}\right) \\ &= 2^{-n} - \frac{(b_{l,n}(x) - a_{l,n}(x))^3}{\pi^2} \sin^2 \frac{\pi(b_{l,n}(x) - l_n(x))}{b_{l,n}(x) - a_{l,n}(x)} \\ &+ \frac{(b_{u,n}(x) - a_{u,n}(x))^3}{\pi^2} \sin^2 \frac{\pi(u_n(x) - a_{u,n}(x))}{b_{u,n}(x) - a_{u,n}(x)}. \qquad \text{(A10)} \end{aligned}$$

where the last equality is derived using $1 - \cos x = 2\sin^2(x/2)$. Then (A7) follows from Lemma 13(i) and (A8) follows from

$$0 \leq \frac{2\pi(b_{l,n}(x) - l_n(x))}{b_{l,n}(x) - a_{l,n}(x)} \leq 2\pi < \sqrt{56}$$

and applying Lemma 13(ii) to the second term in (A9) and Lemma 13(i) to the last term in (A10).

Lemma 15. Consider $b_j \in \mathcal{B}$ and let $\mathcal{B}_n = \{B_n(x) \,|\, x \in \mathcal{X}\}$. Then,

$$\liminf_{n \to \infty} \sum_{B \in \mathcal{B}_n : P(B) \leq P(B_n(b_j))} P(B) \leq \alpha - \beta.$$

Proof: Let $\mathcal{C}_{1,n} = \{B \in \mathcal{B}_n \,|\, B \cap \mathcal{B} \neq \emptyset, P(B) \leq P(B_n(b_j))\}$ and $\mathcal{C}_{2,n} = \{B \in \mathcal{B}_n \,|\, B \cap \mathcal{B} = \emptyset, P(B) \leq P(B_n(b_j))\}$. Then, $\{B \in \mathcal{B}_n \,|\, P(B) \leq P(B_n(b_j))\} = \mathcal{C}_{1,n} \cup \mathcal{C}_{2,n}$. We will prove that (i) $P(\Sigma(\mathcal{C}_{2,n})) \to \alpha - \beta$ as $n \to \infty$ and (ii) $P(\Sigma(\mathcal{C}_{1,n})) \to 0$ as $n \to \infty$, subject to $n \in N(b_j)$, where $\Sigma(\mathcal{C}) = \cup_{C \in \mathcal{C}} C$.
(i) Since $b_j \in A$, we have that $P(B_n(b_j))/\text{volume}(B_n(b_j)) \to 1$, and also

$$P(B_n(x))/\text{volume}(B_n(x)) \to 1 + \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(x - a_i)}{b_i - a_i}$$

when $x \in (a_i, b_i)$ for some i. It is then clear that $\liminf_{n\to\infty} \Sigma(\mathcal{C}_{2,n}) \supset \mathcal{D}_{1,\epsilon} = \{x \in \mathcal{X} \,|\, f(x) < 1 - \epsilon\}$. Further, if $x \in A$, then by (A6) $B_n(x) \cap \mathcal{B} \neq \emptyset$ for any n and so $\limsup_{n\to\infty} \Sigma(\mathcal{C}_{2,n}) \subset \mathcal{D}_{2,\epsilon} = \{x \in \mathcal{X} \,|\, f(x) < 1 + \epsilon, a_i < x < b_i \text{ for some } i\}$ for any $\epsilon > 0$. Hence, $\{x \in \mathcal{X} \,|\, f(x) < 1\} \subset \liminf_{n\to\infty} \Sigma(\mathcal{C}_{2,n}) \subset \limsup_{n\to\infty} \Sigma(\mathcal{C}_{2,n}) \subset \{x \in \mathcal{X} \,|\, f(x) <= 1, a_i < x < b_i \text{ for some } i\}$. Note $P(\{x \in \mathcal{X} \,|\, f(x) = 1, a_i < x < b_i \text{ for some } i\}) = 0$ because this set contains countably many points. Thus,

$$\begin{aligned} P(\Sigma(\mathcal{C}_{2,n})) &\to P(\{x \in \mathcal{X} \,|\, f(x) < 1\}) \\ &= \sum_{i=1}^{\infty} \int_{\frac{a_i+b_i}{2}}^{b_i} 1 + \frac{(b_i - a_i)^2}{\pi} \sin \frac{2\pi(x - a_i)}{b_i - a_i} dx = \alpha - \beta. \end{aligned}$$

(ii) Fix $\epsilon \in (0, 1/4)$. By the comment after (A6) we have that $a_{l,n}(b_j) = a_j$ and $b_{l,n}(b_j) = b_j$. As n increases, $b_{u,n}(b_j) - a_{u,n}(b_j)$ converges to 0, because the length of $B_n(b_j)$ shrinks to 0, and also $b_j - l_n(b_j) \to 0$. Hence, there is a number $N_0 > 0$ such that for all $n \geq N_0$, we have that $\pi^2(2^{-n})^2/3(b_j - a_j)^2 < \min\{1, \epsilon\}$ and $b_{u,n}(b_j) - a_{u,n}(b_j) < \epsilon(b_j - a_j)/4$. Then, for all $n \geq N_0$ in $N(b_j)$ defined in Lemma 11, and using $b_j - l_n(b_j) \leq 2^{-n}$, the upper bound on $P(B_n(b_j))$ in (A8) becomes

$$\begin{aligned} P(B_n(b_j)) &\leq 2^{-n} - (b_j - a_j)(b_j - l_n(b_j))^2 \left(1 - \frac{\pi^2(b_j - l_n(b_j))^2}{3(b_j - a_j)^2}\right) \\ &\quad + (b_{u,n}(b_j) - a_{u,n}(b_j))(u_n(b_j) - a_{u,n}(b_j))^2 \\ &\leq 2^{-n} - (b_j - a_j)(2^{-n}/2)^2 \left(1 - \frac{\pi^2(2^{-n})^2}{3(b_j - a_j)^2}\right) + \epsilon(b_j - a_j)(2^{-n}/2)^2 \\ &< 2^{-n} - (b_j - a_j)(2^{-n}/2)^2(1 - 2\epsilon). \end{aligned}$$

For any $x \in A$, the inequality (A7) becomes

$$\begin{aligned} P(B_n(x)) &\geq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x)(b_{l,n}(x) - l_n(x))^2 \\ &\geq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))2^{-2n}. \end{aligned}$$

Hence, for all $x \in A$ satisfying $(b_{l,n}(x) - a_{l,n}(x)) < (b_j - a_j)/8$,

$$\begin{aligned} P(B_n(x)) &\geq 2^{-n} - (b_{l,n}(x) - a_{l,n}(x))2^{-2n} > 2^{-n} - (b_j - a_j)2^{-2n}/8 \\ &\geq 2^{-n} - (b_j - a_j)2^{-2n}(1 - 2\epsilon)/4 > P(B_n(b_j)). \end{aligned}$$

Thus, for $x \in A, P(B_n(x)) \leq P(B_n(b_j))$ implies $b_{l,n}(x) - b_{l,n}(x) \geq (b_j - a_j)/8$.

Since $b_i - a_i \to 0$ as $i \to \infty$, there is $J \in \mathbb{N}$ such that $b_i - a_i < (b_j - a_j)/8$ for all $i > J$. It then follows that

$$\begin{aligned}\mathcal{C}_{1,n} &= \{B \in \mathcal{B}_n \,|\, B \cap \mathcal{B} \neq \emptyset, P(B) \leq P(B_n(b_j))\} \\ &= \{B_n(x) \,|\, x \in \mathcal{B}, P(B_n(x)) \leq P(B_n(b_j))\} \\ &\subset \{B_n(b_i) \,|\, i \leq J\}.\end{aligned}$$

Then, since $f(x) \leq 2$ for all $x \in \mathcal{X}$, we have that $P(\Sigma(\mathcal{C}_{1,n})) \leq P(\Sigma(\{B_n(b_i) \,|\, i \leq J\})) \leq 2J/2^n \to 0$ as $n \to \infty$ subject to $n \in N_0(b_j)$ and this establishes (ii).

The result follows by combining (i) and (ii).

We now have the final result.

Theorem 16. For the distribution on $\mathcal{X} = (0, 1)$, with continuous density given by f, and for $b_j \in \mathcal{B}$, we have that $P(f(x) \leq f(b_j)) = 1 - \beta - \alpha$ while $\sum_{B \in \mathcal{B}_n : P(B) \leq P(B_n(b_j))} P(B)$ either doesn't converge or has limit equal to $\alpha - \beta < 1 - \beta - \alpha$.

Proof: By Lemma 12, $b_j \in A$ so by Lemma 10(ii) we have that $P(f(x) \leq f(b_j)) = 1 - \beta - \alpha$. By Lemma 15, if $\sum_{B \in \mathcal{B}_n : P(B) \leq P(B_n(b_j))} P(B)$ converges, then it converges to $\alpha - \beta$ since we constructed a subsequence converging to this value in Lemma 15. If $\alpha - \beta \geq 1 - \beta - \alpha$, then $\alpha \geq 1/2$, but by Lemma 10(i) and $\delta \in (0, 1/8)$, we have that $\alpha < \delta \leq 1/8$.

References

Bayarri, M.J. and Berger, J.O. (2000) P-values for composite null models (with discussion). Journal of the American Statistical Association, Vol. 95, 452, p. 1143-1156.

Bayarri, M.J. and Castellanos, M.E. (2007) Bayesian checking of the second levels of hierarchical models (with discussion). Statistical Science, 22, 3, p. 322-343.

Berger, J.O. and Delampady, M. (1987) Testing precise hypotheses. Statistical Science, 3, p. 317-352.

Berger, J.O. and Selke, T. (1987) Testing a point null hypothesis: the irreconcilability of P-values and evidence (with comments). Journal of the American Statistical Association, 82, p. 112-122.

Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, A, 143, p. 383-430.

Evans, M. and Moshonov, H. (2006) Checking for prior-data conflict. Bayesian Analysis, Volume 1, Number 4, p. 893-914.

Evans, M. and Moshonov, H., (2007) Checking for prior-data conflict with hierarchically specified priors. Bayesian Statistics and its Applications, eds. A.K. Upadhyay, U. Singh, D. Dey, Anamaya Publishers, New Delhi, p. 145-159.

Johnson, V.E. (2004) A Bayesian chi-square test for goodness of fit. Annals of Statistics, 32, p. 2361-2384.

Fraser, D.A.S. (1979) Inference and Linear Models. McGraw-Hill.

Gelman, A., Meng, X.-Li, and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, p. 733-807.

Loomis, L.H. and Sternberg, S. (1968) Advanced Calculus. Addison-Wesley Publishing, Reading, Mass.

Meng, X.-Li. Posterior Predictive p-Values (1994) The Annals of Statistics, Vol. 22, No. 3, p. 1142-1160.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, Series 5, Volume 50, No. 302, p. 157-176.

Rudin, W. (1974) Real and Complex Analysis. McGraw-Hill, New York.

Schervish, M.J. (1996) P Values: What they are and what they are not. The American Statistician, Vol. 50, No. 3, p. 203-206.

Tjur, T. (1974) Conditional Probability Distributions. Lecture Notes 2, Institute of Mathematical Statistics, University of Copenhagen.

Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.