Statistica Sinica **18**(2008), 1185-1200

BAYESIAN HYPOTHESIS TESTS USING NONPARAMETRIC STATISTICS

Ying Yuan and Valen E. Johnson

University of Texas and M.D. Anderson Cancer Center

Abstract: Traditionally, the application of Bayesian testing procedures to classical nonparametric settings has been restricted by difficulties associated with prior specification, prohibitively expensive computation, and the absence of sampling densities for data. To overcome these difficulties, we model the sampling distributions of nonparametric test statistics—rather than the sampling distributions of original data—to obtain the Bayes factors required for Bayesian hypothesis tests. We apply this methodology to construct Bayes factors from a wide class of nonparametric test statistics having limiting normal distributions and illustrate these methods with data. Finally, we consider the extension of our methodology to nonparametric test statistics having limiting χ^2 distributions.

Key words and phrases: Bayes factor, Kruskal-Wallis test, Logrank test, Mann-Whitney-Wilcoxon test, nonparametric hypothesis test, Wilcoxon signed rank test.

1. Introduction

In parametric settings, the use of Bayesian methodology for conducting hypothesis tests has been limited by two factors. First, the calculation of Bayes factors often involves the evaluation of high-dimensional integrals. This can be a prohibitively expensive undertaking for non-statisticians, both from a numerical and conceptual perspective. Second, Bayes factors require the specification of informative prior densities on parameters appearing in the parametric statistical models that comprise each hypothesis. And unlike Bayesian estimation procedures, tests based on Bayes factors retain their sensitivity to prior assumptions even when sample sizes become large. Prior specification is therefore an important task and one which can be difficult in models containing many parameters.

In nonparametric hypothesis testing, a third difficulty arises. Namely, sampling distributions for data are not specified. Without sampling distributions for data, Bayesian hypothesis tests cannot be defined.

The goal of this article is to overcome these obstacles to Bayesian testing by extending methodology proposed in Johnson (2005) to the classical nonparametric setting. We accomplish this by using results from the asymptotic theory of U-statistics and linear rank statistics (e.g., Serfling (1980)) to define alternative distributions for test statistics that take the form of Pitman translation alternatives (e.g., Randles and Wolfe (1979)). In so doing, we obtain sampling distributions for non-parametric statistics under alternative models that contain only two unknown parameters: a scale parameter and an asymptotic test-efficacy parameter.

Because the distribution of the test statistics under the null hypothesis is known, by specifying a sampling distribution for nonparametric test statistics under a class of alternative models we are able to eliminate two of the obstacles to Bayesian hypothesis testing. That is, by modeling the distribution of test statistics directly we obtain the sampling distributions required for the definition of Bayes factors. We also obtain closed-form expressions for the resulting Bayes factors, which means that numerical evaluation of their values is unnecessary.

The third obstacle to the use of Bayes factors—the specification of subjective prior densities—is also considerably simplified. Within our framework, alternative models contain only two scalar parameters. One, the test-efficacy parameter, is determined by the choice of the test statistic. In principle this leaves only a scale parameter for which a prior density must be specified. Methods for handling this parameter are discussed below.

To begin, it is worthwhile to review the definition of Bayes factors for the case of nested models. Suppose then that data $\boldsymbol{x}$ arise from a sampling density $p(X|\boldsymbol{\theta}, \boldsymbol{\phi})$ with unknown parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, and suppose we wish to test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ against the alternative hypothesis $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta_0}$ where $\boldsymbol{\theta_0}$ is assumed known. If $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x})$ represent the marginal densities of $\boldsymbol{x}$ under H_0 and H_1, then the Bayes factor between H_0 and H_1 can be expressed

$$B_{01} = \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} = \frac{\int p(\boldsymbol{x}|\boldsymbol{\theta_0}, \boldsymbol{\phi}) p(\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\phi}}{\int p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}, \boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\theta}\, d\boldsymbol{\phi}},$$

where $p(\boldsymbol{\phi})$ and $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ are the prior densities on unknown parameters under H_0 and H_1, respectively. Furthermore, if $p(H_0)$ and $p(H_1)$ denote the prior probabilities assigned to H_0 and H_1, then the posterior odds of H_0 versus H_1 is obtained from B_{01} according to

$$\frac{p(H_0|\boldsymbol{x})}{p(H_1|\boldsymbol{x})} = \frac{p(H_0)}{p(H_1)} B_{01}.$$

Thus, Bayes factors represent the weight of evidence contained in data in support of each hypothesis. When deciding between two simple hypotheses, the Bayes factor is simply the likelihood ratio; in more complicated settings it can be regarded as an integrated likelihood ratio. Further discussion of Bayes factors can be found in, for example, Kass and Raftery (1995).

Our motivation for defining Bayes factors based on nonparametric test statistics is to avoid the many pitfalls inherent to the use of p values in formal test procedures. For a discussion of these issues in the classical setting, interested readers may consult (among many other articles) Berger and Delampady (1987), Berger and Sellke (1987), and Goodman (1999a,b). Similar issues arise in the use of Bayesian p values for testing model adequacy as described in, for example, Gelman, Meng and Stern (1996) who base their definition of Bayesian p values on posterior-predictive distributions. However, because posterior predictive p values also represent tail-area probabilities, they too are subject to many of the pitfalls inherent to classical p values.

As we demonstrate in the sequel, our methodology allows us to transform nonparametric test statistics to an appropriate—and interpretable—probability scale, rather than to what is essentially an uncalibrated and comparatively uninterpretable p-value scale.

Our approach is based on the observation that, although the sampling density $p(x|\theta)$ of the data X may not be specified in nonparametric settings, the distribution of the test statistic $T(X)$ is often known under both null and alternative hypotheses, at least asymptotically. To make this notion more precise, we assume for the remainder of this article that the sampling density of the test statistic $T(X)$ can be expressed as $p(T(x)|\theta)$, where the parameter θ may be either a scalar or vector-valued parameter. Under the null hypothesis H_0, we assume that $\theta = \theta_0$ for a known value θ_0. Under the alternative hypothesis H_1, we assume that the sampling distribution of T is obtained by averaging over a prior density $p(\theta)$ defined on the domain of θ. When these assumptions hold, the Bayes factor based on $t = T(X)$ can be defined as

$$BF_{01}(t) = \frac{p(t|\theta_0)}{\int p(t|\theta)p(\theta)\,\mathrm{d}\theta}.$$

For suitable choices of $p(\theta)$, we find that Bayes factors based on nonparametric test statistics can often be expressed in simple form. From such expressions, it is possible to obtain an upper bound on the weight of evidence against the null hypothesis. Such bounds are often useful and may serve to illustrate the maximum extent to which data provide evidence against the null hypothesis. Their use also eliminates much of the subjectivity associated with the definition of Bayes factors.

2. Theory

Many nonparametric test statistics have limiting distributions that are either normal or χ^2. While our methods can sometimes be applied in finite sample settings, it is generally more straightforward to specify alternative distributions in the large sample setting. For this reason, we restrict attention to this case and begin with nonparametric statistics that have limiting normal distributions.

The asymptotic normality of a variety of nonparametric test statistics has been established by the theory of U-statistics and linear rank statistics (e.g., Serfling (1980)). These results are widely used in practice to approximate the exact sampling distributions of nonparametric test statistics, which often do not have closed forms and have to be computed numerically.

The class of nonparametric test statistics with limiting normal distributions includes many commonly used nonparametric statistics. Among these are the sign test and Wilcoxon signed rank test for one-sample location problems, the Mann-Whitney-Wilcoxon test for two-sample location problems, the Ansari-Bradley test and Mood test for scale problems, Kendall's tau and Spearman test for testing independence, the Theil test for slope parameters in regression problems, the Mantel test (or logrank test), and the Hollander-Proschan test of exponentiality in survival analysis.

In order to describe how statistics from these tests can be used to define Bayes factors, let T_k, $k = 1, 2 \ldots$, denote a sequence of nonparametric test statistics based on n_k observations, and suppose that $n_k \to \infty$ as $k \to \infty$. Consider the test of the null hypothesis

$$H_0 : \theta = \theta_0$$

versus the local alternative

$$H_1(n_k) : \theta_k = \theta_0 + \Delta/\sqrt{n_k},$$

where Δ is a bounded constant. This form of the alternative hypothesis is often called the Pitman translation alternative (e.g., Randles and Wolfe (1979)).

Our attention focuses on the asymptotic distribution of the standardized value of T_k,

$$T_k^* = \frac{T_k - \mu_k(\theta_0)}{\sigma_k(\theta_0)},$$

where μ_k and σ_k are the mean and standard deviation of T_k, respectively. Under H_0, we assume that T_k^* has a limiting standard normal distribution. Under $H_1(n_k)$, the asymptotic distribution of T_k^* is given in the following lemma. The proof follows Noether (1955) and appears in the Appendix.

Lemma 1. *Assume $H_1(n_k)$ and*

(A1) $[T_k - \mu_k(\theta_k)]/\sigma_k(\theta_k) \xrightarrow{\mathcal{L}} N(0, 1)$;

(A2) $\sigma_k(\theta_k)/\sigma_k(\theta_0) \xrightarrow{p} 1$;

(A3) $\mu_k(\theta)$ *is differentiable at* θ_0*;*

(A4) $\mu_k^{'}(\theta_0)/\sqrt{n_k}\sigma_k(\theta_0) \xrightarrow{p} C$ *where* C *is a constant.*
Then $T_k^* \xrightarrow{\mathcal{L}} N(C\Delta, 1)$.

In typical settings where they are used, each of the nonparametric statistics mentioned above satisfies these assumptions, and similar conditions are often required in evaluating Pitman's asymptotic relative efficiency (Noether (1955)). The value of C is the efficacy of the test based on T_k. With the asymptotic distribution of T_k^* under both H_0 and H_1 known, the following result follows from Bayes theorem.

Theorem 1. *If assumptions A1–4 of Lemma 1 are satisfied, and the scalar parameter Δ is assumed* a priori *to follow a $N(0, \tau^2)$ distribution, then the Bayes factor based on T_k^* is given by*

$$BF_{01}(T^*) = (1 + C^2\tau^2)^{\frac{1}{2}} \exp\left\{-\frac{C^2\tau^2 T_k^{*2}}{2(1 + C^2\tau^2)}\right\}.$$

The prior density assumed for Δ in the above theorem centers the distribution of θ on the null value of $\theta = \theta_0$. Such centering is natural under classes of local alternatives and is also consistent with the general philosophy advocated by Jeffreys (1961).

The Bayes factor described in Theorem 1 depends on the values of the constants C and τ. Values of C for several commonly used nonparametric test statistics are listed in Table 1. (For convenience, we have also listed the related constant for the logrank test that requires a slightly different alternative structure; see, for example, Fleming and Harrington (1991)). Unfortunately, the calculation of C generally involves the density function of the data under the null model, which is unknown in nonparametric hypothesis testing. This difficulty may be circumvented in several ways. For example, Lehmann (1975) addresses this problem by approximating the unknown null distribution by a parametric distribution (e.g., a normal distribution) whose parameters are estimated from data. Such parametric approximations are useful for a variety of purposes, including the calculation of statistical power against specified alternatives.

A more direct solution is available if we set the value of the scale parameter τ to its maximum marginal likelihood estimate (MMLE) under the alternative hypothesis. By so doing, we obtain a limit on the Bayes factor that does not depend on the unknown constant C.

Under the conditions of Theorem 1, the MMLE of τ^2 under $H_1(n)$ is given by

$$\tau^2 = \frac{T_k^{*2} - 1}{C^2},$$

Table 1. Some commonly used nonparametric test statistics and associated constants.

	Test Statistic $(T)^a$	Standardized Test Statistic (T^*)	C^b
Sign test	$\sum_{i=1}^{n} \psi_i$	$\frac{T-n/2}{\sqrt{n/4}}$	$2p(0)$
Wilcoxon signed rank test	$\sum_{i=1}^{n} R_i^1 \psi_i$	$\frac{T-n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$	$\sqrt{12}\int_{-\infty}^{\infty} p^2(x)\mathrm{d}x$
Mann-Whitney-Wilcoxon	$\sum_{i=1}^{n_1} R_i^2$	$\frac{T-n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}}$	$\sqrt{12a_1a_2}\int_{-\infty}^{\infty} p^2(x)\mathrm{d}x$
Ansari-Bradley test	$\sum_{i=1}^{n_1} S_i$	$\frac{T-n_1(N+2)/4}{\sqrt{\frac{n_1n_2(N+2)(N-2)}{48(N-1)}}}$, N even $\frac{T-n_1(N+2)/4}{\sqrt{\frac{n_1n_2(N+2)(N-2)}{48(N-1)}}}$, N odd	$\sqrt{48a_1a_2}\int_{-\infty}^{\infty} xp^2(x)\mathrm{d}x$
Kendall's tau	$\sum_{1\le i<j\le n}^{n} Q((X_i,Y_i),(X_j,Y_j))$	$\frac{T}{\sqrt{n(n-1)(2n+5)/18}}$	1.5
Logrank test	$\frac{\sum_{j=1}^{D}(d_{1j}-n_{1j}d_j/n_j)}{\sqrt{\sum_{j=1}^{D}\frac{d_j(n_j-d_j)n_{1j}n_{2j}}{n_j^2(n_j-1)}}}$	T	$\left(\int \frac{\pi_1\pi_2}{a_1\pi_1+a_2\pi_2}\mathrm{d}\Lambda\right)^{-\frac{1}{2}}$

[a] Sample values are $\mathbf{X} = (X_1, \ldots, X_{n_1})$ and $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$. If $X_i > 0$, $\psi_i = 1$, otherwise $\psi_i = 0$. R_i^1 is the rank of X_i among $\mathbf{X}$; R_i^2 is the rank of X_i in the combined sample $(\mathbf{X}, \mathbf{Y})$. S_i is the score assigned to X_i in such a way that a score of 1 is assigned to the smallest and largest observations in the combined sample $(\mathbf{X}, \mathbf{Y})$, a score of 2 is assigned to the second smallest and second largest in the combined sample, and so on. $Q((X_i, Y_i), (X_j, Y_j)) = 1$ if $(Y_j - Y_i)(X_j - X_i) > 0$, otherwise $Q = -1$. For the logrank test, $t_1 < \cdots < t_D$ denote the distinct event times in the pooled sample. At t_j, there are d_{1j} events out of n_{1j} subjects at risk in sample $\mathbf{X}$; similarly for d_{2j} and n_{2j} for sample $\mathbf{Y}$; $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$.

[b] $p(x)$ denotes density function under the null. $a_1 = n_1/N$ and $a_2 = n_2/N$ where $N = n_1 + n_2$. For the logrank test, $(T_i, C_i), i = 1, 2$, are independent failure and censoring time variables for two samples, and $X_i = min(T_i, C_i)$. $\pi_i = p(X_i \le t)$. Λ is the null distribution of cumulative hazard.

provided T_k^{*2} exceeds its expectation under H_0 (i.e. $T_k^{*2} > 1$). The Bayes factor obtained by setting τ to its MMLE is

$$\widetilde{BF}_{01}(T^*) = |T_k^*| \exp\left(\frac{1 - T_k^{*2}}{2}\right).$$

This value represents an upper bound on the weight of evidence against H_0. Note that $\widetilde{BF}_{01}(T^*)$ does not depend on the constant C. If equal probabilities are assigned to H_0 and H_1 *a priori*, the corresponding lower bound on the posterior probability of H_0 satisfies

$$P(H_0|x) \ge \widetilde{P}(H_0|x) = \left(1 + \frac{1}{|T_k^*|} e^{\frac{T_k^{*2}-1}{2}}\right)^{-1}. \tag{1}$$

Here and throughout the remainder of the paper, $\widetilde{BF}$ and $\widetilde{P}(H_0|x)$ indicate values of the Bayes factor and marginal posterior probability of the null hypothesis obtained by setting scale parameters appearing in the alternative model equal to their MMLE; BF and $P(H_0|x)$ refer to the corresponding values for a fixed value of the scale parameter.

Unfortunately, seeking an upper bound on the weight of evidence in favor of the null hypothesis is not practically useful unless constraints are first imposed on the value of τ. Without constraints, the alternative model can usually be assigned negligible probability simply by letting τ become large. Such values correspond to arbitrarily diffuse priors on the space of alternative models. On the other hand, values of τ that are close to 0 make the alternative model indistinguishable from the null, leading to a Bayes factor that is close to 1. Thus, useful upper bounds on the evidence in favor of the null model can only be obtained by deterministically constraining the value of τ, or by imposing a prior distribution on it.

Note that evidence in favor of the null hypothesis is obtained whenever T_k^{*2} is less than its expectation under H_0 (i.e., $T_k^{*2} < 1$). In this case it follows that the maximum value of the Bayes factor in favor of the alternative model is 1 and is achieved for $\tau = 0$. That is, the most likely alternative model is obtained by letting the alternative distribution collapse onto the null distribution. When H_0 and H_1 are equally likely *a priori*, the lower bound on the probability of the null model is then given by $\widetilde{P}(H_0|x) = 0.5$.

3. Applications

In this section, we illustrate our methodology in several examples using common nonparametric test statistics.

Wilcoxon signed rank test for depression data

Figure 1 displays data from a study of the effectiveness of a new therapy for reducing symptoms of depression. The data reflect changes to Hamilton depression scale Factor IV measurements (the suicidal factor) for nine patients with anxiety or depression before and after tranquilizer therapy (Hollander and Wolfe (1999)).

To test for a treatment effect, we apply the Wilcoxon signed rank test. For these data, the standardized value of this statistic is $W^* = -2.07$. This corresponds to an exact p value of 0.0382. The large sample approximation to the p value is 0.0391, which is similar to the exact value. This suggests that the asymptotic approximation to the sampling distribution of the Wilcoxon statistic can be used to obtain an approximate Bayes factors from these data, even though the sample size is only 9. For both values, the hypothesis of no treatment effect is rejected in a 5% significance test.

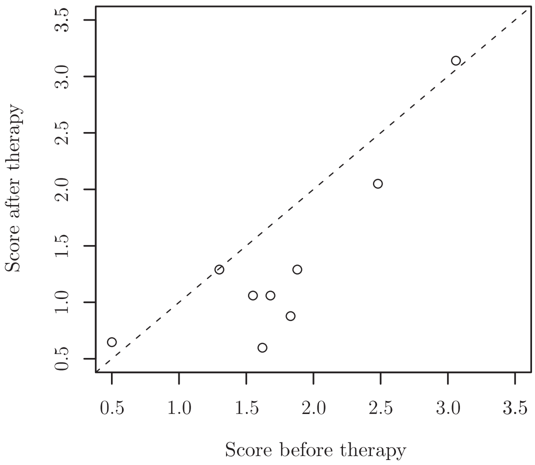

Figure 1. Hamilton depression scale Factor IV measurements (the suicidal factor) for nine patients with anxiety or depression before and after tranquilizer therapy.

In contrast to the p value, the lower bound $\widetilde{BF}_{01}(W^*)$ is 0.399. If H_0 and H_1 are given equal weight *a priori,* this means that the posterior probability that H_0 is true is at least $\widetilde{P}(H_0|W^*) = 0.285$. That is, the posterior odds that the treatment has an effect are then at most 2.5:1, which is a much weaker finding than the p value suggests.

The sensitivity of the posterior probability of H_0 to the choice of τ can be calculated by assuming that the density function for the raw data is approximately normal (e.g., Lehmann (1975)). Using the data provided in Figure 1 to estimate the mean and variance of this normal distribution, and assuming that H_0 and H_1 are equally likely *a priori*, this approximation leads to the posterior probabilities of H_0 displayed in Figure 2.

The posterior probabilities in Figure 2 can be interpreted in several ways. For example, suppose that a change in median depression score equal to the standard deviation of these scores (0.72) is regarded as clinically significant. Then a value of $\tau = 0.72$ might be used to specify the distribution of the test statistic under the alternative hypothesis. This value of τ leads to a Bayes factor of 0.40 and a corresponding posterior probability of H_0 equal to 0.286 (again assuming equal odds *a priori*). More generally, by identifying a range of clinically-important effect sizes, it is possible to obtain explicit Bayes factors instead of a lower bound. Alternatively, one might specify a prior distribution over a range of scale parameter values and execute a one-dimensional numerical integration scheme to obtain the resulting Bayes factor. Of course, both solutions require knowledge of, or an approximation to, the efficacy parameter C.

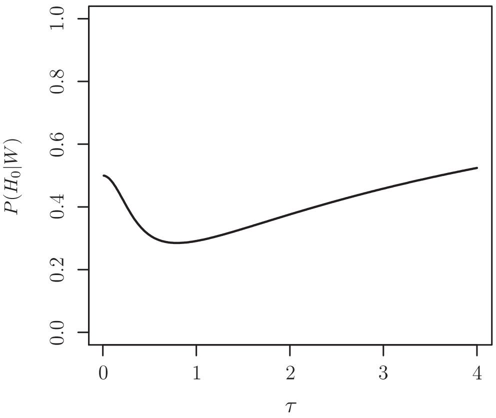

Figure 2. Posterior probability of H_0 under different values of τ for the depression data.

Logrank test for ovarian cancer treatment data

Armstrong et al. (2006) reported results of a randomized, Phase III clinical trial in which a regimen of six cycles of treatment with intravenous paclitaxel followed by intravenous cisplatin (intravenous therapy) was compared to treatment with six cycles of intervenous paclitaxel followed by intraperitoneal cisplatin and intraperitoneal paclitaxel (intraperitoneal therapy) in women with previously untreated stage III ovarian cancer. The primary study endpoints were progression-free survival and overall survival. Two hundred and ten women received intravenous therapy and 205 received intraperitoneal therapy. The logrank test was used to test whether there was a survival difference between therapies; the value of the test statistic resulted in p values of 0.05 and 0.03 for progression-free survival and overall survival, respectively. The authors concluded that intraperitoneal therapy was a more effective treatment for ovarian cancer in this population of patients.

For these data, the upper bounds on the Bayes factors against the null hypothesis of no treatment difference were 0.47 and 0.34 for progression-free survival and overall survival, respectively. If the null and the alternative hypothesis are assumed to be equally likely *a priori*, the posterior odds that intraperitoneal therapy is better than intravenous therapy are at most 2.11 for progression-free survival and 2.94 for overall survival. Thus, the data provide some evidence that intraperitoneal therapy improves survival, but are not as strongly as the p values cited by Armstrong et al. would suggest.

4. Simulation Studies

For one reason or another, Neyman-Pearson tests are typically conducted to bound the Type I error at 5%. In this section, we examine values of Bayes factors obtained from several nonparametric test statistics when those statistics fall on the boundary of their 5% critical region.

Considering first the one-sample Wilcoxon test, we simulated random samples of size $n = 50$ with equal probability from $H_0 : X \sim t_4(0, 1)$ or $H_1 : X \sim t_4(\mu, 1)$, where $t_4(\mu, 1)$ denotes a t-distribution with mean μ, scale parameter 1 and 4 degrees of freedom. Under H_1, we generated the location parameter μ from a $N(0, \tau^2/n)$ distribution for various values of τ. Samples of 50 observations were generated in this way until we obtained 1,000 samples that resulted in p values between 0.045 and 0.055. Based on these 1,000 samples (i.e., conditioning on samples with p values approximately equal to 0.05), we calculated the posterior probability of H_0, $P(H_0|X)$, based on the Wilcoxon test statistic and compared this probability with the true proportion of samples generated from H_0.

Figure 3a depicts $P(H_0|X)$ obtained at the correct (known) value of τ, the lower bound $\widetilde{P}(H_0|X)$ based on (1), and the proportion of samples actually generated from H_0 versus τ. Interestingly, the lower bound $\widetilde{P}(H_0|X)$ is approximately 0.3 for a wide range of alternatives. Thus, a p value of 0.05 corresponds to *at least* a 30% probability that the null hypothesis is true for this class of alternative hypotheses when null and alternative hypotheses are assigned equal weight *a priori*. This result is comparable to those reported by Berger and Delampady (1987), Berger and Sellke (1987) and Johnson (2005). As expected, the posterior probability of H_0 based on the Bayes factor with the correctly specified scale parameter τ is very close to the true proportion of null models that were used to generate the data.

To assess the robustness of our Bayes factor to the misspecification of the prior distribution of μ, we also simulated μ from a $t_4(0, \tau/\sqrt{n})$ distribution rather than a normal distribution. As Figure 3b illustrates, this does not cause significant degradation of the estimate of $P(H_0|X)$.

A similar study was performed to evaluate the performance of our method using the Mann-Whitney-Wilcoxon test statistic. In this simulation, we repeatedly generated two samples of size 50 by drawing independent $t_4(0, 1)$ and $t_4(\mu, 1)$ random variables. Under H_0, the value of μ was fixed at 0 so that all 100 observations were independent and identically distributed. Two models for μ were considered under the alternative model: either the location parameter μ was generated from a $N(0, \tau^2/n)$ distribution, or μ was drawn from a $t_4(0, \tau/\sqrt{n})$ distribution. Under both scenarios 1,000 datasets that had p values between 0.045 and 0.055 were obtained through simulation. We then applied our method to these datasets to compare Bayes factors and marginal posterior model probabilities under various model assumptions. Results from this simulation are displayed in Figure 3c and 3d. As in the case of the one-sample Wilcoxon test, when τ is assumed to be known there is excellent agreement between the exact marginal probabilities

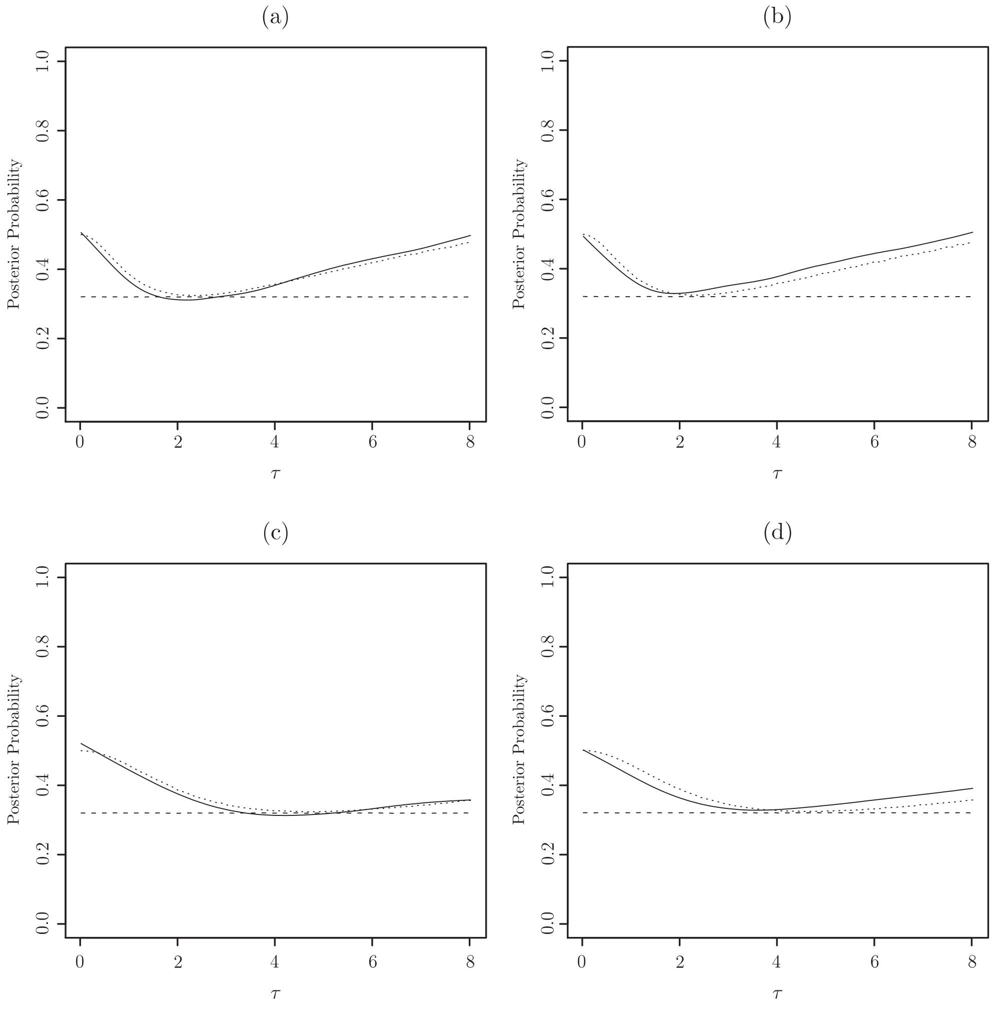

Figure 3. This figure depicts $P(H_0|X)$ as a function of τ, the lower bound $\widetilde{P}(H_0|X)$ (dashed line), and the actual probability that a sample was generated from H_0 (solid line) for data sets yielding p values of approximately 0.05 for the one-sample Wilcoxon test as the prior distribution of μ under H_1 is (a) normally distributed or (b) t distributed, and for the Mann-Whitney-Wilcoxon test as the prior distribution of μ under H_1 is (c) normally distributed or (d) t distributed.

computed from the full data model and the Bayes factors based on the Mann-Whitney-Wilcoxon test statistic. Also, the lower bound on the probability of the null hypothesis provides a reasonable approximation to these marginal probabilities for moderate values of τ.

Finally, we examined the accuracy of our method in small and moderate

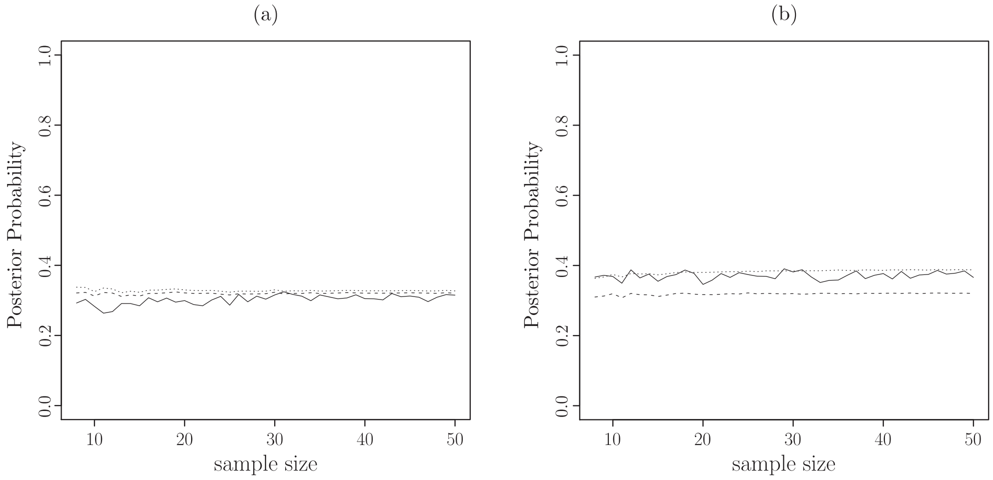

Figure 4. This figure depicts $P(H_0|X)$ evaluated at true value of $\tau = 2$, the lower bound $\widetilde{P}(H_0|X)$ (dashed line), and the actual probability that a sample was generated from H_0 (solid line) for data sets yielding p values of approximately 0.05 for (a) the one-sample Wilcoxon test and (b) the Mann-Whitney-Wilcoxon test as sample sizes are varied.

sample settings. As before, we focused attention on values of the one-sample Wilcoxon statistic and Mann-Whitney-Wilcoxon statistic that led to p values near 0.05. Figure 4 displays the following: (i) the lower bound $\widetilde{P}(H_0|X)$, (ii) $P(H_0|X)$ evaluated at the true value of τ, and (iii) the actual proportion of test statistics drawn from H_0. When sample sizes were larger than approximately 8 or 9, Figure 4 shows that the asymptotic approximation to the sampling distributions of both the one-sample Wilcoxon test and the Mann-Whitney-Wilcoxon test yield marginal model probabilities that are close to their nominal values. In general, however, Bayes factors generated from our approach cannot be expected to provide accurate results whenever the asymptotic approximation to the distribution of the nonparametric test statistic is not accurate. Discussion of the accuracy of such asymptotic approximations can be found in, for example, Gibbons and Chakraborti (2003).

5. Extension to χ^2 Distributed Test Statistics

When testing for differences between the distribution of values obtained from three or more populations, most nonparametric test statistics do not have a limiting normal distribution. Instead, their limiting distribution is often χ^2. Such is the case for the Kruskal-Wallis test in one-way ANOVA problems and Friedman's test in two-way ANOVA settings. We illustrate the extension of our methodology to such settings in the context of the Kruskal-Wallis test (Kruskal and Wallis

(1952)).

The Kruskal-Wallis test is a direct generalization of the two-sided Mann-Whitney-Wilcoxon test to the $k \geq 3$ sample location problem. To fix notation, let $X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k}$ denote k independent samples from continuous distributions $F(x-\theta_1), \ldots, F(x-\theta_k)$, respectively, where $\theta_1, \ldots, \theta_k$ denote medians of the k populations. Also, let the total sample size be $n = \sum_{i=1}^{k} n_i$ and suppose interest focuses on testing $H_0 : \theta_1 = \cdots = \theta_k$ versus $H_1 : \theta_i \neq \theta_j$ for some $i \neq j$.

If R_{ij} denotes the rank of X_{ij} among $X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k}$, then the Kruskal-Wallis statistic W is defined as

$$W = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2 ,$$

where $\bar{R}_i$ is the average of the ranks associated with the ith sample, i.e., $\bar{R}_i = (1/n_i) \sum_{j=1}^{n_i} R_{ij}$.

Under H_0, W has an asymptotic χ^2 distribution with $k-1$ degrees of freedom as all $n_i \to \infty$ simultaneously (Kruskal and Wallis (1952)). Because the test statistic W is consistent against fixed alternatives, we again consider testing H_0 against the sequence of local alternatives

$$H_1(n) : \theta_i = \theta_0 + \Delta_i/\sqrt{n}, \qquad i = 1, \ldots, k,$$

where the $\{\Delta_i\}$ are not all equal.

Assuming $n_i/n \to a_i > 0$ where a_i is a constant for $i = 1, \ldots, k$, Andrews (1954) showed that under $H_1(n)$ the limiting distribution of W is a $\chi^2_{k-1}(\rho)$ distribution with non-centrality parameter $\rho = 12 \left\{ \int_{-\infty}^{\infty} p^2(x)\, \mathrm{d}x \right\}^2 \sum_{i=1}^{k} a_i (\Delta_i - \bar{\Delta})^2$, where $p(\cdot)$ denotes the density function of the data under the null hypothesis and $\bar{\Delta} = \sum_{i=1}^{k} a_i \Delta_i$. The non-centrality parameter ρ can be written as a quadratic form according to

$$\rho = 12 \left\{ \int_{-\infty}^{\infty} p^2(x)\, \mathrm{d}x \right\}^2 \boldsymbol{\Delta}' \boldsymbol{P}' \boldsymbol{Q} \boldsymbol{P} \boldsymbol{\Delta},$$

where

$$\boldsymbol{\Delta} = \begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_k \end{pmatrix}, \boldsymbol{P} = \boldsymbol{I} - \begin{pmatrix} a_1 \cdots a_k \\ \vdots \quad\quad \vdots \\ a_1 \cdots a_k \end{pmatrix}, \boldsymbol{Q} = \begin{pmatrix} a_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_k \end{pmatrix}.$$

Because $\boldsymbol{P}' \boldsymbol{Q} \boldsymbol{P}$ is a non-negative definite matrix with rank $k-1$, there exists a nonsingular $k \times k$ matrix $\boldsymbol{R}$ such that

$$\boldsymbol{P}' \boldsymbol{Q} \boldsymbol{P} = \boldsymbol{R}' \begin{pmatrix} \boldsymbol{I}_{k-1} & 0 \\ 0 & 0 \end{pmatrix} \boldsymbol{R}.$$

To obtain a Bayes factor based on W, it is thus necessary to assume a prior distribution on $\boldsymbol{\Delta}$. A convenient prior for this purpose can be obtained by assuming that $\boldsymbol{\Delta}$ follows a multivariate normal distribution of the form

$$\boldsymbol{\Delta} \sim N_k(\mathbf{0}, c(\boldsymbol{R}^{'}\boldsymbol{R})^{-1}),$$

where c is a scaling constant.

Letting $\tau = 12c\{\int_{-\infty}^{\infty} p^2(x)\,\mathrm{d}x\}^2$, it follows that $\tau^{-1}\rho$ follows a χ^2 distribution with $k-1$ degrees of freedom. The conditional distribution of W given ρ, say $p(W|\rho)$, is thus a $\chi^2_{k-1}(\rho)$ distribution, and the prior distribution on ρ, say $p(\rho)$, is a scaled χ^2 distribution $\tau\chi^2_{k-1}$. It then follows that the marginal distribution of W under $H_1(n)$ can be expressed as

$$p_1(W) = \int_0^{\infty} p(W|\rho)p(\rho)\,\mathrm{d}\rho = Ga\left(W|\frac{k-1}{2}, \frac{1}{2(\tau+1)}\right), \tag{2}$$

where $Ga(\cdot)$ denotes a gamma distribution. Thus, the Bayes factor based on W is

$$BF_{01}(W) = (\tau+1)^{\frac{k-1}{2}} \exp\left\{-\frac{\tau W}{2(\tau+1)}\right\}.$$

The value of τ that maximizes $p_1(W)$ in (2) is $\{W-(k-1)\}/(k-1)$, provided that W exceeds its expectation under H_0. Setting τ at this value, we find that the upper bound of the Bayes factor against H_0 is

$$\widetilde{BF}_{01}(W) = \left(\frac{W}{k-1}\right)^{\frac{k-1}{2}} \exp\left\{-\frac{W-(k-1)}{2}\right\}.$$

Generalizations to other nonparametric test statistics that have limiting χ^2 distributions under the null hypothesis can be derived in a similar way.

To illustrate the application of our method to the Kruskal-Wallis test, we consider the study of Curtin et al. (2005) that compared genome-wide alterations in four types of melanoma based on exposure to ultraviolet light. One hundred and twenty-six patients were included in the study and there was special interest in examining whether the extent of chromosomal aberration varied by melanoma type. To this end, a Kruskal-Wallis test was applied and yielded a p value of 0.004. The hypothesis of no difference in the degree of chromosomal aberrations among the four types of melanoma is therefore rejected in a 5% significance test; in classical jargon this result is "highly significant."

Applying our method, we find that the Bayes factor corresponding to the upper bound on the weight of evidence against H_0 is 0.054, which leads to a value of $\widetilde{P}(H_0|W) = 0.057$ when the null and alternative hypothesis are given equal weight *a priori.* This value suggests strong evidence that these four groups of melanoma have different degrees of chromosomal aberrations. This conclusion is consistent with the p value, but the Bayes factor provides a more natural and well-calibrated measure of evidence in terms of a probability.

6. Conclusion

Traditionally, the application of Bayesian testing procedures to classical nonparametric settings has been hindered by the absence of sampling densities for data. In this article, we have demonstrated how this difficulty can be circumvented by modeling the distribution of test statistics directly. Use of this methodology allows scientists to summarize the results of tests in terms of model probabilities and Bayes factors rather than p values, and thereby represents an important advance in the field of nonparametric statistical hypothesis testing. By reducing the subjectivity typically associated with the use of Bayes factors, we also hope to alleviate objections from those opposed to subjective test procedures.

Methodology proposed in this article relies on asymptotic approximations to the distribution of common nonparametric test statistics. However, numerical evidence presented in Section 4 suggests that Bayes factors based on such approximations are not particularly sensitive to the large sample approximations to the distributions of at least two common test statistics. For both one-sample and two-sample Wilcoxon tests, our method appears to give accurate results when sample sizes are larger than 8.

Acknowledgements

The authors would like to thank two referees and an associate editor for their helpful and detailed comments.

Appendix

Proof of Lemma 1.

$$\begin{aligned} T_k^* &= \frac{T_k - \mu(\theta_0)}{\sigma_k(\theta_0)} = \frac{T_k - \mu(\theta_k) + \mu(\theta_k) - \mu(\theta_0)}{\sigma_k(\theta_0)} \\ &\simeq \frac{T_k - \mu(\theta_k) + \mu^{'}(\theta_0)(\theta_k - \theta_0)}{\sigma_k(\theta_0)} = \frac{T_k - \mu(\theta_k)}{\sigma_k(\theta_0)} + \frac{\mu^{'}(\theta_0)}{\sqrt{n_k}\sigma_k(\theta_0)}\Delta. \end{aligned}$$

Now,

$$\frac{T_k - \mu(\theta_k)}{\sigma_k(\theta_0)} = \frac{T_k - \mu(\theta_k)}{\sigma_k(\theta_k)} \frac{\sigma_k(\theta_k)}{\sigma_k(\theta_0)} \xrightarrow{\mathcal{L}} N(0,1),$$

so application of Slutsky's theorem gives the desired result.

References

Andrews, F. C. (1954). Asymptotic behavior of some rank tests for analysis of variance. *Ann. Statist.* **25**, 724-736.

Armstrong, D. K., Bundy, B. and et al. (2006). Intraperitoneal cisplatin and paclitaxel in ovarian cancer. *New England J. Medicine* **354**, 34-43.

Berger, J. O. and Delampady, M. (1987). Testing precise Hypotheses (with discussions). *Statist. Sci.* **2**, 317-352.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p value and evidence. *J. Amer. Statist. Assoc.* **82**, 112-122.

Curtin, J. A., Fridlyand, J. and et al. (2005). Distinct sets of genetic alterations in melanoma. *New England J.Medicine* **353**, 2135-2147.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis.* Wiley, New York.

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

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference.* Marcel Dekker, Inc.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: the p value fallacy. *Ann. Internal Medicine* **130**, 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: the bayes factor. *Ann. Internal Medicine* **130**, 1005-1013.

Hollander M. and Wolfe D. A. (1999). *Nonparametrics Statistical Methods.* Wiley, New York.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, London.

Johnson, E. V. (2005). Bayes factor based on test statistics. *J. Roy. Statist. Soc. Ser. B* **67**, 689-701.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* **47**, 583-621.

Lehmann E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

Noether, G. E. (1955). On a theorem of Pitman. *Ann. Math. Statisit.* **26**, 64-68.

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, U.S.A.

E-mail: yyuan@mdanderson.org

Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX, U.S.A.

E-mail: vejohnson@mdanderson.org

(Received October 2006; accepted April 2007)