Bayesian Hypothesis Tests Using Nonparametric Statistics

Ying Yuan* Valen E. Johnson†

* M. D. Anderson Cancer Center, yyuan@mdanderson.org
† M.D. Anderson Cancer Center

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/mdandersonbiostat/paper21

Copyright ©2006 by the authors.
Bayesian Hypothesis Tests Using Nonparametric Statistics

Ying Yuan and Valen E. Johnson

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 or chi-square distributions. We also demonstrate how this testing strategy can be extended to simplify meta-analyses in which only p values or the values of test statistics have been reported.
Bayesian Hypothesis Tests Using Nonparametric Statistics

Ying Yuan and Valen E. Johnson

Department of Biostatistics and Applied Mathematics, M.D. Anderson Cancer Center,

Houston, TX, U.S.A.

yyuan@mdanderson.org  vejohnson@mdanderson.org

Summary

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 or \( \chi^2 \) distributions. We also demonstrate how this testing strategy can be extended to simplify meta-analyses in which only \( p \) values or the values of test statistics have been reported.

*Key words:* Bayes factor, nonparametric hypothesis test, sign test, Wilcoxon signed rank test, Mann-Whitney-Wilcoxon test, Kruskal-Wallis test, log-rank test
1 Introduction

In parametric settings, the use of Bayesian methodology for conducting hypothesis tests has been limited by two factors. First, Bayesian test procedures require the specification of informative prior densities on parameters appearing in the parametric statistical models that comprise each hypothesis. Second, even when it is possible to subjectively define these prior densities (or to define “objective” substitutes for them), 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. Together, these two factors have severely limited the application of Bayesian methodology to testing problems.

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

The goal of this article is to overcome all three of these impediments to the application of Bayesian testing procedures by using nonparametric test statistics to define Bayes factors, which represent the factors that multiply the prior odds assigned to the null and alternative hypotheses to obtain the posterior odds. Our approach eliminates the second and third obstacles to calculating these factors, and substantially diminishes the first. Our approach is based on an extension of methodology described in Johnson (2005).

Given the broad application of nonparametric test statistics and the ready availability of software to perform classical significance tests based on these statistics, a few preliminary
comments regarding our motivation for “mixing paradigms” are perhaps warranted.

From a Bayesian perspective, the use of the $p$ value as a measure of evidence in hypothesis testing can be criticized from several perspectives (Berger and Delampady, 1987; Berger and Sellke, 1987; Goodman, 1999a, 1999b, 2001). One important criticism of $p$ values is that they tend to systematically reject the null hypothesis in large samples—the so-called Lindley paradox (1957). This problem is particularly pronounced in, for example, sample surveys or large clinical trials in which thousands of observations are collected. In such settings, the null hypothesis is seldom exactly correct and so is rejected. Bayesian testing methods do not suffer from the same deficiency: If alternative models that differ from the null model in substantively meaningful ways are specified, then the null hypothesis is not necessarily rejected as sample sizes become large.

As proposed by Fisher (1956), the $p$ value represents an informal measure of discrepancy between the data and the null hypothesis. However, it is important to remember that the $p$ value does not provide a formal mechanism for assessing the evidence contained in data from a single experiment. Although the $p$ value is often confused with the Neyman-Pearson Type 1 error rate, which has a valid frequentist interpretation, the data-dependent $p$ value does not. Berger and Delampady (1987) and Goodman (1999a) provide more detailed discussion of this issue.

Even if $p$ values did represent approximate long-run error rates, they would still fail to answer the question of interest in most testing problems: In light of observed data from the current experiment and existing scientific knowledge, is there enough evidence to reject the null hypothesis? Unfortunately, the $p$ value does not have a direct relation to the probability
of observing data under either the null or alternative hypotheses. Instead, it represents only the null probability of observing data as extreme or more extreme than was actually observed. As Jeffreys (1961) pointed out, this means that “a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.”

$P$ values also lack a quantitative interpretation in terms of the amount of evidence against the null hypothesis. For example, $p$ values of both 0.01 and 0.001 result in the rejection of the null hypothesis in a 5% Neyman-Pearson test, but there is no mechanism to compare these values quantitatively. That is, a $p$ value of 0.001 does not present 10 times more evidence against the null hypothesis than does a $p$ value of 0.01. In contrast, Bayes factors and posterior model probabilities naturally lend themselves to such quantitative comparisons.

Another objection to $p$ values is their tendency to overstate evidence against the null hypothesis, especially when testing point null hypotheses (e.g., Dickey, 1977; Berger and Sellke, 1987). Berger and Sellke (1987) show that the actual evidence against a null hypothesis (as measured by its posterior probability) can differ by an order of magnitude from the $p$ value. When testing a normal mean, for instance, if both hypotheses are deemed equally likely before seeing the result of an experiment, data that yield a $p$ value of 0.05 result in a posterior probability of the null hypothesis of at least 0.3 for a broad class of objective alternative hypotheses. In certain cases, like one-sided hypothesis testing, the $p$ value and Bayesian posterior probability of null hypothesis may be reconcilable (Casella and Berger, 1987), but even then their interpretations are quite different.

Finally, it is difficult to combine the evidence from different experiments using $p$ values. Standard meta-analyses weight experiments by their resulting precisions and calculate the
p value based on the combined dataset. Unfortunately, this p value has little relation to the p values obtained from the individual experiments. In contrast, Bayes factors provide a natural and intuitive way to combine evidence from different experiments.

Our motivation for defining Bayes factors based on non-parametric test statistics is to avoid many of the pitfalls inherent to p values. As we demonstrate in the sequel, our methodology allows us to transform test statistics to an appropriate—and interpretable—probability scale, rather than to what is essentially an uncalibrated and comparatively uninterpretable p-value scale.

The role of Bayes factors in parametric hypothesis testing has been studied in great depth. Kass and Raftery (1995) provide a recent review. In contrast, comparatively little research has been reported for computing Bayes factor in the nonparametric hypotheses testing, at least in the classical sense of nonparametric statistics (see Rousseau (2006) for a discussion of methodology for defining and calculating Bayes factors for Bayesian nonparametric models).

Our approach is based on the observation that, although the sampling density \( f(x|\theta) \) of the data \( X \) is not specified in nonparametric tests, the distribution of the test statistic, say \( T(X) \), 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 \( g(T(x)||\theta) \), where \( g \) denotes a probability density function defined with respect to an appropriate underlying measure, and \( \theta \) may be either a scalar or vector-value parameter. Under the null hypothesis \( H_0 \), we assume that \( \theta = \theta_0 \) for a known value \( \theta_0 \). Under the alternative hypothesis \( H_1 \), we assume that the sampling distribution of \( T \) is obtained by averaging over a prior density \( \pi(\theta) \) defined.
on the domain of $\theta$. When these assumptions hold, the Bayes factor based on $t = T(X)$ can be defined as

$$BF_{01}(t) = \frac{g(t|\theta_0)}{\int g(t|\theta)\pi(\theta)\,d\theta}. \quad (1)$$

For suitable choices of $\pi(\theta)$, we find that Bayes factors based on nonparametric test statistics can often be expressed in simple form. As a result, it is possible to obtain lower bounds on the posterior probability of the null hypothesis by maximizing over the marginal likelihood of the test statistic under $H_1$. Taking this approach eliminates much of the subjectivity normally associated with the definition of Bayes factors.

## 2 Methods

Many nonparametric test statistics have limiting distributions that are either normal or $\chi^2$. While our methods can (in some cases) be applied in finite sample settings when suitable distributions for test statistics under the alternative model can be defined, 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.

### 2.1 Bayes factor for nonparametric test statistics having limiting normal distribution

The asymptotic normality of a variety of nonparametric test statistics has been established by the theory of $U$-statistics (Hoeffding, 1948) and linear rank statistics (Hajek, 1961; Chernoff
and Savage, 1985). These results are widely used in practice to approximate exact sampling distributions of nonparametric test statistics which usually do not have closed forms and have to be computed numerically.

The class of nonparametric test statistics with limiting normal distributions includes a large number of 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, the Randall’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-Proshan 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 (or contiguous) alternative

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

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 $\mu_k$ and $\sigma_k$ 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 appears in the Appendix.

**Lemma 1.** Assume $H_1(n_k)$ and conditions A1–4 below apply.

(A1) $\frac{T_k - \mu_k(\theta_k)}{\sigma_k(\theta_k)} \xrightarrow{L} N(0, 1)$,

(A2) $\frac{\sigma_k(\theta_k)}{\sigma_k(\theta_0)} \xrightarrow{P} 1$,

(A3) $\mu_k(\theta)$ is differentiable at $\theta_0$,

(A4) $\frac{\mu'(\theta_0)}{\sqrt{n_k\sigma_k(\theta_0)}} \xrightarrow{P} C$ where $C$ is a constant.

Then $T_k^* \xrightarrow{L} N(C\Delta, 1)$.

We note that each of the nonparametric statistics mentioned above satisfy these assumptions, and that 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 $\Delta$ is assumed a priori to follow a $N(0, \tau^2)$ distribution, then the Bayes factor based on $T_k^*$ is given by

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

The prior density assumed for $\Delta$ in the above theorem centers the distribution of $\theta$ on the null value of $\theta = \theta_0$. Such centering is natural under classes of local alternatives and also is consistent with the general philosophy advocated by Jeffreys (1961).
There is an interesting parallel between the result of Theorem 1 and results cited in Johnson (2005) for $\chi^2$ statistics. Under the conditions of the theorem, the random variable $T_k^{*2}$ has a limiting $\chi^2$ distribution with 1 degree freedom under $H_0$, and a limiting noncentral $\chi^2$ distribution with 1 degree freedom and non-centrality parameter $C^2\Delta^2$ under $H_1(n_k)$. If $\Delta \sim N(0, \tau^2)$ a priori, then the marginal distribution of $T_k^{*2}$ under $H_1(n_k)$ is a gamma distribution with shape parameter 1/2 and scale parameter $1/[2(C^2\tau+1)]$. Letting $Ga(\cdot|\alpha, \beta)$ denote a gamma density with shape parameter $\alpha$ and scale parameter $\beta$, the Bayes factor based on $T_k^{*2}$ is given by

$$BF_{01}(T_k^{*2}) = \frac{Ga(T_k^{*2}|1, 1)}{Ga(T_k^{*2}|1, 2(C^2\tau+1))} \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad$$

$$= (1 + C^2\tau^2)^{1/2} \exp \left\{ -\frac{C^2\tau^2T_k^{*2}}{2(1 + C^2\tau^2)} \right\} \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad$$

$$= BF_{01}(T_k^*)$$

This equation provides a connection to results presented for $\chi^2$ statistics in Johnson (2005) and is explored further in Section 2.2.

Several approaches might be taken for setting the value of the parameter $\tau$ that appears in (2). If possible, prior information or scientific guidance should be used to subjectively specify an appropriate prior distribution on $\tau$. In the absence of such information, $\tau$ can be (objectively) determined by equating its value to its maximum marginal likelihood estimate (MMLE) under the alternative hypothesis. This leads to an upper bound on the Bayes factors against point null hypotheses. This strategy has previously been advocated by Edwards, Lindman and Savage (1963), Dickey (1977), Good (1950, 1958, 1967, 1986), Berger (1985), Berger and Sellke (1987), and Berger and Delampday (1987), among others.
Under the conditions of Theorem 1, the MMLE of \( \tau^2 \) under \( H_1(n) \) is given by

\[
\tau^2 = \frac{T_k^* - 1}{C^2}
\]

provided that the statistic \( T_k^* \) exceeds its expectation under \( H_0 \) (i.e. \( T_k^* > 1 \)). The Bayes factor obtained by setting \( \tau \) to this value is

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

This value represents an upper bound on the weight of evidence against \( H_0 \). Note that \( \tilde{BF}_{01}(T^*) \) does not depend on the constant \( C \), which is fortunate since \( C \) often depends on the unknown null distribution of the data. When \( T_k^* < 1 \), the alternative model collapses onto the null model (\( \tau = 0 \)), making the Bayes factor equal to 1. For \( T_k^* > 1 \), if equal probabilities are assigned to \( H_0 \) and \( H_1 \) a priori, the corresponding lower bound on the posterior probability of \( H_0 \) is

\[
\tilde{P}(H_0|x) = \left( 1 + \frac{1}{|T_k^*|} e^{\frac{T_k^*}{2} - 1} \right)^{-1}.
\]  

(3)

From a philosophical perspective, it is interesting to compare probability (3) with the \( p \) value based on the large sample normal approximation. Given the value of test statistics \( T_k^*(x) \), the \( p \) value under the normal approximation is \( 2(1 - \Phi(T_k^*)) \), where \( \Phi(\cdot) \) denotes the standard normal distribution function. The following theorem shows that this \( p \) value is substantially smaller than the posterior probability of the null hypothesis for this class of tests. The proof of this theorem follows along lines similar to those exposed in Berger and Sellke (1987).
Theorem 2. For $T^*_k > 1.96$, 
\[
\frac{\tilde{P}(H_0|x)}{T_k^2 p} > 1.67
\]

Furthermore, 
\[
\lim_{T_k^* \to \infty} \frac{\tilde{P}(H_0|x)}{T_k^2 p} = \sqrt{e\pi/2} \approx 2.066,
\]

where $p$ denotes the $p$-value based on the large sample normal approximation.

We now turn to an examination of the consistency of the Bayes factor provided in (2). A Bayes factor between null and alternative hypotheses is consistent if the value of the Bayes factor goes to infinity when $H_0$ is true, and goes to 0 when $H_1$ is true. Obviously, the Bayes factor in (2), like most Bayes factors, is not consistent against local alternatives. In practice, however, it is important for a test procedure to be consistent against alternative hypotheses for which the value of $\theta$ is bounded in probability away from its value under the null. To examine the consistency of (2) in this setting, we replace the prior assumption in Theorem 1 with an assumption that

\[
\Delta \sim N(0, n_k \tau^2).
\]

This assumption implies that the variance of $(\theta_k - \theta_0)$ is $O(1)$, and results in a Bayes factor equal to

\[
BF_{01}(T^*_k) = (1 + C^2 n_k \tau^2)^{1/2} \exp \left\{ - \frac{C^2 n_k \tau^2 T_k^*}{2(1 + C^2 n_k \tau^2)} \right\}.
\]

For fixed values of $\tau$, the following theorem applies.

Theorem 3. Under the assumptions stated above, the Bayes factor given in (4) is consistent.
2.1.1 Application to Specific Nonparametric Tests

Wilcoxon signed rank test and sign test

Consider paired replicate data for \( n \) subjects, and suppose that \((X_i, Y_i), i = 1, \ldots, n, \) denote pre- and post-treatment measurements for \( n \) items or subjects. Define \( Z_i = Y_i - X_i \) to be the associated treatment effects, and suppose the treatment effects are independent and distributed according to an unknown distribution function \( F(Z|\theta) \). Assume further that \( F(Z|\theta) \) is symmetric around \( \theta \) and that we wish to test the null hypothesis of no treatment effect against a two-sided alternative using the Wilcoxon signed rank test. Under an appropriate parameterization, we assume that the null hypothesis may be expressed \( H_0 : \theta = 0 \), while the alternative hypothesis may be represented \( H_1 : \theta \neq 0 \).

To define the Wilcoxon signed rank test statistic, order the absolute values \(|Z_1|, \ldots, |Z_n|\) from smallest to largest. Let \( R_i \) denote the rank of \(|Z_i|\) and define the indicator variable \( \delta_i, i = 1, \ldots, n, \) as

\[
\delta_i = \begin{cases} 
1 & \text{if } Z_i > 0 \\
0 & \text{if } Z_i < 0
\end{cases}.
\]

The Wilcoxon signed rank statistics \( W \) is then given by

\[
W = \sum_{i=1}^{n} \delta_i R_i.
\]

The limiting distribution of \( W \), suitably standardized, is a normal distribution under \( H_0 \):

\[
W^* = \frac{W - E(W)}{\sqrt{\text{Var}(W)}} = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathcal{L}} N(0, 1).
\]

Furthermore, it can be shown that conditions \( A1–4 \) of Lemma 1 are satisfied (e.g., Lehmann 1975) with \( C = \sqrt{12} \int_{-\infty}^{\infty} f^2(x) \, dx \) where \( f(\cdot) \) is the density function of \( F \). If the prior
density on $\Delta$ is assumed to be $N(0, \tau^2)$, then by Theorem 1 the Bayes factor between null and alternative hypotheses is given by

$$BF_{01}(W^*) = \left\{ 1 + 12 \left( \int_{-\infty}^{\infty} f^2(x) \, dx \right)^2 \tau^2 \right\}^{1/2} \exp \left\{ -\frac{6 \left( \int_{-\infty}^{\infty} f^2(x) \, dx \right)^2 \tau^2 W^*^2}{1 + 12 \left( \int_{-\infty}^{\infty} f^2(x) \, dx \right)^2 \tau^2} \right\}.$$  

If we set the value of $\tau^2$ to its MMLE, the resulting Bayes factor is

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

This value represents an upper bound on the weight of evidence against $H_0$.

The sign test provides an alternative to the signed rank test. Though typically not as efficient as the signed rank test, it does not require a symmetry assumption on $F(Z|\theta)$. Furthermore, the distribution of the sign test can be derived exactly in finite samples under both null and alternative hypotheses.

Using the same notation as above, the sign statistic $S$ is defined as the number of positive $Z_i$ values, i.e.,

$$S = \sum_{i=1}^{n} \delta_i.$$  

The Bayes factor based on the sign test can be derived from the exact sampling distribution of $S$. Under $H_0$, $Pr(Z_i > 0) = Pr(Z_i < 0) = 1/2$, and the random variable $S$ follows a binomial distribution with success probability $1/2$ and denominator $n$. When $H_0$ is false, $S$ follows a binomial distribution with success probability $p = 1 - F(0|\theta)$. To obtain an exact Bayes factor in this case, it is convenient to assume that the prior distribution for $p$ under the alternative model is a Beta distribution centered at $1/2$, or

$$p \sim Beta(\alpha, \alpha),$$
where $\alpha$ is a known constant. The Bayes factor that results from these assumptions is

$$BF(S) = \frac{\Gamma(\alpha)^2 \Gamma(n + 2\alpha)}{2^s \Gamma(2\alpha) \Gamma(S + \alpha) \Gamma(n + \alpha - S)}.$$ 

Alternatively, the large sample value of the Bayes factor based on the sign test can be based on the approximation of binomially-distributed random variables by normally-distributed random variables. If we assume that $\alpha = n/\tau^2$, where $\tau^2$ is a constant, then the asymptotic value of the Bayes factor is consistent with Theorem 1 and is given by

$$BF(S) = (1 + \tau^2/2)^{1/2} \exp \left\{ -\frac{(S - n/2)^2 \tau^2}{(1 + \tau^2/2)n} \right\}.$$ 

**Two-sample Wilcoxon (or Mann-Whitney-Wilcoxon) test**

The two-sample Wilcoxon rank sum test is a commonly used nonparametric test for two-sample location problems. To define this test statistic, suppose that the data consist of two independent samples. The first sample, $X_1, \ldots, X_m$, is assumed drawn from a control group having distribution function $F$; the second sample $Y_1, \ldots, Y_n$ is assumed to be drawn from a treatment population distributed according to $G$. To complete the specification of the alternative hypothesis, we assume a location-shift model for $G$ (e.g., Lehmann 1975), or that $G(t) = F(t - \theta)$. The null hypothesis corresponding to no treatment effect is then

$$H_0 : \theta = 0,$$

which asserts that the $X$’s and $Y$’s have the same (but unspecified) probability distribution.

The Wilcoxon two-sample rank sum statistic is given by

$$U = \sum_{j=1}^n S_j,$$
where $S_j$ denotes the rank of $Y_j$ in the combined sample of $X$ and $Y$.

The standardized value of $U$, say $U^*$, has a limiting standard normal distribution under $H_0$. Specifically,

$$U^* = \frac{U - n(m + n + 1)/2}{\{mn(m + n + 1)/12\}^{1/2}} \xrightarrow{d} N(0, 1).$$

Assuming $m/(m + n) \to \lambda$ where $0 < \lambda < 1$, it can be shown that conditions A1–4 of Lemma 1 are satisfied (e.g., Randles and Wolfe 1979), and that $C = \{12\lambda(1-\lambda)\}^{1/2} \int_{-\infty}^{\infty} f^2(x) \, dx$ where $f(\cdot)$ is the density function of $F$. Assuming that the prior distribution of $\theta$ under the alternative hypothesis is

$$\Delta \sim N(0, \tau^2),$$

it follows from Theorem 1 that the Bayes factor between null and alternative hypotheses is given by

$$BF_{01}(U^*) = \left\{ 1 + 12\lambda(1-\lambda) \left( \int_{-\infty}^{\infty} f^2(x) \, dx \right)^2 \tau^2 \right\}^{1/2} \exp \left\{ -\frac{6\lambda(1-\lambda) \left( \int_{-\infty}^{\infty} f^2(x) \, dx \right)^2 \tau^2 U^{*2}}{1 + 12\lambda(1-\lambda) \left( \int_{-\infty}^{\infty} f^2(x) \, dx \right)^2 \tau^2} \right\}.$$

If set $\tau^2$ to its MMLE, then the Bayes factor reduces to

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

**Weighted logrank test**

Let $T_{11}, \ldots, T_{1n_1}$ and $T_{21}, \ldots, T_{2n_2}$ be independent samples from continuous lifetime distribution functions $F_1$ and $F_2$, respectively, and let $C_{11}, \ldots, C_{1n_1}$ and $C_{21}, \ldots, C_{2n_2}$ be censoring times that are independent of the $T$’s and drawn from distributions $G_1$ and $G_2$. For $i = 1, 2$ and $j = 1, \ldots, n_i$, we observe lifetimes $X_{ij} = \min(T_{ij}, C_{ij})$ and indicators $\delta_{ij} = I(X_{ij} = T_{ij})$. Let $N_i(t) = \sum_{j=1}^{n_i} I(X_{ij} \leq t, \delta_{ij} = 1)$ and $Y_i(t) = \sum_{j=1}^{n_i} I(X_{ij} \geq t)$. Then a class of weighted
logrank test statistics with exemplar $L$ (Gill’s class “$K$,” Gill (1980)) may be defined according to

$$L = \int_0^\infty K(s) \frac{dN_1(s)}{Y_1(s)} - \int_0^\infty K(s) \frac{dN_2(s)}{Y_2(s)}$$

where

$$K(s) = \left( \frac{n_1 + n_2}{n_1 n_2} \right)^{\frac{1}{2}} W(s) \frac{Y_1(s) Y_2(s)}{Y_1(s) + Y_2(s)}.$$  

Members of this class include a variety of commonly used test statistics. The choice $W(s) = 1$ corresponds to the logrank test. If $W(s)$ is taken to be the Kaplan-Meier estimate of the common survival function across groups, then the Prentice-Wilcoxon test statistic is obtained. The Gehan-Wilcoxon statistic corresponds to setting $W(s) = (n_1 + n_2)^{-1}(Y_1(s) + Y_2(s))$.

Statistics within this class can be used to test the null hypothesis that $F_1 = F_2 = F$, where $F$ is an unknown lifetime distribution. If $n = n_1 + n_2$, $n_1/n \to a_1 > 0$ and $n_2/n \to a_2 > 0$, where $a_1$ and $a_2$ are constant, and

$$\hat{\sigma}^2(L) = \int_0^\infty \left( \frac{K(s)^2}{Y_1(s)} + \frac{K(s)^2}{Y_2(s)} \right) \left( 1 - \frac{N_1^D(s) + N_2^D(s) - 1}{Y_1(s) + Y_2(s) - 1} \right) \frac{d(N_1(s) + N_2(s))}{Y_1(s) + Y_2(s)},$$

where $N_i^D = N(s) - N(s^-)$ for $i = 1, 2$, then the standardized value of $L$ under $H_0$ converges to a standard normal distribution (Gill, 1980); that is,

$$L^* = \frac{L}{\hat{\sigma}(L)^{1/2}} \stackrel{L}{\longrightarrow} N(0, 1).$$

Fleming and Harrington (1991, chapter 7) show that tests based on $L^*$ are consistent against fixed alternatives. Following the same line of reasoning, we consider tests of $H_0$.
against a sequence of contiguous alternatives \( \{F_1^n, F_2^n\} \) which satisfy, for any \( t \) and \( i = 1, 2, \)

\[
\sup|F_i^n(t) - F(t)| \to 0
\]

\[
\int_0^t K(s) \left( \frac{d\Lambda_i^n(s)}{d\Lambda(s)} - 1 \right) d\Lambda(s) \to r_i(t)
\]

as \( n \to \infty \). Define cumulative hazard functions according to

\[
\Lambda_i^n(s) = \int_0^s \{1 - F_i^n(u)\}^{-1} dF_i^n(u)
\]

and

\[
\Lambda(s) = \int_0^s \{1 - F(u)\}^{-1} dF(u).
\]

Letting \( r = r_1(t) - r_2(t) \), it follows that the asymptotic distribution of \( L^* \) under this sequence of alternatives is given by

\[
L^* \xrightarrow{L} N(r/\sigma, 1),
\]

where \( \sigma \) is the standard deviance of \( L \) under \( H_0 \), given by

\[
\sigma = \left\{ \int_0^t \frac{(a_1 \pi_1(s) + a_2 \pi_2(s))}{\pi_1(s) \pi_2(s)} (1 - \Lambda_D(s)) d\Lambda(s) \right\}^{1/2}
\]

with \( \Lambda_D(s) = \Lambda(s) - \Lambda(s^-) \) and \( \pi_i(s) = P(X_{ij} \geq s) \). In this case, a natural prior for the non-centrality parameter \( r \) is a normal distribution of the form

\[
r \sim N(0, \tau^2 \sigma^2).
\]

Under these assumptions, it follows that the Bayes factor between null and alternative hypotheses is given by

\[
BF_{01}(L^*) = (1 + \tau^2)^{1/2} \exp \left\{ -\frac{\tau^2 L^*^2}{2(1 + \tau^2)} \right\}.
\]
In contrast to the non-centrality parameter $\Delta$ in the previous examples, the non-centrality parameter $r$ lacks a simple interpretation in terms of the deviation of $H_1$ from $H_0$. As a consequence, it is difficult to specify a value of $\tau$ that reflects prior information concerning the form of plausible alternatives. We therefore take an objective approach and set $\tau$ to its MMLE under the alternative. As before, this procedure yields an upper bound of the weight of evidence against $H_0$. At the MMLE, the value of $BF_{01}(L^*)$ is

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

This quantity can be computed easily from observed survival data. For example, suppose $t_1 < t_2 < \cdots < t_D$ denote the distinct event times in the pooled sample. At $t_j$, we observe $d_{ij}$ events out of $Y_{ij}$ subjects at risk in the $i$th sample for $i = 1, 2$ and $j = 1, \cdots, D$. Let $d_j = d_{1j} + d_{2j}$ and $Y_j = Y_{1j} + Y_{2j}$ denote the number of events and the number at risk in the pooled sample at time $t_j$. Then the statistic $L^*$ can be computed as

$$L^* = \frac{\sum_{j=1}^{D} a_j (d_{1j} - \frac{Y_{1j}}{Y_j} d_j)}{\left\{ \sum_{j=1}^{D} a_j^2 \frac{Y_{1j}}{Y_j} (1 - \frac{Y_{1j}}{Y_j}) \frac{Y_j - d_j}{Y_j - 1} d_j \right\}^{1/2}}.$$

The logrank, Prentice-Wilcoxon and Gehan-Wilcoxon statistics are obtained from this expression by taking $a_j$ to be 1, $\prod_{k \leq t_j} (1 - d_k/Y_k)$, and $Y_j$, respectively.

### 2.1.2 Example: Wilcoxon signed rank test for depression data

Table 1 presents Hamilton depression scale Factor IV (the suicidal factor) measurements for nine patients with anxiety or depression before and after tranquilizer therapy (Hollander and Wolfe, 1999). These data arise from a study designed to determine the effectiveness of a new therapy in reducing depression.
Table 1: Hamilton Depression Scale Factor for nine patients before and after receiving a therapy

<table>
<thead>
<tr>
<th>Patient</th>
<th>before therapy ($X_i$)</th>
<th>after therapy ($Y_i$)</th>
<th>$Z_i = Y_i - X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.83</td>
<td>0.878</td>
<td>-0.952</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>0.647</td>
<td>0.147</td>
</tr>
<tr>
<td>3</td>
<td>1.62</td>
<td>0.598</td>
<td>-1.022</td>
</tr>
<tr>
<td>4</td>
<td>2.48</td>
<td>2.05</td>
<td>-0.43</td>
</tr>
<tr>
<td>5</td>
<td>1.68</td>
<td>1.06</td>
<td>-0.62</td>
</tr>
<tr>
<td>6</td>
<td>1.88</td>
<td>1.29</td>
<td>-0.59</td>
</tr>
<tr>
<td>7</td>
<td>1.55</td>
<td>1.06</td>
<td>-0.49</td>
</tr>
<tr>
<td>8</td>
<td>3.06</td>
<td>3.14</td>
<td>0.08</td>
</tr>
<tr>
<td>9</td>
<td>1.30</td>
<td>1.29</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

To test for a treatment effect, we apply the Wilcoxon signed rank test. Letting $\theta$ denote the change in the depression factor due to the tranquilizer, we test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The standardized Wilcoxon signed rank statistic is $W^* = -2.07$, which leads to an exact $p$ value of 0.0382; the large sample approximation to the $p$ value is 0.0391. The hypothesis of no treatment effect would be therefore be rejected in a 5% significance test. In contrast, $\tilde{B}F_{01}(W^*)$, the Bayes factor corresponding to the upper bound on the weight of the evidence against $H_0$, is 0.399, which leads to a value of $\tilde{P}(H_0|Z) = 0.285$ if the null and alternative hypothesis are given equal weight a priori. Thus, the posterior odds that the treatment has an effect are approximately 2.5:1. Values of the posterior probability of the null hypothesis obtained for different choices of $\tau$ are presented in Figure 1.

2.2 Bayes factors for nonparametric test statistics having limiting $\chi^2$ distributions

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 $\chi^2$. 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 Wilcoxon two-sample
test to the \( k \geq 3 \) sample location problem. 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. Let the total sample size be \( n = \sum_{i=1}^{k} n_i \). We wish to test
\[ H_0 : \theta_1 = \cdots = \theta_k \]
versus
\[ H_1 : \theta_i \neq \theta_j \quad \text{for some} \quad 1 \leq i, j \leq k. \]

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 \( i \)th sample, i.e.,
\[
\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}.
\]

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

\[
H_1(n) : \theta_i = \theta_0 + \Delta_i / \sqrt{n}, \quad i = 1, \ldots, k \tag{5}
\]
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

\[
\rho = 12 \left\{ \int_{-\infty}^{\infty} f^2(x) \, dx \right\}^2 \sum_{i=1}^{k} a_i (\Delta_i - \bar{\Delta})^2
\]

and

\[
\bar{\Delta} = \sum_{i=1}^{k} a_i \Delta_i.
\]

The non-centrality parameter \( \rho \) can be written as a quadratic form according to

\[
\rho = 12 \left\{ \int_{-\infty}^{\infty} f^2(x) \, dx \right\}^2 \Delta' P' Q P \Delta
\]

where

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

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

\[
P' Q P = R' \begin{pmatrix}
I_{k-1} & 0 \\
0 & 0
\end{pmatrix} R.
\]

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

\[
\Delta \sim N_k(0, c(R' R)^{-1})
\]

\[ (6) \]
where $c$ is a scaling constant.

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

$$
m_1(W) = \int_0^\infty p(W|\rho)\pi(\rho) \, d\rho
= Ga\left[ W | \frac{k-1}{2}, \frac{1}{2(\tau+1)} \right].
$$

(7)

Thus, the Bayes factor based on $W$ is

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

(8)

The value of $\tau$ that maximizes $m_1(W)$ in (7) is

$$
\tau = \frac{W - (k-1)}{k-1},
$$

provided that $W$ exceeds its expectation under $H_0$. Setting $\tau$ at this value, we obtain the upper bound of the Bayes factor against $H_0$,

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

(8)

As before, this Bayes factor (8) is not consistent against local alternatives $H_1(n)$. However, the consistency of Bayes factor can be achieved if the values of two components of $\theta$ are bounded away from each other under the alternative hypothesis. Such a constraint can
be imposed by replacing the prior distribution in (6) by an assumption that

\[ \Delta \sim N_k(0, cn(R'R)^{-1}). \]

In this case, the Bayes factor is given by

\[ BF_{01}(W) = (n\tau + 1)^{\frac{k-1}{2}} \exp \left\{ -\frac{n\tau W}{2(n\tau + 1)} \right\}, \]

which is consistent as \( n \to \infty \).

Generalizations to other nonparametric test statistics that have limiting \( \chi^2 \) distributions under the null hypothesis can be derived in a similar way. For example, Friedman’s test statistic (Friedman, 1937) can be applied to randomized complete block designs when subjects are divided into homogeneous blocks and it is of interest to compare subjects receiving different treatments within blocks. Under the null hypothesis that all treatment effects are equal, Friedman’s statistic has a limiting \( \chi^2 \) distribution. Under local alternatives similar to (5) (where not all treatment effects are equal), Friedman’s statistic also follows a limiting noncentral \( \chi^2 \) distribution (Elteren and Noether, 1959), and so the methods illustrated above for the Kruskal-Wallis test can be applied. The resulting Bayes factor has a form similar to (8).

3 Extensions to Meta-analyses

Our approach to Bayesian hypothesis testing provides an alternative to traditional meta-analysis methods for combining evidence across studies. Our approach is especially useful when study effect sizes are not reported but values of test statistics or \( p \)-values are, or when study designs or treatment levels are so different that combining effect sizes is inappropriate.
A Bayesian meta-analysis using test statistics (or \( p \)-values) proceeds as follows. For simplicity, we restrict attention here to the case in which all studies are based the same test (e.g., the log-rank test) or in which the non-centrality parameters of all tests can be transformed to a scale upon which they have a common interpretation. For concreteness, we consider the case of test statistics that have asymptotic normal distributions; results for test statistics having \( \chi^2 \) distributions follow along similar lines. Let \( T_1, \ldots, T_s \) denote the values of the test statistic from \( s \) independent studies. It follows that the asymptotic distribution of \( T = \{T_1, \ldots, T_s\} \) under \( H_0 \), say \( m_0(T) \), is that of the product of \( s \) independent standard normal random variables. Under \( H_1 \), the marginal distribution of \( T \), say \( m_1(T) \), is obtained by averaging over the common prior distribution on the non-centrality parameter \( \Delta \). Following the reasoning of Section 2.1, we specify that the prior distribution for \( \Delta \) is \( N(0, \tau^2) \). Under the assumptions of Theorem 1, it follows that \( m_1(T) \) can be expressed

\[
m_1(T) = \int p(T_1, \ldots, T_s | \Delta) p(\Delta) d\Delta = \int p(T_1 | \Delta) \cdots p(T_s | \Delta) p(\Delta) d\Delta = (1 + sC^2\tau^2)^{-1} \left( \frac{1}{2\pi} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left[ \sum T_i^2 - \frac{s^2C^2\tau^2}{1 + sC^2\tau^2} (\bar{T})^2 \right] \right\}
\]

where \( \bar{T} = \sum_{i=1}^s T_i/s \).

The Bayes factor arising from the test statistics reported in the \( s \) studies is therefore

\[
BF_{01}(T) = \frac{m_0(T)}{m_1(T)} = (1 + sC^2\tau^2)^{1/2} \exp \left[ -\frac{s^2C^2\tau^2(\bar{T})^2}{2(1 + sC^2\tau^2)} \right].
\]

Minimizing this expression with respect to (a common) \( \tau \), we find that the upper bound of
the Bayes factor against the null hypothesis is
\[
\tilde{BF}_{01}(T) = s^2 |T| \exp \left[ \frac{1 - s|T|^2}{2} \right].
\]  (9)

The next example provides an illustration of this methodology by combining evidence from several studies in which log-rank tests were used to summarize evidence against the null hypothesis.

**Meta-analysis comparing two chemotherapies of ovarian cancer**

The Ovarian Cancer Meta-analysis Project (1991) undertook a meta-analysis of ovarian cancer treatments based on four published randomized clinical trials that compared two regimens of chemotherapy for ovarian carcinoma: cyclophosphamide plus cisplatin (CP), and cyclophosphamide, doxorubicin and cisplatin (CAP). Table 2 displays the logrank statistics \((O - E) / \sqrt{V}\) reported from the four studies. Each statistic was based on a comparison of survival times in the treatment of ovarian carcinoma using CP or CAP. None show a significant survival benefit between CP and CAP in 5% tests. To combine evidence from these four trials, a meta-analysis project was conducted to retrospectively collect information from all 1194 individual patients in all four trials. Information from the four trials was pooled, and a stratified log-rank test in which studies were treated as stratum was conducted. The results from the meta-analysis are summarized in the last line of Table 2. The classical meta-analysis suggested a significant survival benefit for CAP (\(p\) value = .02), concluding that CAP was significantly better than CP. This approach, the so-called meta-analysis of individual patient data (IPD), required the retrieval of individual patient records from trial investigators, and was thus very time consuming and resource intensive. In more typical meta-analysis, only
Table 2: Four clinical trials comparing the survival time in the treatment of ovarian carcinoma by using CP or CAP regimen in chemotherapy. $O$ is the observed number of events; $E$ is the expected number of events if the probability of death were unrelated to treatment; and $V$ is the variance of $O - E$.

<table>
<thead>
<tr>
<th>Study</th>
<th>sample size</th>
<th>$O - E$</th>
<th>$V$</th>
<th>$\frac{O - E}{\sqrt{V}}$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GICOG</td>
<td>382</td>
<td>-7.4</td>
<td>73.9</td>
<td>-0.86</td>
<td>0.39</td>
</tr>
<tr>
<td>GOG</td>
<td>412</td>
<td>-10.5</td>
<td>58.2</td>
<td>-1.38</td>
<td>0.17</td>
</tr>
<tr>
<td>DACOVA</td>
<td>275</td>
<td>-10.8</td>
<td>55.1</td>
<td>-1.45</td>
<td>0.15</td>
</tr>
<tr>
<td>GONO</td>
<td>125</td>
<td>-4.6</td>
<td>22.4</td>
<td>-0.97</td>
<td>0.33</td>
</tr>
<tr>
<td>Meta-Analysis</td>
<td>1194</td>
<td>-33.3</td>
<td>209.6</td>
<td>-2.30</td>
<td>0.02</td>
</tr>
</tbody>
</table>

aggregated results from individual trials are available, and it is then difficult to combine $p$ values or other reported summary statistics together in order to arrive at a combined measure of evidence.

The Bayes factor methodology proposed here provides a second way to combine information across trials. Based on (9), it follows that the value of $\tilde{BF}_{01}(T)$ is 0.253. This value is quite similar to the upper bound of Bayes factor (against the null) derived directly from the single combined logrank test statistics based on IPD meta-analysis. The value of the latter is 0.269. If equal probabilities are assumed for each hypothesis a priori, the minimum posterior probability of the null hypothesis based on (9) is 0.202, or odds of about 4:1 that the tested treatment improved survival time.

Interestingly, the statistically significant result of the IPD meta-analysis by Ovarian Cancer Meta-analysis Project led to a large multi-national random trial with 1526 patients, known as the Second International Collaborative Ovarian Neoplasm Study (ICON2), to further compare CAP with carboplatin. Carboplatin was believed to have an efficacy similar to that of CP, but with less toxicity. Surprisingly, the conclusions of ICON2 indicated no survival difference between CAP and carboplatin based on yet another logrank test. The
p value for the second study was 0.98 (ICON collaborators, 1998). Possible explanations for this apparent discrepancy between studies are discussed by Buyse et al. (2003). Our interpretation is simpler: Evidence contained in the original studies for a treatment effect was equivocal and was, unfortunately, overstated.

Actually, based on the reported logrank test statistics, the evidence contained in the original four studies can be combined easily with ICON2 by our approach. In contrast, IPD meta-analysis cannot be performed unless the individual patient information of ICON2 can be retrieved. Our method yields \(\widetilde{BF}_{01}(T)\) of 0.384, or odds of about 5:2 that CAP improves survival time if equal probabilities are assumed for each hypothesis \textit{a priori}. The combination of studies still supports a treatment effect, although the magnitude of the effect (based on MMLE of \(\tau\)) is likely to be small.

4 Comparison of \(p\) values and Bayes factors

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

We first consider the one-sample Wilcoxon test, and assume that the null and alternative hypotheses are assigned equal odds \textit{a priori}. For this test statistic, Figure 2 displays (a) the lower bound on the probability of the null hypothesis under the class of alternative
hypotheses specified in Section 2, (b) the exact $p$ value, and (c) the $p$ value based on the large sample normal approximation. To the extent possible, the test statistic was chosen to fall exactly on the boundary of the 5% critical region; if such a value could not be achieved, we chose the value of the test statistic that resulted in the $p$ value closest to 0.05. For sample sizes larger than 15, Figure 2 shows that $p$ values based on asymptotic approximations are very close to their nominal values, and that the value of $\tilde{P}(H_0|x)$ is quite stable.

Interestingly, the value of $\tilde{P}(H_0|x)$ is approximately 0.3 for all sample sizes. 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. To achieve “agreement” between the $p$ value and the posterior probability of the null model, i.e., to have $\tilde{P}(H_0|x) = 0.05$ when the $p$ value is 0.05, the prior probability assigned to the null hypothesis must be approximately 0.15. These results are comparable to those reported by Beger and Delampady (1987) and Johnson (2005). The relation between $p$ values and Bayes factors based on the Wilcoxon one-sample test statistic is explored further in Figure 3, where we plot $\tilde{P}(H_0|x)$ versus different exact $p$ values for a sample size of 15. Exact $p$ values were obtained from tables contained in Wilcoxon, Katti and Wilcox (1973).

Figures 4 and 5 show similar comparisons for the Mann-Whitney-Wilcoxon test.

5 Conclusion

Traditionally, the application of Bayesian testing procedures to classical nonparametric settings has been restricted due to the absence of sampling densities for the 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, it should also alleviate objections from those who prefer more classical procedures. Finally, the use of our methods has the potential for greatly simplifying the process of combining evidence from independent studies in favor of various statistical hypotheses.

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 common test statistics. For both one-sample and two-sample Wilcoxon tests, our method appears to give accurate results for sample sizes larger than 15. This figure is consistent with recommendations for the application of large sample approximations cited by Gibbons and Chakraborti (2003).

6 Appendix

6.1 Proof of Theorem 1

\[ T_k^* = \frac{T_k - \mu(\theta_0)}{\sigma_k(\theta_0)} = \frac{T_k - \mu(\theta_k) + \mu(\theta_k) - \mu(\theta_0)}{\sigma_k(\theta_0)} \]
\[ \approx \frac{T_k - \mu(\theta_k) + \mu'(\theta_0)(\theta_k - \theta_0)}{\sigma_k(\theta_0)} \]
\[ = \frac{T_k - \mu(\theta_k)}{\sigma_k(\theta_0)} + \frac{\mu'(\theta_0)}{\sqrt{\Delta}} \]
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.

### 6.2 Proof of Theorem 3

Taking the logarithm of (4) we obtain
\[
\log(BF_{01}) = \frac{1}{2} \log(1 + C^2 \tau^2 n_k) - \frac{C^2 n_k \tau^2 T_k^*}{2(1 + C^2 n_k \tau^2)}.
\]

If $H_0$ is true, $T_k^*$ has an asymptotic $\chi^2$ distribution with 1 degree of freedom, so the second term is bounded in probability. Since the first term is $O(\log(n_k))$, it follows that $\log(BF_{01}) \to \infty$.

If $H_1$ is true, $\log(BF_{01})$ can be rewritten as
\[
\log(BF_{01}) = \frac{1}{2} \log(1 + C^2 \tau^2 n_k) - \frac{1}{2} C^2 n_k \tau^2 \left( \frac{T_k^*}{\sqrt{1 + C^2 n_k \tau^2}} \right)^2.
\]
Because
\[
\left( \frac{T_k^*}{\sqrt{1 + C^2 n_k \tau^2}} \right)^2
\]
has an asymptotic $\chi^2$ distribution with 1 degree of freedom, it is bounded probability. The second term is $O(n_k)$. Since the first term is $O(\log(n_k))$, it follows that $\log(BF_{01}) \to -\infty$. 
References

Annals of Mathematical Statistics, 25, 724-736

[2] Buyse, M., Burzykowski, T. and et al. (2003), Using the expected survival to explain 
differences between the results of randomized trials: a case in advanced ovarian cancer. 
Journal of Clinical Oncology, 21, 1682-1687

York: Springer.

Statistical Science, 2, 317-352

of p value and evidence. Journal of American statistical Association, 82, 112-122

the one-sided testing problem. Journal of American statistical Association, 82, 106-111


American statistical Association, 72, 138-142


Figure 1: Posterior probability of $H_0$ under different values of $\tau$ for the depression data

Figure 2: This figure depicts the posterior probability of $H_0$, the exact $p$ value and the normal approximation to the $p$ value for the one-sample Wilcoxon test as sample sizes are varied when the value of the test statistic corresponding most closely to an exact $p$ value of .05 is observed. The dotted line indicates the value of 0.05.
Figure 3: This figure shows the relation between the posterior probability of $H_0$ and the normal approximation to the $p$ value for the one-sample Wilcoxon test based on a sample size of 15. The dotted line corresponds to the exact $p$ value.

Figure 4: This figure depicts the posterior probability of $H_0$, the exact $p$ value and the normal approximation to the $p$ value for the two-sample Wilcoxon test as sample sizes are varied when the value of the test statistic corresponding most closely to an exact $p$ value of .05 is observed. The dotted line indicates the value of 0.05.
Figure 5: This figure shows the relation between the posterior probability of $H_0$ and the normal approximation to the $p$ value for the two-sample Wilcoxon test based on a sample size of 15. The dotted line corresponds to the exact $p$ value.