Generalized *p*-Values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters

KAM-WAH TSUI and SAMARADASA WEERAHANDI*

This article examines some problems of significance testing for one-sided hypotheses of the form $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, where θ is the parameter of interest. In the usual setting, let x be the observed data and let $T(X)$ be a test statistic such that the family of distributions of $T(X)$ is stochastically increasing in θ. Define C_x as $\{X : T(X) - T(x) \ge 0\}$. The p value is $p(x) = \sup_{\theta \le \theta_0} \Pr(X \in C_x \mid \theta)$. In the presence of a nuisance parameter η, there may not exist a nontrivial C_x with a p value independent of η. We consider tests based on generalized extreme regions of the form $C_x(\theta, \eta) = \{X : T(X; x, \theta, \eta) \ge T(x; x, \theta, \eta)\}$, and conditions on $T(X; x, \theta, \eta)$ are given such that the p value $p(x) = \sup_{\theta \le \theta_0} \Pr(X \in C_x(\theta, \eta))$ is free of the nuisance parameter η, where T is stochastically increasing in θ. We provide a solution to the problem of testing hypotheses about the differences in means of two independent exponential distributions, a problem for which the fixed-level testing approach has not produced a nontrivial solution except in a special case. We also provide an exact solution to the Behrens–Fisher problem. The p value for the Behrens–Fisher problem turns out to be numerically (but not logically) the same as Jeffreys's Bayesian solution and the Behrens–Fisher fiducial solution. Our approach of testing on the basis of p values is especially useful in multiparameter problems where nontrivial tests with a fixed level of significance are difficult or impossible to obtain.

KEY WORDS: Behrens–Fisher problem; Exponential distribution; Invariance; Unbiasedness.

1. INTRODUCTION

This article examines some problems of significance testing for one-sided hypotheses of the form $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, where θ is the parameter of interest. Our study is motivated by the usual lack of a nontrivial continuous family of similar tests [and hence uniformly most powerful (UMP) tests] based on minimal sufficient statistics when nuisance parameters are present.

In the usual setting for significance testing of the one-sided hypotheses just given, a data-based extreme region C_x is typically of the form

$$C_x = \{X : T(X) - T(x) \ge 0\}, \qquad (1.1)$$

where x denotes the observed data, X denotes possible sample points, and $T(\cdot)$ is a function, known as a test statistic, such that large values of $T(\cdot)$ indicate evidence against H_0. The p value, or the observed level of significance, is

$$p = \sup_{\theta \in H_0} \Pr(X \in C_x \mid \theta). \qquad (1.2)$$

A small value of p suggests that the observed x does not support H_0.

When a nuisance parameter η is present, the p value based on a nontrivial and well-behaved test statistic may depend on η and hence cannot be used to judge against H_0. We consider a generalization of the data-based extreme region of a test, which not only depends on the observed data x but also may involve the parameter $\xi = (\theta, \eta)$, provided that the p value is independent of η. We call the resulting extreme region, $C_x(\xi)$, a *generalized extreme region*. We show in this article, through several examples, that with this generalization the p value in (1.2), with C_x replaced by $C_x(\xi)$, is independent of η and hence can be used as a measure of evidence against H_0. Other approaches to significance testing in the presence of a nuisance parameter were described by Kempthorne and Folks (1971, chap. 12), among others. Substantial discussions on significance tests and their applications can be found, for example, in Cox (1977), Kempthorne and Folks (1971), Cox and Hinkley (1974), Gibbons and Pratt (1975), and Thompson (1985).

To illustrate the potential usefulness of our approach, consider the Behrens–Fisher problem, which can be formulated as follows. Suppose $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are two sets of independent observations from normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Assume that $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$ are independent. It is desired to test the null hypothesis $H_0 : \mu_1 - \mu_2 \le 0$ against the alternative $H_1 : \mu_1 - \mu_2 \underline{>} \underline{0}$ on the basis of the independent sufficient statistics $\overline{X}$, $\overline{Y}$, S_1^2, and S_2^2, which are also the maximum likelihood estimators of the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2, respectively. With this notation, the distributions of the underlying random variables are given by

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right), \qquad \overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right),$$

$$\frac{mS_1^2}{\sigma_1^2} \sim \chi^2_{m-1}, \qquad \frac{nS_2^2}{\sigma_2^2} \sim \chi^2_{n-1}, \qquad (1.3)$$

and the random variables $\overline{X}$, $\overline{Y}$, S_1^2, and S_2^2 are all independent.

Let $(\bar{x}, \bar{y}, s_1^2, s_2^2)$, x, and y be the observed values of $(\overline{X}, \overline{Y}, S_1^2, S_2^2)$, X, and Y, respectively. In this problem, let the parameter of interest be $\theta = (\mu_1 - \mu_2)/(\sigma_1^2/m +$

* Kam-Wah Tsui is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706. Samaradasa Weerahandi is Member of Technical Staff, Bell Communications Research, Livingston, NJ 07039. The authors thank the referees for helpful comments and Ella Mae Matsumura for useful suggestions.

Journal of the American Statistical Association
June 1989, Vol. 84, No. 406, Theory and Methods

$\sigma_2^2/n)^{1/2}$ (or, equivalently, $\mu_1 - \mu_2$). The nuisance parameter is $\eta = (\sigma_1^2, \sigma_2^2)$.

To find a random quantity that can be used in place of T in (1.1), consider

$$\tilde{W}(x, Y; x, y, \eta) = (\bar{X} - \bar{Y})\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)^{-1/2} \times \left[\frac{\sigma_1^2 s_1^2}{mS_1^2} + \frac{\sigma_2^2 s_2^2}{nS_2^2}\right]^{1/2}. \quad (1.4)$$

Then, $\tilde{w} = \tilde{W}(x, y; x, y, \eta) = \bar{x} - \bar{y}$. Moreover, for given x and y, the distribution of $\tilde{W}(X, Y; x, y, \eta)$ is free of η and has the same distribution of $Z[s_1^2/U + s_2^2/V]^{1/2}$, where $Z \sim N(\theta, 1)$, $U \sim \chi^2_{m-1}$, $V \sim \chi^2_{n-1}$, and the random variables Z, U, and V are independent. Furthermore, the family of cumulative distributions of $\tilde{W}(X, Y; x, y, \eta)$ for given x and y is stochastically increasing in θ (see Lehmann 1986, p. 84).

We call $\tilde{W}(\cdot)$ a *generalized test variable*. It depends not only on the random variables X and Y, but also on the observed x and y and the parameters. This general form of $\tilde{W}(\cdot)$ enables us to extend the notion of usual data-based extreme regions of (1.1) to generalized extreme regions, such as

$$C_{x,y}(\theta, \eta) = \{(X, Y) : \tilde{W}(X, Y; x, y, \eta) - \tilde{W}(x, y; x, y, \eta) \geq 0\}, \quad (1.5)$$

a set of possible sample points (X, Y) that are viewed as more extreme than or as extreme as (x, y). The flexibility of allowing for the observed data and the parameters in constructing generalized extreme regions enables us to produce a p value that does not depend on the nuisance parameter of the testing problem. In the Behrens–Fisher problem, the p value is

$$\begin{aligned} p &= \Pr(\tilde{W} \geq \tilde{w} \mid \theta = 0) \\ &= \Pr\left[T(m + n - 2)^{-1/2}\left(\frac{s_1^2}{B} + \frac{s_2^2}{1 - B}\right)^{1/2} \geq \bar{x} - \bar{y}\right] \\ &= \mathrm{E}_B\left\{\Psi\left[(\bar{y} - \bar{x})\left(\frac{s_1^2}{B} + \frac{s_2^2}{1 - B}\right)^{-1/2} \times (m + n - 2)^{1/2}\right]\right\}, \end{aligned} \quad (1.6)$$

where $T = Z[(U + V)/(m + n - 2)]^{-1/2}$ has a Student-t distribution with $(m + n - 2)$ degrees of freedom and is independent of $B = U/(U + V) \sim \text{beta}((m - 1)/2, (n - 1)/2)$, $\Psi(\cdot)$ is the cdf of Student's t distribution with $(m + n - 2)$ degrees of freedom, and E_B denotes expectation with respect to B.

Salaevskii (1963) showed that in fixed-level testing there are no nontrivial continuous families of nonrandomized similar tests for the Behrens–Fisher problem (see also Linnik 1968); our extended p value approach, however, produces a nontrivial solution. This is in contrast to problems not involving nuisance parameters, where fixed-level testing and the usual p value approach provide essentially the same result in most cases.

The remainder of the article is organized as follows. In Section 2, we discuss some general theory on constructing generalized extreme regions and examine invariant testing problems. A key result on data-based power functions of invariant tests is derived. The usefulness of this result is illustrated in Section 3 by providing a solution to the problem of testing hypotheses about the differences in means of two exponential distributions and by providing a solution to the problem of testing the mean of a truncated exponential distribution. Section 4 provides a discussion on solutions of the Behrens–Fisher problem. In particular, we provide a justification for considering tests based on the generalized test variable $\tilde{W}$ in (1.5) alone and give conditions under which the p value in (1.6) is unique. Section 5 contains concluding remarks.

2. GENERAL THEORY

Let X be a random quantity having a density function $f(x \mid \xi)$, where $\xi = (\theta, \eta)$ is an unknown vector of parameters assuming values in a parameter space H, θ is the parameter of interest, and η is the vector of nuisance parameters. Let χ be the sample space and x be the observed value of X. The problem of interest is to test the null hypothesis $H_0 : \theta \leq \theta_0$ versus the alternative hypothesis $H_1 : \theta > \theta_0$. A fixed-level test rejects the null hypothesis if the observed x falls in a critical region C, which typically is of the form $C = \{X : T(X) \geq c\}$ for some statistic $T(X)$ and some constant c, such that $\sup_{\theta \in H_0} \Pr(X \in C \mid \theta)$ is equal to a prefixed significance level α. In testing problems involving nuisance parameters η, nontrivial tests with a fixed level of significance are often difficult or impossible to obtain.

An alternative approach is to consider a data-based critical region, a set C_x, consisting of sample points X considered at least as extreme as x according to the ordering of a certain test statistic whose family of distributions is stochastically increasing in θ. A small p value, computed according to (1.2), indicates that the observed x does not support H_0. Although this p value approach appears to be more flexible than fixed-level testing in the sense that C_x is allowed to involve x, the discussion in Section 1 shows that it may encounter difficulty in multiparameter problems; the p value can depend on η, the nuisance parameter, and hence may not be computable.

To overcome this difficulty, we consider *generalized test variables* of the form $T(X; x, \xi)$, which is not a function of X only, but also involves the observed x and the parameter ξ. We impose three requirements on $T(X; x, \xi)$:

Requirement 1. $T(x; x, \xi)$ is free of ξ.

Requirement 2. For fixed x and $\xi = (\theta_0, \eta)$, the distribution of $T(X; x, \xi)$ is free of the nuisance parameter η.

Requirement 3. For fixed x and η, $\Pr(T(X; x, \xi) \geq t \mid \theta)$ is nondecreasing in θ.

Without loss of generality, one may consider Requirement 1 to be redundant, because if it is not satisfied, then we can define a generalized test variable $T'(X; x, \xi)$ as $T'(X; x, \xi) = T(X; x, \xi) - T(x; x, \xi)$ and impose Requirements 2 and 3 on T'. As in the usual significance testing approach, Requirement 3 requires the generalized test variable to be stochastically increasing in θ.

For a generalized statistic $T(X; x, \xi)$ satisfying these requirements, consider the test based on the generalized extreme region

$$C_x(\xi) = \{X : T(X; x, \xi) - T(x; x, \xi) \geq 0\}. \quad (2.1)$$

The p value of this test is defined in (1.2) with C_s replaced by $C_x(\xi)$ and is equal to

$$\Pr(X \in C_x(\xi) \mid \theta = \theta_0). \quad (2.2)$$

This p value is computable, since it is free of the nuisance parameter η. In this article we confine our attention mainly to $T(X; x, \xi)$ satisfying Requirements 1–3.

The problem of finding a generalized test variable can be approached in the usual way. One can first reduce a testing problem by sufficiency. The principle of invariance and possibly the notion of unbiasedness can then be used. The usual definitions and basic material on invariance, maximal invariants, unbiasedness, and other useful concepts in fixed-level testing can be found in Lehmann (1986). When analogous concepts used in the p value approach here differ from those in fixed-level testing, we use *p-invariant,* for example, instead of *invariant* to avoid confusion. Requirement 2 is the counterpart of the notion of "similarity" on the boundary of the hypotheses.

Let $T(X; x, \xi)$ be a generalized test variable, and let $C_x(\xi)$ be the generalized extreme region as defined in (2.1). Given the observed x, the data-based power function of a test based on $C_x(\xi)$ is defined as

$$\pi_\theta(x) = \Pr(X \in C_x(\xi) \mid \theta). \quad (2.3)$$

Suppose that the testing problem is invariant under a group G of transformations on the sample space χ, in the usual sense. Let $\overline{G}$ be the group of transformations induced on the parameter space H. It is natural to require that the p value, $p(x)$, as a function of x, be invariant under G. The invariance of the data-based power function, as defined next, satisfies this requirement.

Definition 2.1. A test based on a generalized extreme region $C_x(\xi)$ is *p-invariant* under G if $\pi_\theta(g(x)) = \pi_\theta(x)$ for all $x \in \chi$ and $g \in G$.

Lehmann (1986, pp. 285, theorem 1, and 289) showed that in fixed-level testing, given a maximal invariant $M(x)$ with respect to G, the class of all invariant tests is the class of tests depending on M. The result in the following theorem provides conditions on a generalized test variable $T(X; x, \xi)$ so that in the p value approach, the data-based power function of any p-invariant test can be obtained through T and a maximal invariant $M(x)$ with respect to G.

Theorem 2.1. Suppose that the testing problem is invariant under a group G of transformations on the sample space χ. Let $M(x)$ be a maximal invariant with respect to G. Suppose $T = T(X; x, \xi)$ is an absolutely continuous random variable (for fixed x and ξ). If the observed value $t(x) = T(x; x; \xi)$ and the distribution of T depends on x only through $m = M(x)$, then the data-based power function $\pi_\theta(x)$ of any p-invariant test can be obtained using only T and m (but not x).

Proof. Since $\pi_\theta(x)$ is invariant, it can be expressed as $\pi_\theta(x) = \psi_\theta(M(x))$; that is, $\pi_\theta(x)$ depends on the data only through $M(x)$ (see Lehmann 1986, p. 285). Now we need to show that $\psi_\theta(m)$ is attainable using T, where $m = M(x)$. To do this, let $F(t; \xi, m)$ be the cdf of T. Since T is an absolutely continuous random variable, $W = F(T; \xi, m)$ has a uniform distribution on [0, 1]. Let $w_\xi(m) = F(t(m); \xi, m)$ be the observed value of W, where $t(m) = T(x; x; \xi)$. Then consider the particular extreme region defined as

$$\begin{aligned} C_m &= w_\xi \leq W \leq w_\xi + \psi_\theta, && \psi_\theta \leq .5,\ w_\chi \leq .5 \\ &= w_\xi - \psi_\theta \leq W \leq w_\xi, && \psi_\theta \leq .5,\ w_\xi > .5 \\ &= 0 \leq W \leq \psi_\theta, && \psi_\theta \geq .5,\ w_\xi \leq .5 \\ &= 1 - \psi_\theta \leq W \leq 1, && \psi_\theta \geq .5,\ w_\xi > .5. \end{aligned}$$

Since W is a uniform random variable, $\Pr(W \in C_m) = \psi_\theta(m)$. Moreover, by this construction, $w_\xi \in C_m$. Hence the generalized extreme region $\{T : F(T; \xi, m) \in C_m\}$ contains $t(m)$ and yields the data-based power function $\pi_\theta(x) = \psi_\theta(m)$.

The next section illustrates the usefulness of Theorem 2.1 in two applications.

3. EXAMPLES OF p-INVARIANT TESTS

3.1 Comparison of Means of Exponential Distributions

Let $G(\alpha, \beta)$ denote the gamma distribution with shape parameter α and scale parameter β. Let $X_1, X_2, \ldots, X_m$ be a random sample from the exponential distribution $G(1, \mu_1)$ with mean μ_1, and let $Y_1, Y_2, \ldots, Y_n$ be a random sample from $G(1, \mu_2)$ with mean μ_2. The X_i's and Y_j's are assumed independent. Suppose the null hypothesis $H_0 : \mu_1 - \mu_2 \leq \delta_0$ is tested against the alternative $H_1 : \mu_1 - \mu_2 > \delta_0$ $(\delta_0 \geq 0)$. By sufficiency, we can confine our attention to tests based on $X = \sum X_i \sim G(m, \mu_1)$ and $Y = \sum Y_i \sim G(n, \mu_2)$.

Here the underlying distribution is invariant under the group of common scale changes $(X, Y) \rightarrow (kX, kY)$, $(\mu_1, \mu_2) \rightarrow (k\mu_1, k\mu_2)$, $k > 0$. This testing problem is considered invariant in the usual sense only when $\delta_0 = 0$. This is not an inherent problem of the exponential distribution, but it is a drawback in the usual formulation of the testing of hypotheses. In practice, one should not be restricted to specifying the hypotheses with a certain scale. For example, suppose μ_i is the mean lifetime of an electronic component of brand i. We may write $H_0 : \mu_1 - \mu_2 \leq 2$ if the units are in years and $H_0 : \mu_1 - \mu_2 \leq 24$ if the units are in months. With the normal distribution one can transform the original problem so that the transformed problem has $\delta_0 = 0$. In this study it is shown that, in testing with the p value, one can more generally make a testing problem invariant in a different manner, as follows.

Suppose x and y are observed values of X and Y, respectively. Let $\lambda_i = \mu_i/x$ ($i = 1, 2$), $\theta_0 = \delta_0/x$, and $\theta = \lambda_1 - \lambda_2$. Then the problem is equivalent to testing

$$H_0 : \theta \le \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0. \tag{3.1}$$

With this reparameterization the testing problem is invariant, regardless of the value of the unit-free quantity θ_0 under the common scale transformation $(X, Y) \rightarrow (kX, kY)$ and the induced transformation on the new parameters $(\theta, \lambda_2) \rightarrow (\theta, \lambda_2)$. Here θ is the parameter of interest and λ_2 is the nuisance parameter. [Other choices such as μ_i/y or $\mu_i/(x + y)$ in place of μ_i/x would perform equally well.]

Consider the generalized test variable

$$\begin{aligned} T((X, Y); (x, y), (\theta, \lambda_2)) &= \lambda_2 y/Y - (\theta + \lambda_2)x/X + \theta \\ &= (y/x)/V - 1/U + \theta, \end{aligned} \tag{3.2}$$

where $U = X/[(\theta + \lambda_2)x] \sim G(m, 1)$, $V = Y/(\lambda_2 x) \sim G(n, 1)$, and U and V are independent. The cdf of T, $F_T(t) = \Pr((y/x)/V - 1/U + \theta \le t)$, and the observed value $T((x, y); (x, y), (\theta, \lambda_2)) = 0$ depend on the data only through the maximal invariant $M(x, y) = y/x$. Hence, by Theorem 2.1 the power of function of any p-invariant test can be obtained using T.

Recall that $\mu_1 = (\theta + \lambda_2)x$ and $\mu_2 = \lambda_2 x$. Consider the p-invariant test based on the generalized extreme region

$$\begin{aligned} &C_{x,y}(\theta, \lambda_2) \\ &\quad = \{(X, Y) : (y/x)(\mu_2/Y) - \mu_1/X + \theta \ge 0\}. \end{aligned} \tag{3.3}$$

Observe that the cdf of T is decreasing in θ and is independent of the nuisance parameter, so when $\theta = \theta_0$ the probability of $C_{x,y}$ serves as a measure of how the data support H_0. The p value of this test is

$$\begin{aligned} p &= \Pr(T \ge 0 \mid \theta = \theta_0) \\ &= \Pr((y/x)/V - 1/U - \theta_0 \ge 0) \\ &= \mathrm{E}_U\{\Gamma_n[(y/x)U(1 + \theta_0 U)^{-1}]\} \\ &= \mathrm{E}_U\{\Gamma_n[yU(x + \delta_0 U)^{-1}]\}, \end{aligned} \tag{3.4}$$

where $\Gamma_n(z)$ is the incomplete gamma function with parameter n and the expectation, E_U, is taken with respect to $U \sim G(m, 1)$. The p value is clearly free of the nuisance parameter.

The usual testing procedures for comparing two exponential means apply only when $\theta_0 = 0$. In this case, the usual F test coincides with ours. To see this, note that when $\theta_0 = 0$ (or $\mu_1 - \mu_2 = 0$), the original problem is invariant in the usual sense under common scale changes. $M(X, Y) = Y/X$ is a maximal invariant, and its density function has the form $\gamma f(\gamma z)$, $z > 0$, where $\gamma = \mu_1/\mu_2$ and $f(\cdot)$ is the density function of an F distribution with degrees of freedom $2n$ and $2m$. The family $\{\gamma f(\gamma z)\}_{\gamma > 0}$ has the monotone likelihood ratio property in z^{-1}. Hence, in fixed-level testing, UMP tests for testing $H_0 : \mu_1/\mu_2 \le 1$ against $H_1 : \mu_1/\mu_2 > 1$ have critical regions of the form $\{(Y/X)^{-1} > c\}$. Moreover, the distribution of X/Y is stochastically increasing in $\gamma = \mu_1/\mu_2$. Therefore, in conventional significance testing with a p value, the extreme region $\{X/Y > x/y\}$ is used. In other words, the p value given in (3.4) when $\theta_0 = 0$, $p = \Pr(y/x \ge V/U) = \Pr(X/Y \ge x/y)$, is the same as the p value of a UMP-invariant test, the usual F test (see also Lawless 1982, p. 112). One can also show that the usual test is UMP-unbiased when $\theta_0 = 0$.

3.2 Testing the Mean of the Truncated Exponential Distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from the truncated exponential distribution with parameters α and β—namely, $f(x; \alpha, \beta) = \beta^{-1}\exp(-(x - \alpha)/\beta)$ for $x > \alpha$. The null hypothesis $H_0 : \alpha + \beta \le \mu_0$ is to be tested against the alternative $H_1 : \alpha + \beta > \mu_0$, where $\mu = \alpha + \beta$ is the mean of the distribution. Sufficiency reduces the problem to tests based on the two independent statistics $U = \min X_i$ and $V = \overline{X} - \min X_i$, where $U - \alpha \sim G(1, \beta/n)$ and $V \sim G(n - 1; \beta/n)$.

Let u and v be the observed values of U and V, respectively. Let $\lambda = \alpha/v$, $\mu = \beta/v$, and $\theta = \lambda + \mu$. Our problem is equivalent to testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$. With this reparameterization, the problem is invariant under the common scale transformation $(V, U) \rightarrow (kV, kU)$, $k > 0$. The parameter of interest is θ, and λ (or, equivalently, $\alpha = \lambda v$) is the nuisance parameter. Consider the generalized test variable

$$\begin{aligned} T((V, U); (v, u), (\theta, \lambda)) &= \theta + (U - \lambda v)/V - (\theta - \lambda)v/V \\ &= \theta + Y_1/Y_2 - 2n/Y_2, \end{aligned} \tag{3.5}$$

where $Y_1 \sim \chi^2_2$ and $Y_2 \sim \chi^2_{2(n-1)}$ are independent. Furthermore, the distribution of T is independent of the nuisance parameter λ and the observed u and v, and the observed value of T, $T((v, u); (v, u), (\theta, \lambda)) = u/v$, is a maximal invariant. From Theorem 2.1, the power function of any p-invariant test can be obtained using T. The p value of the test with the generalized extreme region

$$C_{v,u}(\theta, \lambda) = \{(V, U) : T((V, U); (v, u), (\theta, \lambda)) \ge u/v\}$$

is

$$\begin{aligned} p &= \Pr(\theta_0 + (Y_1 - 2n)/Y_2 \ge u/v) \\ &= \Pr((Y_1 - 2n)/Y_2 \ge (u - \mu_0)/v). \end{aligned} \tag{3.6}$$

This p value depends on μ_0 given in the original null hypothesis and is free of any unknown parameters. Moreover, p has a uniform distribution on [0, 1]. This means that in this case, one can equivalently carry out the test with a fixed level of significance as well.

4. MORE ON THE BEHRENS–FISHER PROBLEM

In this section, we discuss the relationship of the p value solution (1.6) of the Behrens–Fisher problem described in Section 1 to solutions derived via other approaches. We then provide some theoretical properties of the generalized test statistic $\tilde{W}(\cdot)$ in (1.4). For the problem of testing $H_0 : \mu_1 - \mu_2 \leq \delta_0$ versus $H_1 : \mu_1 - \mu_2 > \delta_0$, it is easily deduced from (1.6) that the p value derived by means of $\tilde{W}(\cdot)$ in (1.4) is

$$p = \mathrm{E}_B\{\Psi[(\delta_0 + \bar{y} - \bar{x})(m + n - 2)^{1/2} \times (s_1^2/B + s_2^2/(1 - B))^{-1/2}]\}. \quad (4.1)$$

We also provide conditions under which we can consider tests based on W alone.

Johnson and Weerahandi (1988) provided a Bayesian solution to the multivariate Behrens–Fisher problem. In the univariate case with noninformative reference priors on the parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, the posterior probability that the null hypothesis is true $(\mu_1 - \mu_2 \leq \delta_0)$ was found to be the same as the p value given in (4.1). Moreover, this Bayesian solution is equivalent to Jeffreys's (1961) solution given in terms of the Behrens–Fisher distribution. This means that one can compute the p value in (1.6) using the tables constructed by Sukhatmé (1938) for the Behrens–Fisher distribution. Although these different approaches lead to the same numerical value, the interpretations of this value are very different. In the p value approach presented in this article, p in (1.6) is viewed as a frequency probability. For general one-sided tests when no nuisance parameter is involved, Casella and Berger (1987) provided some results on the relationship of p values and posterior probabilities.

Weerahandi (1987) provided a solution similar to (1.6) in a regression setting, but did not discuss theoretical properties of the solution. Observe that the test based on the generalized extreme region $C_{x,y}(\theta, \eta)$ in (1.5) satisfies the property that the data-based power function at θ always exceeds the p value given in (1.6) for all θ in the alternative hypothesis. We call this property *p-unbiased*. We also use the term *p-similar* to describe a test based on a generalized extreme region satisfying the property that the data-based power function is free of the nuisance parameter η when the parameter of interest, θ, equals θ_0, the boundary of the hypotheses. For a p-similar test, the p value defined by (2.2) is computable and free of the nuisance parameter. Note that a p-unbiased test with a data-based power function continuous in θ is p-similar.

To show that we can consider tests based only on W given in (1.4), we first show how to reduce the Behrens–Fisher problem by invariance. The notation from Section 1 will be used here. By location invariance, the problem can be reduced to tests based on $(\bar{X} - \bar{Y})$, S_1^2, and S_2^2. Let $h = \sigma_1^2/\sigma_2^2$. Common scale invariance reduces the problem to tests based on the variables

$$T_1 = (\bar{X} - \bar{Y})/(mS_1^2 + nS_2^2h)^{1/2}$$

$$T_2 = (mS_1^2 + nS_2^2h)\{s_1^2/mS_1^2 + s_2^2/nS_2^2h\}/s_1^2. \quad (4.2)$$

By independence of the sum and the ratio of the χ^2 variables given in (1.3), T_1 and T_2 are independent. Moreover, T_1 and T_2 can be expressed as

$$T_1 = T(m + n - 2)^{-1/2}(1/m + 1/(nh))^{1/2}$$

$$T_2 = (s_1^2/B + s_2^2/(1 - B))/s_1^2, \quad (4.3)$$

where $T = Z[(U + V)/(m + n - 2)]^{1/2}$ has a noncentral t distribution $t_{m+n-2}(\theta)$, with $m + n - 2$ degrees of freedom and noncentrality parameter θ, and $B \sim \text{beta}((m - 1)/2, (n - 1)/2)$. Since all tests based on $(\bar{X} - \bar{Y})$, S_1^2, and S_2^2 can also be obtained using T_1, S_1^2, and S_2^2, it is sufficient to show that given T_1, any common scale-invariant data-based power function constructed from S_1^2 and S_2^2 can also be obtained using T_2.

Observe that the distribution of $T_2 = 1/B + (s_2^2/s_1^2)/(1 - B)$ depends on the data only through s_1^2/s_2^2, a maximal invariant, and that the observed value of T_2 is a function of s_1^2/s_2^2. Theorem 2.1 implies that the data-based power function of any p-invariant test based on S_1^2 and S_2^2 can be obtained using T_2/s_1^2 or, equivalently, T_2. Therefore, p-invariant tests (location and common scale) based on T_1, S_1^2, and S_2^2 can be found using T_1 and T_2.

From (4.2) and (4.3), the generalized test variable $\tilde{W}$ given in (1.4) can be expressed as $\tilde{W} = s_1W$, where

$$W = (1/m + 1/(nh))^{-1/2}T_1T_2^{1/2}$$

$$= (m + n - 2)^{-1/2}TT_2^{1/2}. \quad (4.4)$$

Let $\beta(h) = (ms_1^2 + ns_2^2h)(m + nh)/mnhs_1^2$. Note that any p-invariant test can be generated using T_2 and W. The observed value of W is $w = (\bar{x} - \bar{y})/s_1$, and the observed value of T_2 is $\beta(h)$. Suppose we only consider p-unbiased tests based on generalized extreme regions of the form

$$C(\theta, h) = \{(T_2, W) : G(\beta(h), T_2, W, \theta, h) \geq 0\}, \quad (4.5)$$

where $G = G_w$ is a measurable and continuous function in all of its arguments. When the family of distributions of G is stochastically increasing in θ, we call these *C continuous* generalized extreme regions. These p-unbiased tests are also p-similar. Hence, the power function of any of these tests is constant for $\theta = 0$ and for all $h > 0$. Let $f_\theta(T_2, W)$ be the joint density of T_2 and W. When $\theta = 0$, $f_0(T_2, W)$ is free of unknown parameters and $f_0 > 0$ for all W and T_2. By p-similarity,

$$\int_{C(0,1)}\int f_0(T_2, W) = \int_{C(0,h)}\int f_0(T_2, W) \quad (4.6)$$

for all $h > 0$. Let $A = \{(T_2, W) : G(\beta(h), T_2, W, 0, h) \neq G(\beta(1), T_2, W, 0, 1)\}$. Then it follows from (4.6) that $\int\int_A f_0 = 0$ and in turn that A is a set of measure 0, because $f_0 > 0$. Hence, the continuity of G implies that $G(\beta(h), T_2, W, 0, h) = G(\beta(1), T_2, W, 0, 1)$ for all $h > 0$. In other words, p-unbiased tests based on C in (4.5) can be obtained using T_2, W, and w only. From (4.3), however, the conditional distribution of T_2 given $W = w$ depends on the data through w. Consequently, any p-unbiased test based on a continuous generalized extreme region using W and T_2 can be obtained using W alone. Hence the p value can be computed as $p = \Pr(W \geq w) = \Pr(\tilde{W} \geq \tilde{w})$.

5. CONCLUDING REMARKS

In this article, we show how in situations where nuisance parameters are present, significance testing based on p values using generalized extreme regions can produce nontrivial exact solutions in testing problems where nontrivial solutions using fixed-level testing may be difficult or impossible to obtain. In these situations, however, the *p* value as a function of the observed data does not necessarily have a uniform distribution (see Kempthorne and Folks 1971, p. 346). Thus our testing procedures often cannot be applied when fixed levels of significance are used. (The example in Sec. 3.2 is an exception.) For this reason, our testing procedure cannot be used to generate confidence sets either. Our focus is mainly on testing of hypotheses. Some results on the relationship of a family of tests using fixed levels of significance and confidence sets were given by Lehmann (1986, chap. 5). Finally, Cox (1977) provided a lengthy discussion on the difference between significance testing using *p* values and fixed-level testing.

[*Received October 1987. Revised November 1988.*]

REFERENCES

Casella, G., and Berger, R. (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Journal of the American Statistical Association,* 82, 106–111.

Cox, D. R. (1977), "The Role of Significance Tests," *Scandinavian Journal of Statistics,* 4, 49–62.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics,* London: Chapman & Hall.

Gibbons, J. D., and Pratt, J. W. (1975), "*P* Values: Interpretation and Methodology," *The American Statistician,* 29, 20–24.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford, U.K.: Oxford University Press.

Johnson, R. A., and Weerahandi, S. (1988), "A Bayesian Solution to the Multivariate Behrens–Fisher Problem," *Journal of the American Statistical Association,* 83, 145–149.

Kempthorne, O., and Folks, L. (1971), *Probability, Statistics and Data Analysis,* Ames: Iowa State University Press.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data,* New York: John Wiley.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: John Wiley.

Linnik, Y. V. (1968), *Statistical Problems With Nuisance Parameters,* New York: American Mathematical Society.

Salaevskii, O. V. (1963), "On the Non-existence of Regularly Varying Tests for the Behrens–Fisher Problem," *Akademija Nauk Ukrainskoi SSR, Doklady,* 151, 509–510.

Sukhatmé, P. V. (1938), "On Fisher and Behrens' Test of Significance for the Difference in Means of Two Normal Samples," *Sankhyā,* 4, 39–48.

Thompson, W. A., Jr. (1985), "Optimal Significance Procedures for Simple Hypotheses," *Biometrika,* 72, 230–232.

Weerahandi, S. (1987), "Testing Regression Equality With Unequal Variances," *Econometrica,* 55, 1211–1215.