Statistical Science
1999, Vol. 14, No. 4, 355–381

The Emperor's New Tests

Michael D. Perlman and Lang Wu

Abstract. In the past two decades, striking examples of allegedly inferior likelihood ratio tests (LRT) have appeared in the statistical literature. These examples, which arise in multiparameter hypothesis testing problems, have several common features. In each case the null hypothesis is composite, the size α LRT is not similar and hence biased, and competing size α tests can be constructed that are less biased, or even unbiased, and that dominate the LRT in the sense of being everywhere more powerful. It is therefore asserted that in these examples and, by implication, many other testing problems, the LR criterion produces "inferior," "deficient," "undesirable," or "flawed" statistical procedures.

This message, which appears to be proliferating, is wrong. In each example it is the allegedly superior test that is flawed, not the LRT. At worst, the "superior" tests provide unwarranted and inappropriate inferences and have been deemed scientifically unacceptable by applied statisticians. This reinforces the well-documented but oft-neglected fact that the Neyman–Pearson theory *desideratum* of a more (or most) powerful size α test may be scientifically inappropriate; the same is true for the criteria of unbiasedness and α-admissibility. Although the LR criterion is not infallible, we believe that it remains a generally reasonable first option for non-Bayesian parametric hypothesis-testing problems.

Key words and phrases: Hypothesis test, significance test, likelihood ratio test, power, size α test, unbiased test, α-admissibility, d-admissibility, order-restricted hypotheses, multiple endpoints in clinical trials, test for qualitative interactions, bioequivalence problem, multivariate one-sided alternatives, Fisher–Neyman debate.

1. A STATISTICAL ALLEGORY

In a distant land, a wise and benign Emperor ruled over his domain. Whenever decisions were to be made, data were gathered and the imperial statisticians applied long-accepted statistical procedures to reach reasonable conclusions. The Emperor, having read Fisher, Neyman, Pearson, Wald, Wilks and Bahadur, was particularly fond of the likelihood ratio test (LRT), for it seemed sensible, reliable and generally robust. His subjects benefitted from his wisdom and were happy and prosperous.

One day a bright young statistician, while seeking a Ph.D. dissertation topic that would hopefully win him a position in the imperial court, noticed that for some hypothesis-testing problems involving two or more parameters, the size α LRT was not similar on the boundary of the null hypothesis and therefore biased. Furthermore, being an extremely clever fellow, he was able to construct New Tests that were also size α, more nearly similar and less biased (or, in some cases, actually unbiased) and that dominated the LRT in the sense of being everywhere more powerful! This was accomplished in a deceptively simple manner: by carefully enlarging the rejection region of the LRT in such a way as to maintain the size at α. Once accomplished, this trivially increased the power at all parameter values.[1]

Excitedly, the young statistician wrote his dissertation, entitled "On the Inferiority of the Likelihood Ratio Criterion" and published several papers in the leading imperial statistical journals. As the Emperor read these papers his eyes opened wider

Michael D. Perlman is Professor, Department of Statistics, University of Washington, Seattle, Washington 98195-4322 (e-mail: michael@ms.washington.edu). Lang Wu is Postdoctoral Fellow, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (e-mail: lwu@hsph.harvard.edu).

and wider, and he thought "I must be a poor emperor indeed to have clung so long to the obviously inferior LRT." The young statistician was appointed to the Imperial Court as Creator of Better New Tests for Multi-Parameter Hypothesis Testing Problems. The Emperor was delighted that his enlightened support of basic research in mathematical statistics had led to such a marvelous advance in statistical science.

But one day, as the young Creator presented yet another example of an "inferior" LRT and a cleverly constructed unbiased and everywhere more powerful New Test, the Emperor became concerned. The testing problem under discussion was that of testing

$$(1) \qquad H_0\colon |\mu| \geq 1 \quad \text{versus} \quad H_1\colon |\mu| < 1$$

with σ^2 unknown, based on $X \sim N(\mu, \sigma^2)$ and $s^2 \sim \sigma^2\chi_n^2$ with X and s^2 independent, the sufficient statistics for a random sample from a univariate normal distribution.

"Do you mean to tell me," the Emperor said, "that if I observe $\hat{\mu} \equiv X = 0$ and $\hat{\sigma}^2 \equiv s^2/n = 10^{10}$, then your New Test regards this as sufficiently strong evidence to reject H_0 in favor of the alternative H_1? With such a large estimated variance, certainly no one can distinguish between $\mu = 1$ and $\mu = -1$, let alone between H_0 and H_1." "Yes, that is what my New Test declares," the young Creator replied. "I agree that this violates statistical intuition and that the LRT would not make this obviously unwarranted assertion, but I assure you that my New Test is indeed size 0.05, unbiased, and everywhere more powerful than the size 0.05 LRT!"

The Emperor pondered this for a while, then said "If that is the case, then your New Test is defective. Therefore the criteria of unbiasedness and more (or most) powerful size-α test must be inappropriate for this problem and probably for the other problems that you have treated as well." The Creator of Better New Tests was dismissed, the New Tests abandoned, and the LRT reinstated as the Test of First Resort for Non-Bayesian Testing Problems. The Emperor decreed that every statistics course should include a lecture on "The Emperor's New Tests: the Case for Common Sense in Statistics."

(We return to this example in Section 7.)

2. COMMON FEATURES OF THE NEW TESTS

This tale is not entirely fictional. In the past two decades, such New Tests have proliferated in the statistical literature. In Sections 4–8 we review a series of five examples of multiparameter hypothesis-testing problems with the following common features:

1. The null hypothesis H_0 is a noncompact composite set.
2. The "least favorable" null distribution is not attained in H_0 but is only approached asymptotically (cf. Sasabuchi, 1980; Robertson, Wright and Dykstra, 1988, Chapter 2; Berger, 1989).
3. The size α LRT is not similar on the boundary of H_0, hence is biased.
4. Competing size α tests can be constructed that are more nearly or exactly similar, hence less biased or unbiased, and that dominate the LRT in the sense of being everywhere[2] more powerful. These New Tests are constructed by carefully enlarging the rejection region to preserve the size at α (again, approached only asymptotically on H_0), which trivially implies their power dominance of the corresponding LRTs and therefore renders the LRTs *α-inadmissible* (cf. Lehmann, 1986, Section 6.7).
5. It is concluded that the LRT is inferior to the New Tests.

We assert that this conclusion is wrong. As some of their creators themselves admit, these New Tests often are of no practical value, for they may fail to properly assess the evidence provided by the data for or against the scientific hypotheses under investigation. This leads to the realization that the Neyman–Pearson (NP) theory *desiderata* of unbiasedness and more (or most) powerful size α test, as well as the related criterion of α-admissibility (which, on H_0, takes into account only the supremum of the power function, not its detailed behavior) may lead to undesirable statistical procedures, hence should not be regarded as sacrosanct for hypothesis testing problems. When these critera violate statistical common sense, it is they, not the LR criterion, that should be abandoned.

These issues are far from new, having been raised by Fisher in his celebrated debate with Neyman on hypothesis testing and since revisited many times; see Section 9.

The LR criterion, while not infallible (see Section 10), is a readily understood and generally useful tool for statistical inference. It should not be casually discarded, certainly not in a manner that obscures the proper role of statistics in scientific inquiry.[3]

Before introducing the likelihood ratio test (LRT) criterion and embarking upon its defense, we emphasize that in most if not all scientific applications, the question to be addressed by a statistical hypothesis test ($\equiv$ significance test) is the following: based on the observed data, does the family of distributions represented by the alternative hypothesis H_1 fit (or support, or explain) the observed data *significantly* better than the family represented by the null hypothesis H_0? Only if the fit is *significantly* better should we reject H_0 in favor of H_1. The tests discussed in this paper are evaluated on the basis of this (we believe) generally accepted criterion.[4]

3. THE GENERALIZED LIKELIHOOD RATIO TEST CRITERION

Consider a parametric statistical model specified by a family $\{p_\theta(x) \mid \theta \in \Theta\}$ of probability density functions, where x denotes the entire set of observations. The general hypothesis-testing problem has the following form: based on a random observation X, test

$$(2) \qquad H_0\colon \theta \in \Theta_0 \quad \text{versus} \quad H_1\colon \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are disjoint subsets of Θ. The (generalized) likelihood ratio test (LRT) rejects H_0 in favor of H_1 if $\Lambda(X) \geq c$, where

$$(3) \qquad \Lambda(x) := \frac{\sup_{\theta \in \Theta_0 \cup \Theta_1} p_\theta(x)}{\sup_{\theta \in \Theta_0} p_\theta(x)} \quad (\geq 1)$$

and $c \geq 1$. Because we require *significantly* better support for H_1 in order to reject H_0, in fact we must take $c > 1$, a small but important distinction; also see Solomon (1975). The generalized LRT was introduced by Neyman and Pearson (1928, 1933); its asymptotic properties, including asymptotic optimality, have been developed by Wilks (1938, 1962), Wald (1941a, b, 1943), Bahadur (1967) and many others.

In Sections 4–6 the statistical models are families of p-variate normal distributions with unknown mean vector μ and *known* covariance matrix equal to the identity matrix I_p ($p \geq 2$); that is, $X \sim N_p(\mu, I_p)$. For such models it is easy to show that the LRT statistic $\Lambda(X)$ is an increasing function of

$$(4) \qquad \|X - \Theta_0\|^2 - \|X - (\Theta_0 \cup \Theta_1)\|^2,$$

the difference of the squared Euclidean distances from X to Θ_0 and $\Theta_0 \cup \Theta_1$, respectively.

4. TESTING A ONE-SIDED OR ORDER-RESTRICTED ALTERNATIVE

We first examine the New Tests that have been proposed for the following bivariate one-sided testing problem; based on the bivariate normal random vector,

$$(5) \qquad X \equiv (X_1, X_2) \sim N_2(\mu \equiv (\mu_1, \mu_2), I_2),$$

where I_2 denotes the 2×2 identity matrix, test

$$(6) \qquad \begin{aligned} &H_0\colon \mu_1 \leq 0 \text{ or } \mu_2 \leq 0 \\ \text{versus} \quad &H_1\colon \mu_1 > 0, \mu_2 > 0. \end{aligned}$$

This problem, and its extensions to higher dimensions and/or to alternatives given by convex cones more general than the quadrant in H_1, occurs frequently in the applied literature; examples include the problems of *testing whether an identified treatment is better than several controls*, or of *testing a treatment in a clinical trial with multiple endpoints*. For further discussion, see Pocock, Geller and Tsiatis (1987), Laska and Meisner (1989), Tang, Geller and Pocock (1993), Tang (1998), Perlman and Wu (2000b).

By (4), for $0 < \alpha \leq 1/2$ the size α LRT for (6) rejects H_0 in favor of H_1 if

$$(7) \qquad \min(X_1^+, X_2^+) \geq z_\alpha,$$

where $x + = \max(x, 0)$ and z_α is the upper α quantile of the standard normal distribution. The rejection region (7) is labelled as R_1 in Figure 1. It is easy to see that the LRT is not similar on the boundary of H_0, hence is biased[5] for (6):

$$(8) \qquad \begin{aligned} &\sup_{\mu \in \partial H_0} P_\mu[\min(X_1^+, X_2^+) \geq z_\alpha] \\ &\quad = \lim_{\mu_1 \to \infty} P_{(\mu_1, 0)}[\min(X_1, X_2) \geq z_\alpha] = \alpha, \end{aligned}$$

$$(9) \qquad \begin{aligned} &\inf_{\mu \in \partial H_0} P_\mu[\min(X_1^+, X_2^+) \geq z_\alpha] \\ &\quad = P_{(0,0)}[\min(X_1, X_2) \geq z_\alpha] = \alpha^2. \end{aligned}$$

The bias is even more pronounced in p dimensions, where (9) becomes α^p, which approaches 0 as $p \to \infty$.

Berger (1989) and Liu and Berger (1995) deemed this bias a "deficiency" of the LRT.[6]

By modifying a construction of Lehmann (1952, pages 542 and 543) and Nomakuchi and Sakata (1987, page 492), Berger (1989) constructed two New Tests whose rejection regions properly contain the rejection region of the size α LRT (see Figure 1 for the forms of these enlarged rejection regions), yet which still have size α for (6). Trivially, the powers of these tests are strictly greater than that of the size α LRT for *every* $\mu \in H_0 \cup \mathrm{H}_1 \equiv \mathbf{R}^2$, yet they remain size α tests with smaller bias than the LRT. Liu and Berger (1995) and McDermott and Wang (2000) also construct New Tests with these properties. Liu and Berger (1995) assert, therefore, that "in a very general class of problems, the LRT

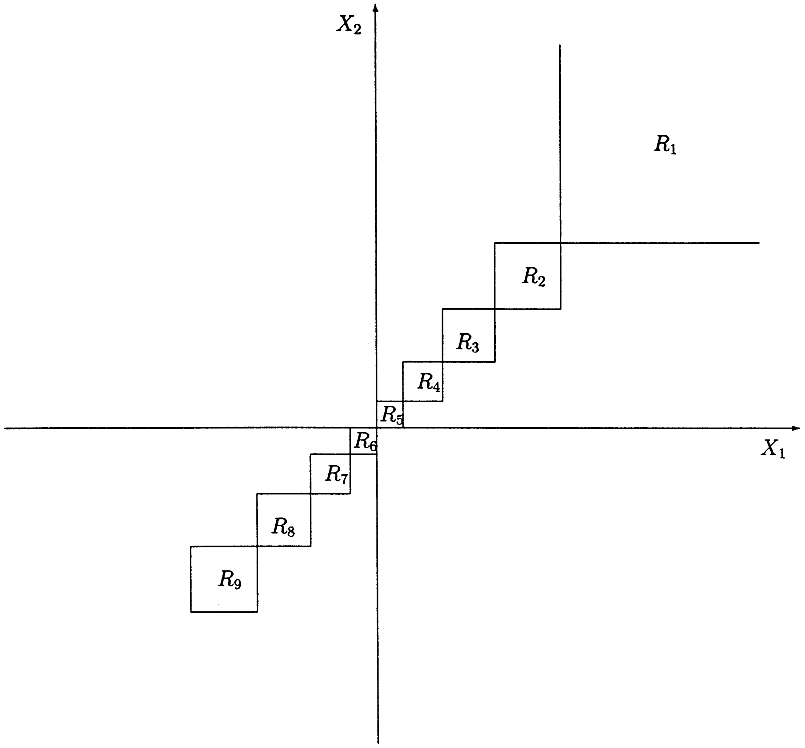

FIG. 1. *The rejection region for the LRT for testing problem* (6) *is* R_1. *The rejection regions for Berger's tests I and II are* $R_1 \cup R_2 \cup \cdots \cup R_5$ *and* $R_1 \cup R_2 \cup \cdots \cup R_9$, *respectively.*

can be uniformly dominated." Are we to abandon the LR criterion?

Of course, the quoted statement does not hold in the decision-theoretic sense, where "dominated" means that the risk function is inferior on *both* the null and alternative, not just the alternative.[7] Since risk = 1 − power for parameter points in the null hypothesis, the risk function of Liu and Berger's test is *larger* than that of the LRT everywhere on the null hypothesis. In fact, it is well known that the LRTs for problems such as (6) are *d-admissible*, that is, admissible in the usual decision-theoretic sense (cf. Cohen, Gatsonis and Marden, 1983; Nomakuchi and Sakata, 1987), so cannot be dominated when *both* Type I and Type II error probabilities are considered.

Nonetheless, the existence of size α tests that are more powerful than the size α LRT everywhere on the alternative has been cited frequently as a serious shortcoming of the LR criterion.[8] We believe that this criticism is invalid.

A glance at Figure 1 shows that the rejection region of Berger's Test I includes sample points (x_1, x_2) arbitrarily close to the origin, which is a member of the null hypothesis. That is, their test would interpret such sample points as providing significant evidence *against* the null hypothesis, yet this is clearly inappropriate. Since $(x_1, x_2) = \hat{\mu}$, the MLE of the true parameter value under $H_0 \cup H_1$, an observation (x_1, x_2) very close to the null hypothesis clearly *cannot* provide significant support for the alternative hypothesis.[9] The same criticism applies to the tests proposed by Liu and Berger (1995) and McDermott and Wang (2000).[10]

Because the rejection region of Berger's Test II is even larger than that of Test I, it is even more powerful and can be shown to be even less biased. However, this is achieved by including in the rejection region sample points for which the corresponding MLEs are both negative, hence actually inside the null hypothesis. Even more incredibly, the p-dimensional version of Test II would have us assert that all p population means are positive for certain outcomes for which all p sample means are negative!

Berger acknowledges that his Test II is "counterintuitive" and "may be primarily of theoretical interest." The logical conclusion should be, however, that the goal of constructing tests that are less biased and everywhere more powerful than the LRT is without intrinsic merit. We agree with Fisher that in scientific investigations, *the purpose of statistical hypothesis testing is to assess the evidence that the data provide about the hypotheses, not necessarily to optimize with respect to size, bias, and/or power* (see Section 9).

Berger (1989, Section 6) also constructed New Tests for the following related testing problem[11] with a "two-sided" alternative: based on X in (5), test

$$H_0\colon \mu \notin \mathcal{O} \cup -\mathcal{O} \quad \text{versus} \quad H_1\colon \mu \in \mathcal{O} \cup -\mathcal{O}, \tag{10}$$

where $\mathcal{O} \equiv \{\mu \mid \mu_1 > 0, \mu_2 > 0\}$ is the positive orthant in $\mathbf{R}^2$. Berger noted that this problem is equivalent to a bivariate case of the problem of *testing for qualitative interactions* (cf. Zelterman, 1990; Russek-Cohen and Simon, 1993). By carefully enlarging the rejection region of the size α LRT, both Berger and Zelterman again obtain size α tests whose power functions are everywhere greater than that of the LRT. Russek-Cohen and Simon (1993, page 467) state: "We believe that these tests have some nonintuitive properties and would not be accepted by nonstatisticians." We urge statisticians to share their skepticism.

Our assertion that the New Tests for (6) and (10) (as well as their multivariate generalizations) lead to inappropriate inferences requires some clarification. This assertion is valid *unless* one implicitly or explicitly adopts a restrictive prior distribution that assigns little or no mass to one or more (possibly large) open regions in the null hypothesis H_0. The corresponding Bayes test may assign sample points in or near these regions to its rejection region, and consequently its rejection and/or acceptance region need not be monotone, convex, or even connected. For example, the New Tests of Berger, Liu and Berger, and McDermott and Wang possibly may be (approximately) Bayes for prior distributions that, under the null hypothesis, assign no mass to some (possibly large) neighborhood of the origin $(0, 0)$. Such prior distributions, however, are unlikely to represent the views of the practitioners.

5. TESTING "OBLIQUE" ORDER-RESTRICTED HYPOTHESES

We next encounter New Tests in the problems considered by Warrack and Robertson (1984), Gutmann (1987), Menendez and Salvador (1991), Menendez, Rueda and Salvador (1992) and Mukerjee and Tu (1995, page 721), where both the null and alternative hypotheses are determined by order restrictions on the mean vector. The following simplified two-dimensional example illustrates the essential ideas: based on X as in (5), test

$$H_0\colon \mu \in C_0 \quad \text{versus} \quad H_1\colon \mu \in C_1 \setminus C_0, \tag{11}$$

where

$$\begin{aligned} C_0 &\equiv \{\mu \mid \mu_1 \le 0, \mu_2 \ge \mu_1\}, \\ C_1 &\equiv \{\mu \mid \mu_1 \ge 0, \mu_2 \ge \mu_1\}, \end{aligned} \tag{12}$$

respectively obtuse and acute, are the closed convex cones depicted in Figure 2. Note that

$$C_0 \cup C_1 = \{\mu \mid \mu_2 \ge \mu_1\}, \tag{13}$$

a closed half-space.

The rejection region of the size α LRT for (11) is the open region (cf. Figure 3)

$$R = \big\{x \equiv (x_1, x_2) \mid \|x - C_0\|^2 - \|x - (C_0 \cup C_1)\|^2 \ge z_\alpha^2\big\}. \tag{14}$$

This test is not similar on the boundary between C_0 and C_1; once again the least favorable distribution in H_0 is not attained but is only approached by $N_2((0, \mu_2), I_2)$ as $\mu_2 \to \infty$. Warrack and Robertson [WR] (1984) show that a less biased and more powerful size α test for (11) can be obtained by ignoring the information that $\mu_1 \ge \mu_2$ in both H_0 and H_1. Specifically, for testing (11) they propose the LRT for the simpler problem

$$H_0'\colon \mu_1 \le 0 \quad \text{versus} \quad H_1'\colon \mu_1 > 0. \tag{15}$$

The rejection region of the size α LRT for (15) is the open half-space (cf. Figure 3)

$$R \cup A \cup B = \{x \mid x_1 \ge z_\alpha\}. \tag{16}$$

When used as a test for (11), this test obviously remains size α and, because $R \cup A \cup B \supset R$, is less biased and everywhere more powerful than the LRT for (11) itself. WR (1984, page 882) call this a "failure" of the LR criterion and wonder whether the LR criterion generally "fails" in such problems, that is, in testing problems where the null and alternative hypotheses are "oblique" rather than orthogonal (cf. Figure 2).

As Mukerjee and Tu (1995, page 721) note, however, using the LRT for (15) as a test for (11) would create a "philosophical dilemma." Its rejection region $R \cup A \cup B$ contains all sample points of the form $\tilde{x}_b = (z_\alpha + \varepsilon, b)$ for $-\infty < b < 0$, with $\varepsilon > 0$

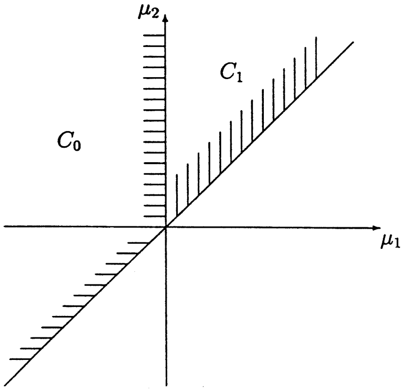

FIG. 2. *The closed convex cones C_0 and C_1 in* (12).

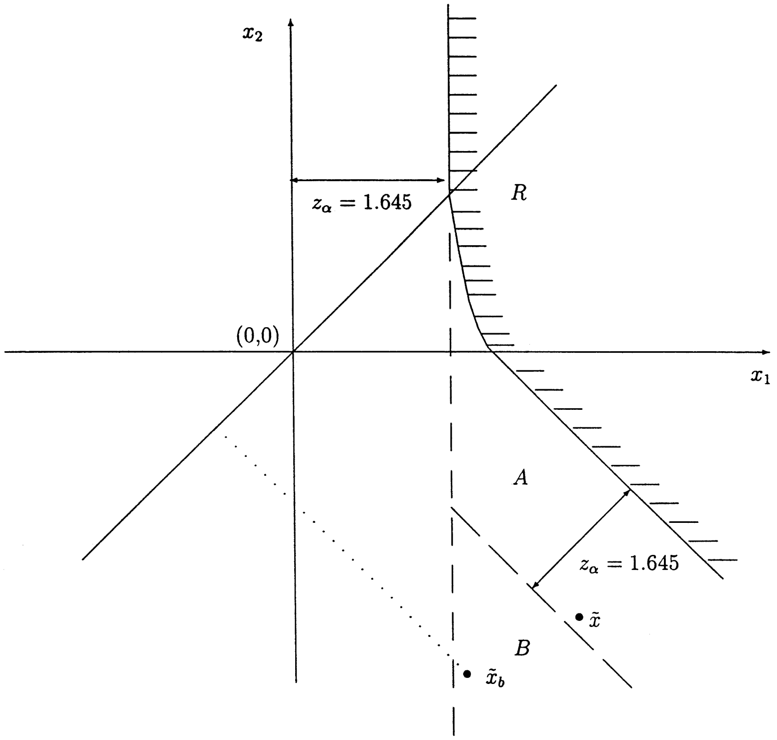

FIG. 3. *The rejection region R of the LRT for* (11). *The rejection regions of the competing tests of Warrack and Robertson* (1984) *and Mukerjee and Tu* (1995) *are $R \cup A \cup B$ and $R \cup A$, respectively* ($\alpha = 0.05$).

sufficiently small and fixed (cf. Figure 3). For all sufficiently large $|b|$, $\tilde{x}_b$ actually lies *closer* to C_0 than to C_1, yet WR's recommended test would *reject* H_0 in favor of H_1 for such sample points. Clearly, such points cannot be interpreted as providing significant evidence in favor of the alternative H_1 relative to H_0; in fact, the opposite is true. Yet WR's favored test is indeed less biased and everywhere more powerful than the original LRT for (11).

Mukerjee and Tu (1995, page 421) propose a modification of WR's test that eases, but does not resolve, their philosophical dilemma. The rejection region of their modified test is $R \cup A$ (cf. Figure 3), which does not contain sample points such as $\tilde{x}_b$ that lie closer to C_0 than to C_1. Like WR's test, their test is less biased and everywhere more powerful than the LRT for (11). Nevertheless, sample points such as $\tilde{x}$ in Figure 3 lie in this modified rejection region yet are almost equidistant from C_0 and C_1, hence still should not be interpreted as providing significant evidence in favor of H_1.

These "dilemmas" are again easy to resolve: we must abandon secondary criteria such as unbiasedness and α-admissibility if these criteria conflict with statistical common sense. The original LRT for (11), although more biased and less powerful on the alternative, does not "fail": it is *not* dominated by WR's test in the decision-theoretic sense and clearly makes more appropriate decisions.[12]

6. TESTING A UNION OF LINEAR SUBSPACES

Berger and Sinclair (1984) proposed a somewhat less unreasonable class of New Tests for testing problems of the following type: based on

$$X \equiv (X_1, X_2, X_3) \sim N_3(\mu \equiv (\mu_1, \mu_2, \mu_3), I_3), \tag{17}$$

test

$$H_0\colon \mu \in L_1 \cup L_{23} \quad \text{versus} \quad H_1\colon \mu \in \mathbf{R}^3 \setminus (L_1 \cup L_{23}), \tag{18}$$

where

$$L_1 \equiv \{\mu \mid \mu_2 = \mu_3 = 0\}, \quad L_{23} \equiv \{\mu \mid \mu_1 = 0\} \tag{19}$$

are orthogonal linear subspaces of dimensions 1 and 2, respectively. The size α LRT for (18) rejects H_0 iff

$$\|X - (L_1 \cup L_{23})\|^2 \equiv \min(X_1^2, X_2^2 + X_3^2) \geq \chi^2_{2,\alpha}, \tag{20}$$

where $\chi^2_{2,\alpha}$ is the upper α quantile of the χ^2_2 distribution. This test is not similar on H_0; the least favorable distribution in H_0 is not attained but is only approached by $N_3((\mu_1, 0, 0), I_3)$ as $\mu_1 \to \infty$.

Berger and Sinclair (1984) recommended the New Test that rejects H_0 iff

$$(21)\qquad \begin{aligned} &\min\left(\frac{\|X - L_1\|^2}{\chi^2_{2,\alpha}}, \frac{\|X - L_{23}\|^2}{\chi^2_{1,\alpha}}\right) \\ &\quad \equiv \min\left(\frac{X_2^2 + X_3^2}{\chi^2_{2,\alpha}}, \frac{X_1^2}{\chi^2_{1,\alpha}}\right) \geq 1. \end{aligned}$$

It is easy to see that this test is also size α and that its rejection region properly contains that of the size α LRT, hence is less biased and everywhere more powerful. Nonetheless, although it does not produce strikingly inappropriate decisions, the test (21) does *not* invalidate the LRT. As above, the risk function of (21) is *larger* than that of the LRT everywhere on H_0. Theorem 3.1 of Nomakuchi and Sakata (1987) can be extended to show that both tests are d-admissible and therefore can be approximated by Bayes tests. The LRT (20) can be approximated by a Bayes test for a normal prior distribution that assigns equal dispersions over L_1 and L_{23}, while the approximating prior for test (21) would assign a larger dispersion over L_{23}.

For example, when $\alpha = 0.05$, then $\chi^2_{1,0.05} = 3.89$ and $\chi^2_{2,0.05} = 5.99$. Thus the Berger–Sinclair test (21) would assign the same p-value 0.05 to the sample points $(\sqrt{3.89}, \sqrt{5.99}, 0)$ and $(\sqrt{5.99}, \sqrt{5.99}, 0)$, even though the former point is closer to H_0 than is the latter. By contrast, the LRT assigns the p-values 0.14 and 0.05, respectively, to these two points. In the absence of prior information assigning unequal dispersions over L_1 and L_{23}, the LRT yields the more appropriate inference.

7. THE BIOEQUIVALENCE PROBLEM

We now confront the New Tests that occur in the Emperor's testing problem (1), a special case of the so-called bioequivalence problem that has attracted the attention of both statisticians and biologists. The papers by Berger and Hsu (1996) and Brown, Hwang and Munk (1997) nicely review the applied and theoretical background. Wang (1997), Munk (1999) and Wang, Hwang and Dasgupta (1999) are recent papers on this topic.

The rejection region of the size α LRT for (1) is the open triangle R given by

$$(22)\qquad R = \{(X, s) \mid |X| + t_{n,\alpha} s/\sqrt{n} \leq 1\},$$

where $t_{n,\alpha}$ is the upper α quantile of the t_n distribution (cf. Figure 4).[13] It is straightforward to show that the LRT given by (22) is not similar on the boundary between H_0 and H_1, hence is biased for (1):

$$(23)\qquad \begin{aligned} &\sup_{(\mu,\sigma)\in\partial H_0} P_{(\mu,\sigma)}[(X, s) \in R] \\ &\quad = \lim_{\sigma\to 0} P_{(\pm 1,\sigma)}[(X, s) \in R] = \alpha, \end{aligned}$$

$$(24)\qquad \begin{aligned} &\inf_{(\mu,\sigma)\in\partial H_0} P_{(\mu,\sigma)}[(X, s) \in R] \\ &\quad = \lim_{\sigma\to \infty} P_{(\pm 1,\sigma)}[(X, s) \in R] = 0. \end{aligned}$$

After declaring that the LRT "suffers from a lack of power," Berger and Hsu (1996) construct a New Test that is also size α but less biased and everywhere more powerful than the LRT, improving upon earlier such tests proposed by Anderson and Hauck (1983), Patel and Gupta (1984), and Rocke (1984). Brown, Hwang and Munk (1997) offer a further refinement, constructing a New Test that is actually unbiased for (1); the form of its rejection region $R \cup Q$ is shown in Figure 4. Because $R \cup Q \supset R$, once again the New Test trivially dominates the LRT on H_1. The LR criterion appears to have failed yet again.

However, a glance at Figure 4 shows that the unbiased New Test achieves its dominance by extending the rejection region to include *arbitrarily large values of* s/n, the estimated standard deviation. As the Emperor and any student of Imperial Statistics 101 can see, for a fixed sample size, one cannot hope to declare a significant difference between any two specified values of the mean μ if the

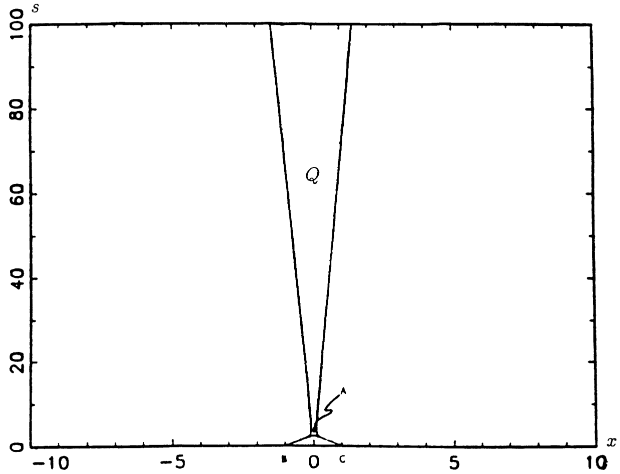

FIG. 4. *The triangle ABC is the rejection region R of the LRT (= TOST) for the bioequivalence testing problem* (1). *The rejection region for the unbiased, everywhere more powerful test of* Brown, *Hwang and Munk* (1997) *is* $R \cup Q$; *note that Q is unbounded in both x and s.*

standard deviation is unbounded. In fact, Theorem 5.2(b) of Hoeffding and Wolfowitz (1958) guarantees that no fixed-sample-size test can adequately distinguish between H_0 and H_1': $\mu = 0$, let alone between H_0 and H_1; sequential sampling schemes such as Stein's two-stage procedure (cf. Lehmann, 1986, pages 258 and 259) are inherently required.[14]

We are not the first to offer such criticism. In his discussion accompanying Berger and Hsu (1996), Schuirmann states: "In my personal opinion, this property [that the rejection region is unbounded in s] renders all similar or approximately similar tests of the hypothesis [(1)] unacceptable."[15] Brown, Hwang and Munk (1997, page 2348) acknowledge Schuirmann's criticism and a second as well. They write: "If s is quite large, [our proposed test] can thus leave the statistician in the embarrassing position of rejecting the null hypothesis that $[|\mu| \geq 1]$ while at the same time estimating a value $\hat{\mu} \equiv x$ for which $[|\hat{\mu}| > 1]$". Here, $\hat{\mu}$ is the MLE under $H_0 \cup H_1$. They respond by proposing several modifications of their test where Q, the addition to the rejection region, is truncated either horizontally or vertically or both.

The need for such mathematical acrobatics reinforces our contention that the quest for (nearly) unbiased tests more powerful than the LRT is misguided. As Schuirmann implies, the LRT ≡ TOST already makes perfectly appropriate inferences: if s and/or $|x|$ is too large, it correctly declares that the evidence is insufficient to reject H_0 in favor of H_1. If this is deemed unsatisfactory, then the solution is very simple: more observations are needed, not Better New Tests.[16]

8. COMPLETELY UNKNOWN COVARIANCE MATRIX: *Nostra Culpa*

We have vigorously criticized those who espouse the notion that the LRT is necessarily rendered inferior if a New Test of the same size, smaller bias, and everywhere greater power can be constructed. We now confess that we have, very recently, committed the same sin ourselves.

Perlman (1969) derived the LRT for a testing problem with a multivariate one-sided alternative and *completely unknown* covariance matrix. This problem takes the following canonical form: based on $X \sim N_p(\mu, \Sigma)$ and $S \sim W_p(n; \Sigma)$, respectively the p-variate normal distribution and Wishart distribution with n degrees of freedom ($n \geq p$), with X and S independent and μ and Σ both unknown, test

$$(25)\quad H_0\colon \mu = 0 \quad \text{versus} \quad H_1(C)\colon \mu \in C \setminus \{0\},$$

where $C \subset \mathbf{R}^p$ is a pointed[17] closed convex cone of full dimension p. Because Σ is unknown, (25) is more accurately stated as

$$(26)\quad \begin{aligned} &H_0\colon (\mu, \Sigma) \in \{0\} \times \mathscr{P} \quad \text{versus} \\ &H_1(C)\colon (\mu, \Sigma) \in (C \setminus \{0\}) \times \mathscr{P}, \end{aligned}$$

where $\mathscr{P}$ is the set of all $p \times p$ positive definite matrices, so H_0 is in fact composite.

Perlman (1969) showed that the size α LRT for (25) ≡ (26) rejects H_0 if

$$(27)\quad U(C) \equiv \frac{\|\pi_S(X; C)\|_S^2}{1 + \|X - \pi_S(X; C)\|_S^2} \geq c_\alpha,$$

where $\pi_S(X; C)$, the projection of X onto C with respect to the norm $\|x\|_S^2 \equiv xS^{-1}x'$, is the MLE $\hat{\mu}$ of μ under $H_0 \cup H_1(C)$, where $c_\alpha \equiv c_{\alpha, p, n}$ satisfies

$$(28)\quad \begin{aligned} \alpha = {} & \frac{1}{2} P\left[\frac{\chi_{p-1}^2}{\chi_{n-p+1}^2} \geq c_\alpha\right] \\ & + \frac{1}{2} P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} \geq c_\alpha\right] \end{aligned}$$

and where the chi-square variates are independent. Note that c_α does *not* depend on the cone C.

Perlman (1969) noted that the LRT is not similar on H_0; in fact,

$$(29)\quad \begin{aligned} & \inf_{\Sigma > 0} P_{0, \Sigma}[U(C) \geq c_\alpha] \\ & \quad = \frac{1}{2} P\left[\frac{\chi_1^2}{\chi_{n-p+1}^2} \geq c_\alpha\right], \end{aligned}$$

hence the LRT is biased for (25) and this bias becomes substantial as p increases (cf. Table 1 of Perlman and Wu, 2000a). For this reason, the LRT has been considered inadequate for (25) (cf. Robertson, Wright and Dykstra, 1988, page 223) and for other order-restricted testing problems when Σ is completely unknown (e.g., Laska, Tang and Meisner, 1992). The recent papers of Tang (1994), Wang and McDermott [WM] (1998a) and, we confess, Perlman and Wu [PW] (2000a) are predicated on this viewpoint.

Tang (1994) and WM (1998a) have constructed New Tests for (25) that are actually similar size α, (apparently)[18] unbiased, and everywhere more powerful than the size α LRT. Tang's test for (25) is just the LRT for testing

$$(30)\quad H_0\colon \mu = 0 \quad \text{versus} \quad H_1(D)\colon \mu \in D \setminus \{0\}$$

with Σ unknown, where D is any half-space in $\mathbf{R}^p$ that properly contains C. The size α LRT for (30) rejects $H_1(D)$ [and hence rejects $H_1(C)$] if $U(D) > c_\alpha$, where $U(D)$ is given by (27) with C replaced by D and where c_α is again given by (28). Because $\Sigma^{-1/2}D$ is again a half-space for every Σ, Tang's test is similar size α on H_0. Furthermore, because $U(D) \geq U(C)$ (since $D \supset C$) with strict inequality holding when $X \in D \setminus C$, the rejection region of Tang's test properly contains the rejection region of the LRT, hence trivially is everywhere more powerful than the LRT.

As in our previous examples, however, Tang's test does *not* dominate the LRT in the decision-theoretic sense. Furthermore, WM (1998a) point out that Tang's test can reject H_0 for certain sample points (x, s) such that the MLE $\hat{\mu} \equiv \pi_s(x; C) = 0$, clearly an inappropriate decision.

WM (1998a) propose a New Test that is also similar but does not share this deficiency. Their "conditional LRT"[19] for (25) rejects H_0 if $U(C) > c_\alpha(V)$, where $V \equiv XX' + S$ and where $c_\alpha(\cdot) \equiv c_{\alpha, p, n}(\cdot)$ satisfies

$$P_{0,\Sigma}[U(C) \geq c_\alpha(v) \mid V = v] = \alpha, \tag{31}$$

so their test has exact conditional size α, hence is an unconditionally similar size α for H_0.

Because V is a complete and sufficient statistic for Σ under H_0 and because Tang's test is similar for H_0, WM note that Tang's test must have Neyman structure (Lehmann, 1986, Theorem 4.2); that is,

$$P_{0,\Sigma}[U(D) \geq c_\alpha \mid V = v] = \alpha. \tag{32}$$

Because $U(D) \geq U(C)$ with strict inequality holding when $X \in D \setminus C$, it follows from (31) and (32) that $c_\alpha(v) < c_\alpha$ for almost every v, hence the rejection region of WM's test again properly contains that of the unconditional LRT and is therefore everywhere more powerful. Furthermore, because $U(C) = 0$ when the MLE $\hat{\mu} \equiv \pi_S(X; C) = 0$, WM's test does *not* reject H_0 for such sample points.[20]

Although WM's test does not have the same deficiency as Tang's test, once again it does not dominate the unconditional LRT in the decision-theoretic sense and, we assert, achieves its increased power at the expense of making possibly inappropriate decisions. To demonstrate this, we consider the special case actually treated by WM (1998a), where $C = \mathscr{O}$, the nonnegative orthant in $\mathbf{R}^p$. In this case (31) becomes

$$\begin{aligned} \alpha &= P_{0,\Sigma}[U(\mathscr{O}) \geq c_\alpha(v) \mid V = v] \\ &= \sum_{k=1}^{p} P\left[\frac{\chi_k^2}{\chi_{n-p+1}^2} \geq c_\alpha(v)\right] \\ &\qquad \times P_{0,\Sigma}[K = k \mid V = v] \\ &= \sum_{k=1}^{p} P\left[\frac{\chi_k^2}{\chi_{n-p+1}^2} \geq c_\alpha(v)\right] P_{0,\Sigma=v}[K' = k], \end{aligned} \tag{33}$$

where $K \equiv K(X, S)$ and $K' \equiv K'(X, \Sigma)$ denote the number of positive components of $\pi_S(X; \mathscr{O})$ and $\pi_\Sigma(X; \mathscr{O})$, respectively. The third equality follows from Proposition 2.1 of WM (1998b). Whereas WM's test is based on the conditional distribution of $U(\mathscr{O})$ given V, if we introduce a finer conditioning based on both V and $K \equiv K(X, S)$, the conditional distribution of $U(\mathscr{O})$ can change considerably, and, conditionally, WM's test can become severely anticonservative.

This effect is strongest when $K = p$, in which case $U(\mathscr{O}) = \|X\|_S^2$. By arguments similar to those used in the proof of Proposition 2.1 of WM (1998b), it can be shown that $\|X\|_S^2$ is independent of $(V, \{K = p\})$ when $\mu = 0$, hence

$$\begin{aligned} &P_{0,\Sigma}[U(\mathscr{O}) \geq c_\alpha(v) \mid V = v, K = p] \\ &\quad = P\left[\|X\|_S^2 \geq c_\alpha(v)\right] \\ &\quad = P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} \geq c_\alpha(v)\right]. \end{aligned} \tag{34}$$

It is shown in the Appendix that (34) is nearly 1, rather than the nominal value α, if p is moderately large and v has the form

$$v_{\beta,\gamma} = \beta(I_p - \gamma e_p' e_p), \tag{35}$$

where $e_p = (1/\sqrt{p})(1, \ldots, 1)$ is a unit vector along the central ray in $\mathscr{O}$, $\beta > 0$ is fixed, and $0 < \gamma < 1$ with $\gamma \uparrow 1$. For such sample points, therefore, WM's test is indeed severely anticonservative.

For example, it follows from (55) that when $\alpha = 0.05$, $p = 10$ and $n - p + 1 = 10$, (34) is nearly 0.95 for such sample points, while for $p = 10$ and $n - p + 1 = 60$ it exceeds 0.999.

By contrast, for the unconditional LRT, the conditional probability corresponding to (34) is, for $v = v_{\beta,\gamma}$,

$$\begin{aligned} &P_{0,\Sigma}\left[U(\mathscr{O}) \geq c_\alpha \mid V = v_{\beta,\gamma}, K = p\right] \\ &\quad = P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} \geq c_\alpha\right], \end{aligned} \tag{36}$$

which is much closer to the nominal value α. In fact, (28) yields the bounds

$$\begin{aligned} \alpha &< P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} \geq c_\alpha\right] \\ &< P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} \geq c_{p-1,\alpha}\right], \end{aligned} \tag{37}$$

where, for $k = 1, \ldots, p$, $c_{k,\alpha} \equiv c_{k,n-p,\alpha}$ satisfies

$$\alpha = P\left[\frac{\chi_k^2}{\chi_{n-p+1}^2} \ge c_{k,\alpha}\right]. \tag{38}$$

For example, when $\alpha = 0.05$, $p = 10$ and $n - p + 1 = 10$, the right-hand probability in (36) is 0.06, while for $p = 10$ and $n - p + 1 = 60$ it is 0.05.

For the special case $C = \mathcal{O}$ considered by WM (1998a), PW (2000a) also construct a competing test that is less biased than the LRT for (25) and (based on simulations) also more powerful in most (but not all) regions of the alternative. This test, which is based on the conditional distribution of the LRT statistic $U(\mathcal{O})$ given K, rejects H_0 if $U(\mathcal{O}) \ge c_{K,\alpha}$, where $c_{K,\alpha}$ is given by (38). For $k = 1, \ldots, p$, PW show that this test is conditionally similar size α given $K = k$, but, since it never rejects H_0 when $K = 0$, its conditional size is 0 when $k = 0$. Thus PW's test is *not* unconditionally a similar size α on H_0, but they advocate it nevertheless on the grounds that it is much more nearly similar than the unconditional LRT.

Unlike the New Tests of Tang (1994) and WM (1998a), the rejection region of PW's test does not strictly contain that of the LRT but nearly does so, since

$$c_{1,\alpha} < \cdots < c_{p-1,\alpha} < c_\alpha < c_{p,\alpha} \tag{39}$$

by (37). Thus, the conditional rejection region of PW's test given $K = k$ properly contains that of the unconditional LRT for $k = 1, \ldots, p - 1$ but is a proper subset for $k = p$. In fact, for every v,

$$\begin{aligned} &P_{0,\Sigma}\left[U(\mathcal{O}) \ge c_{p,\alpha} \mid V = v, K = p\right] \\ &\quad = P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} \ge c_{p,\alpha}\right] = \alpha, \end{aligned} \tag{40}$$

so PW's test does not exhibit the strongly anticonservative behavior noted for WM's test.

Nonetheless, fairness compels us to apply the same criticism to PW's test that we have applied to the other competitors to the LRT described in this section. That is, PW's test does not dominate the LRT in the decision-theoretic sense, and the mere fact of being less biased and often more powerful does not necessarily render it superior to the LRT in the sense of making appropriate inferences. We now believe that the unconditional LRT remains the preferred test for (25) ≡ (26).

Little is known about the d-admissibility of the tests considered in this section. PW's test is probably inadmissible due to the discontinuous nature of the boundary of its acceptance and rejection regions. See Section 10 for further comments.

9. THE FISHER–NEYMAN DEBATE ON HYPOTHESIS TESTING

The issues raised in this paper echo the celebrated Fisher–Neyman debate concerning the proper formulation of statistical hypothesis testing (cf. Lehmann, 1993; Royall, 1997) and, we believe, strongly support Fisher's position.

Three fundamental and generally accepted tenets of the Neyman–Pearson (NP) testing theory are its emphasis upon:

(a) Explicit formulation of *both* the null and alternative hypotheses.

(b) The role of the power function for evaluating tests *as decision rules*.

(c) Use of the LR for constructing most powerful size α tests for simple hypotheses and intuitively reasonable tests for composite hypotheses.

It is not surprising, therefore, that the NP school came to focus its efforts on the quest for most powerful tests size α tests and, when this proved unattainable for composite hypotheses, introduced secondary criteria such as unbiasedness in order to obtain optimal tests within restricted classes.

Fisher, however, viewed hypothesis testing, or significance testing, as a means of interpreting data as evidence concerning a scientific hypothesis and regarded such NP notions as a fixed significance level, power and unbiasedness as "merely mathematical consideration[s]" (cf. Lehmann, 1993, pages 1244 and 1245). Dawid (1991, page 80) writes

> For Fisher, inference involves the subtle teasing from the data at hand of the information that they contain . . . and any suggested method of inference is to be judged on how well it succeeds in extracting their secret. Fisher dismissed Neyman's preoccupation with the behaviour of inference rules in repeated sampling as being founded on a misguided analogy with 'acceptance procedures,' such as those used in industry to control the quality of incoming material, and as having no relevance to the task of advancing scientific understanding.

Views similar to Fisher's have been frequently expressed:

"But this book, by its very excellence, its thoroughness, lucidity, and precision, intensifies my growing feeling that nevertheless [NP] theory is arbitrary, be it however 'objective,' and the problems it solves, however precisely it may solve them, are not even simplified theoretical counterparts of the real problems to which it is applied." (Pratt, 1961, page 164, reviewing the first edition of Lehmann's *Testing Statistical Hypotheses*.)

"There is no statistical sense to significance levels." (Rubin, in Cornfield, 1969, page 655)

"The difficulty is that the solution to this problem [finding the best rejection region of size α] has no relevance per se to the problems of applied statistics ..." (Kempthorne, in Kiefer, 1977, page 817)

"I think that the attacks of the past 20 years (or actually 40 years, since R. A. Fisher should be included) on [the NP paradigm] have been largely successful." (Dempster, in Kiefer, 1977, page 815)

"... the familiar optimality criteria of statistics are in fact in conflict with scientific principles ..." (Fraser and Reid, in Brown, 1990, page 503)

"[NP theory] does not address the problem of representing and interpreting statistical evidence, and the decision rules derived from NP theory are not appropriate tools for interpreting data as evidence." (Royall, 1997, page 58)

"This points to the difference between statistics as an effort to learn, to get at the truth, and decision theory—a difference that was emphasized by Fisher in some of his disputes with Neyman." (Lehmann, 1998, after noting the appropriateness of the NP formulation in a hypothetical commercial application.)

Such views are often supported by simple one-parameter examples showing that a most powerful size α test may be inappropriate even for the elementary problem of testing a simple hypothesis versus a simple alternative. We briefly review two familiar examples.

EXAMPLE 1.[21] Based on a single observation X, consider the problem of testing

$$\begin{aligned} &H_0\colon X \sim \text{Uniform}[0,1] \\ \text{versus}\quad &H_1\colon X \sim \text{Uniform}[0.99,1]. \end{aligned} \tag{41}$$

The test T with rejection region $[0.95, 1]$ is a most powerful size 0.05 test (its power is 1), but inappropriately interprets an observation $X \in [0.95, 0.99)$ as evidence supporting H_1 over H_0. Furthermore, although its rejection region strictly contains the rejection region $[0.99, 1]$ of the size 0.01 LRT T^* and hence, like the New Tests, is more (or at least as) powerful on both H_1 and H_0, T is d-inadmissible: it is dominated by T^*, which also has power 1. Thus the most powerful size 0.05 test T fails *both* the Fisherian requirement to appropriately interpret the data as evidence for or against the hypotheses under consideration and the decision-theoretic requirement of admissibility.

EXAMPLE 2. *A composite alternative*. Based on one observation $X \sim \text{Uniform}[\theta, \theta+1]$, consider the problem of testing

$$H_0\colon \theta = 0 \quad \text{versus} \quad H_1\colon \theta > 0. \tag{42}$$

The test T with rejection region $[0.95, \infty]$ is the uniformly most powerful size 0.05 test and is unbiased and d-admissible, but wrongly interprets an observation $x \in [0.95, 1]$ as supporting the alternative H_1 more strongly than H_0. This interpretation is inappropriate because the LR $p_\theta(x)/p_0(x)$ for such an observation never exceeds 1 for any $\theta > 0$. Like the New Tests, the rejection region of T strictly contains the rejection region $(1, \infty]$ of the LRT T^* hence is everywhere more (or at least as) powerful than T^*, yet T^* (also d-admissible) does not make inappropriate inferences.

The failures of the New Tests in the multiparameter examples of Sections 4–8 demonstrate even more dramatically that the NP criteria of unbiasedness and more (or most) powerful size α test can lead to scientifically inappropriate statistical procedures.

10. SOME INADMISSIBLE LIKELIHOOD RATIO TESTS

We do know of a natural hypothesis-testing problem for normal means where the LRT is actually d-inadmissible. This problem has the following canonical form: based on $X \sim N_p(\mu, \Sigma)$ and $S \sim W_p(n; \Sigma)$ as in Section 8, test

$$H_0\colon \mu \in L_0 \quad \text{versus} \quad H_1\colon \mu \in L_1 \setminus L_0, \tag{43}$$

where $\{0\} \subseteq L_0 \subset L_1$ are linear subspaces of $\mathbf{R}^p$ and where Σ is completely unknown. The size α LRT for (43) rejects H_0 if

$$\begin{aligned} &U(L_0; L_1) \\ &\quad \equiv \frac{\|\pi_S(X; L_1) - \pi_S(X; L_0)\|_S^2}{1 + \|X - \pi_S(X; L_1)\|_S^2} \ge c_\alpha. \end{aligned} \tag{44}$$

The LRT is similar, unbiased, and conditionally most powerful and admissible given the ancillary statistic $\|X - \pi_S(X; L_1)\|_S^2$. Marden and Perlman (1980) showed, however, that for the usual significance levels, the LRT is *unconditionally d-inadmissible* for this testing problem. (The proof is nonconstructive; no test that dominates the LRT is known.)

Brown (1990, page 489), citing Fisher, Savage, and Cox, notes, "It is widely held that statistical inference should be carried out conditional on the value of any ancillary statistic." Gleser (cf. Brown, 1990, page 508) writes that "except for some very specialized applications, unconditional [d-]admissibility is generally only of interest to a statistician seeking to do well in many similar problems, but not to users of the inference ... presented by the statistician in any particular problem." (See Gleser's convincing baseball example on page 512.) Thus it is arguable that the LRT (44) is appropriate despite its unconditional [d-]inadmissibility.

Because of the similarity of the LRT statistics $U(C)$ and $U(\mathscr{O})$ in Section 8 to $U(L_0; L_1)$ here (consider especially the case $L_0 = \{0\}$), we conjecture that each of the tests discussed in Section 8 (the LRT and the tests of Tang, WM, and PW) will be unconditionally d-inadmissible for (25). Again, however, a case can be made for the LRT.

A well-known example[22] due to Stein can be used to exhibit a d-inadmissible LRT for normal covariance matrices. Based on the independent bivariate normal observations

$$\begin{aligned} &X \equiv (X_1, X_2) \sim N_2(0, \Sigma) \\ \text{and} \quad &Y \equiv (Y_1, Y_2) \sim N_2(0, \delta\Sigma), \end{aligned} \tag{45}$$

where Σ is an unknown positive definite covariance matrix and δ is an unknown scalar, test

$$H_0\colon \delta = 1 \quad \text{versus} \quad H_1\colon \delta > 1. \tag{46}$$

It is straightforward to show that the LRT statistic is identically 1, so the size α LRT is the trivial randomized test that ignores X, Y and accepts H_0 with probability $1 - \alpha$ and rejects H_0 with probability α. But $Y_1^2/X_1^2 \sim \delta F_{1,1}$, so the test that rejects H_0 if $Y_1^2/X_1^2 \geq F_{1,1;1-\alpha}$ has size α and power strictly greater than α, hence dominates the LRT.

Another well-known example[23] due to Stein shows that the LRT may be not only d-inadmissible, but actually "worse than useless" in the sense that its power may be strictly smaller than that of the trivial randomized test with the same size. In a simplified version of this example, a single observation X with range $\{0, 1, 2\}$ and corresponding probabilities p_0, p_1, p_2 is obtained, and it is desired to test H_0 versus $H \equiv H_1 \cup H_2$, where

$$\begin{aligned} &H_0\colon p_0 = 0.50, p_1 = 0.25, p_2 = 0.25, \\ &H_1\colon p_0 = 0.60, p_1 = 0.40, p_2 = 0.00, \\ &H_2\colon p_0 = 0.60, p_1 = 0.00, p_2 = 0.40. \end{aligned} \tag{47}$$

The size $\alpha = 0.50$ LRT, which rejects H_0 in favor of H iff $X = 1$ or $X = 2$, has power $0.40 < 0.50$, hence is d-inadmissible and "worse than useless." In fact, the "reversed LRT," which rejects H_0 in favor of H iff $X = 0$, also has size alpha 0.50 but power $0.60 > 0.50$.[24]

Aitken (1991, page 140), responding to a similar example of Goldstein, remarks that the "worse than useless" behavior of the LRTs in such examples disappears if two or more observations are taken rather than one and notes a similarity to the well-known difficulty encountered by likelihood methods when attempting inference in a two-parameter family with only one observation.

The examples noted in this section, not the New Tests examined above, show that the LRT is not universally satisfactory for hypothesis-testing problems. It would be of interest to characterize those problems where the LRT is or is not successful. We believe that the LR criterion remains a generally reasonable first option for non-Bayesian parametric hypothesis-testing problems.

11. A CAUTIONARY NOTE

In their criticism of the LRT and advocacy of the New Tests for the bioequivalence problem, Berger and Hsu (1996, page 292) make the following statement: "We believe that notions of size, power, and unbiasedness are more fundamental than 'intuition' ..." In our opinion, such a statement places the credibility of statistical science at serious risk within the scientific community. If we are indeed teaching our students to disregard intuition in scientific inquiry, then a fundamental reassessment of the mission of mathematical statistics is urgently needed.

APPENDIX: VERIFICATION OF THE ASSERTION FOLLOWING (34)

As $\gamma \uparrow 1$,

$$\beta^{-1} v_{\beta,\gamma} \to I_p - e_p' e_p, \tag{48}$$

a symmetric idempotent ($\equiv$ projection) matrix, hence also

$$\beta^{-1/2} v_{\beta,\gamma}^{1/2} \to I_p - e_p' e_p. \tag{49}$$

It follows that there exist positive constants $b_{\beta,\gamma} \equiv b_{\beta,\gamma,p}$ such that as $\gamma \uparrow 1$,

$$b_{\beta,\gamma} v_{\beta,\gamma}^{-1/2} \to e_p' e_p, \tag{50}$$

the projection matrix onto the one-dimensional subspace spanned by e_p. Thus, as $\gamma \uparrow 1$,

$$v_{\beta,\gamma}^{-1/2} \mathscr{O} \to \{\delta e_p \mid 0 \leq \delta < \infty\}, \tag{51}$$

the ray through e_p, hence

$$\begin{aligned} &P_{0,\Sigma = v_{\beta,\gamma}}[K' = k] \\ &\quad \to \begin{cases} 1/2, & \text{if } k = 0, 1, \\ 0, & \text{if } k = 2, \ldots, p. \end{cases} \end{aligned} \tag{52}$$

It follows from (33) that

$$c_\alpha(v_{\beta,\gamma}) \to c_\alpha' \text{ as } \gamma \uparrow 1, \tag{53}$$

where $c_\alpha' \equiv c_{\alpha,n-p}'$ satisfies

$$\alpha = \frac{1}{2} P\left[\frac{\chi_1^2}{\chi_{n-p+1}^2} > c_\alpha'\right]. \tag{54}$$

By (34), therefore, for fixed $\beta > 0$ and $\gamma \equiv \gamma(\beta)$ sufficiently near 1,

$$
\begin{aligned}
&P_{0,\Sigma}\left[U(\mathscr{O}) > c_\alpha(v_{\beta,\gamma}) \mid V = v_{\beta,\gamma}, K = p\right] \\
(55)\qquad &\approx P\left[\frac{\chi_p^2}{\chi_{n-p+1}^2} > c_\alpha'\right].
\end{aligned}
$$

By (54), however, the right-hand expression in (55) approaches 1 rather than α as p increases (provided that n also increases so that $n - p + 1 > 0$), as asserted after (34).

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation. We warmly thank Alan Agresti, Steen Andersson, Roger Berger, Julian Besag, Larry Brown, James Calvin, Brent Coull, Arthur Cohen, David Cox, Persi Diaconis, David Draper, Morris Eaton, Brad Efron, Joe Felsenstein, Tom Ferguson, Clark Glymour, Peter Guttorp, Lucien Le Cam, Erich Lehmann, John Marden, Michael McDermott, Hari Mukerjee, Jaxk Reeves, Thomas Richardson, Richard Royall, Harold Sackrowitz, Fritz Scholz, Julie Shaffer, Mervyn Silvapulle, Muni Srivastava, Charles Stein, Steve Stigler, Dei-In Tang, Elizabeth Thompson, Ming-Tien Tsai, Yining Wang and the Editor and referees for their interest and helpful comments.

NOTES

1. "This does also increase the power everywhere on the null hypothesis," he thought, "thereby increasing the probability of a Type I error, but as long as the size is maintained at α, my new procedures are valid competitors to the size α LRT."

2. Including H_0, where the power is the probability of Type I error; hence these tests do *not* dominate the LRT in the decision-theoretic sense of having everywhere smaller risk function.

3. The LR criterion has recently been criticized for allegedly anomalous behavior of a different sort in parametric testing problems with one-sided or order-restricted alternatives (cf. Cohen and Sackrowitz, 1998 and Cohen, Kemperman and Sackrowitz, 1997). Perlman and Wu (2000b) assert that this criticism is also unwarranted.

4. See, for example, Hacking (1965), Edwards (1972), Dempster (1997, page 250) and Royall (1997), especially their discussions of the interpretation of data as *evidence* and *support* for statistical hypotheses.

5. It should be noted, however, that *no* nontrivial unbiased tests exist for problems of this type (cf. Lehmann, 1952, Section 3; Nomakuchi and Sakata, 1987, Section 2).

6. We disagree. By (9), the least favorable distributions in H_0 are not attained but are approached by $N_2(\mu, I_2)$ with $\mu = (\mu_1, 0)$ or $(0, \mu_2)$ as $\mu_1 \to \infty$ or $\mu_2 \to \infty$ (cf. Sasabuchi, 1980). These are the distributions in H_0 that are the most difficult to distinguish from H_1. Because $\mu = (0, 0)$ is easier to distinguish from H_1, it is entirely appropriate that the power of the LRT, that is, the probability of a Type I error, be less than α at $\mu = (0, 0)$. This reasoning is incontrovertible in the classical univariate problem of testing a one-sided hypothesis about a normal mean ν with known variance: H_0: $\nu \le 0$ versus H_1: $\nu > 0$. Here $\nu = 0$ determines the least favorable distribution in H_0 and the power of the LRT is less than α whenever $\nu < 0$.

7. Berger (1989) uses "dominated" in the decision-theoretic sense in the first column on page 193, noting that the LRT is not dominated, but reverts to the quoted usage in the second column on page 193.

8. This renders the LRT α-inadmissible (Lehmann, 1986, Section 6.7), a serious-sounding indictment. However, our defense of the LRT suggests that the guilt lies with the criterion of α-admissibility itself.

9. This is incontrovertible in the classical one-dimensional testing problem in Note 6. Exactly the same reasoning applies in the multivariate problems discussed here.

10. Furthermore, each of these allegedly superior tests has acceptance and rejection regions that are neither convex, monotone, nor, in some cases, connected. Laska and Meisner (1989, pages 1140 and 1141) make a convincing case that the acceptance and rejection regions be monotone. Note, however, that their argument is implicitly predicated on the assumption that, in Bayesian terminology, substantial prior mass is assigned to the vicinity of the origin $(0, 0)$ within the null hypothesis H_0. If this is not the case, then the corresponding optimal (Bayes) test need not be monotone.

11. Berger's results for (10) also apply to the multivariate extension of the bivariate problem (10), and both Berger's and Zelterman's results should extend to the case where $\mathscr{O}$ is replaced by any proper convex cone $\mathscr{C}$ in R^p, as well as the problem where H_0 and H_1 are interchanged. Our criticism applies to these cases as well.

12. Once again we are assuming a nonBayesian framework. WR's test *does* arise as a Bayes test for restrictive priors, for example, for a prior that concentrates its mass on any horizontal line above the μ_1-axis, hence is d-admissible for (11). We conjecture that the LRT is also d-admissible and is approximately Bayes for some prior that spreads its mass more uniformly on $C_0 \cup C_1$.

13. This is often called the "two one-sided tests" ($\equiv$ TOST) procedure. See Berger and Hsu (1996, Sections 3 and 4) for a simple proof that it has size α for (1).

14. The same criticism applies to the New Tests proposed by Anderson and Hauck (1983), Patel and Gupta (1984), Rocke (1984) and Berger and Hsu (1996).

15. Berger and Hsu's rebuttal (1996, page 317) is off the point. They deflect Schuirmann's criticism by discussing a different testing problem; namely, the paired difference problem. Here, unlike the bioequivalence problem (1), the distance from the alternative hypothesis H_1: $\mu_1 \neq \mu_2$ to the null hypothesis H_0: $\mu_1 = \mu_2$, measured in standard units, is unbounded, so it is *not* unreasonable that a rejection region be unbounded in s.

16. Our assertion that these New Tests make inappropriate inferences again rests on the implicit assumption that no unusual prior information is available. If, for example, one assumes a prior distribution over H_1 under which $\mu = 0$ and which assigns all its mass to the unbounded interval $\sigma \in (b, \infty)$ on the σ-axis according to some heavy-tailed density, then for sufficiently large b, the Bayes test might be approximated by a test more nearly resembling the unbiased test of Brown, Hwang and Munk (1997).

17. A closed convex cone is *pointed* if it contains no nonzero linear subspace.

18. The proof of Proposition 5.1 in WM (1998a), asserting the unbiasedness of their New Test, is incorrect; the inequality $\mu'\Sigma^{-1}\mu_0 > 0$ on page 385, column 1, does not hold.

19. Their test is based on the conditional distribution of the unconditional LRT statistic $U(C)$ given V, hence is not the actual conditional LRT. The latter would be obtained from the conditional distribution of (X, S) given V, which seems difficult to obtain.

20. Nor, for that matter, does the unconditional LRT.

21. Similar examples have been frequently noted (e.g., Dempster, 1997, pages 249 and 250; Berger and Wolpert, 1988, Example 4a, page 8; Aitken, 1991, page 112; Royall, 1997, pages 16 and 17).

22. See Cox and Hinkley (1974, Example 5.22); Lehmann (1986, Examples 6.11 and 9.9); Eaton (1989, Example 6.5).

23. See Lehmann (1950, page 2); Lehmann (1986, Problem 6.18); Hacking (1965, pages 97–99). Also see Basu (1975, page 34).

24. Hacking (1965, pages 98 and 99) argues on Fisherian grounds that the LRT, although "worse than useless," is actually more appropriate as a measure of evidence than the reversed LRT, but we do not find his argument convincing.

25. To address Berger's uncertainty about our use of "distance" and "fit," "close" refers to the information metric which, for normal distributions with known covariance, is the same as the covariance metric.

26. Not even Neyman himself. Erich Lehmann has written to us that for problem (6) of Section 4, "I believe Neyman would have preferred the LRT, as I do."

27. Neither does the LRT, but as we shall see, it exhibits unreasonable behavior of a different sort.

28. This example suggests a partial answer to the question that we posed about the LR criterion at the end of Section 10; in general the LRT is unsuccessful for a null hypothesis consisting of several regions of differing dimensionalities. The same is true of the Wald and Rao score tests.

REFERENCES

AITKEN, M. (1991). Posterior Bayes factors (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 111–142.

ANDERSON, S. and HAUCK, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Comm. Statist. Theory Methods* **12** 2662–2692.

BAHADUR, R. R. (1967). An optimal property of the likelihood ratio statistic. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 13–26. Univ. California Press, Berkeley.

BASU, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā Ser. A* **37** 1–71.

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. IMS, Hayward, CA.

BERGER, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *J. Amer. Statist. Assoc.* **84** 192–199.

BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets (with discussion). *Statist. Sci.* **11** 283–319.

BERGER, R. L. and SINCLAIR, D. (1984). Testing hypotheses concerning unions of linear subspaces. *J. Amer. Statist. Assoc.* **79** 158–163.

BROWN, L. D. (1990). An ancillarity paradox which appears in multiple linear regression (with discussion). *Ann. Statist.* **18** 471–538.

BROWN, L. D., HWANG, J. T. G. and MUNK, A. (1997). An unbiased test for the bioequivalence problem. *Ann. Statist.* **25** 2345–2367.

COHEN, A., GATSONIS, C. and MARDEN, J. I. (1983). Hypothesis tests and optimality properties in discrete multivariate analysis. In *Studies in Econometrics, Time Series, and Multivariate Statistics* (S. Karzin, T. Amemiya and L. A. Goodman, eds.) 379–405. Academic Press, New York.

COHEN, A., KEMPERMAN, J. H. B. and SACKROWITZ, H. B. (1997). A critique of likelihood inference for order restricted models. Technical Report 97-010, Dept. Statistics, Rutgers Univ.

COHEN, A. and SACKROWITZ, H. B. (1998). Directional tests for one-sided alternatives in multivariate models. *Ann. Statist.* **26** 2321–2338.

CORNFIELD, J. (1969). The Bayesian outlook and its applications (with discussion). *Biometrics* **25** 617–657.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

DAWID, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *J. Roy. Statist. Soc. Ser. B* **53** 79–109.

DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Statist. Comput.* **7** 247–252. (Originally published in 1973).

EATON, M. L. (1989). *Group Invariance Applications in Statistics. Regional Conference Series in Probability and Statistics* **1**. IMS, Hayward, CA.

EDWARDS, A. W. F. (1972). *Likelihood*. Cambridge Univ. Press.

GUTMANN, S. (1987). Tests uniformly more powerful than uniformly most powerful monotone tests. *J. Statist. Plann. Inference* **17** 279–292.

HACKING, I. (1965). *Logic of Statistical Inference*. Cambridge Univ. Press.

HOEFFDING, W. and WOLFOWITZ, J. (1958). Distinguishability of sets of distributions. *Ann. Math. Statist.* **29** 700–718.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

LASKA, E. M. and MEISNER, M. J. (1989). Testing whether an identified treatment is best. *Biometrics* **45** 1139–1151.

LASKA, E. M., TANG, D.-I. and MEISNER, M. J. (1992). Testing hypotheses about an identified treatment when there are multiple endpoints. *J. Amer. Statist. Assoc.* **87** 825–831.

LEHMANN, E. L. (1950). Some principles of the theory of testing hypotheses. *Ann. Math. Statist.* **21** 1–26.

LEHMANN, E. L. (1952). Testing multiparameter hypotheses. *Ann. Math. Statist.* **23** 541–562.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*. Wiley, New York.

LEHMANN, E. L. (1993). The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or two? *J. Amer. Statist. Assoc.* **88** 1242–1249.

LEHMANN, E. L. (1998). Letter to M. D. Perlman, 8 November 1998.

LIU, H. and BERGER, R. L., (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *Ann. Statist.* **23** 55–72.

MARDEN, J. I. and PERLMAN, M. D. (1980). Invariant tests for means with covariates. *Ann. Statist.* **8** 25–63.

MCDERMOTT, M. P. and WANG, Y. (2000). Construction of uniformly more powerful tests for hypotheses about linear inequalities. *J. Statist. Plann. Inference*. To appear.

MENENDEZ, J. A., RUEDA, C. and SALVADOR, B. (1992). Dominance of likelihood ratio tests under cone constraints. *Ann. Statist.* **20** 2087–2099.

MENENDEZ, J. A. and SALVADOR, B. (1991). Anomalies of the likelihood ratio test for testing restricted hypotheses. *Ann. Statist.* **19** 889–898.

MUKERJEE, H. and TU, R. (1995). Order-restricted inferences in linear regression. *J. Amer. Statist. Assoc.* **90** 717–728.

MUNK, A. (1999). A note on unbiased testing for the equivalence problem. *Statist. Probab. Lett.* **41** 401–406.

NEYMAN, J. and PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference I, II. *Biometrika* **20A** 175–240, 263–294.

NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A* **231** 289–337.

NOMAKUCHI, K. and SAKATA, T. (1987). A note on testing two-dimensional normal mean. *Ann. Inst. Statist. Math.* **39** 489–495.

PATEL, H. I. and GUPTA, G. D. (1984). A problem of equivalence in clinical trials. *Biometrical J.* **26** 471–474.

PERLMAN, M. D. (1969). One-sided testing problems in multivariate analysis. *Ann. Math. Statist.* **40** 549–567 (Correction: *Ann. Math. Statist.* **41** 1777).

PERLMAN, M. D. and WU, L. (2000a). A class of conditional tests for multivariate one-sided alternatives. *J. Statist. Plann. Inference*. To appear.

PERLMAN, M. D. and WU, L. (2000b). A defense of the likelihood ratio test for one-sided and order-restricted alternatives. *J. Statist. Plann. Inference*. To appear.

POCOCK, S., GELLER, N. L. and TSIATIS, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43** 465–472.

PRATT, J. W. (1961). Review of *Testing Statistical Hypotheses* (1959) by E. L. Lehmann. *J. Amer. Statist. Assoc.* **56** 153–156.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order-Restricted Statistical Inference*. Wiley, New York.

ROCKE, D. M. (1984). On testing for bioequivalence. *Biometrics* **40** 225–230.

ROYALL, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, London.

RUSSEK-COHEN, E. and SIMON, R. (1993). Qualitative interactions in multifactor studies. *Biometrics* **49** 467–477.

SASABUCHI, S. (1980). A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika* **67** 429–439.

SOLOMON, D. L. (1975). A note on the non-equivalence of the Neyman–Pearson and generalized likelihood ratio tests for testing a simple null versus a simple alternative hypothesis. *Amer. Statist.* **29** 101–102.

TANG, D.-I. (1994). Uniformly more powerful tests in a one-sided multivariate problem. *J. Amer. Statist. Assoc.* **89** 1006–1011.

TANG, D.-I. (1998). Testing the hypothesis of a normal mean lying outside a convex cone. *Comm. Statist. Theory Methods* **27** 1517–1534.

TANG, D.-I., GELLER, N. L. and POCOCK, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49** 23–30.

WALD, A. (1941a). Asymptotically most powerful tests of statistical hypotheses. *Ann. Math. Statist.* **12** 1–19.

WALD, A. (1941b). Some examples of asymptotically most powerful tests. *Ann. Math. Statist.* **12** 396–408.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

WANG, W. (1997). Optimal unbiased tests for equivalence intrasubject variability. *J. Amer. Statist. Assoc.* **92** 1163–1170.

WANG, W., HWANG, J. T. G. and DASGUPTA, A. (1999). Statistical tests for multivariate bioequivalence. *Biometrika* **86** 395–402.

WANG, Y. and MCDERMOTT, M. P. (1998a). Conditional likelihood ratio test for a nonnegative normal mean vector. *J. Amer. Statist. Assoc.* **93** 380–386.

WANG, Y. and MCDERMOTT, M. P. (1998b). A conditional test for a nonnegative mean vector based on a Hotelling's T^2-type statistic. *J. Multivariate Anal.* **66** 64–70.

WARRACK, G. and ROBERTSON, T. (1984). A likelihood ratio test regarding two nested but oblique order-restricted hypotheses. *J. Amer. Statist. Assoc.* **79** 881–886.

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9** 60–62.

WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

ZELTERMAN, D. (1990). On tests for qualitative interactions. *Statist. Probab. Lett.* **10** 59–63.