More Powerful Tests for the Sign Testing Problem

Khalil G. Saikali
ClinTrials, Incorporated
Morrisville, NC 27560

Roger L. Berger
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

July, 1998
NCSU Institute of Statistics Mimeo Series No. 2511

Abstract

For $i = 1, \ldots, p$, let $X_{i1}, \ldots, X_{in_i}$ denote independent random samples from normal populations. The ith population has unknown mean μ_i and unknown variance σ_i^2. We consider the sign testing problem of testing $\mathrm{H}_0 : \mu_i \leq a_i$, for some $i = 1, \ldots, p$, versus $\mathrm{H}_1 : \mu_i > a_i$, for all $i = 1, \ldots, p$, where $a_1, \ldots, a_p$ are fixed constants. Here, H_1 might represent p different standards that a product must meet before it is considered acceptable. For $0 < \alpha < 1/2$, we first derive the size-α likelihood ratio test (LRT) for this problem, and then we describe an intersection-union test (IUT) that is uniformly more powerful than the likelihood ratio test if the sample sizes are not all equal. For a more general model than the normal, we describe two intersection-union tests that maximize the size of the rejection region formed by intersection. Applying these tests to the normal problem yields two tests that are uniformly more powerful than both the LRT and IUT described above. A small power comparison of these tests is given.

Keywords: Independence, intersection-union test, likelihood ratio test, normal population, t distribution.

1 Introduction

Let X_{ij}, $i = 1, \ldots, p$ $(p \geq 2)$, $j = 1, \ldots, n_i$ (each $n_i \geq 2$), denote independent normal random variables. For each $i = 1, \ldots, p$, $X_{i1}, \ldots, X_{in_i}$ represents a random sample from a normal population with unknown mean μ_i and unknown variance σ_i^2. Let $a_1, \ldots, a_p$ be fixed constants. We consider the "sign testing problem" of testing

$$\begin{aligned} &\mathrm{H}_0 : \mu_i \leq a_i, \quad \text{for some } i = 1, \ldots, p, \\ \text{versus} \quad & \\ &\mathrm{H}_1 : \mu_i > a_i, \quad \text{for all } i = 1, \ldots, p. \end{aligned} \tag{1}$$

The inequalities in H_1 might represent various efficacy and safety standards that a product must meet before it is acceptable. When H_0 and H_1 are stated in this way, the consumer's risk will be protected as discussed by Berger (1982). Or the different inequalities could refer to different methods of measuring the same variable, and we wish to determine if all the methods are giving consistently positive results.

Sign testing has been considered in various contexts by many authors such as Lehmann (1952), Berger (1982), Gutmann (1987), Berger (1989), Shirley (1992), and Liu and Berger (1995). Sasabuchi (1980, 1988a, 1988b) derived the LRT for the cases of known variances, unknown but equal variances, and completely unknown covariance matrix, respectively. But, apparently no one has considered our model with independent samples but unknown, possibly unequal, variances.

First, for $0 < \alpha < 1/2$, we derive the LRT for testing (1). Then, we show that a simple intersection-union test (IUT), constructed from one-sample t tests for each μ_i, is uniformly more powerful than the LRT if the sample sizes are not all equal. For a more general model than the normal, we describe two methods of constructing IUTs that maximize the size of the resulting rejection region. Using these methods, we describe two more tests of (1) that are both uniformly more powerful than both the LRT and the simple IUT. We present a small power comparison of these tests.

Our notation will be simplified by considering these transformed data. Let $Y_{ij} = X_{ij} - a_i$, $i = 1, \ldots, p$, $j = 1, \ldots, n_i$. Then, $Y_{ij} \sim \mathrm{n}(\theta_i, \sigma_i^2)$, where $\theta_i = \mu_i - a_i$, and testing (1) is equivalent to testing

$$
\begin{aligned}
& \mathrm{H}_0 : \theta_i \leq 0, \quad \text{for some } i = 1, \ldots, p, \\
\text{versus} & \\
& \mathrm{H}_1 : \theta_i > 0, \quad \text{for all } i = 1, \ldots, p.
\end{aligned} \tag{2}
$$

In this form, the origin of the term "sign testing" is apparent. These hypotheses can also be written as

$$
\begin{aligned}
& \mathrm{H}_0 : \min_{1 \leq i \leq p} \{\theta_i\} \leq 0 \\
\text{versus} & \\
& \mathrm{H}_1 : \min_{1 \leq i \leq p} \{\theta_i\} > 0.
\end{aligned}
$$

Throughout the remainder of this paper, we will express our results in terms of the Y_{ij}s and express the hypotheses as (2). All of the tests we consider will depend on the data only through the sufficient statistics $\overline{Y}_1, \ldots, \overline{Y}_p$, the sample means, and $S_1^2, \ldots, S_p^2$, the sample variances defined by $S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2/(n_i - 1)$. The complete data vector will be denoted by $\boldsymbol{Y}$; $\boldsymbol{y}$ will denote an observed value.

2 LRT and a Simple IUT

The hypotheses (2) can be expressed as

$$
\begin{aligned}
& \mathrm{H}_0 : \cup_{i=1}^{p} \{\theta_i \leq 0\} \\
\text{versus} & \\
& \mathrm{H}_1 : \cap_{i=1}^{p} \{\theta_i > 0\}.
\end{aligned}
$$

This type of testing problem, in which the null hypothesis is conveniently expressed as a union and the alternative hypothesis is expressed as an intersection, is the type for which it is natural to use an IUT. IUTs were first described by Gleser (1973) and Berger (1982).

2.1 LRT of H_0 versus H_1

The results in Berger (1997) can be used to find both the LRT and a simple IUT for problems of this type. The LRT statistic for testing (2) is defined to be

$$\lambda(\boldsymbol{Y}) = \frac{\sup_{\mathrm{H}_0} L(\theta_1,\ldots,\theta_p,\sigma_1^2,\ldots,\sigma_p^2|\boldsymbol{Y})}{\sup_{\mathrm{H}_0\cup\mathrm{H}_1} L(\theta_1,\ldots,\theta_p,\sigma_1^2,\ldots,\sigma_p^2|\boldsymbol{Y})}$$

where $L(\cdot|\boldsymbol{Y})$ is the normal likelihood function.

For each $i = 1,\ldots,p$, consider testing the individual hypotheses

$$\begin{array}{c} \mathrm{H}_{i0} : \theta_i \leq 0 \\ \text{versus} \\ \mathrm{H}_{i1} : \theta_i > 0. \end{array} \tag{3}$$

Standard calculations yield that the LRT statistic for testing the individual hypotheses (3) is

$$\begin{aligned} \lambda_i(\boldsymbol{Y}) &= \frac{\sup_{\mathrm{H}_{i0}} L(\theta_1,\ldots,\theta_p,\sigma_1^2,\ldots,\sigma_p^2|\boldsymbol{Y})}{\sup_{\mathrm{H}_{i0}\cup\mathrm{H}_{i1}} L(\theta_1,\ldots,\theta_p,\sigma_1^2,\ldots,\sigma_p^2|\boldsymbol{Y})} \\ &= \begin{cases} 1, & \text{if } \overline{Y}_i \leq 0, \\ \left(1+\dfrac{T_i^2}{n_i-1}\right)^{-n_i/2}, & \text{if } \overline{Y}_i > 0, \end{cases} \end{aligned}$$

where

$$T_i = \frac{\overline{Y}_i}{S_i/\sqrt{n_i}}$$

is the usual t statistic for testing H_{i0}. Note that λ_i is computed from all the data. But, because of the independence, the likelihood factors, and the parts of the likelihood that do not depend on the ith sample cancel out of the LRT statistic. The size-α LRT rejects H_{i0} in favor of H_{i1} if $\lambda_i(\boldsymbol{Y}) \leq c_{i\alpha}$, where $c_{i\alpha}$ is an appropriately chosen cutoff value. When $\theta_i = 0$, T_i has a Student's t distribution. Because $0 < \alpha < 1/2$,

$$c_{i\alpha} = \left(1+\frac{t^2_{\alpha,n_i-1}}{n_i-1}\right)^{-n_i/2}, \tag{4}$$

where t_{α,n_i-1} is the upper 100α percentile of a t distribution with n_i-1 degrees of freedom. Thus, the LRT of H_{i0} versus H_{i1} rejects H_{i0} if $T_i \geq t_{\alpha,n_i-1}$, the usual one-sided t test.

Using (15.3) in Berger (1997), the LRT statistic for testing H_0 is $\lambda(\boldsymbol{Y}) = \max_{1\leq i\leq p}\lambda_i(\boldsymbol{Y})$, and the size-$\alpha$ LRT rejects H_0 if $\lambda(\boldsymbol{Y}) \leq c_\alpha$. By Theorem 15.2.2 in Berger (1997), the cutoff value c_α that yields a size-α test is $c_\alpha = \min_{1\leq i\leq p} c_{i\alpha}$. For each $\alpha = .01$, $.05$, and $.10$, and for all sample sizes $n_i = 2,\ldots,100$, the constants $c_{i\alpha}$ from (4) were computed. It was found that for fixed α, $c_{i\alpha}$ is increasing in n_i. Thus, at least on this range, c_α is the $c_{i\alpha}$ in (4) that corresponds to the smallest sample size. Let $n_{(1)} = \min_{1\leq i\leq p} n_i$. Then the LRT can be expressed in terms of the t statistics thusly. The size-α LRT of (2) rejects H_0 if, for every $i = 1,\ldots,p$,

$$T_i \geq \left\{\left[\left(1+\frac{t^2_{\alpha,n_{(1)}-1}}{n_{(1)}-1}\right)^{n_{(1)}/n_i} - 1\right](n_i-1)\right\}^{1/2}. \tag{5}$$

If $n_i = n_{(1)}$, the cutoff value for T_i is t_{α,n_i-1}, but, if $n_i > n_{(1)}$, the cutoff value for T_i is greater than t_{α,n_i-1}.

2.2 Simple IUT that is more powerful

The size-α LRTs of the individual hypotheses can be combined directly, using the intersection-union method, to obtain a size-α test of H_0 versus H_1. By Theorems 15.1.1 and 15.1.2 of Berger (1997) (cf., Berger (1982)), the test that rejects H_0 if, for every $i = 1, \ldots, p$,

$$T_i \geq t_{\alpha,n_i-1} \tag{6}$$

is a size-α test of H_0 versus H_1. We will call this test the simple IUT (SIUT).

If $n_1 = \cdots = n_p$, then all the right-hand sides in (5) and (6) are equal to t_{α,n_1-1}, and the LRT and SIUT are the same test. This test is also the LRT found by Sasabuchi (1988a) for the model in which the covariance matrix is completely unknown. So, as far as the LRT is concerned, when the sample sizes are all equal, no advantage is gained by assuming the populations are independent.

But, if the sample sizes are not all equal, then for any i with $n_i > n_{(1)}$, the cutoff value in (5) is greater than the corresponding cutoff value in (6). In this case, the rejection region in (5) is a proper subset of the rejection region in (6). Both tests are size-α tests, and the SIUT is uniformly more powerful than the LRT. For the case $p = 2$, $n_1 = 40$, $n_2 = 2$, and $\alpha = .05$, the rejection regions for the LRT and the SIUT are shown in Figure 1. In either the left or right graph, the rejection region of the SIUT is the rectangle in the upper right corner, bounded by solid lines, defined by $\{t_1 \geq 1.68, t_2 \geq 6.31\}$. The rejection region of the LRT is the smaller rectangle in the upper right corner, bounded by solid and dashed lines, defined by $\{t_1 \geq 2.82, t_2 \geq 6.31\}$. (Figure 1 will be discussed more in Section 4.)

The SIUT is the uniformly most powerful (UMP), size-α, monotone test based on the t statistics $T_1, \ldots, T_p$. A test is monotone if the sample point $(t_1, \ldots, t_p)$ is in the rejection region and $t_i' \geq t_i$ for all $i = 1, \ldots, p$ imply the sample point $(t_1', \ldots, t_p')$ is in the rejection region. The statistic T_i has a noncentral t distribution with noncentrality parameter $\nu_i = \sqrt{n_i}\theta_i/\sigma_i$. Testing (2) is equivalent to testing

$$\begin{aligned} &\mathrm{H}_0 : \cup_{i=1}^{p}\{\nu_i \leq 0\} \\ \text{versus}\quad & \\ &\mathrm{H}_1 : \cap_{i=1}^{p}\{\nu_i > 0\}. \end{aligned} \tag{7}$$

Because the noncentral t distribution has monotone likelihood ratio in the noncentrality parameter, Theorem 3.3 in Cohen, Gatsonis, and Marden (1983), yields that the SIUT is the UMP, size-α, monotone test. Lehmann (1952) and Gutmann (1987) also discussed UMP monotone tests.

If we consider nonmonotone tests, there are size-α tests of H_0 versus H_1 that are uniformly more powerful than the SIUT (and, of course, the LRT). In the remainder of this paper, we will describe two such tests. First, in the next section, we will discuss two general results about the construction of IUTs.

3 Two General IUT Results

In this section we consider two results useful in the construction of more powerful IUTs. The first concerns combining tests for one intersection ($p = 2$) into a test for a multiple intersection ($p > 2$). The second concerns constructing IUTs for one intersection in an "optimal" way.

3.1 Combining tests for one intersection

Consider a general testing problem in which we might consider using an IUT. Suppose the data $\boldsymbol{X}$ has a distribution that depends on a (possibly vector valued) parameter $\boldsymbol{\theta}$. Suppose the hypotheses to be tested are expressed as

$$\text{(8)} \qquad \begin{array}{c} \mathrm{H}_0 : \boldsymbol{\theta} \in \cup_{i=1}^{p} \Theta_{i0} \\ \text{versus} \\ \mathrm{H}_1 : \boldsymbol{\theta} \in \cap_{i=1}^{p} \Theta_{i0}^{c}, \end{array}$$

where $\Theta_{10}, \ldots, \Theta_{p0}$ are some subsets of the parameter space. The simplest use of the intersection-union method to construct a level-α test of (8) is as we did in (6). Let R_i be a level-α rejection region for testing $\mathrm{H}_{i0} : \boldsymbol{\theta} \in \Theta_{i0}$ versus $\mathrm{H}_{i1} : \boldsymbol{\theta} \in \Theta_{i0}^{c}$. Then, by Theorem 15.1.1 in Berger (1997) (cf., Theorem 1, Berger (1982)), the test with rejection region $R = \cap_{i=1}^{p} R_i$ is a level-α test of (8). If conditions like those in Theorem 15.2.2 of Berger (1997) are met, this test will be a size-α test. But, although tests constructed in this way might have the correct size, they often have poor power.

Many authors have found that by considering two individual hypotheses at a time, say H_{i0} and H_{j0}, more powerful tests can be constructed. Authors that have done this include Berger (1989), Liu and Berger (1995), and Wang and McDermott (1996) . They have proposed tests for problems of the form $\mathrm{H}_{(ij)0} : \boldsymbol{\theta} \in \Theta_i \cup \Theta_j$ versus $\mathrm{H}_{(ij)1} : \boldsymbol{\theta} \in \Theta_i^c \cap \Theta_j^c$. Their tests have often been IUTs, with rejection region $R_{ij} = R_i \cap R_j$, where R_i and R_j are size-α rejection regions of H_{i0} and H_{j0}, respectively, but R_i and R_j have been specifically chosen so that their intersection is a "large" set and the test is powerful. Shirley (1992) and Saikali (1996) specifically considered alternative hypotheses with more than two intersections. But, these constructions tend to be more complicated and we will not consider them herein.

Tests of two individual hypotheses at a time can be combined to obtain a test of (8). Liu and Berger (1995) pointed out that, if p is even, we can pair the hypotheses and express (8) as

$$\begin{array}{c} \mathrm{H}_0 : \boldsymbol{\theta} \in \cup_{i=1}^{p/2} \{\Theta_{(2i-1)0} \cup \Theta_{(2i)0}\} \\ \text{versus} \\ \mathrm{H}_1 : \boldsymbol{\theta} \in \cap_{i=1}^{p/2} \{\Theta_{(2i-1)0}^{c} \cap \Theta_{(2i)0}^{c}\}. \end{array}$$

If $R_{2i-1,2i}$ is the rejection region of a size-α test of $\mathrm{H}_{2i-1,2i,0} : \boldsymbol{\theta} \in \{\Theta_{(2i-1)0} \cup \Theta_{(2i)0}\}$, then $R = \cap_{i=1}^{p/2} R_{2i-1,2i}$ is a level-α rejection region of (8). If the $R_{2i-1,2i}$s are chosen carefully, this test should be more powerful than an IUT constructed from p tests of the individual H_{i0}s.

If p is odd, Liu and Berger (1995) suggested that H_{p0} be paired with another hypothesis say H_{10}, the size-α rejection region $R_{p,1}$ for $\mathrm{H}_{p,1}$ be chosen and used in forming an IUT.

But, this is unnecessarily complicated. The null hypothesis in (8) can be expressed as

$$\mathrm{H}_0 : \boldsymbol{\theta} \in \left(\cup_{i=1}^{(p-1)/2}\{\Theta_{(2i-1)0} \cup \Theta_{(2i)0}\}\right) \cup \Theta_p.$$

From this is is clear that any size-α rejection region for testing $\mathrm{H}_{p0} : \boldsymbol{\theta} \in \Theta_p$ can be used to intersect with the other rejection regions to yield a level-α IUT of H_0. As always, consideration of how these rejection regions will intersect might yield more powerful tests.

In the remainder of this paper, we will concentrate on tests for one intersection.

3.2 IUT construction for one intersection ($p = 2$)

In this section we will propose two methods of constructing an IUT for the case that the null and alternative hypotheses consist of one union and one intersection, respectively. These methods have a certain optimality property that should help ensure that the resulting IUTs have good power.

Whenever an IUT is used, there should be concern that the resulting rejection region is too small and the test has poor power. If the rejection regions that are intersected do not have many points in common, the resulting rejection region will be small. If an IUT rejection region is formed as $R = R_1 \cap R_2$, it would be ideal if $R_1 = R_2$ $(= R)$. Then no sample points at all are lost in the intersection, and R is as large as possible, in some sense. The constants k_1 and k_2 in Section 2 of Wang and McDermott (1996) and d in Liu and Berger (1995) are "tuning constants" that are used to adjust R_1 and R_2 so the their intersection is as large as possible. In this section, we give a construction that, under general conditions, always yields $R_1 = R_2 = R$.

One particular advantage of having $R_1 = R_2 = R$ is that the size of R, as a test for H_0, is the maximum of the size of $R = R_1$ as a test of H_{10} and the size of $R = R_2$ as a test of H_{20}, because

$$\sup_{\mathrm{H}_0} P(R) = \sup_{\mathrm{H}_{10}\cup\mathrm{H}_{20}} P(R) = \max\{\sup_{\mathrm{H}_{10}} P(R), \sup_{\mathrm{H}_{20}} P(R)\}.$$

The two values in the "max" are the sizes of R as a test of the two individual hypotheses.

We consider this general model. Let ν_1 and ν_2 be two real-valued parameters. We consider testing

$$\mathrm{H}_0 : \{\nu_1 \leq 0\} \cup \{\nu_2 \leq 0\}$$

versus

$$\mathrm{H}_1 : \{\nu_1 > 0\} \cap \{\nu_2 > 0\}.$$

We assume the parameter space is a rectangle of the form $\{(\nu_1, \nu_2) : \nu_1^L < \nu_1 < \nu_1^U, \nu_2^L < \nu_2 < \nu_2^U\}$. The endpoints ν_i^L and ν_i^U can be $\pm\infty$.

Let T_1 and T_2 denote two statistics. We make the following assumptions. The distribution of T_i is continuous and depends only on ν_i. T_1 and T_2 are independent. The statistic T_i is a test statistic for testing $\mathrm{H}_{i0} : \nu_i \leq 0$, and large values of T_i give evidence that $\mathrm{H}_{i1} : \nu_i > 0$ is true. Let $F_i(t)$ denote the cumulative distribution function of T_i when $\nu_i = 0$, and let $f_i(t|\nu_i)$ denote the density of T_i. Let $m_i = F_i^{-1}(1/2)$ denote the median of T_i when $\nu_i = 0$. We assume that the support of T_i is an interval (t_i^L, t_i^U); the endpoints t_i^L and t_i^U can be $\pm\infty$. We also assume

(9) $$T_i \text{ converges in probability to } t_i^U \text{ as } \nu_i \to \nu_i^U.$$

We also assume that the distribution of T_i has the property that, if A is a subset of $\{t : t \geq m_i\}$, then

$$P_{\nu_i}(T_i \in A) \leq P_0(T_i \in A), \quad \text{for all } \nu_i \leq 0. \tag{10}$$

This property allows us to check the size of a test by considering only the boundary value $\nu_i = 0$ in this way. Suppose R is a set of (t_1, t_2) values such that $t_1 \geq m_1$ for all $(t_1, t_2) \in R$. Define $R(t_2) = \{t_1 : (t_1, t_2) \in R\}$. Note that $R(t_2) \subset \{t_1 : t_1 \geq m_1\}$ Then for for any $(\nu_1, \nu_2) \in \mathrm{H}_{10}$, that is with $\nu_1 \leq 0$,

$$\begin{aligned} P_{\nu_1,\nu_2}((T_1, T_2) \in R) &= \int_{-\infty}^{\infty} \int_{R(t_2)} f_1(t_1|\nu_1)\, dt_1\, f_2(t_2|\nu_2)\, dt_2 \\ &= \int_{-\infty}^{\infty} P_{\nu_1}(T_1 \in R(t_2)) f_2(t_2|\nu_2)\, dt_2 \\ &\leq \int_{-\infty}^{\infty} P_0(T_1 \in R(t_2)) f_2(t_2|\nu_2)\, dt_2 \qquad (11) \\ &= P_{0,\nu_2}((T_1, T_2) \in R). \end{aligned}$$

Thus, if $P_{0,\nu_2}((T_1, T_2) \in R) \leq \alpha$, for all ν_2, R is a level-α rejection region for testing H_{10}. By a similar argument, if $R \subset \{(t_1, t_2) : t_2 \geq m_2\}$, then we need check only parameter values of the form $(\nu_1, 0)$ to determine if R is a level-α rejection region for testing H_{20}.

Let $c_{i,\alpha} = F_i^{-1}(1 - \alpha)$. The test that rejects H_{i0} if $T_i \geq c_{i,\alpha}$ is a size-α test of H_{i0}. Hence, the test that rejects H_0 if $T_1 \geq c_{1,\alpha}$ and $T_2 \geq c_{2,\alpha}$ is a level-α test of H_0. This test is analogous to the SIUT in (6). In fact, by results in Lehmann (1952) and (9), this test is the UMP, size-α, monotone test of H_0. Now we will describe two other size-α tests of H_0 that are uniformly more powerful than this test.

3.2.1 Rectangle test

The test we describe in this section we will call Test R (for Rectangles). It is a generalization of a test in Berger (1989). Recalling that $0 < \alpha < 1/2$, define the integer J by the inequality $J - 1 < (2\alpha)^{-1} \leq J$. Then, for $i = 1$ and 2, define constants $c_0^i, \ldots, c_J^i$ by $c_0^i = t_i^U = F_i^{-1}(1 - 0\alpha)$, $c_j^i = F_i^{-1}(1 - j\alpha)$, $j = 1, \ldots, J-1$, and $c_J^i = F_i^{-1}(1/2) = m_i$. (Note, if $(2\alpha)^{-1}$ is an integer, as it is for $\alpha = .10$, $.05$, and $.01$, then $c_J^i = F_i^{-1}(1 - J\alpha)$.) For $j = 1, \ldots, J$, define

$$R^j = \{(t_1, t_2) : c_j^1 \leq t_1 < c_{j-1}^1, c_j^2 \leq t_2 < c_{j-1}^2\}.$$

Then the set $R = R^1 \cup \cdots \cup R^J$ is a size-α rejection region for H_{10}, H_{20}, and H_0, and this test is uniformly more powerful than the UMP, size-α, monotone test. The rectangle R^1 is the rejection region of the UMP, size-α, monotone test. So, obviously the test based on R is uniformly more powerful. It remains to show that the size of R is α.

To verify that R has size at most α for testing H_{10}, by (11) we need consider only parameter points of the form $(0, \nu_2)$. Using the notation in (11) we have

$$R(t_2) = \begin{cases} [c_j^1, c_{j-1}^1) = [F_1^{-1}(1 - j\alpha), F_1^{-1}(1 - (j-1)\alpha)), & c_j^2 \leq t_2 < c_{j-1}^2, \\ & \quad j = 1, \ldots, J-1, \\ [c_J^1, c_{J-1}^1) \subset [F_1^{-1}(1 - J\alpha), F_1^{-1}(1 - (J-1)\alpha)), & c_J^2 \leq t_2 < c_{J-1}^2, \\ \emptyset, & t_2 < c_J^2. \end{cases}$$

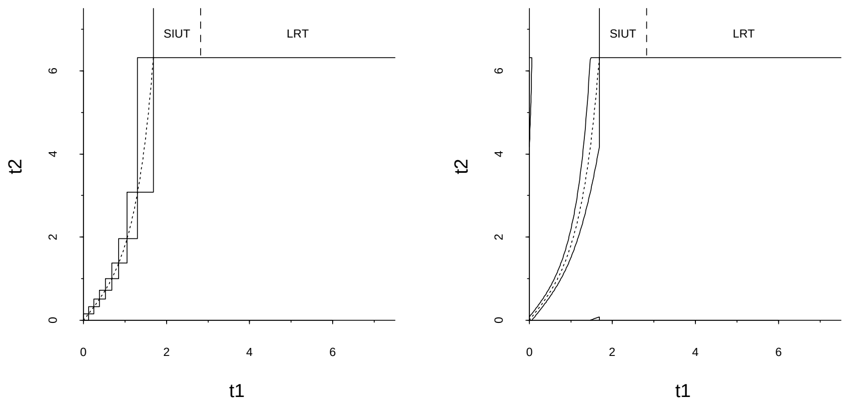

Figure 1: Rejection regions for Tests R (left graph) and S (right graph) for testing normal means, $n_1 = 40$, $n_2 = 2$, $\alpha = .05$.

Therfore,

$$P_0(T_1 \in R(t_2)) \begin{cases} = \alpha, & c_{J-1}^2 \leq t_2 \\ \leq \alpha, & c_J^2 \leq t_2 < c_{J-1}^2, \\ = 0, & t_2 < c_J^2. \end{cases}$$

Calculating as in (11), for any value of ν_2 we have

$$\begin{aligned} P_{0,\nu_2}((T_1, T_2) \in R) &= \int_{-\infty}^{\infty} P_0(T_1 \in R(t_2)) f_2(t_2|\nu_2)\, dt_2 \\ &= \int_{c_J^2}^{\infty} P_0(T_1 \in R(t_2)) f_2(t_2|\nu_2)\, dt_2 \\ &\leq \int_{c_J^2}^{\infty} \alpha f_2(t_2|\nu_2)\, dt_2 \\ &= \alpha P_{\nu_2}(T_2 \geq c_J^2) \\ &\leq \alpha. \end{aligned}$$

(Note, the first inequality is an equality if $(2\alpha)^{-1}$ is an integer.) Therefore, by the argument at (11), R is a level-α rejection region for testing H_{10}. Because $R^1 \subset R$ and R^1 is a size-α rejection region for testing H_{10}, R is, in fact, a size-α rejection region for testing H_{10}. Reversing the roles of T_1 and T_2 in this argument shows that R is a size-α rejection region for testing H_{20}. Thus, R is a size-α rejection region for testing H_0.

It is very easy to check if a sample point is in R because of the simple definition of R in terms of the rectangles R^j. An example of such an R is shown in the left graph in Figure 1 (to be discussed more later). But some have criticized the "jagged edges" on R. Next we describe another test that has smoother edges.

3.2.2 Smoother test

The test we describe in this section we will call Test S (for Smooth). It is a generalization of a test in Wang and McDermott (1996). Wang and McDermott define a certain subset of the unit square (Figure 2 in their paper) that has three parts, $A = A_0 \cup A_a \cup A_b$. Using simpler notation than theirs, these sets are

$$
\begin{aligned}
A_0 &= \{(u_1,u_2) : 1-\alpha \le u_1 \le 1, 1-\alpha \le u_2 \le 1\} \\
A_a &= \{(u_1,u_2) : |u_1-u_2| \le \alpha/2, 1/2 \le u_1 < 1-\alpha, 1/2 \le u_2 < 1-\alpha\} \\
A_b &= \{(u_1,u_2) : 1/2 \le u_2 \le u_1 - 1/2 + 3\alpha/2, u_1 < 1-\alpha\} \\
&\quad \cup \{(u_1,u_2) : 1/2 \le u_1 \le u_2 - 1/2 + 3\alpha/2, u_2 < 1-\alpha\}.
\end{aligned} \tag{12}
$$

The set A has this important property. Let $A(u_2) = \{u_1 : (u_1,u_2) \in A\}$. If $u_2 < 1/2$, $A(u_2) = \emptyset$. If $1/2 \le u_2 \le 1$, $A(u_2)$ consists of one or two intervals whose total length is exactly α. Thus, if U_1 and U_2 are independent random variables and U_1 has a uniform$(0,1)$ distribution, then, regardless of the distribution of U_2,

$$
P((U_1,U_2) \in A) = \int_0^1 P(U_1 \in A(u_2)) f(u_2)\, du_2 = \alpha P(1/2 \le U_2 \le 1) \le \alpha. \tag{13}
$$

The set A is symmetric in u_1 and u_2. Thus, it is also true that, if $U_2 \sim$ uniform$(0,1)$ then $P((U_1,U_2) \in A) = \alpha P(1/2 \le U_1 \le 1) \le \alpha$.

Now, we define Test S for testing H_0. Let $U_1 = F_1(T_1)$ and $U_2 = F_2(T_2)$. Test S is the test that rejects H_0 if $(U_1,U_2) \in A$. Let R_S denote the rejection region of Test S. Note that

$$
\begin{aligned}
\{(U_1,U_2) \in A_0\} &= \{1-\alpha \le F_1(T_1), 1-\alpha \le F_2(T_2)\} \\
&= \{F_1^{-1}(1-\alpha) \le T_1, F_2^{-1}(1-\alpha) \le T_2\} \\
&= \{c_{1,\alpha} \le T_1, c_{2,\alpha} \le T_2\}
\end{aligned}
$$

Thus, the rejection region of Test S contains the rejection region of the UMP, size-α, monotone test, and, hence, Test S is uniformly more powerful. To see that Test S is a size-α test of H_0, note that for every $(u_1,u_2) \in A$, $u_1 \ge 1/2$ and $u_2 \ge 1/2$. Thus, for every $(t_1,t_2) \in R_S$, $F_1(t_1) \ge 1/2$ and $F_2(t_2) \ge 1/2$, that is, $t_1 \ge m_1$ and $t_2 \ge m_2$, and we can use (10). Finally, if $\nu_i = 0$, then $U_i = F_i(T_i) \sim$ uniform$(0,1)$. Thus, by (11) and (13), R_S is a level-α rejection region for testing H_{10} and H_{20}. Because $R^1 \subset R_S$ and R^1 is a size-α rejection region for testing H_{10} and H_{20}, R_S is, in fact, a size-α rejection region for testing H_{10} and H_{20}. Thus, R_S is a size-α rejection region for testing H_0. An example of a rejection region R_S is shown in the right graph in Figure 1.

If $(2\alpha)^{-1}$ is an integer, Test R and Test S both have the same power on the boundary of H_0, namely,

$$
\begin{aligned}
P_{0,\nu_2}((T_1,T_2) \in R) &= \alpha P_{\nu_2}(T_2 \ge m_2) = P_{0,\nu_2}((T_1,T_2) \in R_S), \\
P_{\nu_1,0}((T_1,T_2) \in R) &= \alpha P_{\nu_1}(T_1 \ge m_1) = P_{\nu_1,0}((T_1,T_2) \in R_S).
\end{aligned} \tag{14}
$$

If $(2\alpha)^{-1}$ is not an integer, the power of Test R is slightly less than the power of Test S on the boundary of H_0 because the cross-sectional probabilities for Test R are less than α near zero. A power comparison of these two tests for a normal model is given in Section 5.

3.2.3 Relationships with p-values

Tests R and S can both be described in terms of p-values for testing H_{10} and H_{20}. Note that $P_i = 1 - U_i = 1 - F_i(T_i)$ is the p-value for testing H_{i0} for a test that rejects for large values of T_i. The UMP, size-α, monotone test rejects H_0 if $P_1 \leq \alpha$ and $P_2 \leq \alpha$. Test S rejects H_0 if $(1 - P_1, 1 - P_2) \in A$. The rectangles that define Test R can be expressed in terms of the p-values as, for $j = 1, \ldots, J-1$,

$$R^j = \{(j-1)\alpha < P_1 \leq j\alpha, (j-1)\alpha < P_2 \leq j\alpha\}$$

and

$$R^J = \{(J-1)\alpha < P_1 \leq 1/2, (J-1)\alpha < P_2 \leq 1/2\}.$$

Tests R and S give effective ways of combining the p-values from simple tests of H_{10} and H_{20} into a test of H_0, when the p-values are independent. Further study of combining p-values in these kinds of problems, when the p-values are not independent, would be of interest.

4 New Tests for Normal Sign Testing

We now return to the problem of sign testing for normal populations. That is, we consider the problem of testing (2). We will discuss the new Tests R and S from Section 3.2 for this model. In this discussion we consider only $p = 2$ populations, realizing that tests for two populations could be combined for more populations as discussed in Section 3.1.

Recall that we will base our tests on the two statistics T_1 and T_2. T_i has a noncentral t distribution with $n_i - 1$ degrees of freedom and noncentrality parameter $\nu_i = \sqrt{n_i}\theta_i/\sigma_i$. Hypotheses (2) about the normal means can be equivalently express in terms of the noncentrality parameters as (7).

The test constructions in Section 3.2 can be used for this normal model because the noncentral t distributions satisfy all the required conditions. The distributions F_1 and F_2 are central t distributions, so $m_1 = m_2 = 0$. The only condition that requires verification is (10). This theorem verifies (10).

Theorem 1 *Let T be a random variable with a noncentral t distribution with r degrees of freedom and noncentrality parameter ν. Let A be a subset of $\{t : t \geq 0\}$. Then for any $\nu \leq 0$, $P_\nu(T \in A) \leq P_0(T \in A)$.*

Proof: The random variable $X/\sqrt{V/r}$ has the same distribution as T if X and V are independent, $X \sim \mathrm{n}(\nu, 1)$, and V has a chi-squared distribution with r degrees of freedom. Let $f(v)$ denote the density of V, and, for every $v > 0$, let $A(v) = \{x : x/\sqrt{v/r} \in A\}$. Then,

$$P_\nu(T \in A) = P_\nu(X/\sqrt{V/r} \in A) = \int_0^\infty P_\nu(X \in A(v)) f(v)\, dv.$$

For every $v > 0$, $A(v) \subset \{x : x \geq 0\}$. This implies, by Theorem 2.2 in Berger (1989), if $\nu \leq 0$, $P_\nu(X \in A(v)) \leq P_0(X \in A(v))$. Thus,

$$P_\nu(T \in A) \leq \int_0^\infty P_0(X \in A(v)) f(v)\, dv = P_0(T \in A),$$

as was to be shown. $\square$

Thus, the Tests R and S from Section 3.2 are size-α tests that are uniformly more powerful than the LRT and the UMP monotone test for sign testing about normal means. We now express R and S in terms of t distribution percentiles and the central t cdfs F_1 and F_2. (Recall, F_i is the cdf of a central t distribution with $n_i - 1$ degrees of freedom.)

Test R for normal model

For $j = 1, \ldots, J-1$,

$$R^j = \{(t_1, t_2) : t_{j\alpha, n_1-1} \le t_1 < t_{(j-1)\alpha, n_1-1}, t_{j\alpha, n_2-1} \le t_2 < t_{(j-1)\alpha, n_2-1}\}$$

and

$$R^J = \{(t_1, t_2) : 0 \le t_1 < t_{(J-1)\alpha, n_1-1}, 0 \le t_2 < t_{(J-1)\alpha, n_2-1}\}.$$

Test R rejects H_0 if $(T_1, T_2) \in R^j$, for some $j = 1, \ldots, J$.

The rejection region for Test R, for the case of $\alpha = .05$, $n_1 = 2$, and $n_2 = 40$, is shown in Figure 1. This might be compared with Figure 2 in Berger (1989). In that figure, the two marginal distributions are the same, and the rectangles (squares in that case) are centered on a straight line from $(0,0)$ to (z_α, z_α). In this example the two distributions, F_1 and F_2 are different and the rectangles are centered on a curved line from $(0,0)$ to $(t_{\alpha,n_1-1}, t_{\alpha,n_2-1})$. The curved line is defined by $t_2 = F_2^{-1}(F_1(t_1))$.

Test S for normal model

For the normal model, the three sets that define Test S can be expressed as

$$\begin{aligned}
A_0 &= \{(t_1, t_2) : t_{\alpha,n_1-1} \le t_1, t_{\alpha,n_2-1} \le t_2\} \\
A_a &= \{(t_1, t_2) : F_2^{-1}(F_1(t_1) - \alpha/2) \le t_2 \le F_2^{-1}(F_1(t_1) + \alpha/2), \\
&\qquad 0 \le t_1 < t_{\alpha,n_1-1}, 0 \le t_2 < t_{\alpha,n_2-1}\} \\
A_b &= \{(t_1, t_2) : 0 \le t_2 \le F_2^{-1}(F_1(t_1) - 1/2 + 3\alpha/2), t_1 < t_{\alpha,n_1-1}\} \\
&\quad \cup \{(t_1, t_2) : F_2^{-1}(F_1(t_1) + 1/2 - 3\alpha/2) \le t_2 < t_{\alpha,n_2-1}, 0 \le t_1\}.
\end{aligned}$$

Test S rejects H_0 if $(T_1, T_2) \in A_0 \cup A_a \cup A_b$. Or, it may be simpler to let $U_i = F_i(T_i)$ and say Test S rejects H_0 if $(U_1, U_2) \in A_0 \cup A_a \cup A_b$, where A_0, A_a, and A_b are defined in (12).

The rejection region for Test S, for the case of $\alpha = .05$, $n_1 = 2$, and $n_2 = 40$, is shown in Figure 1. This might be compared with Figure 3a in Liu and Berger (1995) (although that figure is not exactly for this hypothesis). In that figure, the rejection region is centered on a straight line; here it is centered on the curved line $t_2 = F_2^{-1}(F_1(t_1))$.

In their Table 3, Liu and Berger (1995) discussed adjusting their rejection regions so the they had the biggest possible intersection. the point of the constructions of Test R and S is that this is not necessary. The same rejection region is used to test both H_{10} and H_{20}, and no points are lost in the intersection.

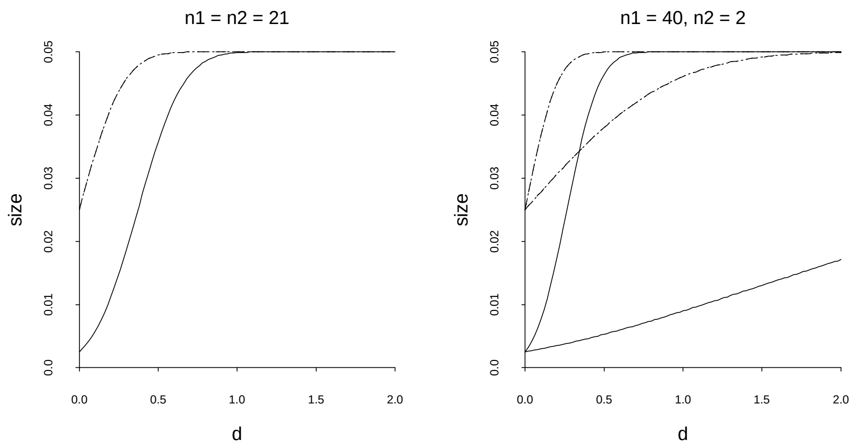

Figure 2: Size functions for three tests, Test R (dashed), Test S (dotted), and SIUT (solid). In left graph (equal sample sizes), size functions are the same on both boundaries. In right graph, upper lines are on $(\theta_1, 0)$ boundary, and lower lines are on $(0, \theta_2)$ boundary.

5 Power Comparison

We now compare the powers of three normal sign tests. Specifically we compare Tests R and S from Section 4 and the SIUT from Section 2.2. In our comparisons, $\alpha = .05$. We consider two sets of sample sizes, $n_1 = n_2 = 21$, and $n_1 = 40$, $n_2 = 2$. In both cases, the total sample size is 42. Thus, we will be able to see the effect of balanced versus unbalanced sample allocation on the powers. First we compare the size functions of the tests, then we compare the power functions on the alternative. All of these calculations were performed using IMSL Fortran routines for probability calculations and numeric integration.

5.1 Size comparison

The *size function* of a test is the power function on the boundary of H_0. Figure 2 shows the size function for Tests R, S, and SIUT. In these graphs, the size function is a dashed line for Test R, dotted line for Test S, and solid line for the SIUT. By (14), the size functions for R and S are, in fact, exactly equal in these cases because $(2\alpha)^{-1}$ is an integer. Indeed, for this normal problem, the size functions for both R and S are

$$\begin{aligned} P_{0,\theta_2}(\text{reject}) &= \alpha P_{\theta_2}(T_2 \geq 0) = P(Z \geq -\nu_2) = P(Z \geq -\sqrt{n_2}\theta_2/\sigma_2), \\ P_{\theta_1,0}(\text{reject}) &= \alpha P_{\theta_1}(T_1 \geq 0) = P(Z \geq -\nu_1) = P(Z \geq -\sqrt{n_1}\theta_1/\sigma_1), \end{aligned} \tag{15}$$

where Z has a standard normal distribution. We see that the size function for Tests R and S on the axis where θ_i is nonzero depends only on n_i, θ_i, and σ_i; it does not depend on the sample size or variance of the other population. Also, the size function is strictly increasing in n_i.

These graphs in Figure 2 show the size functions on the two boundaries of H_0, the $(\theta_1, 0)$ axis and the $(0, \theta_2)$ axis. In the equal sample sizes case, the size functions are the same on both axes, so there is just one set of graphs. In the unequal sample sizes case, The top dotted/dashed and solid lines are for the $(\theta_1, 0)$ axis, and the bottom lines are for the $(0, \theta_2)$ axis. Recall, θ_1 corresponds to the larger sample size, $n_1 = 40$, and θ_2 to the smaller sample size, $n_2 = 2$. In these graphs and in Figure 3, the horizontal axis is $d = (\theta_1^2 + \theta_2^2)^{1/2}$, the distance of the parameter point from the origin.

An unbiased test would have size function identically equal to α on the boundary. There is probably not a nonrandomized, unbiased test for this problem. (See Lehmann (1952) for a similar argument.) But, we would like the size function to rise to the value α as quickly as possible as d increases. We see that the size functions of Tests R and S increase much more rapidly than the size function of the SIUT.

5.2 Power comparison

Portions of the power functions for Tests R, S, and SIUT are shown in Figure 3. The left column contains graphs for the equal sample sizes case, and the right column contains graphs for the unequal sample sizes case. The graphs show the power functions on the three lines extending from the origin that make angles of $3\pi/8$, $\pi/4$, and $\pi/8$ with the θ_1 axis. Thus, the parameter points are of the form $(\theta_1, 2.41\theta_1)$ in the top graphs, of the form (θ_1, θ_1) in the middle graphs, and of the form $(\theta_1, .41\theta_1)$ in the bottom graphs. Again, the horizontal axis is $d = (\theta_1^2 + \theta_2^2)^{1/2}$, the distance of the parameter point from the origin.

In all cases, the power functions of Tests R and S are almost indistinguishable. There seems no practical reason to choose one of these tests over the other, based on these power functions. But, for every parameter point we considered, the computed power for Test R was equal to or greater than the computed power for Test S. Thus, Test R may have a theoretical advantage over Test S. This possibility deserves further investigation.

The powers of Tests R and S are seen to be much superior to the power of the SIUT, and, hence, also much superior to the power of the LRT, because the SIUT is uniformly more powerful than the LRT. The improvement is biggest near the origin, where the SIUT has poor power. For example, in the unequal sample size case, when $(\theta_1, \theta_2) = (.19, .46)$ ($d = .5$, top graph), the powers for Tests R and S are about .078, while the power for the SIUT is only .033, an improvement of 136%.

To compare the power functions in the equal and unequal sample sizes cases, note that different scales are used in the left and right columns of Figure 3. For all the parameter points shown and all three tests, the power function of a test appears to be the same or higher with equal sample sizes than with unequal sample sizes. For example, for $(\theta_1, \theta_2) = (.19, .46)$ again, Test R has power .170 with equal sample sizes and has power .078 with unequal sample sizes. It appears that for all these tests, choosing equal sample sizes will, in general, yield the best power. However, this power domination is not uniform over the whole alternative parameter space. The size functions for the unequal sample size Tests R and S are strictly bigger than the size functions for the equal sample size Tests R and S on the θ_1 axis. See Figure 2 and (15). So, because the power functions are continuous, there is a region in the alternative, near the θ_1 axis, where the powers of the unequal sample size tests are higher than the powers of the equal sample size tests.

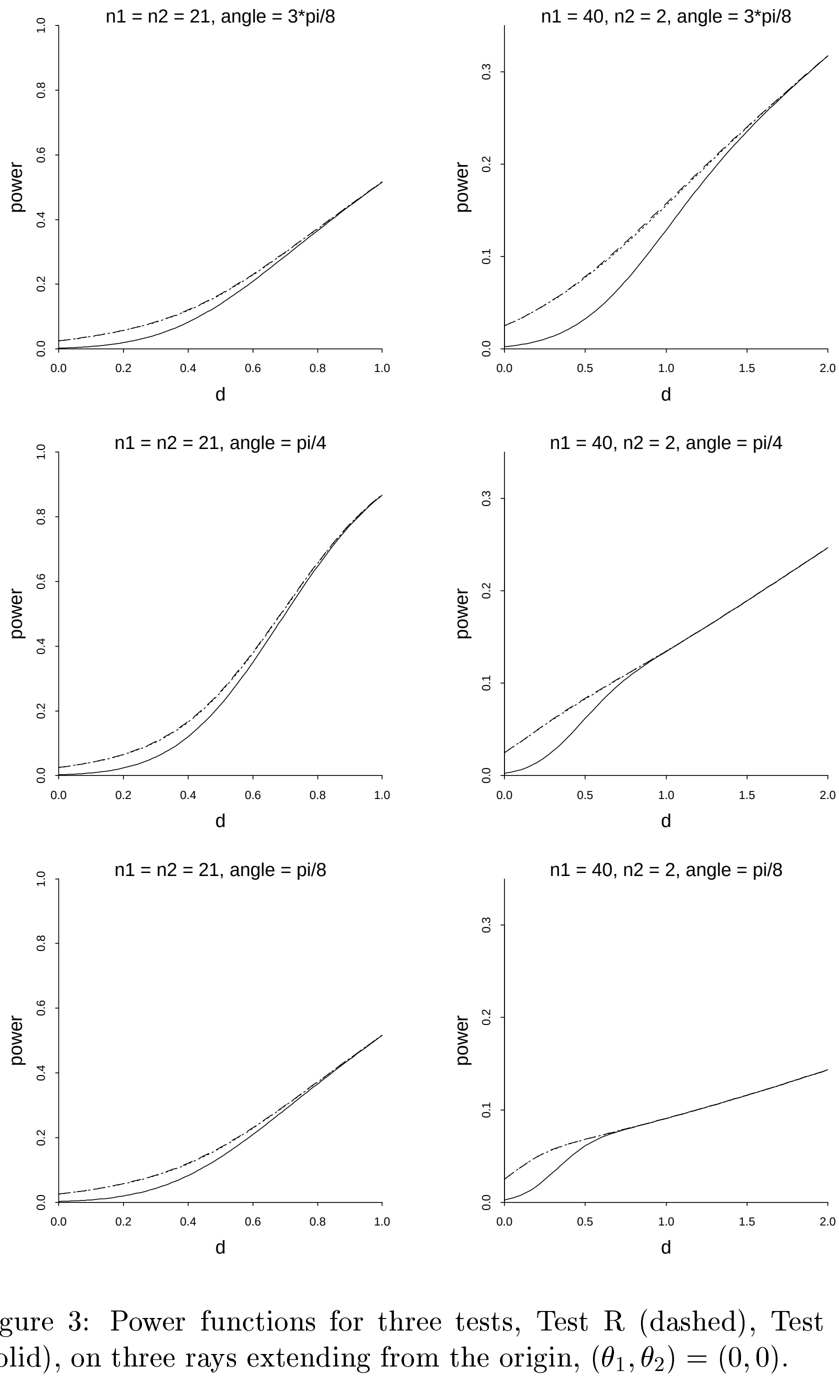

Figure 3: Power functions for three tests, Test R (dashed), Test S (dotted), and SIUT (solid), on three rays extending from the origin, $(\theta_1, \theta_2) = (0, 0)$.

6 Conclusion

In this paper, we have presented some general theory regarding improved tests for the sign testing problem. For normal populations, the new Tests R and S are shown to have much better power than the LRT and the simple intersection-union test. In the examples we considered, Tests R and S had very similar power.

References

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300.

Berger, R. L. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal means. *Journal of the American Statistical Association*, 84:192–199.

Berger, R. L. (1997). Likelihood ratio tests and intersection-union tests. In Panchapakesan, S. and Balakrishnan, N., editors, *Advances in Statistical Decision Theory and Applications*, pages 225–237. Birkhäuser, Boston.

Cohen, A., Gatsonis, C., and Marden, J. (1983). Hypothesis tests and optimality properties in discrete multivariate analysis. In Karlin, S., Amemiya, T., and Goodman, L. A., editors, *Studies in Econometrics, Time Series, and Multivariate Statistics*, pages 379–405. Academic Press, New York.

Gleser, L. J. (1973). On a theory of intersection-union tests (preliminary report). *IMS Bulletin*, 2:233.

Gutmann, S. (1987). Tests uniformly more powerful than uniformly most powerful monotone tests. *Journal of Statistical Planning and Inference*, 17:279–292.

IMSL (1995). *STAT/LIBRARY, FORTRAN Subroutines for Statistical Analysis, Version 3.0*. Visual Numerics, Houston.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Annals of Mathematical Statistics*, 23:541–552.

Liu, H. and Berger, R. L. (1995). Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. *Annals of Statistics*, 23:55–72.

Saikali, K. G. (1996). *Uniformly more powerful tests for linear inequalities*. PhD thesis, North Carolina State University Statistics Department.

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

Sasabuchi, S. (1988a). A multivariate one-sided test with composite hypotheses when the covariance matrix is completely unknown. *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, 42:37–46.

Sasabuchi, S. (1988b). A multivariate test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor. *Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics*, 42:9–19.

Shirley, A. G. (1992). Is the minimum of several location parameters positive? *Journal of Statistical Planning and Inference*, 31:67–79.

Wang, Y. and McDermott, M. P. (1996). Construction of uniformly more powerful tests for hypotheses about linear inequalities. Technical Report 96/05, University of Rochester Departments of Biostatistics and Statistics.