ELSEVIER

Journal of Statistical Planning and Inference 107 (2002) 187–205

journal of statistical planning and inference

www.elsevier.com/locate/jspi

More powerful tests for the sign testing problem

Khalil G. Saikali[a], Roger L. Berger[b,*]

[a] *SUGEN, 230 E. Grand Avenue, South San Francisco, CA 94080-4811, USA*
[b] *Department of Statistics, North Carolina State University, Box 8203 Raleigh, NC 27695-8203, USA*

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 H_0: $\mu_i \leqslant a_i$, for some $i = 1, \ldots, p$, versus 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 $0 < \alpha < \frac{1}{2}$, 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; Min test; Normal population; t distribution

1. Introduction

Let X_{ij}, $i = 1, \ldots, p$ ($p \geqslant 2$), $j = 1, \ldots, n_i$ (each $n_i \geqslant 2$), 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

$$H_0: \mu_i \leqslant a_i \quad \text{for some } i = 1, \ldots, p$$

* Corresponding author. Tel.: +1-919-515-1914; fax: +1-919-515-1169.
E-mail address: berger@stat.ncsu.edu (R.L. Berger).

0378-3758/02/$ - see front matter
PII: S0378-3758(02)00252-5

versus

$$H_1: \mu_i > a_i \quad \text{for all } i = 1, \ldots, p. \tag{1}$$

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,b) 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 $0 < \alpha < \frac{1}{2}$, 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. Finally, some of the new tests we describe have some counterintuitive properties. We conclude by summarizing the advantages and disadvantages of the various 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, $Y_{ij} \sim n(\theta_i, \sigma_i^2)$, where $\theta_i = \mu_i - a_i$, and testing (1) is equivalent to testing

$$H_0: \theta_i \leqslant 0 \quad \text{for some } i = 1, \ldots, p$$

versus

$$H_1: \theta_i > 0 \quad \text{for all } i = 1, \ldots, p. \tag{2}$$

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

$$H_0: \min_{1 \leqslant i \leqslant p} \{\theta_i\} \leqslant 0,$$

versus

$$H_1: \min_{1 \leqslant i \leqslant p} \{\theta_i\} > 0.$$

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 $\bar{Y}_1, \ldots, \bar{Y}_p$, the sample means, and $S_1^2, \ldots, S_p^2$, the sample variances defined by $S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/(n_i - 1)$. 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

$$\mathrm{H}_0: \bigcup_{i=1}^{p} \{\theta_i \leqslant 0\}$$

versus

$$\mathrm{H}_1: \bigcap_{i=1}^{p} \{\theta_i > 0\}.$$

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

$$\mathrm{H}_{i0}: \theta_i \leqslant 0$$

versus

$$\mathrm{H}_{i1}: \theta_i > 0. \tag{3}$$

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

$$\lambda_i(\boldsymbol{Y}) = \frac{\sup_{\mathrm{H}_{i0}} L(\theta_1, \ldots, \theta_p, \sigma_1^2, \ldots, \sigma_p^2 | \boldsymbol{Y})}{\sup_{\mathrm{H}_{i0} \cup \mathrm{H}_{i1}} L(\theta_1, \ldots, \theta_p, \sigma_1^2, \ldots, \sigma_p^2 | \boldsymbol{Y})}$$

$$= \begin{cases} 1 & \text{if } \bar{Y}_i \leqslant 0, \\ \left(1 + \dfrac{T_i^2}{n_i - 1}\right)^{-n_i/2} & \text{if } \bar{Y}_i > 0, \end{cases}$$

where

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

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 $\lambda_i(\boldsymbol{Y}) \leqslant c_{i\alpha}$, where $c_{i\alpha}$ is an appropriately chosen cutoff value. When $\theta_i = 0$, T_i has a Student's t distribution. Because $0 < \alpha < \frac{1}{2}$,

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

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 $T_i \geqslant t_{\alpha,n_i-1}$, the usual one-sided t test.

Using (15.3) in Berger (1997), the LRT statistic for testing H_0 is $\lambda(\boldsymbol{Y}) = \max_{1\leqslant i\leqslant p} \lambda_i(\boldsymbol{Y})$, and the size-$\alpha$ LRT rejects H_0 if $\lambda(\boldsymbol{Y}) \leqslant c_\alpha$. By Theorem 15.2.2 in Berger (1997), the cutoff value c_α that yields a size-α test is $c_\alpha = \min_{1\leqslant i\leqslant p} c_{i\alpha}$. For each $\alpha = 0.01, 0.05$, and 0.10, 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 $\alpha, c_{i\alpha}$ 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 $n_{(1)} = \min_{1\leqslant i\leqslant p} n_i$. Then the size-$\alpha$ LRT of (2) rejects H_0 if, for every $i = 1, \ldots, p$,

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

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, 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 $i = 1, \ldots, p$,

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

is a size-α test of H_0 versus H_1. We will call this test the simple IUT (SIUT). Theorems 15.1.1 and 15.1.2 of Berger (1997) in fact show that the SIUT is a size-α test of H_0 versus H_1 even if the statistics $T_1, \ldots, T_p$ are not independent. Thus the SIUT can be used for much more general models than those considered here. The two new tests we will describe in Section 4, however, do need the independence assumption. Authors such as Laska and Meisner (1989) have called the SIUT the "min test".

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

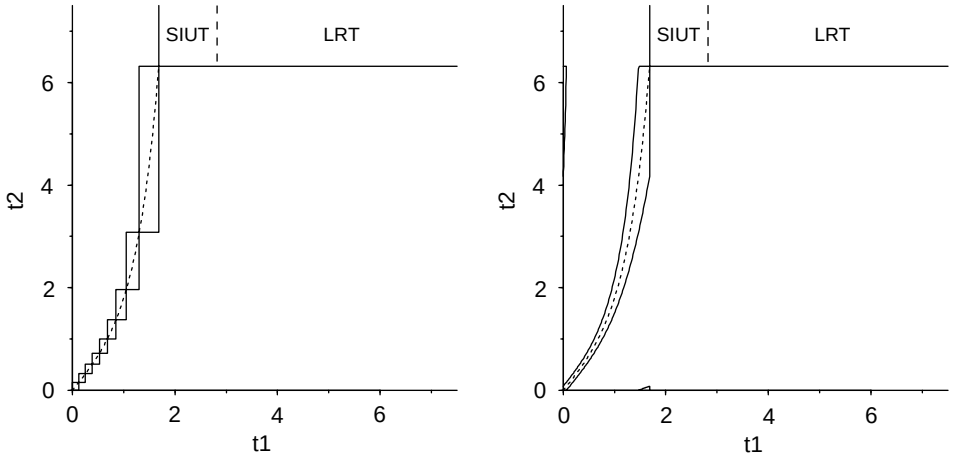

Fig. 1. Rejection regions for Tests R (left graph) and S (right graph) for testing normal means, $n_1 = 40$, $n_2 = 2$, $\alpha = 0.05$.

case $p = 2$, $n_1 = 40$, $n_2 = 2$, and $\alpha = 0.05$, the rejection regions for the LRT and the SIUT are shown in Fig. 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 $\{t_1 \geqslant 1.68, t_2 \geqslant 6.31\}$. The rejection region of the LRT is the smaller rectangle in the upper right corner, bounded by solid and dashed lines, defined by $\{t_1 \geqslant 2.82, t_2 \geqslant 6.31\}$. (Fig. 1 will be discussed more in Section 4.) For the models considered by Sasabuchi (1980, 1988a, 1988b), Laska and Meisner (1986, 1989), and Laska et al. (1992), the LRT and the SIUT are the same test. This is verified in the latter three articles. So, it is important to note that in our simple model, if the sample sizes are unequal, these two tests differ, and the SIUT is uniformly more powerful. In our opinion, the SIUT is also easier to implement and easier to understand.

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 $t_i' \geqslant t_i$ 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 $v_i = \sqrt{n_i}\theta_i/\sigma_i$. Testing (2) is equivalent to testing

$$\mathrm{H}_0: \bigcup_{i=1}^{p} \{v_i \leqslant 0\}$$

versus

$$\mathrm{H}_1: \bigcap_{i=1}^{p} \{v_i > 0\}. \tag{7}$$

Because the noncentral t distribution has monotone likelihood ratio in the noncentrality parameter, Theorem 3.3 in Cohen et al. (1983), yields that the SIUT is the UMP, size-α, monotone test. Lehmann (1952), Laska and Meisner (1986), Gutmann (1987), Laska

and Meisner (1989), and Laska et al. (1992) have all discussed UMP monotone tests for various models.

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

$$H_0: \boldsymbol{\theta} \in \bigcup_{i=1}^{p} \Theta_{i0}$$

versus

$$H_1: \boldsymbol{\theta} \in \bigcap_{i=1}^{p} \Theta_{i0}^c, \tag{8}$$

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 $H_{i0}: \boldsymbol{\theta} \in \Theta_{i0}$ versus $H_{i1}: \boldsymbol{\theta} \in \Theta_{i0}^c$. Then, by Theorem 15.1.1 in Berger (1997) (cf., Theorem 1, Berger (1982)), the test with rejection region $R = \bigcap_{i=1}^{p} R_i$ 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 $H_{(ij)0}: \boldsymbol{\theta} \in \Theta_i \cup \Theta_j$ versus $H_{(ij)1}: \boldsymbol{\theta} \in \Theta_i^c \cap \Theta_j^c$. 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

$$\mathrm{H}_0\colon \boldsymbol{\theta} \in \bigcup_{i=1}^{p/2} \{\Theta_{(2i-1)0} \cup \Theta_{(2i)0}\}$$

versus

$$\mathrm{H}_1\colon \boldsymbol{\theta} \in \bigcap_{i=1}^{p/2} \left\{\Theta^c_{(2i-1)0} \cap \Theta^c_{(2i)0}\right\}.$$

If $R_{2i-1,2i}$ is the rejection region of a size-α test of $\mathrm{H}_{2i-1,2i,0}\colon \boldsymbol{\theta} \in \{\Theta_{(2i-1)0} \cup \Theta_{(2i)0}\}$, then $R = \bigcap_{i=1}^{p/2} R_{2i-1,2i}$ 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

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

From this it is clear that any size-α rejection region for testing $\mathrm{H}_{p0}\colon \boldsymbol{\theta} \in \Theta_p$ 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

$$\sup_{H_0} P(R) = \sup_{H_{10} \cup H_{20}} P(R) = \max \left\{ \sup_{H_{10}} P(R), \sup_{H_{20}} P(R) \right\}.$$

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 v_1 and v_2 be two real-valued parameters. We consider testing

$$H_0\colon \{v_1 \leqslant 0\} \cup \{v_2 \leqslant 0\}$$

versus

$$H_1\colon \{v_1 > 0\} \cap \{v_2 > 0\}.$$

We assume the parameter space is a rectangle of the form $\{(v_1, v_2)\colon v_1^L < v_1 < v_1^U, v_2^L < v_2 < v_2^U\}$. The endpoints v_i^L and v_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 v_i. T_1 and T_2 are independent. The statistic T_i is a test statistic for testing $H_{i0}\colon v_i \leqslant 0$, and large values of T_i give evidence that $H_{i1}\colon v_i > 0$ is true. Let $F_i(t)$ denote the cumulative distribution function of T_i when $v_i = 0$, and let $f_i(t|v_i)$ denote the density of T_i. Let $m_i = F_i^{-1}(1/2)$ denote the median of T_i when $v_i = 0$. 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

$$T_i \text{ converges in probability to } t_i^U \text{ as } v_i \to v_i^U. \tag{9}$$

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

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

This property allows us to check the size of a test by considering only the boundary value $v_i = 0$ in this way. Suppose R is a set of (t_1, t_2) values such that $t_1 \geqslant m_1$ for all $(t_1, t_2) \in R$. Define $R(t_2) = \{t_1\colon (t_1, t_2) \in R\}$. Note that $R(t_2) \subset \{t_1\colon t_1 \geqslant m_1\}$. Then for any $(v_1, v_2) \in H_{10}$, that is with $v_1 \leqslant 0$,

$$\begin{aligned} P_{v_1,v_2}((T_1, T_2) \in R) &= \int_{-\infty}^{\infty} \int_{R(t_2)} f_1(t_1|v_1)\,dt_1\, f_2(t_2|v_2)\,dt_2 \\ &= \int_{-\infty}^{\infty} P_{v_1}(T_1 \in R(t_2)) f_2(t_2|v_2)\,dt_2 \\ &\leqslant \int_{-\infty}^{\infty} P_0(T_1 \in R(t_2)) f_2(t_2|v_2)\,dt_2 \\ &= P_{0,v_2}((T_1, T_2) \in R). \end{aligned} \tag{11}$$

Thus, if $P_{0,v_2}((T_1, T_2) \in R) \leqslant \alpha$, for all v_2, R is a level-α rejection region for testing H_{10}. By a similar argument, if $R \subset \{(t_1, t_2)\colon t_2 \geqslant m_2\}$, then we need check only parameter values of the form $(v_1, 0)$ 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 $T_i \geqslant c_{i,\alpha}$ is a size-α test of H_{i0}. Hence, the test that rejects H_0 if $T_1 \geqslant c_{1,\alpha}$ and $T_2 \geqslant c_{2,\alpha}$ 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 $0 < \alpha < \frac{1}{2}$, define the integer J by the inequality $J-1 < (2\alpha)^{-1} \leqslant J$. 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 $c_J^i = F_i^{-1}(\frac{1}{2}) = m_i$. (Note, if $(2\alpha)^{-1}$ is an integer, as it is for $\alpha = 0.10$, 0.05, and 0.01, then $c_J^i = F_i^{-1}(1-J\alpha)$.) For $j=1,\ldots,J$, define

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

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 $(0, v_2)$. Using the notation in (11) we have

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

Therefore,

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

Calculating as in (11), for any value of v_2 we have

$$\begin{aligned} P_{0,v_2}((T_1,T_2) \in R) &= \int_{-\infty}^{\infty} P_0(T_1 \in R(t_2)) f_2(t_2|v_2)\,\mathrm{d}t_2 \\ &= \int_{c_J^2}^{\infty} P_0(T_1 \in R(t_2)) f_2(t_2|v_2)\,\mathrm{d}t_2 \\ &\leqslant \int_{c_J^2}^{\infty} \alpha f_2(t_2|v_2)\,\mathrm{d}t_2 \\ &= \alpha P_{v_2}(T_2 \geqslant c_J^2) \\ &\leqslant \alpha. \end{aligned}$$

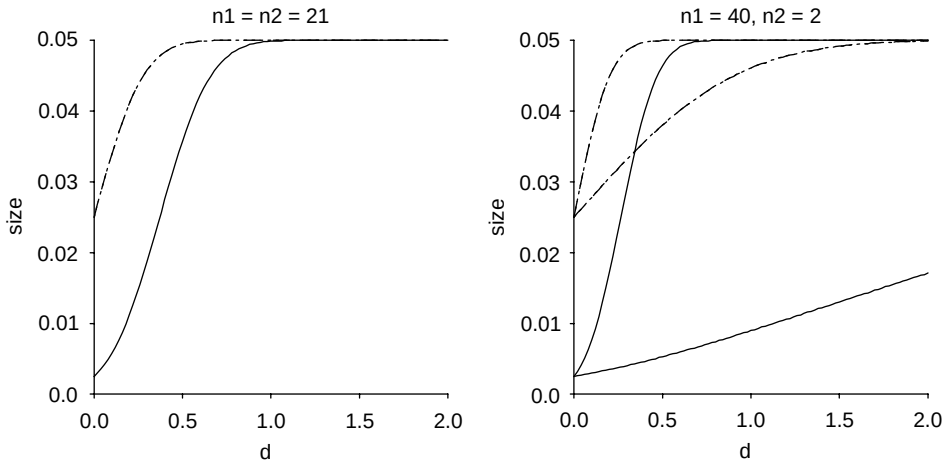

Fig. 2. Size functions for three tests, Test R (dashed), Test S (dotted), and SIUT (solid). In left graph (equal sample sizes), size functions are the same on both boundaries. In right graph, upper lines are on $(\theta_1, 0)$ boundary, and lower lines are on $(0, \theta_2)$ boundary.

(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 Fig. 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 (Fig. 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

$$A_0 = \{(u_1, u_2)\colon 1 - \alpha \leqslant u_1 \leqslant 1, 1 - \alpha \leqslant u_2 \leqslant 1\},$$

$$A_a = \{(u_1, u_2)\colon |u_1 - u_2| \leqslant \alpha/2, 1/2 \leqslant u_1 < 1 - \alpha, 1/2 \leqslant u_2 < 1 - \alpha\},$$

$$A_b = \{(u_1, u_2)\colon 1/2 \leqslant u_2 \leqslant u_1 - 1/2 + 3\alpha/2, u_1 < 1 - \alpha\}$$
$$\cup \{(u_1, u_2)\colon 1/2 \leqslant u_1 \leqslant u_2 - 1/2 + 3\alpha/2, u_2 < 1 - \alpha\}. \quad (12)$$

The set A has this important property. Let $A(u_2) = \{u_1\colon (u_1, u_2) \in A\}$. If $u_2 < \frac{1}{2}$, $A(u_2) = \emptyset$. If $\frac{1}{2} \leqslant u_2 \leqslant 1$, $A(u_2)$ 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,

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

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 $P((U_1,U_2)\in A)=\alpha P(1/2\leqslant U_1\leqslant 1)\leqslant\alpha$.

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

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

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 $(u_1,u_2)\in A$, $u_1\geqslant\frac{1}{2}$ and $u_2\geqslant\frac{1}{2}$. Thus, for every $(t_1,t_2)\in R_S$, $F_1(t_1)\geqslant\frac{1}{2}$ and $F_2(t_2)\geqslant 1/2$, that is, $t_1\geqslant m_1$ and $t_2\geqslant m_2$, and we can use (10). Finally, if $v_i=0$, 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 Fig. 1.

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

$$P_{0,v_2}((T_1,T_2)\in R)=\alpha P_{v_2}(T_2\geqslant m_2)=P_{0,v_2}((T_1,T_2)\in R_S),$$

$$P_{v_1,0}((T_1,T_2)\in R)=\alpha P_{v_1}(T_1\geqslant m_1)=P_{v_1,0}((T_1,T_2)\in R_S). \tag{14}$$

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 $P_1\leqslant\alpha$ and $P_2\leqslant\alpha$. 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 $j=1,\ldots,J-1$,

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

and

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

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 $v_i = \sqrt{n_i}\theta_i/\sigma_i$. Hypotheses (2) about the normal means can be equivalently expressed 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 v. Let A be a subset of $\{t: t \geqslant 0\}$. Then for any $v \leqslant 0$, $P_v(T \in A) \leqslant P_0(T \in A)$.*

Proof. The random variable $X/\sqrt{V/r}$ has the same distribution as T if X and V are independent, $X \sim n(v, 1)$, 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,

$$P_v(T \in A) = P_v(X/\sqrt{V/r} \in A) = \int_0^\infty P_v(X \in A(v)) f(v)\,\mathrm{d}v.$$

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

$$P_v(T \in A) \leqslant \int_0^\infty P_0(X \in A(v)) f(v)\,\mathrm{d}v = P_0(T \in A),$$

as was to be shown. □

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

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

and

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

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 $\alpha = 0.05$, $n_1 = 2$, and $n_2 = 40$, is shown in Fig. 1. This might be compared with Fig. 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

$$A_0 = \{(t_1, t_2)\colon t_{\alpha, n_1 - 1} \leqslant t_1, t_{\alpha, n_2 - 1} \leqslant t_2\},$$

$$A_a = \{(t_1, t_2)\colon F_2^{-1}(F_1(t_1) - \alpha/2) \leqslant t_2 \leqslant F_2^{-1}(F_1(t_1) + \alpha/2), \\ 0 \leqslant t_1 < t_{\alpha, n_1 - 1}, 0 \leqslant t_2 < t_{\alpha, n_2 - 1}\},$$

$$A_b = \{(t_1, t_2)\colon 0 \leqslant t_2 \leqslant F_2^{-1}(F_1(t_1) - 1/2 + 3\alpha/2), t_1 < t_{\alpha, n_1 - 1}\} \\ \cup \{(t_1, t_2)\colon F_2^{-1}(F_1(t_1) + 1/2 - 3\alpha/2) \leqslant t_2 < t_{\alpha, n_2 - 1}, 0 \leqslant t_1\}.$$

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 $\alpha = 0.05$, $n_1 = 2$, and $n_2 = 40$, is shown in Fig. 1. This might be compared with Fig. 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 $t_2 = F_2^{-1}(F_1(t_1))$. The portion of the rejection region corresponding to the set A_b consists of the two tiny triangles near $(1.6, 0)$ and $(0, 5.5)$ in Fig. 1. These two regions are needed to make all the cross-sectional probabilities exactly equal to α. But, these two disconnected regions could be deleted from the rejection region without appreciably changing the power comparisons in Section 5.

In their Table 3, Liu and Berger (1995) discussed adjusting their rejection regions so that 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.

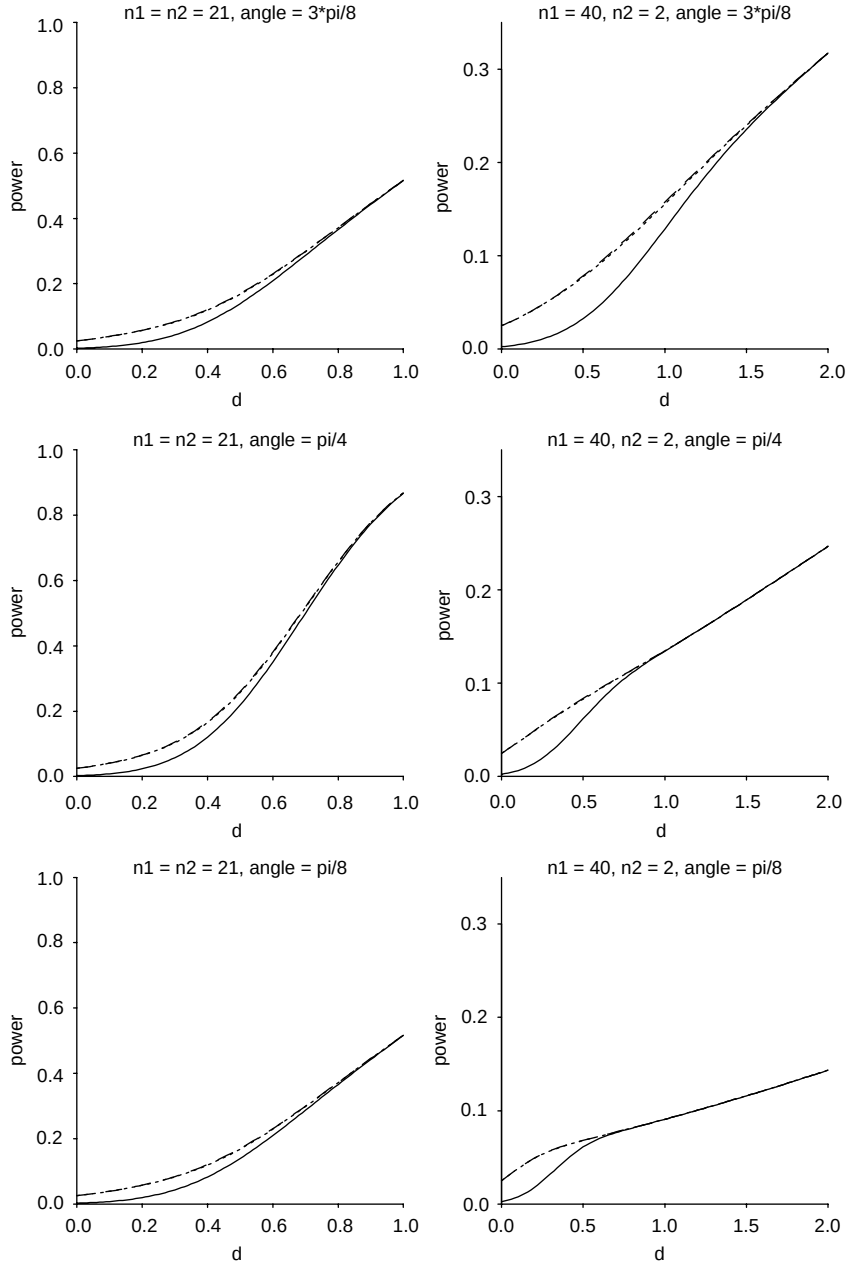

Fig. 3. Power functions for three tests, Test R (dashed), Test S (dotted), and SIUT (solid), on three rays extending from the origin, $(\theta_1, \theta_2) = (0, 0)$.

5. Power comparison

We now compare the powers of three normal sign tests, Tests R and S from Section 4 and the SIUT from Section 2.2. In our comparisons, $\alpha = 0.05$ and $\sigma_1^2 = \sigma_2^2 = 1$. 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. Note, all calculations and comparisons in this section are only for the equal variance case. Separate calculations would be needed for unequal variance cases.

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 1995 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. Fig. 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

$$P_{0,\theta_2}(\text{reject}) = \alpha P_{\theta_2}(T_2 \geqslant 0) = \alpha P(Z \geqslant -v_2) = \alpha P(Z \geqslant -\sqrt{n_2}\theta_2/\sigma_2),$$

$$P_{\theta_1,0}(\text{reject}) = \alpha P_{\theta_1}(T_1 \geqslant 0) = \alpha P(Z \geqslant -v_1) = \alpha P(Z \geqslant -\sqrt{n_1}\theta_1/\sigma_1), \qquad (15)$$

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 Fig. 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 Fig. 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 Fig. 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 $(\theta_1, 0.41\theta_1)$ 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 $(\theta_1, \theta_2) = (0.19, 0.46)$ ($d = 0.5$, top graph), the powers for Tests R and S are about 0.078, while the power for the SIUT is only 0.033, 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 Fig. 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 $(\theta_1, \theta_2) = (0.19, 0.46)$ again, Test R has power 0.170 with equal sample sizes and has power 0.078 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 Fig. 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.

6. Discussion of tests

For the problem of testing (1) with normal data, we have discussed the LRT, the SIUT, and Tests R and S. The SIUT is the same as or uniformly more powerful than the LRT. Tests R and S are uniformly more powerful than the SIUT. Nevertheless, some authors believe that the SIUT or even the LRT should be used. We will now briefly discuss some of the advantages and disadvantages of these four tests. The discussion naturally divides into two comparisons, the SIUT versus the LRT and Tests R and S versus the first two.

If the sample sizes are all equal, the LRT and the SIUT are the same test. But, if the sample sizes are not all equal, these tests are different. Both tests have the same size, α, but the SIUT is uniformly more powerful than the LRT. The SIUT

is an intuitive combination of the univariate t tests and is the uniformly most powerful, level-α monotone test. (A further advantage of the SIUT, which we have not explored in this article, is that it can be conducted in a stepwise manner, with some of the individual hypotheses H_{i0} being rejected while others are not. See Koch and Gansky (1996) for a discussion of this stepwise implementation of the SIUT.). Perlman and Wu (1999b) agree that the SIUT is preferable to LRT. Their reason for this is that the SIUT represents the statistical evidence better as indicated by Berger (1999, discussion of Perlman and Wu), but not because the SIUT is more powerful than the LRT.

Tests R and S provide a more substantial increase in power over the LRT and SIUT. However, their rejection regions have features that some find objectionable. The advantage of these tests is, of course, their increased power over the SIUT and the LRT. Having controlled the Type I error probability at α, we think a major concern should be in increasing the power on the alternative. As Fig. 3 indicates, the major increase in power is at parameter points where all tests have low power. Some claim that such increases are inconsequential. It is true that in some experiments it is possible to choose the sample sizes so that the LRT has power of 80% or 90% at alternatives of interest. At these alternatives the other tests will not have much higher power. But, there are certainly situations in which constraints of time and/or money preclude obtaining such large samples. When small samples must be used, the LRT may have low power (even less than α) at alternatives of interest, and we would think the increase in power provided by Tests R and S would be valuable and welcomed.

A major objection offered to tests like R and S is that their rejection regions are not monotone. For example, Laska and Meisner (1989) discuss difficulties in using nonmonotone tests in a regulatory environment. On the one hand, we can appreciate a regulatory difficulty such as Laska and Meisner describe. On the other hand, every nonmonotone test does not have a monotone test that dominates it in terms of standard criteria such as power. Thus, the two criteria are in conflict, and a user must decide which is more important in a given setting.

Perlman and Wu (1999a, b) criticize tests like S and R in another way. They assert that these tests do not properly assess "the fit of the data to the null hypothesis relative to the alternative hypothesis", because the rejection regions of these tests include points that are arbitrarily close to $(0,0)$. But, it is by adding such points that Tests R and S achieve higher power. So, again, the Perlman and Wu criterion and the power criterion are in conflict, and a user must decide which is more important in a given setting.

Another feature of Test S is that its rejection region is not connected. This may seem undesirable or unacceptable. However, we are not aware of any results that suggest that nonconnected rejection regions should not be used. But, as mentioned in Section 4, the set A_b could be deleted from the rejection region without any appreciable loss of power. Then the rejection region of Test S would be connected.

We think the existence of tests like R and S that have the same size and are uniformly more powerful than the LRT or SIUT should at least challenge statisticians to elucidate a theory that justifies requiring that a test of (1) to be monotone or connected or to satisfy other criteria.

7. Conclusion

In this paper, we have presented some general theory regarding improved tests for the sign testing problem. For normal populations, the SIUT is an intuitive and uniformly most powerful monotone test that has higher power than the LRT. The new Tests R and S are shown to have much better power than the LRT and the SIUT. In the examples we considered, Tests R and S had very similar power. But, Tests R and S have unusually shaped rejection regions. Thus, in considering these tests, one must weigh the advantage of their increased power versus what some consider the nonintuitive shapes of their rejection regions.

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. J. Amer. Statist. Assoc. 84, 192–199.

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

Berger, R.L., 1999. Comment on "The emperor's new tests". Statist. Sci. 14, 370–373.

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

Gleser, L.J., 1973. On a theory of intersection–union tests (preliminary report) IMS Bull. 2, 233.

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

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

Koch, G.G., Gansky, S.A., 1996. Statistical considerations for multiplicity in confirmatory protocols. Drug Inform. Journal 30, 523–534.

Laska, E.M., Meisner, M.J., 1986. Testing whether an identified treatment is best: the combination problem. In: American Statistical Association Proceedings of the Biopharmaceutical Section, American Statistical Association, Washington, DC, pp. 163–170.

Laska, E.M., Meisner, M.J., 1989. Testing whether an identified treatment is best. Biometrics 45, 1139–1151.

Laska, E.M., Tang, D.-I., 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., 1952. Testing multiparameter hypotheses. Ann. Math. Statist. 23, 541–552.

Liu, H., Berger, R.L., 1995. Uniformly more powerful, one-sided tests for hypotheses about linear inequalities. Ann. Statist. 23, 55–72.

Perlman, M.D., Wu, L., 1999a. The emperor's new tests. Statist. Sci. 14, 355–369.

Perlman, M.D., Wu, L., 1999b. Rejoinder to the discussion of The emperor's new tests. Statist. Sci. 14, 377–381.

Saikali, K.G., 1996. Uniformly more powerful tests for linear inequalities. Ph.D. 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. Mem. Fac. Sci. Kyushu Univ. Ser. A Math. 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. Mem. Fac. Sci. Kyushu Univ. Ser. A Math. 42, 9–19.

Shirley, A.G., 1992. Is the minimum of several location parameters positive? J. Statist. Plann. Inference 31, 67–79.

Wang, Y., 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.