The Annals of Statistics
1991, Vol. 19, No. 2, 918–934

SENSITIVE AND STURDY p-VALUES[1]

BY JOHN I. MARDEN

University of Illinois at Urbana-Champaign

We introduce new criteria for evaluating test statistics based on the p-values of the statistics. Given a set of test statistics, a good statistic is one which is robust in being reasonably sensitive to all departures from the null implied by that set. We present a constructive approach to finding the optimal statistic. We apply the criteria to two-sided problems; combining independent tests; testing that the mean of a spherical normal distribution is 0, and extensions to other spherically symmetric and exponential distributions; Bartlett's problem of testing the equality of several normal variances; and testing for one outlier in a normal linear model. For the most part, the optimal statistic is quite easy to use. Often, but not always, it is the likelihood ratio statistic.

1. Introduction. In most hypothesis testing situations, especially multidimensional ones, there are many plausible testing procedures from which to choose. Popular criteria used to judge procedures include consistency, admissibility, Bayesness, robustness, unbiasedness, minimaxity, local optimality, Bahadur exact slope and Pitman efficiency. Often, even after applying several of the above criteria, there remains a number of attractive procedures. In this paper we introduce new criteria which depend only on the null distribution, distinguishing them from the criteria mentioned above, whose use can substantially narrow the search for a single procedure.

We start with a set $\mathbf{T}$ of desirable test statistics, each one sensitive to particular alternatives. This set can be chosen on objective (e.g., admissibility) or heuristic grounds. Our goal is to find a statistic which will be robust among those in $\mathbf{T}$ in the sense of being reasonably sensitive to all departures from the null as implied by $\mathbf{T}$.

We have in mind certain types of hypothesis testing situations. To simplify, we think of three basic types. In the first, the result of the hypothesis test is the main conclusion of the study, for example, whether a drug works or whether a company should sponsor employee support groups. It is reasonable to expect the researcher to work hard to design a test procedure appropriate for the situation, being careful to incorporate whatever prior information about the structure of the alternative space is at hand. Bayesian methods in particular can work well. In the second situation, there is a large but finite-dimensional set of possible effects, and one wishes an overall test to decide whether it is worthwhile investigating these effects more closely. An example

Received October 1989; revised June 1990.
[1]Partial support provided by NSF DMS-88-02556 and AFOSR-87-0041.
AMS 1980 *subject classifications*. Primary 62F03, 62C20, 62F04; secondary 62H15, 62C15.
Key words and phrases. Hypothesis tests, p-values, robustness, meta-analysis, Fisher's procedure, normal distribution, spherical symmetry, exponential family, outliers, Bartlett's problem.

is in analysis of variance, where one tests for high-order interaction, or equality of variances, and would not explore further unless a test showed significance. Goodness-of-fit tests encompass the third situation, where one desires an assessment of whether the model in use gives a reasonable fit to the data. The alternative tends to be very diffuse, often infinite dimensional. There may be some general directions of lack of fit to which one wishes sensitivity, but little knowledge of the exact alternative structure.

This paper is geared toward the latter two situations. We evaluate test statistics via their *p-values*. We prefer the p-value approach to the usual accept–reject paradigm for such situations since the test statistics are to be used to provide directions during a statistical analysis and assurances after the analysis, rather than carrying the weight of the main conclusion of the study. Furthermore, when analyzing a set of data, one is confronted with having to iterate among testing hypotheses, estimating parameters and checking models. It is not unusual to have testing situations of the second and third kind presenting themselves at the rate of several per hour. Thus it is not necessary or appropriate to prespecify a level for each test.

Frequentist notions in general, and p-values in particular, have generated quite a bit of controversy lately. Casella and Berger (1987) and Berger and Sellke (1987) have debated the appropriateness of p-values as evidence against the null. It is clear that it is dangerous to misinterpret a p-value as the probability that the null hypothesis is true. As all frequentist measures, its properties refer to results occurring as the experiment is repeated. In addition, an outcome which is very unlikely under the null might be even more unlikely under the alternative; hence low p-values should not automatically lead one to embrace the alternative. Despite these objections, the p-value does have a precise meaning, being the probability of obtaining, under the null, a result as or more extreme. And when choosing which procedures to use, it is important to balance philosophical comfort with temporal constraints. For example, the Bayesian approach provides a coherent formulation for inference, but for many aspects of a data analysis, it is not worth the time and trouble to conjure up a prior or class of priors and grind through to the posterior. The p-values provide valuable and convenient guidance, and careful researchers will keep in mind their deficiencies.

To present our definitions, we assume we have the set $\mathbf{T}$, a set of real-valued test statistics, along with the knowledge of whether to reject the null hypothesis for large or for small values of $T \in \mathbf{T}$. We will make the simplifying assumptions that each statistic has a continuous distribution F_T under the null, and that this distribution does not depend on which element of the null obtains. Then the p-value for T in $\mathbf{T}$, $p_T(x)$, is simply $1 - F_T(T(x))$ if rejecting for large values of T, or $F_T(T(x))$ if rejecting for small values, where x is the observation. Under the null, the p-value has a uniform $(0, 1)$ distribution.

Consider two test statistics U and V, both functions of the data x, and their respective p-values $p_U(x)$ and $p_V(x)$. Suppose the p-value for U is observed to be α, and consider the possible values p_V could have. If, generally, these values tend to hover around α, then one would be equally comfortable with using U or V. However, if p_V tends to be quite a bit larger than α, then one would prefer to use U, or if p_V tends to be much smaller than α, one would prefer V. There are many ways to quantify the tendencies, such as means, standard deviations, quantiles, etc. We consider two values, the smallest and largest possible values, respectively,

$$i(U; V, \alpha) = \inf_x \{p_U(x) | p_V(x) \geq \alpha\} \tag{1.1}$$

and

$$s(U; V, \alpha) = \sup_x \{p_U(x) | p_V(x) < \alpha\}. \tag{1.2}$$

[For technical reasons, we have " $\geq$ " and " $<$ " in the definitions rather than equalities. In many applications, the extrema occur on the sets where $p_V(x) = \alpha$.] These measures are admittedly rather coarse, but other measures depend in a complicated way on which alternative obtains. We expect that statistics which perform well according to (1.1) and (1.2) will have good overall behavior.

We have two related measures of robustness. They are both functions of α and $\mathbf{T}$, as well as the statistic under consideration.

DEFINITION 1.1. The *sensitivity* of the statistic T is $\mathrm{Se}(T; \alpha) = \inf\{i(U; T, \alpha) | U \in \mathbf{T}\}$, and the *sturdiness* is $\mathrm{St}(T; \alpha) = \sup\{s(T; U, \alpha) | U \in \mathbf{T}\}$.

Thus Se gives the smallest possible p-value which could have arisen from the set $\mathbf{T}$ when the p-value for T is α, and St gives the largest p-value T could yield given that some statistic in $\mathbf{T}$ has p-value α. A statistic is sensitive if Se is not too much smaller than α, so that when $p_T(x) = \alpha$, one can feel confident that no other statistic would be much more significant. A statistic is sturdy if, when it is possible to achieve a p-value of α, $p_T(x)$ remains fairly close to α. These two notions are almost inverses. In Section 2 we show that

$$\mathrm{Se}(T, \alpha) = \beta \Rightarrow \lim_{\beta' \uparrow \beta} \mathrm{St}(T, \beta') \leq \alpha \leq \lim_{\beta' \downarrow \beta} \mathrm{St}(T, \beta') \tag{1.3}$$

and

$$\mathrm{St}(T, \beta) = \alpha \Rightarrow \lim_{\alpha' \uparrow \alpha} \mathrm{Se}(T, \alpha') \leq \beta \leq \lim_{\alpha' \downarrow \alpha} \mathrm{Se}(T, \alpha'). \tag{1.4}$$

In many cases, one can use (1.3) and (1.4) to determine Se from St or vice versa.

For any T,

$$\mathrm{Se}(T, \alpha) \leq \alpha \quad \text{and} \quad \mathrm{St}(T, \alpha) \geq \alpha. \tag{1.5}$$

We are interested in the robustnesses both qualitatively and quantitatively. We will say that a statistic T is *insensitive* at α if $\mathrm{Se}(T, \alpha) = 0$, and *fragile* at α if $\mathrm{St}(T, \alpha) = 1$. On the positive side, $T \in \mathbf{T}$ is *most sensitive* (*most sturdy*) if it maximizes (minimizes) Se (St) among statistics in $\mathbf{T}$.

In Section 2 we present some preliminary results which facilitate calculation of the quantities and provide a constructive approach for finding the optimal procedures. The form of the optimal tests is very reminiscent of the "intersection" part of the union–intersection tests of Roy (1953). He decomposes a multivariate alternative hypothesis into a number of component hypotheses, obtains an acceptance region for each and then takes their intersection. The acceptance regions of our optimal tests are also intersections of a collection of acceptance regions. If the acceptance regions of the statistics in $\mathbf{T}$ are generated by arbitrary intersections of Roy's component tests, then under mild conditions his test is indeed the most sensitive and sturdy.

Section 3 is a short section on two-sided tests, showing that if $\mathbf{T}$ contains all possible two-sided versions of a particular test statistic, then the optimal statistic is the one which provides equal tail areas for each level α. In Section 4 we consider combining independent tests. Letting $\mathbf{T}$ be the set of monotone test statistics, we show that Fisher's procedure is optimal. Section 5 treats problems which test the mean in spherically normal models, as well as some problems in other spherically symmetric and exponential family models. In the normal case, the set $\mathbf{T}$ is the class of admissible tests, and we show that the likelihood ratio statistic (LRS) (extended, if necessary, to be defined for all levels) is optimal. Since Fisher's procedure is also the LRS for its problem, one might wonder whether the LRS is always optimal. It is not in general true. Exceptions are found in Section 2 when testing a normal variance is 1, in Section 5 for certain exponential family models and in Section 6. In that last section, we consider two more examples. In Bartlett's problem of testing the equality of several normal variances, we show that the LRS is not optimal with respect to the admissible (among invariant) procedures. The most sensitive and sturdy statistic is indicated, and a large sample approximation to it is seen to be a weighted sum of squares of the normalized sample variances. In testing for one outlier in linear regression, we show that the LRS is most sensitive and sturdy when $\mathbf{T}$ is the class of statistics generated by the individual tests for targeted observations. However, when $\mathbf{T}$ is the set of admissible statistics, we find that there is no optimal procedure, and in fact all statistics are insensitive and fragile.

REMARK. A Bayesian analog of our criteria can also be imagined. We start with a set of prior distributions on the alternative. For a given such prior π, we obtain a prior on the whole space by setting the probability of the null at $1/2$ and the conditional probability given the alternative at π. Then for an observation x and prior π, we calculate the posterior probability of the null. Analogs to i and s in (1.1) and (1.2) are obtained by inserting priors for statistics, and posterior probabilities of the null for p-values. Then priors optimal according to the analogous criteria to those in Definition 1.1 will be those which are robust with respect to the posteriors. In the case of testing that a spherically symmetric bivariate normal mean is 0 versus that it is in the nonnegative quadrant (as in Section 5), we found that if all priors are allowed, then they are all insensitive and fragile. We did not explore this approach further.

2. Preliminaries. In this section we give some approaches to calculating the robustness in Definition 1.1 without having to deal explicitly with the statistics U. For a statistic T and level α, denote the acceptance region of the corresponding test by

$$A(T,\alpha) = \{x | p_T(x) \geq \alpha\}. \tag{2.1}$$

The intersection of all level α acceptance regions will be denoted

$$A_\alpha = \bigcap_{T \in \mathbf{T}} A(T,\alpha). \tag{2.2}$$

Our first result gives geometric definitions for the robustnesses.

LEMMA 2.1. *For $T \in \mathbf{T}$ and $0 \leq \alpha \leq 1$,*

$$\mathrm{Se}(T,\alpha) = \sup_\beta \{\beta | A(T,\alpha) \subset A_\beta\} \tag{2.3}$$

and

$$\mathrm{St}(T,\alpha) = \inf_\beta \{\beta | A(T,\beta) \subset A_\alpha\}. \tag{2.4}$$

PROOF. First note that for any $U \in \mathbf{T}$, $p_U(x) = \inf\{\beta | x \notin A(U,\beta)\}$, which is to say that the p-value for given U and x is the infimum of the levels β for which the level β test based on U rejects H_0. Thus from Definition 1.1 we have

$$\mathrm{Se}(T,\alpha) = \inf_{U \in \mathbf{T}} \inf_{x \in A(T,\alpha)} \inf_\beta \{\beta | x \notin A(U,\beta)\}. \tag{2.5}$$

Interchange the first two infima in (2.5) and then combine the resulting second and third, so that with

$$\mathbf{A} = \{A(U,\beta) | U \in \mathbf{T} \text{ and } 0 \leq \beta \leq 1\}, \tag{2.6}$$

we obtain

$$\mathrm{Se}(T,\alpha) = \inf_{A \in \mathbf{A}} \{1 - P(A) | A^c \cap A(T,\alpha) \neq \varnothing\} \equiv \alpha^*, \tag{2.7}$$

where P denotes the null probability.

Let $\beta^* = \sup\{\beta | A(T,\alpha) \subset A_\beta\}$, so that (2.3) will follow upon showing that $\alpha^* = \beta^*$. Start by taking $\beta > \beta^*$. Then $A(U,\alpha) \not\subset A_\beta$, so that $A_\beta^c \cap A(T,\alpha) \neq \varnothing$. Hence by (2.2) there exists some A with $1 - P(A) = \beta$ such that $A^c \cap A(T,\alpha) \neq \varnothing$. Thus by (2.7) $\alpha^* \leq \beta$. Since β is an arbitrary number greater than β^*, we have $\alpha^* \leq \beta^*$. Next, take $A \in \mathbf{A}$ with $A^c \cap A(T,\alpha) \neq \varnothing$. Then $A(T,\alpha) \not\subset A$; hence $A(T,\alpha) \not\subset A_\gamma$ where $\gamma = 1 - P(A)$. Thus $1 - P(A) \geq \beta^*$; hence from (2.7) $\alpha^* \geq \beta^*$. We have shown (2.3).

For (2.4) we have that $p_T(x) = \sup\{\beta | x \in A(T,\beta)\}$; hence

$$\mathrm{St}(T,\alpha) = \sup_{U \in \mathbf{T}} \sup_{x \in A(U,\alpha)^c} \sup_\beta \{\beta | x \in A(T,\beta)\}. \tag{2.8}$$

Now combining the first two suprema in (2.8) yields

$$\begin{aligned}\mathrm{St}(T,\alpha) &= \sup_{x \in A_\alpha^c} \sup_\beta \{\beta \mid x \in A(T,\beta)\} \\ &= \sup_\beta \{\beta \mid A_\alpha^c \cap A(T,\beta) \neq \varnothing\} \equiv \alpha'.\end{aligned}$$

Let $\beta' = \inf\{\beta \mid A(T,\beta) \subset A_\alpha\}$. We want to show that $\alpha' = \beta'$. First, if $\beta < \alpha'$, then $A_\alpha^c \cap A(T,\beta) \neq \varnothing$; hence $A(T,\beta) \not\subset A_\alpha$ and $\beta \le \beta'$. Thus $\alpha' \le \beta'$. For the reverse, note that $\beta > \alpha'$ implies that $A(T,\beta) \subset A_\alpha$; hence $\beta \ge \beta'$, proving that $\alpha' \ge \beta'$. Thus we have (2.4). □

Using Lemma 2.1, it is fairly easy to obtain bounds on Se and St and to find the optimal values.

LEMMA 2.2. *Let* $\alpha^* = \sup\{\beta \mid 1 - \alpha \le P(A_\beta)\}$. *Then for any* $T \in \mathbf{T}$ *and* $0 \le \alpha \le 1$,

$$\mathrm{Se}(T,\alpha) \le \alpha^* \quad \textit{and} \quad \mathrm{St}(T,\alpha) \ge 1 - P(A_\alpha). \tag{2.9}$$

PROOF. First

$$\{\beta \mid A(T,\alpha) \subset A_\beta\} \subset \{\beta \mid 1 - \alpha \le P(A_\beta)\} \tag{2.10}$$

since $P(A(T,\alpha)) = 1 - \alpha$. Thus the supremum over the former set in (2.10), which is $\mathrm{Se}(T,\alpha)$ by (2.3), is less than or equal to the supremum over the latter set, which is α^*, proving the first part of (2.9). Next, note that

$$\{\beta \mid A(T,\beta) \subset A_\alpha\} \subset \{\beta \mid 1 - \beta \le P(A_\alpha)\}. \tag{2.11}$$

The infimum over the former set in (2.11) is $\mathrm{St}(T,\alpha)$ by (2.4), and it is no smaller than the infimum over the latter set which is clearly $1 - P(A_\alpha)$, completing the proof of (2.9). □

LEMMA 2.3. (a) *If for* $T^* \in \mathbf{T}$ *and some* α, $A(T^*,\alpha) = A_{\alpha^*}$, *then* $\mathrm{Se}(T^*,\alpha) = \alpha^*$; *hence* T^* *is most sensitive at* α. *If, in addition,*

$$\lim_{\beta \uparrow \alpha^*} P(A_\beta) = P(A_{\alpha^*}), \tag{2.12}$$

then T^* *is essentially uniquely most sensitive at* α *in the sense that for any other* U, $\mathrm{Se}(U,\alpha) = \alpha^*$ *implies that*

$$P[A(T^*,\alpha) \,\Delta\, A(U,\alpha)] = 0. \tag{2.13}$$

(b) *If* $A(T^*,\alpha') = A_\alpha$, *where* $\alpha' = 1 - P(A_\alpha)$, *then* $\mathrm{St}(T^*,\alpha) = \alpha'$; *hence* T^* *is most sturdy for that* α. *Also,* T^* *is the essentially unique most sturdy statistic at* α *in the sense that for any other* U, $\mathrm{St}(U,\alpha) = \alpha'$ *implies that* (2.13) *holds with* α' *instead of* α.

PROOF. For (a), since $P[A(T^*,\alpha)] = 1 - \alpha$, (2.3) and $A(T^*,\alpha) = A_{\alpha^*}$ show that $\mathrm{Se}(T^*,\alpha) \ge \alpha^*$, which by (2.9) proves equality. If $\mathrm{Se}(U,\alpha) = \alpha^*$, then by

(2.3), for any $\beta < \alpha^*$, $A(U, \alpha) \subset A_\beta$; hence

$$A(U,\alpha) \subset \bigcap_{\beta \uparrow \alpha^*} A_\beta \equiv B_{\alpha^*}. \tag{2.14}$$

Now (2.12) shows that $P(B_{\alpha^*}) = 1 - \alpha$; hence with (2.14) we have that $P[B_{\alpha^*} - A(U, \alpha)] = 0$. Equation (2.12) also shows that $P(B_{\alpha^*} - A_{\alpha^*}) = 0$. Thus (2.13) follows. For (b), we use (2.4) and (2.9) to show that $\text{St}(T^*, \alpha) = \alpha'$. If $\text{St}(U, \alpha) = \alpha'$, then by (2.4), for any $\beta > \alpha'$, $A(U, \alpha) \subset A_\alpha$; hence $\bar{A} \equiv \bigcup_{\beta \downarrow \alpha'} A(U, \beta) \subset A_\alpha$. Since $P(\{x | p_U(x) = \alpha'\}) = 0$, $P[A(U, \alpha') - \bar{A}] = 0$, and since $P[A(U, \alpha')] = 1 - \alpha' = P(A_\alpha)$, we have essential uniqueness. □

We now show that sensitivity and sturdiness are almost inverses.

LEMMA 2.4. *Equations* (1.3) *and* (1.4) *hold.*

PROOF. Suppose $\text{Se}(T, \alpha) = \beta$. Then by (2.3), for any $\beta' < \beta$, $A(T, \alpha) \subset A_{\beta'}$; hence by (2.4) $\text{St}(T, \beta') \le \alpha$. Thus the first limit in (1.3) holds. Similarly, if $\beta' > \beta$, then $A(T, \alpha) \not\subset A_{\beta'}$; hence $\text{St}(T, \beta') \ge \alpha$. Thus (1.3) holds. Equation (1.4) can be proven in the same way. □

From (1.3) and (1.4), we have that if $\text{Se}(T, \alpha)$ as a function of α is continuous and strictly increasing on (0, 1), then St is its inverse. Also, if T is insensitive (fragile) for all α, then it is fragile (insensitive) for all α.

3. Two-sided tests. We assume that we have a two-sided testing problem and a single preferred one-dimensional statistic W with continuous distribution function F_W. The desirable tests are those with acceptance regions of the form $\{x | a \le W(x) \le b\}$ for some (possibly infinite) constants a and b. Then $\mathbf{T}$ consists of all statistics of the form $T(W(x))$ such that for each level α, its acceptance region in terms of W is a closed interval. Thus T is a pseudoconvex function.

Using (2.2), we have that

$$A_\alpha = \{x | F_W^{-1}(\alpha) \le W(x) \le F_W^{-1}(1 - \alpha)\} \tag{3.1}$$

since the right-hand side of (3.1) is contained in any level α acceptance region of the form $\{x | a \le W(x) \le b\}$, and the two regions $\{x | W(x) \le F_W^{-1}(1 - \alpha)\}$ and $\{x | F_W^{-1}(\alpha) \le W(x)\}$ are level α acceptance regions. Since the A_α's can be generated by the statistic $T^* \in \mathbf{T}$,

$$T^*(W(x)) \equiv 2\,\text{Min}\{F_W(W(x)), 1 - F_W(W(x))\}, \tag{3.2}$$

and $P(A_\beta) = \text{Min}\{2\beta, 1\}$ is continuous, Lemma 2.3 can be used to prove that T^* is essentially uniquely most sensitive and most sturdy.

From a practical point of view, T^* is very attractive. To find its p-value, upon observing $W(x)$, one need only find the smaller of the two tail areas and double it. In some problems, such as testing that the mean of a normal distribution is 0, there is symmetry about 0, and the T^*-test is the uniformly most powerful unbiased and/or invariant test, and also the likelihood ratio test. In some other problems, such as testing that the variance of a normal distribution is 1, we do not have symmetry, and the test (3.2), uniformly most powerful unbiased, and likelihood ratio test will be different.

Lemma 2.1 can be used to find Se and St for given $T \in \mathbf{T}$ and α. Let the left and right tail probabilities for the acceptance region of the level α test based on T be $l(\alpha)$ and $r(\alpha)$, respectively. Then $\mathrm{Se}(T, \alpha) = \mathrm{Min}\{l(\alpha), r(\alpha)\}$ and $\mathrm{St}(T,\alpha) = \mathrm{Min}\{l(\beta) + r(\beta) | l(\beta) \geq \alpha \text{ and } r(\beta) \geq \alpha\}$.

4. Combining independent tests. We investigate one of the aspects of meta-analysis, specifically, combining independent tests. See Hedges and Olkin (1985). We have n independent hypothesis testing problems based on the independent statistics $X_1, \dots, X_n$, respectively, and wish to combine the statistics into one overall statistic to test the null hypothesis that all n individual null hypotheses hold. We assume that each X_i has a single absolutely continuous distribution under its null hypothesis, and that it is appropriate to reject the ith null hypothesis for small values of X_i. Then, without loss of generality, we can take the X_i's to be Uniform (0, 1) under the null and each X_i to be its own p-value. Letting f_i be the density of X_i, the combined problem tests

$$H_0\colon X_1, X_2, \dots, X_n \text{ are independent } U(0,1) \tag{4.1}$$

versus

$$H_A\colon f_i(x_i) \text{ is nondecreasing in } x_i \text{ on } (0,1) \text{ for each } i,$$

and not all f_i are Uniform(0, 1). The results quoted below will still apply if we take as the alternative all joint densities which are componentwise nondecreasing.

There have been many methods proposed for combining the n statistics into one. We will focus on a select five, chosen for their ease of implementation as well as their spanning the range of admissible tests. Their rejection regions for a given α are

$$\begin{aligned}
&\text{Fisher:} && \prod_{i=1}^{n} X_i < \exp\{-\chi^2_{2n,\alpha}/2\},\\
&\text{Tippett:} && \mathrm{Min}\{X_1, \dots, X_n\} < 1 - (1-\alpha)^{1/n},\\
&\text{Normal:} && \sum_{i=1}^{n} \Phi^{-1}(X_i) < \sqrt{n}\,\Phi^{-1}(\alpha),\\
&\text{Sum:} && \sum_{i=1}^{n} X_i < c_\alpha,\\
&\text{Maximum:} && \mathrm{Max}\{X_1, \dots, X_n\} < \alpha^{1/n}.
\end{aligned} \tag{4.2}$$

Here, $\chi^2_{\nu,\alpha}$ is the upper α point of the chi-squared distribution on ν degrees of freedom, and Φ is the standard normal distribution function. For the sum

TABLE 1
Pairwise comparison of statistics (4.2)

$U \downarrow \| V \rightarrow$	Fisher	Tippett	Normal	Sum	Maximum
			$n = 2$		
Fisher		0.005/0.017	0/0.031	0/0.017	0/0.009
Tippett	0.178/0.118		0/0.003	0/0.001	0/0.001
Normal	1/0.078	1/0.230		0/0.030	0/0.015
Sum	1.000/0.117	1.000/0.291	0.500/0.078		0.025/0.025
Maximum	1/0.200	1/0.397	1/0.141	0.100/0.100	
			$n = 5$		
$U \downarrow \| V \rightarrow$	Fisher	Tippett	Normal	Sum	Maximum
Fisher		$1.5e-6/5.3e-4$	$0/1.3e-2$	$0/2.8e-3$	$0/1.1e-4$
Tippett	0.582/0.516		$0/1.1e-7$	$0/2.9e-9$	$0/1.1e-1$
Normal	1/0.145	1/0.731		$0/8.3e-3$	$0/6.6e-4$
Sum	1.000/0.254	1.000/0.816	0.989/1		$4.2e-4/1.9e-3$
Maximum	1/0.816	1/0.981	1/0.609	1.000/0.644	

Above diagonal: $i(V; U, 0.05)/i(U; V, 0.05)$.
Below diagonal: $s(U; V, 0.05)/s(V; U, 0.05)$.
In each case, a ratio below 1 favors U, and a ratio above 1 favors V.

test, $c_\alpha = (n!\,\alpha)^{1/n}$ if $c_\alpha \le 1$. We also consider other procedures, including those using the logistic and χ^2 distribution functions.

Objections have been raised to the maximum and sum tests since if one x_i is reasonably high, the test will fail to reject even when all the other x_i's are small. Se and St provide a way of quantifying such objections.

We first compare the five tests in (4.2) pairwise. For any pair of statistics, one can imagine the joint distribution of their p-values when X has a given alternative distribution. The range of the pairs of p-values is independent of the distribution on $\mathbf{X}$, the limits of the range being given by i and s in (1.1) and (1.2). See Table 1 for values when $\alpha = 0.05$. A perfect statistic, according to our criteria, would be one for which equalities hold in (1.5). From the table, one can see that Fisher's procedure comes closest to perfection, although it is still far from perfect, especially when compared to the maximum statistic. For the most part, the ordering from best to worst is

$$\text{Fisher} > \text{Tippett} > \text{normal} > \text{sum} > \text{maximum}.$$

One should also notice that all statistics perform worse as n increases.

We turn to Se and St relative to the set of statistics

$$\mathbf{M} \equiv \{T(x) \,|\, T \text{ is continuous and nondecreasing in each } x_i\}.$$

As shown in Birnbaum (1954), the statistics in $\mathbf{M}$ are exactly those which give rise to the admissible tests for problem (4.1). Now for any $U \in \mathbf{M}$ and $\alpha \in (0,1)$, $A(U,\alpha) \subset \mathbf{X} \equiv (0,1)^n$ is closed and nondecreasing in the sense that for $x, y \in \mathbf{X}$,

$$x \in A(U,\alpha) \text{ and } y_i \geq x_i \quad \text{for each } i \Rightarrow y \in A(U,\alpha). \tag{4.3}$$

LEMMA 4.1. *The set* (2.2) *is*

$$A_\alpha = \left\{ x \,\middle|\, \prod_{i=1}^{n} x_i \geq \alpha \right\}. \tag{4.4}$$

PROOF. For any point $z \in \mathbf{X}$, define $A(z) = \{x \in \mathbf{X} | x_i \geq z_i \text{ for some } i\}$. We claim that

$$A_\alpha = \bigcap_{z | \prod z_i = \alpha} A(z), \tag{4.5}$$

from which (4.4) can be derived. To see (4.5), first note that for each z with $\prod z_i = \alpha$, $A(z) = A(T,\alpha)$ for some $T \in \mathbf{M}$: Take $T_z(x) \equiv \max\{x_i/z_i\}$. Thus, from (2.2), we have "$\subset$" in (4.5). On the other hand, suppose that $x \notin A_\alpha$. Then $x \notin A(U,\alpha)$ for some $U \in \mathbf{M}$. By the monotonicity (4.3), we have that $A(U,\alpha) \subset A(x)$, and since T is continuous, $A(T,\alpha)$ is closed; hence $P[A(x) - A(T,\alpha)] > 0$. Thus $\prod x_i \equiv 1 - P[A(x)] < 1 - P[A(U,\alpha)] \equiv \alpha$; hence there exists a z with $\prod z_i = \alpha$ and $x \notin A(z)$, proving (4.5). □

From Lemmas 2.1 and 4.1, it can be seen that

$$\mathrm{Se}(T,\alpha) = \inf_{x \in A(T,\alpha)} \prod_{i=1}^{n} x_i \tag{4.6}$$

and

$$\mathrm{St}(T,\alpha) = F_T\left(\sup_{\prod x_i = b(\alpha)} T(x) \right), \tag{4.7}$$

where $b(\alpha) = \exp(-\chi^2_{2n,\alpha}/2)$. Note, in particular, that a statistic will be insensitive or fragile if and only if there are points in the rejection region for which at least one x_i becomes arbitrarily close to 0. Using (4.6) and (4.7), we can make the following calculations:

Statistic	Sensitivity	Sturdiness
Fisher	$\exp(-\chi^2_{2n,\alpha}/2)$	$P[\chi^2_{2n} > -2\log(\alpha)]$
Tippett	$\left[1-(1-\alpha)^{1/n}\right]^n$	$1-(1-\alpha^{1/n})^n$
Normal	0	1
Sum	0*	$1-(1-\alpha)^n/n!$
Maximum	0	1

The * in the display above is to indicate that the result only holds for $\alpha \le 1 - 1/n!$. Otherwise, $\text{Se}(\text{Sum}, \alpha) = 1 - [n!(1-\alpha)]^{1/n}$. The next theorem summarizes the qualitative results for our statistics.

THEOREM 4.2. *Relative to* $\mathbf{M}$, *for* $0 < \alpha < 1$, *we have that*

(a) *Fisher's statistic is the essentially unique most sensitive and most sturdy statistic for all* α;
(b) *The normal and maximum statistics are insensitive and fragile for all* α;
(c) *The sum statistic is insensitive if and only if* $\alpha \le 1 - 1/n!$;
(d) *Tippett's statistic is neither insensitive nor fragile for any* α, *and the sum statistic is not fragile for any* α.

PROOF. Parts (b), (c) and (d) follow from the calculations above. Consider part (a). Since the acceptance region for Fisher's statistic (4.2) for any level is of the form A_β (4.4), Lemma 2.3 shows that Fisher is essentially uniquely most sturdy for any level. Also, (2.12) holds since $P(A_\beta) = P(\chi^2_{2n} > -2\log(\beta))$ is continuous in β, so that Fisher is essentially uniquely most sensitive for any level. □

Comparing the statistics in (4.2), we have that Fisher is quite a bit more sensitive and sturdy than Tippett, and Tippett is much more sensitive than the remaining three, much more sturdy than the normal and maximum, and somewhat more sturdy than the sum. As n increases, sensitivity and sturdiness of a given statistic necessarily become worse since the number of statistics considered increases exponentially. From (4.6) it can be seen that the sensitivity is a volume of an n-dimensional rectangle. Thus it may be more informative to find the sensitivity "per dimension," that is, look at the normalized limit $\lim_{n\to\infty}[\text{Se}(T, \alpha)]^{1/n}$. For $\alpha \in (0, 1)$, all the statistics in (4.2) have limit 0, except for Fisher which has limit $1/e$. For sturdiness, (4.7) shows that $1 - \text{St}$ is the probability of an n-dimensional object. When T is sum-like, we might expect this object to be a corner of an n-dimensional rectangle. Thus we include a factor of $n!$ in the normalization: $\lim_{n\to\infty}[n!(1 - \text{St}(T, \alpha))]^{1/n}$. The limits are $-\log(\alpha)$ for Fisher, $-\log(\alpha)/e$ for Tippett, $1 - \alpha$ for the sum, and 0 for the normal and maximum statistics. Tables 2 and 3 contain some numerical values of the normalized quantities.

TABLE 2
$\text{Se}(T, 0.05)^{1/n}$

$T \downarrow \mid n \rightarrow$	2	5	10	20	50	∞
Fisher	0.093	0.160	0.208	0.248	0.288	0.368
Tippett	0.025	0.010	0.005	0.003	0.001	0
Normal	0	0	0	0	0	0
Sum	0	0	0	0	0	0
Maximum	0	0	0	0	0	0

TABLE 3
$[n!(1 - \mathrm{St}(T, 0.05)]^{1/n}$

$T \downarrow \mid n \rightarrow$	2	5	10	20	50	∞
Fisher	1.265	1.857	2.290	2.599	2.825	2.996
Tippett	1.098	1.174	1.172	1.155	1.133	1.102
Normal	0	0	0	0	0	0
Sum	0.950	0.950	0.950	0.950	0.950	0.950
Maximum	0	0	0	0	0	0

Finally, we look at some statistics other than those in (4.2). Many popular tests are of the form

$$T_H(x) \equiv \sum_{i=1}^{n} H^{-1}(x_i) < H_n^{-1}(\alpha) \tag{4.8}$$

for some continuous distribution function H, where H_n is the distribution function of a sum of n independent variables with distribution H. (It is also common to have H depend on i.) The Fisher, normal and sum tests (4.2) are of this form, with H being the distribution function of, respectively, $-\chi_2^2$, normal and uniform. Some general points can be made based on the support of H:

1. If H has support $(-\infty, \infty)$ or $(0, \infty)$, then the statistic in (4.8) is insensitive and fragile for all $\alpha \in (0, 1)$.
2. If H has support $(-\infty, 0)$, then the statistic is neither insensitive nor fragile for any $\alpha \in (0, 1)$. In fact, $H(H_n^{-1}(\alpha))^n \le \mathrm{Se}(T_H, \alpha) \le H(H_n^{-1}(\alpha))$ and $H_n(H^{-1}(\alpha)) \le \mathrm{St}(T_H, \alpha) \le H_n(H^{-1}(\alpha^{1/n}))$.

Thus, for example, since the logistic distribution $H(x) = (1 + \exp(-x))^{-1}$ has support $(-\infty, \infty)$, the corresponding statistic (4.8) is insensitive and fragile.

One popular set of choice for H is the χ^2 family, where in our setup we take H_ν to be the distribution of $-\chi_\nu^2$. We allow nonintegral values of ν, so that "χ_ν^2" means gamma with scale parameter $1/2$ and shape parameter $\nu/2$. Letting $T_\nu = T_{H_\nu}$, it can be shown that

$$\mathrm{Se}(T_\nu, \alpha) = \begin{cases} P\left(\chi_\nu^2 \ge \chi_{\nu n, \alpha}^2/n\right)^n, & \text{if } \nu \le 2, \\ P\left(\chi_\nu^2 \ge \chi_{\nu, \alpha}^2/n\right), & \text{if } \nu > 2, \end{cases}$$

and

$$\mathrm{St}(T_\nu, \alpha) = \begin{cases} P\left(\chi_{\nu n}^2 \ge n\chi_{\nu, \alpha^{1/n}}^2\right), & \text{if } \nu \le 2, \\ P\left(\chi_{\nu n}^2 \ge \chi_{\nu, \alpha}^2\right), & \text{if } \nu > 2. \end{cases}$$

We also have that

$$\lim_{n \to \infty} \mathrm{Se}(T_\nu, \alpha)^{1/n} = \begin{cases} P\left(\chi_\nu^2 \ge \nu\right), & \text{if } \nu \le 2, \\ e^{\nu/2}, & \text{if } \nu > 2, \end{cases}$$

and

$$\lim_{n\to\infty}\left[n!\,(1-\mathrm{St}(T_\nu,\alpha))\right]^{1/n}=\begin{cases}\dfrac{-\log(\alpha)\Gamma(\nu/2+1)e^{\nu/2}}{(\nu/2)^{\nu/2}}, & \text{if } \nu\le 2,\\ 0, & \text{if } \nu>2.\end{cases}$$

5. Spherical normal variables. Let $X \sim N_n(\theta, I_n)$, where X and θ are n-vectors, I_n is the $n \times n$ identity matrix and N_n represents the n-dimensional normal distribution. We are interested in testing

$$H_0\colon \theta = 0 \quad \text{versus} \quad H_A\colon \theta \in V - \{0\}, \tag{5.1}$$

where $V \subset R^n$ is a closed convex cone with vertex 0. The book by Barlow, Bartholomew, Bremner and Brunk (1972) contains a thorough discussion of such problems. From Eaton (1970), we have that the set of admissible tests consists of those with acceptance regions (essentially) in $\mathbf{A}(V)$, the set of closed, convex and monotone $[V]$ subsets of R^n, where a set A is monotone $[V]$ if

$$x \in A \text{ and } y - x \in D(V) \Rightarrow y \in V, \tag{5.2}$$

$D(V) \equiv \{y | y'x \le 0 \text{ for all } x \in V\}$ is the dual of V and $y'x = \sum y_i x_i$. Thus, for example, if $V = R^n$, $D(V) = \{0\}$, so that (5.2) is vacuous, and if V is the nonnegative orthant, the condition (5.2) is the same as (4.3).

We take $\mathbf{T}$ to be $\mathbf{T}(V)$, the set of all statistics T such that for each α, $A(T, \alpha) \in \mathbf{A}(V)$. See (2.1). For $\alpha \le 1 - P(D(V))$, the level α likelihood ratio test (LRT) has acceptance region

$$A^*(\alpha) \equiv \bigcap_{\gamma \in V_1} H(\gamma, c_\alpha), \tag{5.3}$$

where

$$H(\gamma, c) \equiv \{x | \gamma' x \le c\}, \tag{5.4}$$

$V_1 = \{\gamma \in V |\, \|\gamma\| = 1\}$ and c_α is the constant which yields the level α. The LRT does not exist for $\alpha \ge 1 - P(D(V))$, but the set $A^*(\alpha)$ exists for all α. Note that the statistic corresponding to the sets in (5.3) is

$$T^*(x) \equiv \sup_{\gamma \in V_1} \gamma' x. \tag{5.5}$$

The next lemma is of use in this and the following sections.

LEMMA 5.1. *Suppose the sample space* $\mathbf{X}$ *is an open convex subset of* R^n *and* P *has support* $\mathbf{X}$. *If all sets in* $\mathbf{A}$ *of* (2.6) *are convex and monotone* $[V]$ (5.2), *and* $\mathbf{A}$ *includes all half-spaces* (5.4) *with* $\gamma \in V_1$ *and* $c \in R$, *then for* (2.2),

$$A_\beta = \bigcap_{\gamma \in V_1} H(\gamma, c_\beta(\gamma)), \tag{5.6}$$

where $c_\beta(\gamma)$ *is given by*

$$P(\gamma' X \geq c_\beta(\gamma)) = \beta. \tag{5.7}$$

PROOF. Since by assumption all sets (5.4) are in $\mathbf{A}$, we have " $\subset$ " in (5.6). If $x \notin A_\beta$, then by (2.2) there exists some $T \in \mathbf{T}$ such that $x \notin A(T, \beta)$. Since $A(T, \beta)$ is closed, convex and monotone [V], the separating hyperplane theorem guarantees that there exist $\gamma \in V_1$ and constant c such that $A(T, \beta) \subset H(\gamma, c)$ strictly and $x \notin H(\gamma, c)$. Thus $1 - \beta \equiv P(A(T, \beta)) < P(H(\gamma, c))$ since P has support $\mathbf{X}$, so that $c > c(\gamma)$. Thus $x \notin H(\gamma, c(\gamma))$; hence x is not in the right-hand side of (5.6), so that we have (5.6). □

Our main result for this section follows.

THEOREM 5.2. *The statistic* T^* *in* (5.5) *is essentially uniquely most sensitive and sturdy in* $\mathbf{T}(V)$ *for all* α. *In addition, for* $T \in \mathbf{T}(V)$,

$$\mathrm{Se}(T, \alpha) = 1 - \Phi\Big(\sup_{\gamma \in V_1} \sup_{x \in A(T, \alpha)} \gamma' x \Big) \tag{5.8}$$

and

$$\mathrm{St}(T, \alpha) = 1 - F_T\Big(\inf_{x \in \partial A_\alpha} T(x) \Big), \tag{5.9}$$

where "∂" *indicates boundary, and* A_α *is as in* (5.6) *with* $c_\alpha(\gamma) \equiv 1 - \Phi(\alpha)$.

PROOF. The result about A_α follows from Lemma 5.1 since for $\gamma \in V_1$, $\gamma' X \sim N(0, 1)$. Equations (5.8) and (5.9) then follow from Lemma 2.1. That T^* has the stated properties follows from Lemma 2.3 by comparing (5.3) and A_α, and noting that $P(A_\alpha)$ is continuous in α. □

Note that (5.8) implies that a statistic is insensitive if and only if the acceptance region is unbounded in some direction $\gamma \in V_1$. In the special case that $V = R^n$, the set $\mathbf{A}(V)$ consists of all convex sets, $T^*(x) = \|x\|$ by (5.5), and (5.8) and (5.9) become

$$\mathrm{Se}(T, \alpha) = 1 - \Phi\Big(\sup_{x \in A(T, \alpha)} \|x\| \Big) \tag{5.10}$$

and

$$\mathrm{St}(T, \alpha) = 1 - F_T\Big(\inf_{\|x\| = c_\alpha} T(x) \Big),$$

respectively, where $c_\alpha = \sqrt{\chi^2_{n, \alpha}}$. Also, T is insensitive and fragile if and only if $A(T, \alpha)$ is unbounded.

The above results can be easily extended to any spherically symmetric null density for X. The only change will be that the "Φ" in (5.8) and (5.10) needs to be replaced by the distribution function of X_1. Of course, the set $\mathbf{T}(V)$ may no longer consist of the acceptance regions of admissible tests. A sufficient condition for $\mathbf{T}(V)$ to contain all the admissible regions is that the density of X be $q(\|x - \theta\|^2)$ for some convex function q. For another example, let $X \sim N_n(\theta, \sigma^2 I_n)$ and $W \sim \sigma^2 \chi^2_\mu$, with X and W independent. Problem (5.1) based on (X, W) is scale invariant, the maximal invariant statistic and parameter being, respectively, $Y \equiv X/\sqrt{W/\mu}$ and $\delta \equiv \theta/\sigma$. Now $\mathbf{T}(V)$ defined on the Y-space yields the set of tests which are admissible among invariant tests. [See Oosterhoff (1969) for the case that V is the nonnegative orthant. The general case can be shown using the ideas in Eaton (1970) and Oosterhoff.] The results in Theorem 5.2 then follow for this problem with X replaced by Y and Φ replaced by the Student's t distribution on μ degrees of freedom.

Finally, suppose that X has a regular exponential family density $f_\theta(x) = \phi(x)\exp\{\theta' x - \psi(\theta)\}$, where ψ is the normalizing constant. Eaton (1970) shows that $\mathbf{T}(V)$ consists of the admissible acceptance regions for (5.1) based on X. Then Lemma 5.1 gives the set (2.2), and Theorem 5.2 shows that the most sensitive and sturdy statistic has acceptance region A_β for β satisfying $P(A_\beta) = \alpha$. The LRT has acceptance region $B_\alpha \equiv \bigcap_{\gamma \in V} H(\gamma, K_\alpha + \psi(\gamma))$, where K_α is chosen to provide level α. Thus both tests are intersections of half-spaces, the difference being that the most sensitive and sturdy test takes all the half-spaces to have the same null probability, while the LRT takes half-spaces determined by the likelihood function. In the normal case, the sets of half-spaces turn out to be the same.

6. Other examples. We look at two further examples which are interesting in their own right and shed some more light on the most sensitive and sturdy criteria.

6.1. *Bartlett's problem.* Suppose $W_1, W_2, \ldots, W_p$ are independent, $W_i \sim \sigma_i^2 \chi^2_{\nu_i}$. We wish to test H_0: $\sigma_1^2 = \cdots = \sigma_p^2$ versus H_A: $\sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$. This problem is scale invariant. We will restrict consideration to scale-invariant statistics, which is equivalent to considering statistics based on the maximal invariant statistic $x \equiv (x_1, \ldots, x_{p-1})'$, where $x_i = y_i - y_p$ and $y_i = w_i/\sum_{j=1}^p w_j$. Cohen and Marden (1989), Corollary 2.2, show that a necessary condition for a test to be admissible among invariant tests is that its acceptance region be essentially convex in x. Thus we take $\mathbf{T}$ to contain all the statistics $T(x)$ such that the corresponding acceptance regions are among the admissible convex sets. It is also clear from their Theorem 2.1 that all half-spaces (5.4) yield admissible (among invariant) tests. Lemma 5.1 then shows that A_α is as in (5.6) and (5.7), where now V_1 is the unit sphere. It is not easy to find the constants $c_\alpha(\gamma)$. We will instead present a large sample approximation to them, and thus derive an approximate most sensitive and sturdy statistic. We will assume that $\nu_i/n \to \rho_i > 0$ for each i, where $n = \sum \nu_i$. It is then straightforward to show that $\sqrt{n}(X - \rho) \to N(0, 2(R + \rho_p J_{p-1}))$ in distribution, where $\rho = (\rho_1, \ldots, \rho_{p-1})'$, R is the diagonal matrix with entries $\rho_1, \ldots, \rho_{p-1}$ and J_{p-1} is the $(p-1) \times (p-1)$ matrix of 1's. Thus the set in (5.7) contains, approximately, the points x such that

$$\frac{\gamma'(x-\rho)}{\sqrt{2}\left[\gamma'(R+\rho_p J_{p-1})\gamma\right]^{1/2}} > z_\alpha. \tag{6.1}$$

The corresponding approximate statistic is then the supremum over γ of the left-hand side of (6.1), which is the square root of

$$\begin{aligned}\tilde{T}(x) &= \frac{1}{2}(x-\rho)'(R+\rho_p J_{p-1})^{-1}(x-\rho)\\ &= \sum_{i=1}^{p}\frac{\left[(y_i-\rho_i)-(\bar{y}-\bar{\rho})\right]^2}{2\rho_i},\end{aligned} \tag{6.2}$$

where $\bar{y}$ and $\bar{\rho}$ are weighted averages with weights $1/\rho_i$. That is, the statistic is the weighted sample sum of squares of the differences $(y_i-\rho_i)$. Also, asymptotically, the null distribution of $\tilde{T}$ approaches χ^2_{p-1}, so the test is easy to implement.

Neither the exact most sensitive and sturdy statistic nor our approximation to it (6.2) is equivalent to the LRS, which is $\prod y_i^{\nu_i}$. The LRT is admissible, and the test (6.2) is admissible among invariant tests. It is open whether the exact optimal test is admissible.

6.2. *Testing for one outlier.* We assume a standard normal linear model with a possible mean-shift outlier, so that we observe $Y = D\beta + \delta + E$, where D is a fixed $n \times p$ matrix of rank $p < n$, β is a $p \times 1$ parameter, $E \sim N_n(0, \sigma^2 I_n)$ for $\sigma^2 > 0$ and δ is an $n \times 1$ parameter vector,

$$\delta \in \Omega \equiv \{w \in R^n | w_i \neq 0 \text{ for at most one } i\}.$$

The vector δ models a possible outlier, where $\delta = 0$ means there are no outliers. Thus we test H_0: $\delta = 0$ versus H_A: $\delta \in \Omega - \{0\}$. The problem is invariant under the affine group $(0,\infty) \times R^p$ acting on Y via $(a,b)\cdot Y \to aY + Db$. The maximal invariant statistic is the vector of normed residuals, $x \equiv r/\|r\|$, where $r \equiv Qy$ and $Q \equiv I_n - D(D'D)^{-1}D'$, and the maximal invariant statistic is δ/σ.

The level α likelihood ratio test has rejection region

$$A^*_\alpha \equiv \bigcap_{i=1}^{n}\left\{|r_i|/\sqrt{Q_{ii}} \le c_\alpha\right\}. \tag{6.3}$$

Equivalently, the LRS is the maximum of the $|r_i|/\sqrt{Q_{ii}}$'s.

First, take $\mathbf{T}$ to be $\mathbf{T}_I$, the set of statistics T such that

$$A(T,\alpha) = \bigcap_{i=1}^{n}\{c_i \le r_i \le d_i\} \tag{6.4}$$

for possibly infinite constants c_i and d_i. Since the distribution under H_0 of r is invariant under sign changes, the set in (2.2) is as in (6.4) with $c_i = -d_i$, and

$$P(|r_i| \le d_i) = 1 - \alpha \quad \text{for each } i. \tag{6.5}$$

It can be shown that $(\gamma'X)^2/\gamma'Q\gamma \sim \text{Beta}(1/2, (n-p-1)/2)$. Thus with $\gamma = (1, 0, \dots, 0)$, we have that d_i in (6.5) is $\sqrt{Q_{ii} b_\alpha}$, where b_α is the upper α point of a $\text{Beta}(1/2, (n-p-1)/2)$. Hence (6.4) becomes (6.3), which with Lemma 2.3 can be used to show that the LRS is most sensitive and sturdy among those in $\mathbf{T}_I$.

Next, take $\mathbf{T}$ to be $\mathbf{T}_A$, the set of admissible statistics. Brown and Marden (1989) can be used to show that the conditions of Lemma 5.1 hold again; hence (5.6) holds. Thus

$$A_\alpha = \left\{ x \,\middle|\, \sup_{\|\gamma\|=1} \frac{(\gamma'x)^2}{\gamma'Q\gamma} \le b_\alpha \right\}. \tag{6.6}$$

Since Q is idempotent, $\|x\| = 1$ and $Qx = x$, the supremum in (6.6) is 1. Thus, unless $\alpha = 0$, A_α is empty. From Lemma 2.1 we obtain that for $\alpha > 0$, every statistic is insensitive and fragile. The dimensionality of the problem may provide an explanation for what has happened. The vector x is restricted to the unit sphere on the $(n-p)$-dimensional subspace given by Q. There are also $n-p$ directions for admissible tests to protect against. Thus there are too many tests, in contrast to $\mathbf{T}_I$ for which there is a finite number of directions.

Acknowledgments. I would like to thank an Associate Editor and referees whose comments led to many improvements.

REFERENCES

BARLOW, R., BARTHOLOMEW, D., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*. Wiley, New York.

BERGER, J. and SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence. *J. Amer. Statist. Assoc.* **82** 112–122.

BIRNBAUM, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49** 559–574.

BROWN, L. D. and MARDEN, J. I. (1989). Complete class results for hypothesis testing problems with simple null hypotheses. *Ann. Statist.* **17** 209–235.

CASELLA, G. and BERGER, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82** 106–111.

COHEN, A. and MARDEN, J. I. (1989). On the admissibility and consistency of tests for homogeneity of variances. *Ann. Statist.* **17** 236–251.

EATON, M. L. (1970). A complete class theorem for multidimensional one-sided alternatives. *Ann. Math. Statist.* **41** 1884–1888.

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic, San Diego.

OOSTERHOFF, J. (1969). *Combination of One-Sided Statistical Tests*. Mathematisch Centrum, Amsterdam.

ROY, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24** 220–238.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
CHAMPAIGN, ILLINOIS 61820