COMBINING INDEPENDENT TESTS OF SIGNIFICANCE*

ALLAN BIRNBAUM
Columbia University

It is shown that no single method of combining independent tests of significance is optimal in general, and hence that the kinds of tests to be combined should be considered in selecting a method of combination. A number of proposed methods of combination are applied to a particular common testing problem. It is shown that for such problems Fisher's method and a method proposed by Tippett have an optimal property.

1. THE PROBLEM AND SOME PROPOSED SOLUTIONS

THE problem of combining independent tests of significance has been discussed and illustrated by a number of writers, including Fisher [2], Karl Pearson (cf. [4]), Wallis [7], and E. S. Pearson [4], to which the reader is referred for general discussions to supplement the present brief section. The formal statistical problem may be stated as follows: A hypothesis H_0 is to be tested. An observed value t_1 of a statistic has been obtained; the best test of H_0 based on this statistic would indicate rejection at the u_1 significance level. That is, u_1 is the "probability level" corresponding to the observed value t_1; for example, if large values of the statistic are critical for H_0, then u_1 is the probability that a value as large as or larger than that observed will occur under H_0. Similarly, independent values of statistics, $t_2, \cdots, t_k$, have been obtained, and in the respective best tests of H_0 based on these statistics the corresponding "probability levels" are $u_2, \cdots, u_k$. The essential requirement of independence of the t_i's will be satisfied if each t_i is based on a separate and independent set of data; if each t_i is based on the same set of observations, the t_i's must be known to be statistically independent functions of the observations.

The problem of "combining these independent tests of significance" then is the problem of giving a test of H_0 on the basis of a set of observed values (probability levels) $u_1, u_2, \ldots, u_k$. The test is not to utilize the observed values $t_1, t_2, \ldots, t_k$; in general, it is assumed that

(a) either these values or else the forms of the distributions of $t_1, t_2, \ldots, t_k$, are unknown to the statistician confronted with the present problem; or

* Work sponsored by the Office of Naval Research.

The writer is grateful to Professor Henry Scheffé for helpful comments on the first draft of this paper. Responsibility for any remaining deficiencies is the writer's.

(b) this information is available but the distributions are such that there is no known or reasonably convenient method available for constructing a single appropriate test of H_0 based on $(t_1, t_2, \ldots, t_k)$.

Procedures for combining independent significance tests may be of practical use even in some situations in which the statistician has complete freedom to determine the design of a complex experiment. Suppose, for example, that a scientific hypothesis asserts that a change in the value of an independent experimental variable will alter the distribution of one or more of k observable variables $t_1, t_2, \ldots, t_k$. For example, an hypothesis to be tested may assert that administration to subjects of a certain drug will have at least one of the three effects:

(a) an increase in the mean of a certain measurable physiological quantity,

(b) an increase in the variance (within a subject) of a second measurable physiological quantity, and

(c) a decrease in the probability of a subject's correctly making a certain sensory discrimination.

Suppose that optimal tests for each of these effects separately could be based respectively on statistics t_1, t_2, and t_3. In such situations construction of a single optimal test for the presence of one or more of the effects may be difficult or impossible. However, combining statistically independent tests based on t_1, t_2, and t_3, a single test at a desired significance level can be given. With appropriate design, this test will also meet given requirements of power to detect one or more of the three effects. It is shown in Section 3 that for some problems such a test will even have certain efficiency properties.

To avoid technical complications not of direct interest here, let us assume that the t_i's have continuous distributions (densities). (See [7] for a discussion of the important discrete case.) Since u_i is the probability when H_0 is true of observing a value of our ith statistic at least as large as t_i, we may write

$$u_i = u_i(t_i) = \int_{t_i}^{\infty} p_i(t_i)dt_i \tag{1}$$

where $p_i(t_i)$ is the probability density function of t_i under H_0. Then the probability that u_i lies in any interval, say $u' \leqq u_i \leqq u''$, equals $u''-u'$, or in other words u_i has a uniform distribution on the unit interval under H_0 with density

$$f(u_i) = \begin{cases} 1, & 0 \leqq u_i \leqq 1, \\ 0, & u_i < 0,\ u_i > 1, \end{cases} \tag{2}$$

for each i, and the u_i's are mutually independent.

Each method of combining tests is a rule prescribing that H_0 should be rejected whenever the set of values $(u_i, \cdots, u_k)$ falls in a certain critical region. Intuitively speaking, small values of the u_i's are indicative of rejection; to discuss satisfactorily the problem of constructing a critical region of values of $(u_i, \cdots, u_k)$, we must consider the possible distributions of the u_i's when H_0 is false. We shall assume here that whenever a u_i has a non-uniform distribution, it is distributed on the unit interval according to some (unknown) density function $g_i(u_i)$ which is non-increasing.[1]

Depending on the nature of the experimental situations in which the u_i's are obtained, the appropriate alternative hypothesis would be either:

H_A: All of the u_i's have the same (unknown) non-uniform, non-increasing density $g(u)$.

or:

H_B: One or more of the u_i's have (unknown) non-uniform densities $g_i(u_i)$.

Under H_A, the t_i's are statistics of the same kind obtained from k replications of an experiment, in which the underlying conditions are assumed to remain constant with H_0 false. Under H_B, the t_i's may be statistics of different kinds (for example, a normal mean and a normal variance), and the conditions under which the t_i's are obtained need not be the same; it is assumed only that H_0 is false in the case of at least one of the t_i's. H_A is seen to be a special case of H_B. Probably in the majority of applications, H_B is the appropriate alternative hypothesis.[2]

---

[1] This assumption is not a strong one for our purposes: Suppose large values of the statistic t are critical for testing H_0 against H_1, and the probability densities of t under H_0 and H_1, are $p(t)$ and $p'(t)$, respectively. Then the definition of the statistic u is

$$u = u(t) = \int_t^\infty p(t)dt, \tag{3}$$

so $du/dt = -p(t)$. If the probability density of t is $p'(t)$ then that of u is

$$g(u) = p'(t)/\left|du/dt\right| = p'(t)/p(t). \tag{4}$$

Hence $g(u)$ will be a non-increasing function of u if and only if $p'(t)/p(t)$ is a non-decreasing function of t. The latter condition is satisfied for most distributions commonly encountered in applied statistics, including those of normal, binomial, and Poisson means, normal variances, and all other distributions of the Koopman form described in Section 3 below.

[2] In some papers cited above, the distinction between the alternatives H_A and H_B seems not to have been made sufficiently clear. Problems corresponding to H_A are considered by Wallis on p. 238 of [7] and by Pearson on p. 142 of [4]. Problems corresponding to H_B are considered by Wallis on pp. 245–56 of [7] and by Pearson on p. 138 of [4].

Some of the methods which have been proposed for combining independent tests of significance (i.e., for constructing critical regions of values of $(u_1, u_2, \cdots, u_k)$) are the following:

(1) Fisher's [2] method: reject H_0 if and only if $u_1 \cdot u_2 \cdot \cdots \cdot u_k \leqq c$, where c is a predetermined constant corresponding to the desired significance level. Wallis, on pp. 231–34 of [7], discusses in detail Fisher's method of appropriately determining c. It turns out that $-2 \log u_1 \cdot u_2 \cdot \cdots \cdot u_k$ is distributed as chi-square with $2k$ degrees of freedom when H_0 is true. If d is such that

$$\text{Prob } \{\chi_{2k}^2 \geqq d\} = \alpha \tag{5}$$

where $1-\alpha$ is the desired significance level, then setting $-2 \log c = d$, we obtain $c = e^{-d/2}$.

(2) Karl Pearson's method: reject H_0 if and only if $(1-u_1)(1-u_2) \cdots (1-u_k) \geqq c$, where c is a predetermined constant corresponding to the desired significance level. In applications, c can be computed by a direct adaptation of the method used to calculate the c used in Fisher's method.

(3) Wilkinson's [8] methods: reject H_0 if and only if $u_i \leqq c$ for r or more of the u_i's, where r is a predetermined integer, $1 \leqq r \leqq k$, and c is a predetermined constant corresponding to the desired significance level. The k possible choices of r give k different procedures which we shall refer to as case 1 $(r=1)$, case 2 $(r=2)$, etc. For example, if $k=2$ and a test at the .95 significance level is desired, the case 1 procedure is: reject H_0 if either u_1 or u_2 or both equal or exceed $c = (.95)^{1/2} = .974$; the case 2 procedure is: reject H_0 if both u_1 and u_2 equal or exceed $c = 1-(.05)^{1/2} = .776$. Case 1 was proposed earlier by Tippett [5].

In the following sections, certain bases for selecting methods of combination for particular problems will be developed.

2. A GENERAL CONDITION FOR ADMISSIBILITY OF METHODS OF COMBINATION

The following condition is readily seen to be satisfied by each of the proposed methods described above:

Condition 1: If H_0 is rejected for any given set of u_i's, then it will also be rejected for all sets of u_i^*'s such that $u_i^* \leqq u_i$ for each i.

Any method of combination which failed to satisfy this condition would seem unreasonable. In fact, it is not difficult to prove that the best test of H_0 against any particular alternative H_B of the kind described above satisfies Condition 1. (A proof is given in the Appendix.)

Since Condition 1 is satisfied by so many possible methods of combination, the question arises whether any further reasonable condition can be imposed to narrow still further the class of methods from which we must choose. The answer is no: So long as we consider the problem in the present generality, and not with reference to a particular kind of testing problem, there are no restrictions on the possible forms of the density functions $g_i(u_i)$ except that they be non-increasing. And it can be shown (see the Appendix) that for *each* method of combination satisfying Condition 1, we can find *some* alternative H_B represented by non-increasing functions $g_1(u_1), \cdots, g_k(u_k)$ against which that method of combination gives a best test of H_0.

These considerations prove that to find useful bases for choosing methods of combination, we must consider further the particular kinds of tests to be combined in any given problem. In the following sections, it is shown that certain methods are optimal for certain important categories of testing problems.

3. DISTRIBUTIONS OF THE KOOPMAN FORM

Nearly all of the density functions and discrete probability distribution functions encountered frequently in applied statistics can be written in the so-called Koopman form, which is

$$f(x, \theta) = c(\theta)a(\theta)^{t(x)}b(x) \tag{6}$$

where θ is a parameter of the distribution and x is an observed value, and a, b, c, and t denote arbitrary functions. Examples are

1. The binomial:

$$f(x, \theta) = \binom{n}{x}\theta^x(1 - \theta)^{n-x} = (1 - \theta)^n\left(\frac{\theta}{1 - \theta}\right)^x\binom{n}{x}. \tag{7}$$

2. The normal, with known (say unit) variance and mean θ:

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2} = \frac{1}{\sqrt{2\pi}}e^{-\theta^2/2}e^{\theta x}e^{-x^2/2}. \tag{8}$$

Other examples are the Poisson and exponential distributions and the normal distribution with known mean and unknown variance.

Consider a problem of combining k independent significance tests, each of which is a test on a distribution of the Koopman form (all k distributions need not be of the same sort however; for example, one might be on a normal mean and another on a binomial mean, as in the illustration in Section 2 above). A method of combining these tests

will be equivalent to a test of a hypothesis specifying the values of k parameters,

$$H_0: \theta_1 = \theta_1^0, \cdots, \theta_k = \theta_k^0,$$

on the basis of the observed values of the statistics $t_1, \cdots, t_k$. For such problems, a minimal criterion for the reasonableness of a test is known. "Reasonableness" is used here in the sense of admissibility of a test, which may be defined as follows: A test is admissible if there is no other test with the same significance level which, without ever being less sensitive to possible alternative hypotheses, is more sensitive to at least one alternative. In other words, an admissible test is one which cannot be strictly (that is, uniformly) improved upon. A necessary condition for admissibility of a test of H_0 in our problem is that the acceptance region of the test (that is, the values of $(t_1, \cdots, t_k)$ for which the test accepts H_0) be convex. (A region is convex if the line segment connecting each pair of points in the region lies entirely in the region.) This is shown in [1].

We may illustrate both this condition for admissibility and its application to methods of combination by considering the problem of combining two tests on means of normal distributions with known (say unit) variances. (The performance of Fisher's method when applied to such a problem has been considered by Wallis (pp. 237–39 of [7]) and by Pearson (p. 142 of [4]). Let $\bar{x}_1$ denote the mean of a sample of n_1 observations obtained in an experiment in which the underlying population mean had the unknown value μ_1; let $\bar{x}_2$ be the mean of a sample of n_2 observations in a similar experiment in which the unknown population mean was μ_2. In this case any method of combining tests of the two hypotheses $\mu_1=0$ and $\mu_2=0$ is equivalent to a test of H_0: $\mu_1=\mu_2=0$; then H_A would specify $\mu_1=\mu_2\neq 0$; and H_B would specify that either μ_1 or μ_2 or both are not zero. Let $t_1=\sqrt{n_1}\bar{x}_1$ and $t_2=\sqrt{n_2}\bar{x}_2$. Then any method of combining the tests on μ_1 and μ_2 can be represented as a test of H_0 by its critical region in the (t_1, t_2) plane. Each of the methods of combination described above has been applied to the present problem, and the critical region corresponding to each method is illustrated in the figures below. The significance level $\alpha=0.5$ was used throughout. The tests on μ_1 and μ_2 to be combined were taken first to be against two-sided alternatives (Figures 1–4) and then against one-sided alternatives (Figures 6–9). In each case the critical region was obtained by first determining the values of u_1 and u_2 for which the method of combination considered would reject H_0 at the .05 significance level, and then plotting the corresponding values of t_1 and t_2 by

use of the equations relating the t_i's to the u_i's. These equations are, for the two-sided tests,

$$u_i = \frac{2}{\sqrt{2\pi}} \int_{|t_i|}^{\infty} e^{-v^2/2} dv, \qquad \text{for } i = 1, 2, \tag{9}$$

and for the one-sided tests,

$$u_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t_i} e^{-v^2/2} dv, \qquad \text{for } i = 1, 2. \tag{10}$$

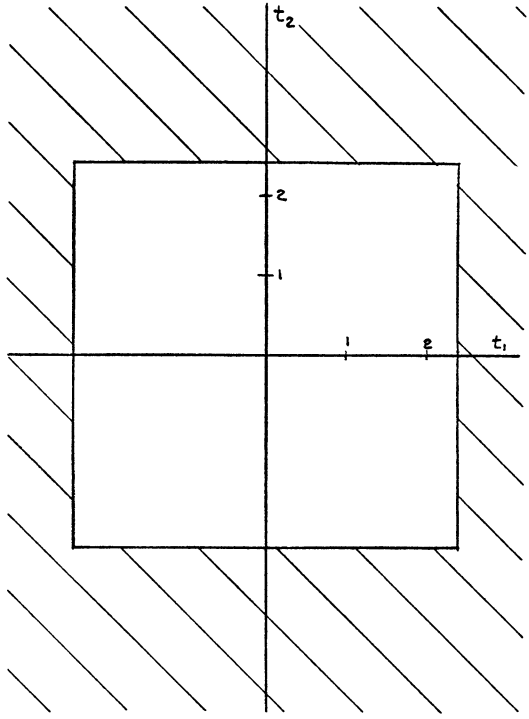

FIG. 1. Wilkinson's Method, Case 1.

We can now apply the condition for admissibility of a test described above: The acceptance regions obtained by Wilkinson's method, case 2, and by Pearson's method are not convex. Hence they represent tests of H_0, and corresponding methods of combination for the present problem, which can be strictly improved upon by other tests and corresponding methods of combination. Present knowledge does not provide

methods of finding tests which actually do strictly improve upon a given inadmissible test in problems like the present one. However, it seems advisable in selecting tests to restrict consideration to the class of admissible tests, and to select from this class a test which seems to have relatively good sensitivity (power) against the range of alternatives of interest.

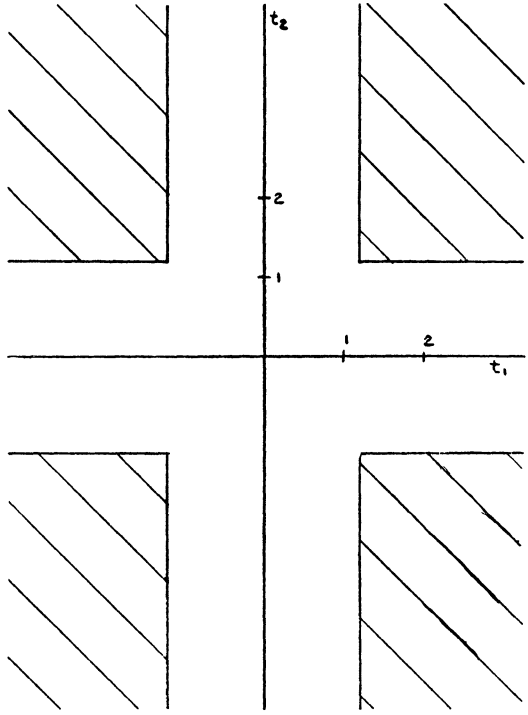

FIG. 2. Wilkinson's Method, Case 2.

It is shown in [1] that, for a category of problems including the present one, convexity of the acceptance region is a sufficient as well as necessary condition for admissibility.

The remaining two methods of combination, Wilkinson's case 1 and Fisher's, correspond to admissible tests. Inspection of Figures 1 and 3 suggests that each is fairly sensitive to departures from H_0 in all directions; that Fisher's method comes close to that test of H_0 (represented in Figure 5) which, if $n_1=n_2$ and if the seriousness of a departure from H_0 is measured by $\mu_1^2+\mu_2^2$, is the best test at the .05 level (as Wallis noted in [5]); and finally that Wilkinson's method, case 1,

gives a relative concentration of sensitivity to alternatives in which the departure from H_0 occurs in just one of the parameters. Similar observations can be made in Figures 6 and 8. Hence, it seems warranted to make a choice between the two methods remaining under consideration on the basis of a subjective appraisal of the context in which a problem like the present one actually occurs; probably in most cases Fisher's method would be preferred.

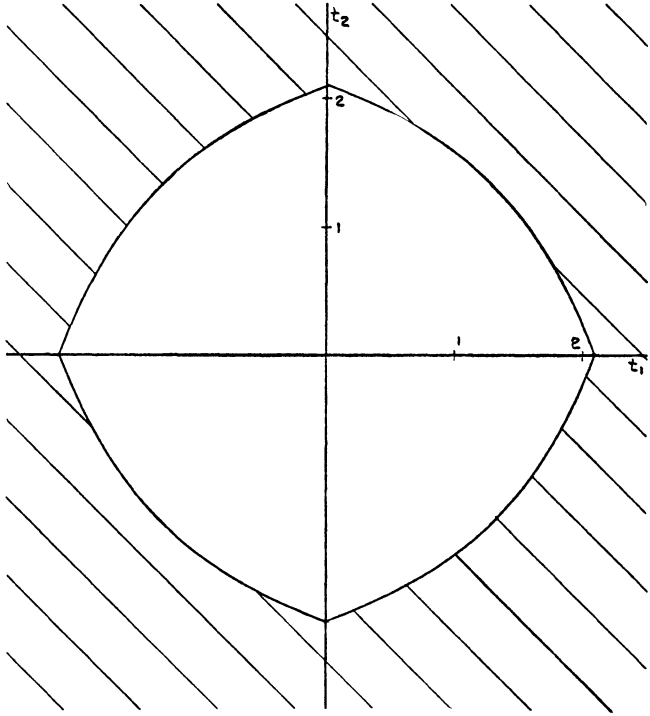

FIG. 3. Fisher's Method.

Having considered in detail a problem involving a particular distribution of the Koopman form, we proceed now to show that similar considerations apply to the whole class of such distributions. It can be verified easily that if Wilkinson's methods are used to combine tests on any such distributions, the result corresponds to a test whose acceptance region has a rectangular boundary like those in Figures 1, 2, 6, and 7, and is convex only in case 1. Hence, only case 1 of Wilkinson's method corresponds to an admissible test for certain of the Koopman-

form distributions being considered. The remaining cases of the method correspond to inadmissible tests for all Koopman-form distributions.

With little more difficulty it can be verified that Pearson's method does not give a test of H_0 with convex acceptance region for any Koopman-form distributions (consider the three points in the (t_1, t_2) plane corresponding to $(u_1, u_2) = (1-c, 0)$, to $(u_1, u_2) = (0, 1-c)$, and to

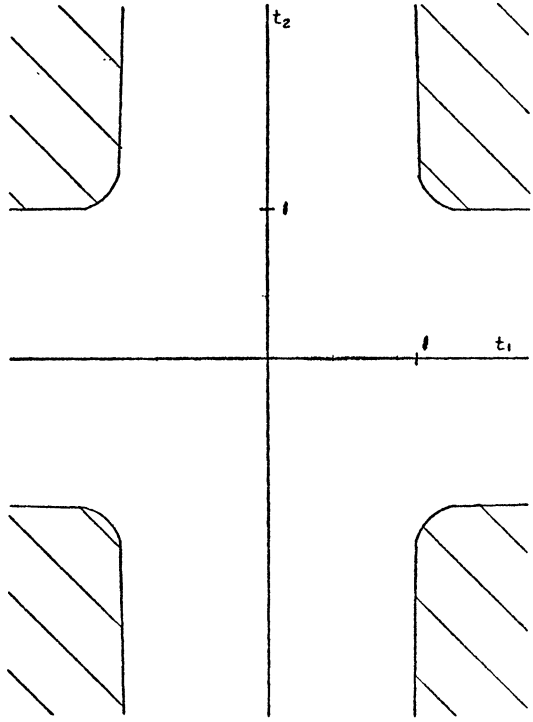

FIG. 4. Pearson's Method.

$(u_1, u_2) = (1-\sqrt{c}, 1-\sqrt{c})$, for the case $(k=2)$. Thus, Pearson's method also may be removed from consideration as inadmissible for Koopman-form distributions. Fisher's method does seem to give tests of H_0 with convex acceptance regions for Koopman-form distributions; consideration of the points in the (t_1, t_2) plane corresponding to $(u_1, u_2) = (1, c)$, to $(u_1, u_2) = (c, 1)$, and to $(u_1, u_2) = (\sqrt{c}, \sqrt{c})$ suggests this, and for particular distributions it may be possible to verify it fully without too much difficulty. For example, for

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}, \qquad x > 0, \tag{11}$$

to combine two one-sided tests on θ based on one observation each, we have

$$u_i = e^{-x_i/\theta_0}, \qquad \text{for } i = 1, 2, \tag{12}$$

and the critical region $u_1 u_2 \leqq c$ corresponds to a test with the convex acceptance region $x_1 + x_2 \leqq -\theta_0 \log c$.

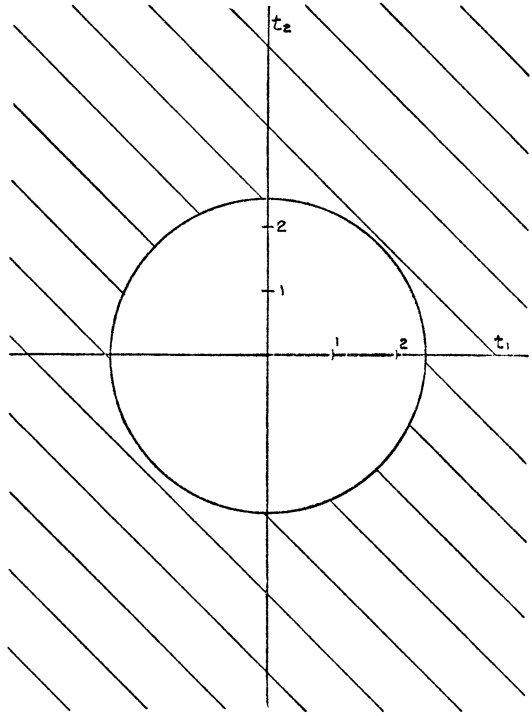

FIG. 5. Best Symmetric Test Against H_B.

4. CONCLUSIONS

While there is no single method of combining tests which is best for all problems, it appears that to combine independent tests on Koopman-form distributions (these include most distributions commonly occurring in applied statistics) one should choose between Fisher's

method and Wilkinson's method, case 1 Fisher's method appears to have somewhat more uniform sensitivity to the alternatives of interest in most problems. For any particular distributions, investigations may be made paralleling those above to obtain a still more conclusive basis for choice of a method of combination.

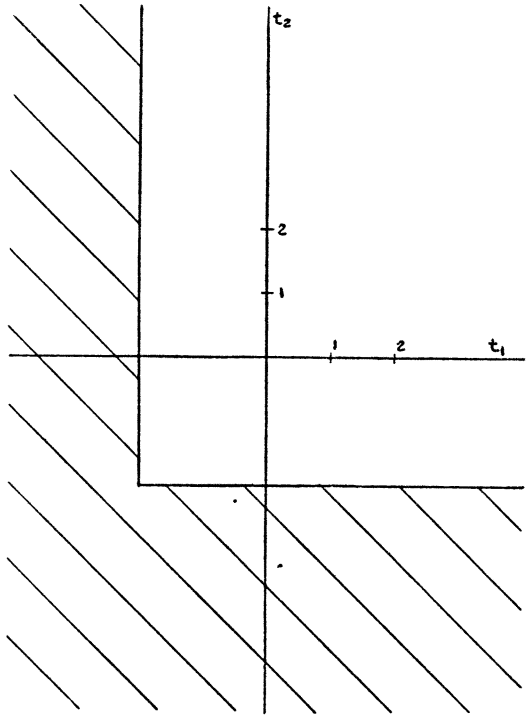

FIG. 6. Wilkinson's Method, Case 1. One-sided Alternatives.

APPENDIX

To prove that, as stated in Section 2 above, every test of H_0 which is best against some particular alternative specifying non-increasing densities, satisfies Condition 1, we use the well-known fact, proved for example in [3], that any best critical region consists of points satisfying

$$\lambda = \frac{g_1(u_1) \cdots g_k(u_k)}{f_1(u_1) \cdots f_k(u_k)} \geqq c, \qquad c \text{ some constant.} \tag{13}$$

Now $f_i(u_i) = 1$ for $0 \leqq u_i \leqq 1$, $i = 1, \cdots, k$. Hence, $\lambda = g_1(u_2) \cdots g_k(u_k)$. As the $g_i(u_i)$'s are non-increasing, $g_1(u_1') \cdots g_k(u_k') \geqq g_1(u_1) \cdots g_k(u_k) \geqq c$ if $(u_1, \cdots u_k)$ is in the best critical region and if $u_i' \leqq u_i$ for $i = 1, \cdots, k$. Thus Condition 1 is satisfied.

However, in general H_B (and even H_A) will include a whole set of possible forms of the $g_i(u_i)$'s, and it is not true in general that there will

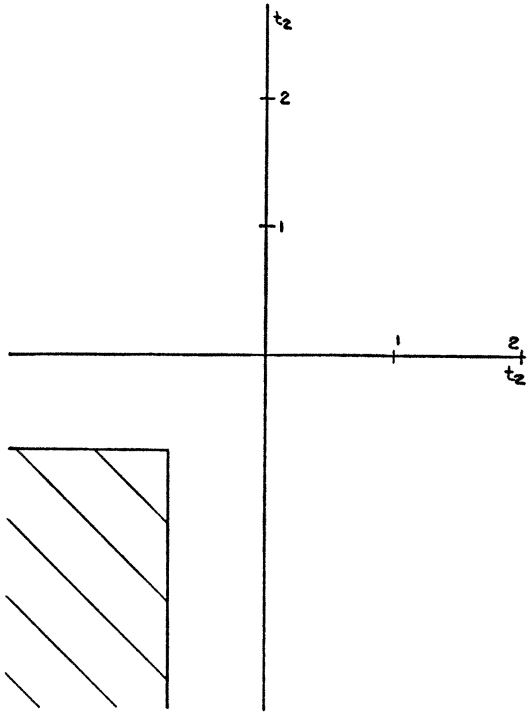

FIG. 7. Wilkinson's Method, Case 2. One-sided Alternatives.

exist a single test of H_0 which is uniformly best against all possibilities. This is illustrated most simply in the following case under H_B: If $g_1(u_1)$ is uniform and $g_2(u_2)$ is nonuniform, a best critical region consists of all (u_1, u_2) such that $u_2 \leqq c$, c some constant; if $g_2(u_2)$ is uniform and $g_1(u_1)$ nonuniform, a best critical region consists of all (u_1, u_2) such that $u_1 \leqq c'$, c' some constant; thus, there is not a single critical region which is best against each alternative. It can be verified directly that every best test of H_0 against a "Bayes mixture" of simple alternatives under

H_B also satisfies Condition 1. It follows, as shown by Wald in [6], that under general assumptions Condition 1 is a necessary condition for admissibility of a test of H_0 against a composite alternative H_B.

We shall show next that, as stated in Section 2 above, *each* method of combination meeting Condition 1 is *best* against *some* particular alternative hypothesis H_B. Taking $k=2$ for simplicity, any critical

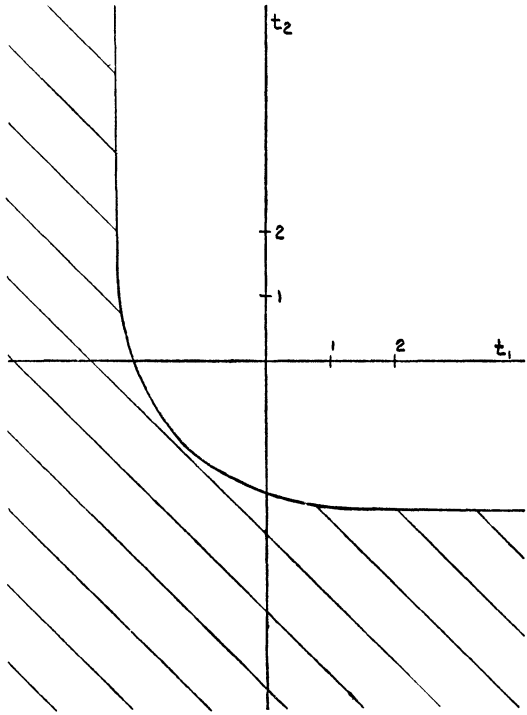

FIG. 8. Fisher's Method. One-sided Alternatives.

region w of values of (u_1, u_2), if it satisfies Condition 1, can be characterized by giving its boundary function $u_2(u_1)$, a non-increasing function such that w consists of all points (u_1, u_2) in the unit square $0 \leqq u_1 \leqq 1$, $0 \leqq u_2 \leqq 1$ for which $u_2 < u_2(u_1)$. Let $u_2(u_1)$ be any such boundary junction. Let $g_2(u_2) = \frac{2}{3}(2-u_2)$ for $0 \leqq u_2 \leqq 1$, and let $g_1(u_1) = \frac{3}{2}c(2-u_2(u_1))^{-1}$ for $0 \leqq u_1 \leqq 1$, where c is determined by the condition that $\int_0^1 g_1(u_1)du_1 = 1$. A best critical region for testing H_0 against the alternative $g_1(u_1)$, $g_2(u_2)$ is the set w' on which $g_1(u_1)g_2(u_2) > c$. But

$g_1(u_1)g_2(u_2) \equiv c(2-u_2)/(2-u_2(u_1)) > c$ if and only if $u_2 < u_2(u_1)$. Thus the arbitrarily given boundary function $u_2(u_1)$ characterizes a best critical region w'.

(Similar methods give analogous results for the problem of testing H_0 against H_A, with Condition 1 now strengthened by the requirement that the boundary function $u_2(u_1)$ be symmetric about the line $u_1 = u_2$.)

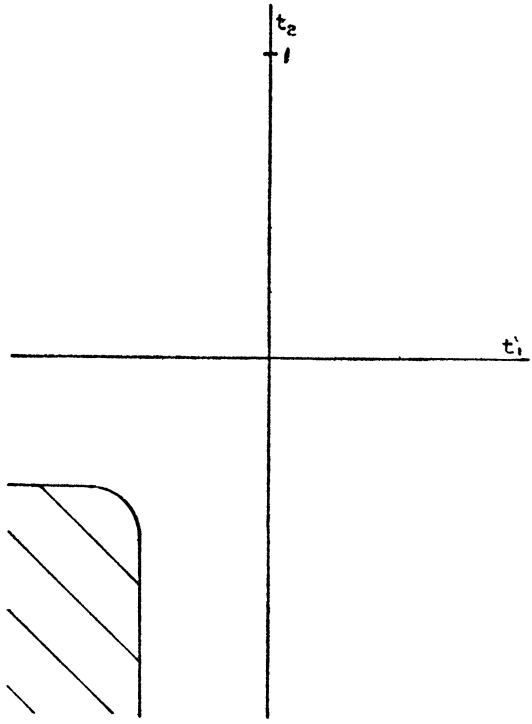

FIG. 9. Pearson's Method. One-sided Alternatives.

REFERENCES

[1] Birnbaum, Allan, "Characterizations of Complete Classes of Tests of Some Multiparametric Hypotheses, with Applications to Likelihood Ratio Tests," to be published in *Annals of Mathematical Statistics*.

[2] Fisher, R. A., *Statistical Methods for Research Workers*, Fourth and later Editions, Edinburgh and London, Oliver and Boyd, 1932 and later, Section 21.1.

[3] Neyman, J., *First Course in Probability and Statistics*, New York, Henry Holt and Company, 1950, Section 5.3.1, 304–8.

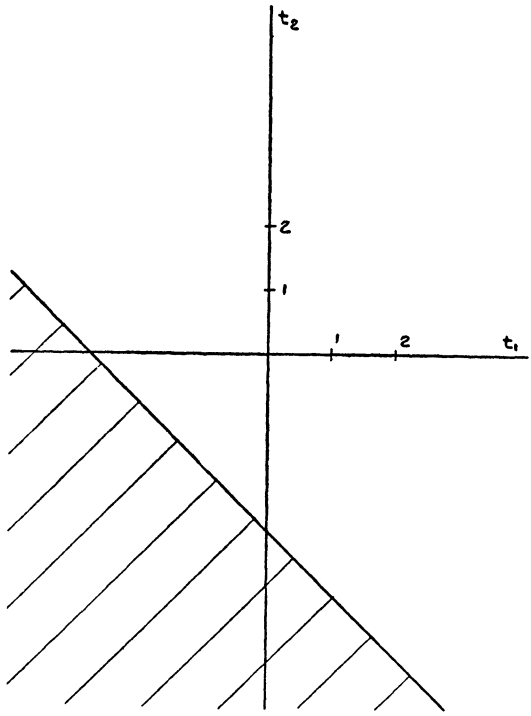

FIG. 10. Best Test of H_0 Against H_A. One-sided Alternatives.

[4] Pearson, E. S., "The Probability Integral Transformation for Testing Goodness of Fit and Combining Independent Tests of Significance," *Biometrika*, 30 (1938), 134–48.
[5] Tippett, L. H. C., *The Methods of Statistics*, First Edition, London, Williams and Norgate, Ltd., 1931, Section 3.5, 53–6.
[6] Wald, Abraham, *Statistical Decision Functions*, New York, J. Wiley and Sons, Inc., 1950, Theorem 3.20, 101.
[7] Wallis, W. Allen, "Compounding Probabilities from Independent Significance Tests," *Econometrica*, 10 (1942), 229–48.
[8] Wilkinson, B., "A Statistical Consideration in Psychological Research," *Psychological Bulletin*, 48 (1951) 156–7.