(2)

ON COMBINING INDEPENDENT SIGNIFICANCE TESTS

BY

F. J. O'REILLY and MICHAEL A. STEPHENS

TECHNICAL REPORT NO. 444

JULY 24, 1991

PREPARED UNDER CONTRACT

N00014-89-J-1627 (NR-042-267)

FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

ON COMBINING INDEPENDENT SIGNIFICANCE TESTS

BY

F. J. O'REILLY and MICHAEL A. STEPHENS

TECHNICAL REPORT NO. 444
JULY 24, 1991

Prepared Under Contract
N00014-89-J-1627 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

1. INTRODUCTION.

How to combine the results of k independent tests of significance has long been an important problem in statistics. The problem can arise, for example, in such diverse situations as when tests on the mean survival time after diagnosis of a terminal disease are made on k groups of patients in different hospitals, or when the sets of observations in k cells of an ANOVA table are separately tested for normality. An important feature of such tests is that often the individual sample sizes will be small, so that asymptotic results will not necessarily be valid. We suggest below that they might even be misleading in some situations.

Although work was done on this problem many years earlier, we base our discussion on a very comprehensive examination of both test situations and techniques by Birnbaum (1954). The techniques studied use the significance levels, or p-levels, of the individual tests, and they include the following. Suppose $p_1, p_2, \ldots, p_k$ are the individual p-levels, of test statistics $t_1, t_2, \ldots, t_k$. Fisher's statistic (perhaps the most widely used) for combining the test results is $T^F = -2\Sigma \log p_i$ (all logarithms are natural logarithms, and all sums or maxima or minima are over i from 1 to k). A similar statistic is Pearson's $T^P = -2\Sigma \log q_i$ where $q_i = 1 - p_i$. Two other statistics are $T^M = \max p_i$ and $T^m = \min p_i$; these are special cases of Statistics derived from methods developed by Wilkinson (Birnbaum, 1954, p. 562, with $r = 1$ or $r = k$ respectively). Birnbaum gives a very careful discussion of these statistics, in particular against two types of alternative to the overall null hypothesis, which we call H_{oo}. If the individual null hypotheses are H_{oi}, $i = 1, \ldots, k$, H_{oo} is

H_{oo} all H_{oi} are true.

It is well known that when all H_{oi} are true, the p_i will be uniformly distributed between 0 and 1, written $U(0,1)$; then T^F and T^P each have the χ^2_{2k} distribution. Birnbaum's alternatives to H_{oo} are

H_A: all H_{oi} are false together, or

H_B: one or more of the H_{oi} are false.

Birnbaum expressed H_A and H_B in terms of the alternative densities $g_i(p_i)$ of the p-values. On H_A, all p_i have the same (unknown), non-increasing densities $g(p_i)$, and on H_B, one or more of the p_i have (unknown) non-uniform densities $g_i(p_i)$. It is probably true that in almost all test situations, H_A or H_B will express the expected alternative to H_{oo}. To quote Birnbaum:

"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_{oo} 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_{oo} 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." (In this quote and again below, we use H_{oo}, our notation).

Birnbaum goes on to prove that the best test of H_{oo} against any particular H_B satisfies his Condition 1; that if H_{oo} is rejected for any given set of p_i, then it must be rejected for p-values p_i^* such that

$p_i^* \leq p_i$, for all i. Then Birnbaum shows that for each method of combination of p-values satisfying Condition 1, "we can find some alternative H_B ... against which that method of combination gives a best test of H_{oo}".

Subsequently, Littell and Folks (1971, 1973) have studied statistics T^F, T^M, and T^m from the point of view of Bahadur efficiency. They included also the statistic $T^N = \Sigma\Phi^{-1}(q_i)$, where Φ^{-1} is the inverse of the standard normal distribution function. Fisher's statistic performs well by the criteria of both Birnbaum (admissibility) and Littell and Folks (asymptotic Bahadur efficiency). However, there are two important reasons why asymptotic considerations may not be valid. Firstly, under alternative H_B, the conditions for examining Bahadur efficiency may not be realized; and secondly, combinations of tests will often be done with relatively small samples, so that asymptotic results will not apply.

In this article we concentrate on a comparison of the Fisher statistic T^F (which has become well-established) and the Pearson T^P. We first discuss asymptotics in Section 2. Since, as far as we know, Bahadur efficiency has not been examined for T^P, we fill in that gap, so that T^P may be compared from this point of view with the other statistics. T^P is shown to be equivalent to T^M and, by these asymptotic considerations, will be inferior to T^F.

In Sections 3 and 4 we turn to small-sample issues. Section 3 contains an example from Fisher's works, for which T^P is more sensitive than T^F. This is, of course only one example, and in Section 4 we discuss small sample results more thoroughly. Three families of alternative distributions for p are suggested, which might well be reasonable representations of situations H_A and H_B. For one of these T^F will be the statistic of choice, and for the other, T^P. This prompts further investigation by Monte Carlo methods, of the important small-sample situation, and the results (which include other statistics) are given in Tables 1 and 2. They support the view that there will be occasions when the Pearson statistic, and also T^N, should be calculated.

2. ASYMPTOTIC RESULTS ON EFFICIENCY.

2.1 Bahadur efficiency for Pearson's method.

In this section we introduce more detailed notation to examine the various methods of combining k tests. Suppose now that $t(1;n_1), t(2;n_2), \ldots, t(k;n_k)$ are independent test statistics, based on $n_1, n_2, \ldots, n_k$ observations, for testing hypotheses $H_{o1}, H_{o2}, \ldots, H_{ok}$. To simplify ideas, we assume for the present that all H_{oi} are the same, for instance, they might be hypotheses concerning a parameter θ, that $\theta \in \theta_o$, where θ_o is a set of possible values. This could occur, for example, in the example of testing mean survival time using results from k hospitals. Assume without loss of generalit that all tests are so defined that they reject for large values of the statistics. Let $F(i;n_i)$ denote the null distribution functions of $t(i;n_i)$, so that $p(i;n_i) = 1 - F(i;n_i)$ is the observed significance level, or p-level, of the i-th test. To obtain Bahadur efficiency asymptotic results, we shall suppose all n_i to grow steadily larger, and also suppose that the $p(i;n_i)$ converge to zero exponentially fast, that is,

$$\lim_{n_i \to \infty} \{\log p(i;n_i)\}/n_i = - c_i(\theta)/2 \quad \text{whenever} \quad \theta \notin \Theta_o .$$

The value $c_i(\theta)$ is called the exact (asymptotic) slope of the i-th test. Informally, we can say that the larger the slope, the better the test statistic from this point of view of efficiency. For a thorough discussion of these ideas, see Bahadur (1967). The more precise definitions of T^F, T^P, T^N, T^M, and T^m will now be as follows. Suppose $n = \Sigma n_i$, and let $q(i;n_i) = 1 - p(i;n_i)$. We add a subscript to indicate the dependence on n, and have

$$T_n^F = - 2\Sigma \log p(i;n_i) \qquad \text{(large)}$$

$$T_n^P = - 2\Sigma \log q(i;n_i) \qquad \text{(small)}$$

$$T_n^N = \frac{1}{\sqrt{k}} \Sigma \Phi^{-1}(q(i;n_i)) \qquad \text{(large)}$$

$$T_n^M = \max p(i;n_i) \qquad \text{(small)}$$

$$T_n^m = \min p(i;n_i) \qquad \text{(small)}$$

The words large or small indicate for which values of the test statistic the overall null hypothesis H_{oo} will be rejected. In order to derive overall exact slopes, statistics T_n^P, T_n^M and T_n^M must be converted to reject for large values, as will be done for T_n^P below. Also, suppose each $n_i \to \infty$, and let $\lambda_i = \lim n_i/n$, $i = 1,\ldots,k$. Littell and Folks (1971) show that the exact slopes for four of the above tests based on the combined statistics are respectively:

$$C^F(\theta) = \Sigma_i \lambda_i c_i(\theta); \quad C^N(\theta) = \frac{1}{k} [\Sigma_i (\lambda_i c_i(\theta))^{\frac{1}{2}}]^2;$$

$$C^M(\theta) = k \min_i \lambda_i c_i(\theta); \quad C^m(\theta) = \max_i \lambda_i c_i(\theta).$$

Note that if all λ_i are the same, implying all sample sizes n_i are the same, and if the slopes of the individual tests are the same ($= C(\theta)$ say) the first three exact slopes quoted are all equal to $C(\theta)$; thus in the sense of Bahadur efficiency these methods would be equivalent. We now examine the efficiency of T_n^P.

In order to make use of available theory, we define the new statistic $T_n^{*P} = \{-\log T_n^P\}^{\frac{1}{2}}$. Thus we have T_n^{*P} significant for large values. We then have the result:

Theorem 1. Under the previous assumptions on the asymptotic behaviour of the independent sequences of test statistics, T_n^{*P} has exact slope $k \min_i \lambda_i c_i(\theta)$.

In order to prove Theorem 1, we first obtain 2 lemmas.

Lemma 1.1. Let the n_i, n, λ_i be defined as in Theorem 1 and let $\{x_{n_1}\}, \{y_{n_2}\}, \ldots, \{z_{n_k}\}$ be k sequences of numbers in $(0,1)$ converging exponentially fast to 1, that is, as the n_i, $n \to \infty$,

$$-\frac{1}{n_1}\log(1 - x_{n_1}) \to c_1/2, \ldots, -\frac{1}{n_k}\log(1 - z_{n_k}) \to c_k/2. \qquad (1)$$

Then

$$\lim_{n, n_i \to \infty} \left[-\frac{1}{n}\log\{-2\log x_{n_1} - 2\log y_{n_2} - \ldots - \log z_{n_k}\}\right] = \min_{1 \le i \le k} \frac{\lambda_i c_i}{2}. \qquad (2)$$

Proof of Lemma 1.1.

We show the result for $k = 2$ since the extension can be made directly by induction.

Let $B_n = -\frac{1}{n}\log\{-2\log x_{n_1} - 2\log y_{n_2}\}$ and observe that, for large n,

$B_n \approx -\frac{1}{n}\log\{-\log x_{n_1}y_{n_2}\}$.

Applying the inequality $1-u < -\log u < (1-u)/u$ for $u \in (0,1)$ to $u = x_{n_1}y_{n_2}$, it is readily seen that $B_n \approx -\frac{1}{n}\log(1 - x_{n_1}y_{n_2})$. Let Z be the minimum of x_{n_1} and y_{n_2}. Then $Z^2 \le x_{n_1}y_{n_2} \le Z$, and so $(1-Z)(1+Z) \ge 1 - x_{n_1}y_{n_2} \ge 1 - Z$. Take logarithms, and multiply by $-1/n$; for large n it follows that $-\frac{1}{n}\log(1 - x_{n_1}y_{n_2}) \approx -\frac{1}{n}\log(1-Z)$. Thus

$B_n = -\frac{1}{n}\log\{-2\log x_{n_1} - 2\log y_{n_2}\} \approx -\frac{1}{n}\log(1 - x_{n_1}y_{n_2}) \approx -\frac{1}{n}\log(1-Z)$;

as $n \to \infty$, the limit of B_n is $\min\{\frac{\lambda_1 c_1}{2}, \frac{\lambda_2 c_2}{2}\}$ from (1). This completes the proof.

Lemma 1.2. Let z have the χ^2_{2k} distribution. Then, as $n \to \infty$,

$$\text{Lim}\{-\frac{1}{n}\log P(z < e^{-nt})\} = kt .$$

Proof of Lemma 1.2. From Johnson and Kotz (1971), p. 179, we have $\log P(z < e^{-nt}) = U_n + V_n$, where

$$U_n = \log\left(\frac{e^{-nkt}}{2^k\Gamma(k)/2}\right) \quad \text{and} \quad V_n = \log \sum_{j=0}^{\infty} A_{nj} ,$$

with $A_{nj} = \frac{(-1)^j (e^{-nt})^j}{2^{j+1}(k+j)j!}$.

$S = \sum_{j=0}^{\infty} A_{nj}$ is an alternating series with terms descending in absolute value.

Thus $$\frac{1}{2k} - \frac{e^{-nt}}{2} \leq S \leq \frac{1}{2k},$$

so $\lim S = \frac{1}{2k}$ and hence $\lim(V_n/n) = 0$.

Also, $\lim(U_n/n) = -kt$, so $\lim\{-\frac{1}{n} \log P(z < e^{-nt})\} = \lim\{-(U_n + V_n)/n\} = kt$.

This completes the proof.

We now proceed to:

Proof of Theorem 1.

In order to obtain the exact slope for T_n^{*P}, Bahadur's results are used; see Bahadur (1967), p. 309.

Let $F_n(\cdot)$ denote the null distribution for T_n^{*P}; we first show that for each $\theta \notin \Theta_o$, as $n \to \infty$,

$$\text{Lim}(T_n^{*P}/\sqrt{n}) = (\min \frac{\lambda_i c_i}{2})^{\frac{1}{2}} = b(\theta), \text{ a.s., and} \tag{3}$$

$$\text{Lim}\{-\frac{1}{n} \log[1 - F_n(\sqrt{n}\, t)]\} = kt^2 = f(t), \; t > 0 . \tag{4}$$

Observe that (3) follows directly from (2) in Lemma 1.1, by setting $x_{n_1} = 1 - p(1;n_1)$, $y_{n_2} = 1 - p(2;n_2)$, etc., and observing that the LHS of (2) is then $(T_n^{*P})^2/n$. In order to show (4), note that

$$-\frac{1}{n}\log\{1 - F_n(\sqrt{nt})\} = -\frac{1}{n}\log P(T_n^{*P} > \sqrt{nt}) = -\frac{1}{n}\log P(T_n^P < e^{-nt^2});$$

since T_n^P has the χ^2_{2k} distribution, application of Lemma 1.2 proves (4).

The notation $b(\theta)$ and $f(t)$ is used by Bahadur (1967), who then shows that the exact slope is $C(\theta) = 2f\{b(\theta)\}$. Thus, in this application, the exact slope is

$$C(\theta) = k \min \lambda_i c_i(\theta).$$

This completes the proof of Theorem 1.

Comment. Theorem 1 shows that the exact slopes of T_n^{*P} and T_n^M are the same; thus, from the point of view of Bahadur efficiency, the statistics T_n^P and T_n^M are asymptotically equivalent.

3. AN EXAMPLE.

In Stephens (1986, Examples 8.15.1 and 8.15.2) two examples are quoted of combining test results. One of these is taken directly from Fisher's first illustration of his test method, and we here examine this example in greater detail. The subscript n will now be dropped from the test statistics. Fisher quotes 3 tests of significance which gave p-values of 0.145, 0.263, and 0.087. Then T^F is 11.42 and in the upper tail of χ^2_6 the overall p-level is 0.076. The q-values are 0.855, 0.737 and 0.913, giving $T^P = 1.105$, with a p-level in the lower tail of χ^2_6 equal to 0.018. The values of $\Phi^{-1}(q)$ are 1.059, 0.634, and 1.360, so that

$T^N = 1.763$. The null distribution of T^N is $N(0,1)$, and the p-value is 0.038. Also, $T^M = 0.263$, and $T^m = 0.087$; the null distributions are respectively $F_M(t) = t^k$ and $F_m(t) = 1 - (1-t)^k$, and give p-values of 0.018 and 0.239 respectively. Thus the p-levels can be summarized in the following small table:

T^F	T^P	T^N	T^M	T^m
0.076	0.018	0.038	0.018	0.239

Both the Pearson statistic and T^M are more sensitive than T^F in this example. It is also interesting, in view of the results of Section 2, that T^P and T^M give almost equal p-values.

4. COMBINATIONS OF TESTS BASED ON FINITE SAMPLES.

4.1 Densities for p-levels on H_A or H_B.

We have seen that asymptotic results cannot be applied to alternatives to H_{oo} of type H_B, where some H_{oi} might be true. It may also be the case that p-densities for H_A alternatives would not always satisfy the conditions necessary to discuss Bahadur efficiency. In this section we therefore examine two models for p-densities, f_1 and f_2 below, which have been chosen because they might reasonably model alternatives of type H_A and H_B. For alternative H_A, all p-levels are supposed to become small together; a model for the density could then be

$$f_1(p;\gamma) = (\gamma+1)(1-p)^\gamma, \quad \gamma > 0, \quad 0 < p < 1.$$

This density approaches zero as $p \to 1$, and gives high probability to small p.

For alternative H_B, where some H_{oi} are true, some p-values will remain uniformly distributed $U(0,1)$, and the overall p-density can be modelled as

$$f_2(p;\gamma) = \frac{1}{\gamma} p^{-(1-1/\gamma)}, \quad \gamma > 1, \quad 0 < p < 1.$$

This density also gives high probability for small p, but nevertheless is non-zero as $p \to 1$. It is well known that T^P is the likelihood ratio test statistic for alternative f_1, and T^F is the likelihood ratio test statistic for f_2.

To complete the study, we decided also to construct an alternative non-uniform family that would allow the p-values to converge exponentially fast to zero. We call this alternative $f_3(p;\gamma)$. It corresponds to a p-value constructed as follows. Let $p = e^{-\frac{\gamma}{2}} \cdot x$, where x is taken from a standard exponential distribution, but conditional on $x < e^{\gamma/2}$, so that $p < 1$. For large γ, $f_3(p;\gamma)$ has the following properties, where E and V denote mean and variance.

$$E\{-\frac{1}{n} \log p\} = \frac{\gamma}{2n} - \frac{1}{n} E(\log x)$$

$$V\{-\frac{1}{n} \log p\} = \frac{1}{n^2} V(\log x) .$$

$E(\log$ and $V(\log$ depend on γ ; as $\gamma \to \infty$ their values are $\Gamma'(1) = -.5772$ and $\Gamma''(1) - [\Gamma'(1)]^2 = 1.6449$, where $\Gamma(x)$ denotes the gamma function and $\Gamma'(x)$, $\Gamma''(x)$ its first two derivatives.

So, if $\{x_1^{(1)},\ldots,x_k^{(1)}\}$; $\{x_1^{(2)},\ldots,x_k^{(2)}\},\ldots,\{x_1^{(m)},\ldots,x_k^{(m)}\},\ldots$

is an infinite sequence of standard exponential samples, then by setting

$n_i = n/k$, $\lambda_i = \frac{1}{k}$, $\gamma = Cn_i$, and defining

$$p(i;n_i) = e^{-\frac{\gamma}{2}} x_i^{(n_i)}, \text{ we have}$$

$$\{\log p(i;n_i)\}/n_i = Z(n_i), \text{ say, } = -r/2n + \frac{\log x_i^{(n_i)}}{n_i} .$$

Then $E\{Z(n_i)\} = -\frac{C}{2} + \frac{E(\log x)}{n_i}$ and $V\{Z(n_i)\} = \frac{1}{n_i^2} V(\log x)$.

As $n_i \to \infty$, so that $\gamma \to \infty$, we have

$$\lim_{n_i \to \infty} E\{Z(n_i)\} = -\frac{C}{2} \text{ and}$$

and $V\{z(n_i)\} \to 0$ as $1/n_i^2$; thus by the Borel-Cantelli lemma, $Z(n_i) \to C/2$ almost surely.

4.2 Monte Carlo results.

Tables 1 and 2 report a Monte Carlo study with 10,000 samples, for each of which $k = 5$ or $k = 10$ values of p are used in the overall test statistic. For alternatives f_1 (Table 1a) and f_2 (Table 1b), the power of four of the test statistics is given for 2 different test levels α and for various values of γ. (T^m does badly throughout and is not reported). The table shows that, as expected, T^P does better against alternative f_1 and T^F against f_2. An interesting feature is the relatively good performance of the other statistics in Table 1a (alternative H_A), and especially of T^N in Table 1b (alternative H_B).

In table 2, γ must approach ∞ for asymptotic comparisons to apply. Since $\gamma = cn_i$, and all n_i are equal then, as mentioned in Section 2.1, all tests have the same Bahadur efficiency, a fact which can be seen from the table. For smaller values of γ the table indicates the approach of each of the tests to equal efficiency. It is clear, however that for smaller values of γ, T^P and T^N are again more effective than T^F or T^M in this equi-sample situation.

These results point to the following conclusions. Asymptotic considerations of Bahadur efficiency, which indicate that Fisher's T^F will be the best statistic for considering test results, can be misleading when (a) the individual tests are based on relatively small samples and (b) the alternatives to H_{00} are H_A or H_B above; then Pearson's statistic T^P, and the "normal scores" T^N can often be superior to T^F. Such situations can quite possibly occur, so that it seems wise, in the practical analysis of data sets, to calculate T^P and T^N, and possibly T^M, along with T^F. Although tests using all the statistics will then change the overall α-level, the statistics themselves will throw light on the possible alternatives to H_{00}.

REFERENCES

[1] Birnbaum, A., (1954). Combining independent tests of significance. Journal of the American Statistical Association, 49, 559-74.

[2] Johnson, N.L. and Kotz, S., (1970). Continuous Univariate Distributions I. New York: Wiley.

[3] Littell, R.C. and Folks, J.L., (1971). Asymptotic Optimality of Fisher's Method of Combining Independent Tests. Journal of the American Statistical Association, 66, p. 802.

[4] Littell, R.C. and Folks, J.L., (1973). Asymptotic Optimality of Fisher's Method of Combining Independent Tests II. Journal of the American Statistical Association, 68, p. 193.

[5] Pearson, E.S., (1938). The Probability Integral Transformation for Testing Goodness-of-Fit and combining independent tests of significance. Biometrika 30, 134-148.

[6] Stephens, M.A., (1986). Chapter 8 in Goodness-of-Fit Techniques, (Eds. R. d'Agostino and M.A. Stephens). New York: Marcel Dekker.

Table 1

Power of 4 combined test statistics against alternatives $f_1(p;\gamma)$ or $f_2(p;\gamma)$.

The table gives the percentage of 10,000 samples significant when $k = 5$ or $k = 10$ values p are taken from f_1 (Table 1a) or f_2 (Table 1b) and combined by the test statistics. The two test levels are $\alpha = 0.01$ and $\alpha = 0.05$.

Table 1a. $\alpha = 0.01$

	$k = 5$				$\alpha = 0.05$			
γ	T^F	T^M	T^P	T^N	T^F	T^M	T^P	T^N
1	7	8	10	9	23	30	36	32
2	16	27	34	28	49	61	70	63
3	31	48	58	50	66	81	89	83
4	46	66	77	66	81	91	97	93
5	57	79	88	79	90	95	100	98
6	68	86	94	88	94	98	100	100
7	77	92	97	93	98	99	100	100
8	83	95	100	96	99	99	100	100
9	88	97	100	98	99	100	100	100
10	93	98	100	99	100	100	100	100
20	100	100	100	100	100	100	100	100
	$k = 10$							
1	14	22	30	26	40	51	66	58
2	42	61	81	69	78	84	97	94
3	70	83	98	93	95	96	100	100
4	89	94	100	99	100	99	100	100
5	97	98	100	100	100	100	100	100
6	99	99	100	100	"	"	"	"
7	100	100	100	100	"	"	"	"

Table 1b. $\alpha = 0.01$

	k = 5				$\alpha = 0.05$			
γ	T^F	T^M	T^P	T^N	T^F	T^M	T^P	T^N
2	32	8	14	26	52	19	30	46
3	66	19	31	58	82	34	51	75
4	85	28	45	77	93	46	65	89
5	93	38	57	88	97	55	73	95
6	96	45	64	94	99	61	78	98
7	98	51	70	97	100	65	81	99
8	99	56	73	98	100	69	85	99
9	100	60	76	99	100	72	87	100
10	100	63	78	99	100	74	89	100
20	100	79	91	100	100	86	95	100
	k = 10							
2	53	8	21	45	74	22	44	67
3	89	20	50	84	96	37	71	93
4	98	32	68	96	100	47	83	99
5	100	39	77	99	100	55	89	100
6	100	46	83	100	100	62	92	100
7	"	52	88	"	100	66	94	100
8	"	57	91	"	100	69	96	100
9	"	60	92	"	100	73	97	100
10	"	64	93	"	100	75	97	100
20	"	81	98	"	100	86	99	100

Table 2

Power of 4 combined test statistics against alternative $f_3(p;\gamma)$.

The table gives the percentage of 10,000 samples of significant when $k = 5$ or $k = 10$ values p are taken from f_3 (Table 2) and combined by the test statistics. The two test levels are $\alpha = .01$ and $\alpha = .05$.

$\alpha = .01$ $\qquad\qquad$ $\alpha = .05$

γ	T^F	T^M	T^P	T^N	T^F	T^M	T^P	T^N
1	6	7	9	8	21	21	25	25
2	13	17	21	19	37	40	48	47
3	32	42	52	47	67	68	80	79
4	66	77	88	83	93	92	98	98
5	92	96	99	98	100	99	100	100

$k = 10$

γ	T^F	T^M	T^P	T^N	T^F	T^M	T^P	T^N
1	12	11	17	17	33	25	40	41
2	31	27	45	45	62	47	40	41
3	70	61	85	86	92	77	96	97
4	97	91	100	100	100	96	100	100
5	100	100	100	100	100	100	100	100

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 444	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) On Combining Independent Significance Tests		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) F. J. O'Reilly and Michael A. Stephens		8. CONTRACT OR GRANT NUMBER(s) N00014-89-J-1627
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE July 24, 1991
		13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Combination of tests of significance, Fisher's method, probability integral transformation		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) PLEASE SEE FOLLOWING PAGE.		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

TECHNICAL REPORT NO. 444

20. ABSTRACT

How to combine the results of k independent tests of significance has long been an important problem in statistics. The problem can arise, for example, in such diverse situations as when tests on the mean survival time after diagnosis of a terminal disease are made on k groups of patients in different hospitals, or when the sets of observations in k cells of an ANOVA table are separately tested for normality. An important feature of such tests is that often the individual sample sizes will be small, so that asymptotic results will not necessarily be valid. We suggest below that they might even be misleading in some situations.