A DISTRIBUTION-FREE TEST FOR REGRESSION PARAMETERS

By H. E. Daniels

Cambridge University, England, and University of Chicago[1]

1. Introduction and Summary. Brown and Mood [1], [4] have recently given convenient distribution-free methods of testing and setting up confidence regions for the parameters of a linear regression model. Their technique, which is based on the use of medians, allows the parameters to be considered singly or simultaneously as required. Theil [5] gives two methods of constructing confidence intervals for single parameters, a "complete" method using rank correlation which is valid under the conditions assumed by Brown and Mood, and an "incomplete" method valid under wider conditions but not making full use of the data. For several parameters simultaneously, he obtains confidence regions in the weak sense of covering the true parameter point with probability not less than an assigned value.

In the present paper we give a new distribution-free test for the hypothesis that all regression parameters have specified values, assuming only that the residuals are independent and have probability $\frac{1}{2}$ of being positive or negative. It can be used to set up exact confidence regions for the true parameter point. The new test avoids a defect which is shown to appear in the corresponding Brown and Mood test when the sample is not large. The distribution of the test statistic is found explicitly only for the case of two parameters, though in principle the idea extends to any number of parameters. The presence of repeated values of the independent variable necessitates certain modifications in the test, and a method of computing the appropriate distributions in such cases is described.

2. The m test. Suppose we have n pairs of observations (x_i, y_i), $i = 1, 2, \cdots, n$, from a bivariate population such that

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2.1}$$

where α, β are unknown parameters and the ϵ_i's are independently distributed errors such that $\Pr(\epsilon_i > 0) = \Pr(\epsilon_i < 0) = \frac{1}{2}$ for all i. The x_i's are assumed to have assigned values, supposed for the present all to be different and arranged in increasing order. By the usual argument the test we obtain will still be valid if the x_i's have any joint distribution provided that for every set of x_i's the conditional distribution of the ϵ_i's satisfies the above conditions. We wish to test the hypothesis that $\alpha = \alpha_0$, and $\beta = \beta_0$.

Rewrite (2.1) in the form

$$\alpha = (-x_i)\beta + y_i - \epsilon_i . \tag{2.2}$$

Received 9/9/53.

[1] Research carried out at the Statistical Research Center, University of Chicago, under sponsorship of the Statistics Branch, Office of Naval Research.

In the (β, α) plane (2.2) defines n straight lines with successively decreasing gradients $-x_1, -x_2, \cdots, -x_n$. On the null hypothesis all these lines must pass through the point (β_0, α_0). The ϵ_i's are, however, not known and we consider instead the n lines

$$\alpha = (-x_i)\beta + y_i \tag{2.3}$$

which are parallel to the corresponding lines of (2.2). In general they are not concurrent but partition the plane into $\frac{1}{2}(n^2 + n + 2)$ polygonal regions of which $2n$ are open regions extending to infinity and the remaining $\frac{1}{2}(n - 1)(n - 2)$ form a set of contiguous closed regions (see Sec. 7). Each line passes above or below the point (β_0, α_0) according as the corresponding ϵ_i is positive or negative; under the null hypothesis either event is equally likely.

So, speaking crudely, one expects that for typical samples the point (β_0, α_0) will be situated somewhere near the middle of the set of closed regions rather than in or near one of the open regions. This idea motivates the following test procedure. Assign a score to each region equal to the minimum number m of lines which have to be crossed to escape from it into one of the open regions. Reject the hypothesis $\alpha = \alpha_0, \beta = \beta_0$ if the score m for the region containing (β_0, α_0) is significantly low. We shall refer to this test as the m test.

3. Characterization of regions. Let $s_i = \operatorname{sgn} \epsilon_1 = \operatorname{sgn} (y_i - \alpha - \beta x_i)$. We call the ordered array of signs $s_1, s_2, \cdots, s_n$ the *signature of the sample* under the hy-

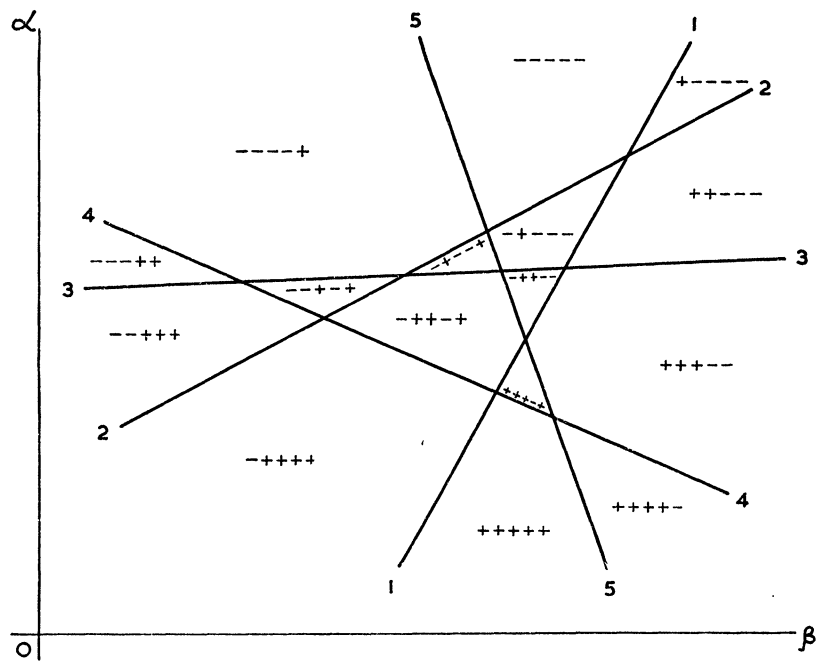

FIG. 1

pothesis (β, α). Each of the polygonal regions in the plane is characterized by the signature which the sample would have if the point (β, α) lay within it. We call this the *signature of the region.*

All 2^n possible combinations of signs are equally likely, though only $\frac{1}{2}(n^2 + n + 2)$ of these appear as signatures of regions for a particular sample. But the $2n$ open regions must each bear a characteristic signature whatever the sample chosen, since a point (β, α) can always be found sufficiently far out in an open region for the signs of the corresponding ϵ_i's to be unaffected by any given parallel displacements of the lines. In particular there is one open region lying between lines 1 and n for which the ϵ_i's are all positive, and a conjugate open region on the other side of the figure lying between the same two lines for which the ϵ_i's are all negative. Starting from each of these regions we can perform a clockwise tour of the open regions, changing the appropriate sign whenever a line is crossed. In this way we obtain the following n pairs of conjugate signatures characterizing the open regions, numbered according to the last line crossed:

$$
\begin{array}{rl|l}
 & 1 & - + + \cdots + + +; \quad + - - \cdots - - -. \\
 & 2 & - - + \cdots + + +; \quad + + - \cdots - - -. \\
 & 3 & - - - \cdots + + +; \quad + + + \cdots - - -. \\
(3.1) & \cdots & \cdots \quad \cdots \quad \cdots \; ; \quad \cdots \quad \cdots \quad \cdots \; . \\
 & n-2 & - - - \cdots - + +; \quad + + + \cdots + - -. \\
 & n-1 & - - - \cdots - - +; \quad + + + \cdots + + -. \\
 & n & - - - \cdots - - -; \quad + + + \cdots + + +.
\end{array}
$$

The signatures of the closed regions can be filled in similarly. Fig. 1 shows a particular set of signatures for $n = 5$. As each successive line is crossed in escaping to an open region from the point (β_0, α_0), the signature of the region traversed changes by one sign at a time. The score m for the given sample must therefore be the minimum number of sign disagreements between the sample signature under (β_0, α_0) and any of the $2n$ signatures (3.1).

4. Distribution of m. We now derive the distribution of sample scores. Let t_i, t_i' be the numbers of sign disagreements with the ith pair of conjugate signatures in (3.1). Obviously $t_i' = n - t_i$ and the sample score is $m = \min_i \min(t_i, n - t_i)$. This cannot exceed $\frac{1}{2}n$, and in fact is proved later not to exceed $[\frac{1}{2}(n - 1)]$. We require $P_n(m_0) = \Pr(m \leqq m_0)$ where

$$
1 - P_n(m_0) = \Pr(m > m_0) = \Pr(m_0 < t_i < n - m_0, \text{ all } i). \tag{4.1}
$$

The method of arriving at the distribution is most easily understood from an example. Suppose $n = 11$ and the sample signature is $+ + - + - - + + + - +$. Comparing it in turn with the signatures (3.1) we get the values of t_i, t_i'

shown in the following table

i	1	2	3	4	5	6	7	8	9	10	11
s_i	+	+	−	+	−	−	+	+	+	−	+
t_i	5	6	5	6	5	4	5	6	7	6	7
t_i'	6	5	6	5	6	7	6	5	4	5	4
w_i	1	2	1	2	1	0	1	2	3	2	3.

The score is $m = 4$. The last row of the table gives the values of the cumulative sum $w_i = s_1 + s_2 + \cdots + s_i$, where s_i is interpreted as ± 1.

In the general case it is evident that if $t(= t_n')$ is the total number of negative signs in the sample signature, $t_i = t + w_i$. In particular, $t_n = n - t = t + w_n$. Since $m_0 < t_n < n - m_0$ is equivalent to $m_0 < t < n - m_0$, (4.1) may be rewritten in the form

$$\Pr(m > m_0) = \sum_{t=m_0+1}^{n-m_0-1} p_n(t, m_0)$$

(4.2)

$$p_n(t, m_0) = \Pr\{m_0 < t + w_i < n - m_0, \quad i = 1, 2, \cdots, n-1; \quad t + w_n = n - t\}.$$

Now $p_n(t, m_0)$ is just the probability that, starting at the point t and proceeding by independent equally likely random steps of ± 1, one arrives after n steps at the point $n - t$, having avoided absorption on boundaries at the points m_0, $n - m_0$. Solving the random walk problem in the usual manner [3] we find

$$p_n(t, m_0) = \frac{1}{2^n} \sum_{j=-\infty}^{\infty} \left\{ \binom{n}{t + j(n - 2m_0)} - \binom{n}{m_0 + j(n - 2m_0)} \right\} \tag{4.3}$$

Hence, after some reduction,

$$p_n(m_0) = \frac{(n - 2m_0)}{2^{n-1}} \sum_{j \geqq 0} \binom{w}{(n - m_0) + j(n - 2m_0)}, \tag{4.4}$$

the series terminating at $j = [m_0/(n - 2m_0)]$.

Confidence regions for (β, α) with confidence coefficients $1 - P_n(m_0)$ are provided by the polygons made up of all regions for which $m > m_0$. In particular $P_n(0) = n/2^{n-1}$ is the chance that (β, α) lies in an open region, as is otherwise evident from the fact that the open regions account for $2n$ signatures out of the 2^n equally likely possibilities. So the largest closed polygon formed by the lines is a confidence region for (β, α) with coefficient[2] $1 - n/2^{n-1}$.

The maximum possible value[3] of m is $[\frac{1}{2}(n - 1)]$. For when n is odd, say

[2] This result was given in [2] for a particular case of the model (7.1) below.

[3] L. J. Savage points out that this is an immediate consequence of the fact that any transversal parallel to one of the lines crosses the $n - 1$ remaining lines.

$n = 2k + 1$, we have

$$p_{2k+1}(k) = \frac{1}{2^{2k}} \sum_{j \geqq 0} \binom{2k+1}{k+j+1} = \frac{1}{2^{2k+1}} \sum_{j=0}^{2k+1} \binom{2k+1}{j} = 1,$$

so that $m \leqq k$. On the other hand when n is even, say $n = 2k$,

$$p_{2k}(k-1) = \frac{1}{2^{2k-2}} \sum_{j \geqq 0} \binom{2k}{k+2j+1} = \frac{1}{2^{2k-2}} \sum_{j \geqq 0} \left\{ \binom{2k-1}{k+2j} + \binom{2k-1}{k+2j+1} \right\}$$

$$= \frac{1}{2^{2k-2}} \sum_{j \geqq 0} \binom{2k-1}{k+j} = \frac{1}{2^{2k-1}} \sum_{j=0}^{2k-1} \binom{2k-1}{j} = 1,$$

so that $m \leqq k - 1$. Hence $m \leqq [\frac{1}{2}(n-1)]$ in all cases. Equality is, however, not necessarily attained in every sample. In the example of Fig. 1, $n = 5$ but $m = 0$ or 1 only. By moving line 4 parallel to itself until it passes beyond the intersection of lines 3 and 5, a region is formed for which $m = 2$.

Values of $P_n(m_0)$ for n ranging from 3 to 30 are given in Table I. When n is

TABLE I

$P_n(m_0) = \Pr(m \leqq m_0)$, Cf. Equation (4.1)

n \ m_0	0	1	2	3	4	5	6	7	8	9	10	11	12	13
3	.750	1.0												
4	.500	1.0												
5	.322	.938	1.0											
6	.188	.750	1.0											
7	.109	.547	.984	1.0										
8	.063	.375	.875	1.0										
9	.035	.246	.703	.996	1.0									
10	.020	.156	.527	.938	1.0									
11	.011	.097	.376	.806	.999	1.0								
12	.006	.059	.258	.645	.969	1.0								
13	.003	.035	.171	.489	.873	1.0	1.0							
14	.002	.021	.111	.356	.733	.984	1.0							
15	.001	.012	.071	.250	.583	.917	.999	1.0						
16	0	.007	.044	.171	.444	.800	.992	1.0						
17	0	.004	.027	.114	.327	.661	.944	1.0	1.0					
18		.002	.016	.075	.233	.523	.850	.996	1.0					
19		.001	.010	.048	.163	.399	.725	.964	1.0	1.0				
20		.001	.006	.030	.111	.296	.591	.887	.998	1.0				
21		0	.003	.019	.074	.214	.658	.776	.977	1.0	1.0			
22		0	.002	.012	.049	.151	.356	.651	.916	.999	1.0			
23			.001	.007	.032	.104	.265	.526	.818	.985	1.0	1.0		
24			.001	.004	.020	.071	.193	.413	.701	.937	1.0	1.0		
25			0	.003	.013	.048	.137	.315	.580	.853	.990	1.0	1.0	
26			0	.002	.008	.031	.096	.235	.466	.745	.952	1.0	1.0	
27				.001	.005	.021	.066	.172	.364	.629	.880	.993	1.0	1.0
28				.001	.003	.032	.045	.124	.278	.515	.782	.964	1.0	1.0
29				0	.002	.008	.030	.087	.208	.410	.672	.903	.996	1.0
30				0	.001	.005	.020	.061	.153	.320	.560	.814	.973	1.0

large (4.4) approximates to

$$p_n(m_0) \sim 4z_0 \sum_{j=0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\tfrac{1}{2}(2j+1)^2 z_0^2] \tag{4.5}$$

where $z_0 = (n - 2m_0)/\sqrt{n}$. The 5 per cent and 1 per cent values of z_0 are 3.023 and 3.562, respectively.

5. Comparison with the Brown and Mood test. It is of interest to compare the m test with the corresponding Brown and Mood test ([1], p. 407) which is valid under the same assumptions. For convenience we suppose n to be even. Brown and Mood separate the observations into two groups of $\frac{1}{2}n$ according as the x_i's lie below or above the median. Putting $\alpha = \alpha_0$, and $\beta = \beta_0$, they count the numbers r_1, r_2 of positive ϵ_i's in the first and second groups, respectively, and reject the hypothesis when

$$A = (8/n)\{(r_1 - n/4)^2 + (r_2 - n/4)^2\} \tag{5.1}$$

is significantly large. For moderately large n, A is approximately a χ^2 variable with 2 degrees of freedom.

The greatest possible value of A is n, which it attains only if r_1 and r_2 are each either 0 or $\frac{1}{2}n$. From the viewpoint of the present paper A can therefore be regarded as measuring the closeness of agreement of the sample signature with any of the four signatures,

$$\begin{array}{cccccccccc} & \overbrace{\qquad}^{\frac{1}{2}n} & & \overbrace{\qquad}^{\frac{1}{2}n} & & \overbrace{\qquad}^{\frac{1}{2}n} & & \overbrace{\qquad}^{\frac{1}{2}n} & & \\ (5.2) & +++ & \cdots & ++ & +++ \cdots ++ & --- & \cdots & -- & --- & \cdots -- \\ & --- & \cdots & -- & +++ \cdots ++ & +++ & \cdots & ++ & --- & \cdots -- \end{array}$$

Since the minimum number of sign discrepancies is $\frac{1}{2}n - |r_1 - n/4| - |r_2 - n/4|$ an alternative statistic more in the spirit of the m test is

$$B = (2/\sqrt{n})(|r_1 - n/4| + |r_2 - n/4|).$$

which for moderately large n is such that

$$\Pr(B \geqq B_0) \sim 4\Phi(B_0)(1 - \Phi(B_0))$$

where $\Phi(x)$ is the cumulative normal distribution function. The 5 per cent and 1 per cent values of B_0 are 2.237 and 2.806, respectively.

But whether A or B is used, the fact that the remaining signatures of (3.1) are not considered makes the test inadequate in the following respect. Suppose the x_i's have assigned values. Let $\frac{1}{2}n$ be even and consider the power of the test with respect to alternatives (β, α) such that either

$$x_{n/4} < (\alpha_0 - \alpha)/(\beta_0 - \beta) < x_{n/4+1}, \tag{5.3}$$

$$x_{3n/4} < (\alpha_0 - \alpha)/(\beta_0 - \beta) < x_{3n/4+1}. \tag{5.4}$$

The vector $(\beta_0 - \beta, \alpha_0 - \alpha)$ is directed towards one of the four open regions

$$(5.5)\qquad \begin{array}{llll} \overbrace{++\cdots+}^{\frac{1}{4}n} & \overbrace{--\cdots-}^{\frac{3}{4}n} & \overbrace{--\cdots-}^{\frac{1}{4}n} & \overbrace{++\cdots+}^{\frac{3}{4}n} \\ \overbrace{++\cdots+}^{\frac{3}{4}n} & \overbrace{--\cdots-}^{\frac{1}{4}n} & \overbrace{--\cdots-}^{\frac{3}{4}n} & \overbrace{++\cdots+}^{\frac{1}{4}n} \end{array}.$$

If the true parameter point (β, α) is sufficiently distant from (β_0, α_0), the value $A = \frac{1}{2}n$ will be practically certain to occur every time. So, when the significance level is less than $\Pr(A \geqq \frac{1}{2}n)$, the power of the test against such alternatives actually tends to zero as (β, α) recedes indefinitely from (β_0, α_0). While for large

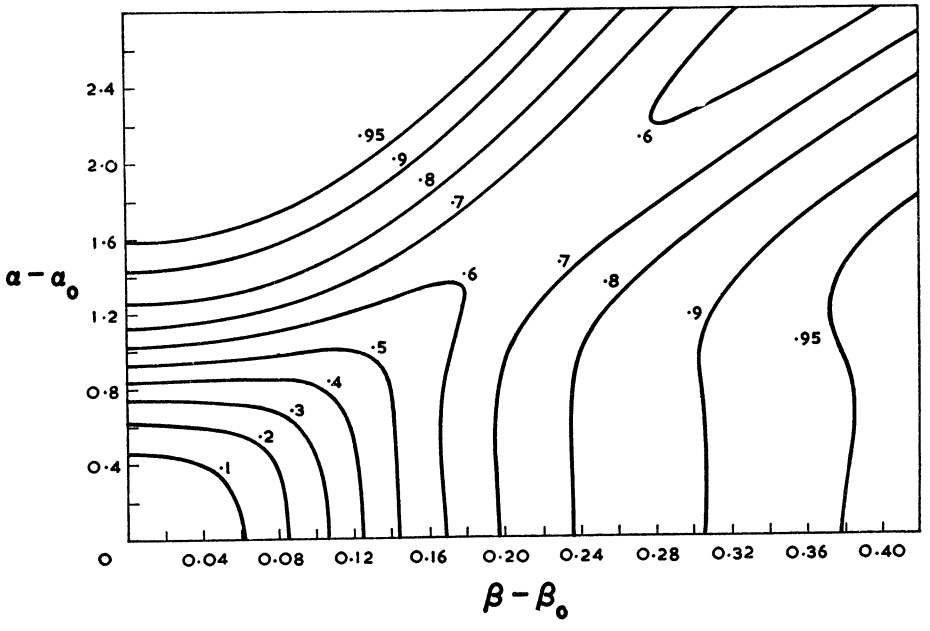

FIG. 2

samples $\Pr(A \geqq \frac{1}{2}n)$ is too small to matter, the effect may be important for moderate values of n. For example, when $n = 16$ we find, using the exact distribution of r_1 and r_2, that the first significance level satisfying the above condition is $\Pr(A \geqq 8.5) = 0.0152$, which is not unduly small. The phenomenon is illustrated in Fig. 2, which shows contours of the power function of the A test in this particular case, when the residuals are assumed normal with unit variance.

6. Power of the m test. The power of the m test is now discussed. Under the alternative (β, α), $\epsilon_i = y_i - \alpha - \beta x_i$ has probability $\frac{1}{2}$ of being positive or negative and

$$y_i - \alpha_0 - \beta_0 x_i = \epsilon_i - (\alpha_0 - \alpha) - (\beta_0 - \beta)x_i . \tag{6.1}$$

Let

$$p_i = \Pr\{\epsilon_i > (\alpha_0 - \alpha) + (\beta_0 - \beta)x_i\}, \qquad q_i = 1 - p_i . \tag{6.2}$$

The probability $P_n(m_0 \mid \alpha, \beta)$ of rejection on the alternative hypothesis is still given by (4.2), but the random walk defining $p_n(t, m_0)$ is now such that the ith step takes values $+1$ or -1 with probabilities p_i and q_i, respectively. A convenient solution of the random walk problem with general p_i, q_i is not known, though in particular cases $p_n(t, m_0)$ and hence $P_n(m_0 \mid \alpha, \beta)$ can be computed by the direct step by step procedure which is not too arduous for moderate values of n.

The problem can be solved simply when $p_i = p$ for all i. This is the case if the alternatives are (β_0, α) and all the ϵ_i's have the same distribution. Then for all i,

$$p_i = \Pr(\epsilon_i > \alpha_0 - \alpha) = p \tag{6.3}$$

and by standard methods we find

$$\begin{aligned} p_n(t, m_0) &= p^t q^{n-t} \sum_{j=-\infty}^{\infty} \left\{ \binom{n}{t + j(n - 2m_0)} - \binom{n}{m_0 + j(n - 2m_0)} \right\} \\ P_n(m_0 \mid \alpha, \beta_0) &= 1 - \sum_{t=m_0+1}^{n-m_0-1} p_n(t, m_0). \end{aligned} \tag{6.4}$$

When n is large write $z_0 = (n - 2m_0)/\sqrt{n}$ as before, and put

$$p_i = \tfrac{1}{2}(1 - \mu/\sqrt{n}), \qquad q_i = \tfrac{1}{2}(1 + \mu/\sqrt{n}).$$

The limiting form of the power function turns out to be

$$\begin{aligned} P_n(m_0 \mid \alpha, \beta_0) = 1 &- \sum_{j=-\infty}^{\infty} e^{2j\mu z_0} [\Phi((2j + 1)z_0 + \mu) - \Phi((2j - 1)z_0 + \mu)] \\ &+ 2e^{-\mu^2/2} \frac{[e^{\mu z_0} - e^{-\mu z_0}]}{\mu} \sum_{j=0}^{\infty} \frac{e^{-(2j+1)^2 z_0^2/2}}{\sqrt{2\mu}} \end{aligned} \tag{6.5}$$

which reduces to (4.5) when $\mu = 0$. If the ϵ_i's have a common density function $f(\epsilon)$, $p_i = \int_{\alpha_0 - \alpha}^{\infty} f(\epsilon)\, d\epsilon \sim \frac{1}{2} - (\alpha_0 - \alpha) f(0)$, and

$$\mu = 2(\alpha_0 - \alpha) f(0) \sqrt{n}. \tag{6.6}$$

It is easy to find the limiting form for large n of the power function of the corresponding Brown and Mood test against the same alternatives, using either the A or B statistic. The A distribution has the noncentral χ^2 form with 2 degrees of freedom and parameter μ^2, while

$$\Pr(B \geqq B_0 \mid \alpha, \beta_0) \sim 1 - \{2\Phi(B_0) - 1\}\{\Phi(B_0 + \mu) + \Phi(B_0 - \mu) - 1\}. \tag{6.7}$$

TABLE II

Asymptotic power functions of four tests, at .05 level, for alternatives $\alpha \neq \alpha_0$, $\beta = \beta_0$. Here $\mu = \sqrt{n}\{1 - 2 \Pr(\epsilon_i > \alpha - \alpha_0)\}$

μ	A	B	m	Normal
0	0.05	0.05	0.05	0.05
0.790	0.10	0.10	0.09	0.13
1.316	0.20	0.20	0.16	0.29
1.666	0.30	0.30	0.25	0.45
1.958	0.40	0.41	0.33	0.59
2.226	0.50	0.51	0.42	0.71
2.493	0.60	0.61	0.52	0.81
2.775	0.70	0.71	0.62	0.89
3.104	0.80	0.81	0.73	0.95
3.557	0.90	0.91	0.85	0.99

Columns 1 and 4 were computed from *Table of Noncentral* χ^2, by Evelyn Fix, University of California Press, 1949, and "Charts of the power function for analysis of variance tests, derived from the noncentral F distribution," by E. S. Pearson and H. O. Hartley, *Biometrika*, Vol. 38, Parts I and II (June, 1951).

In Table II the large sample power functions of the m, A and B tests are compared for such alternatives. The m test turns out to be somewhat less powerful than the A or B tests against these alternatives, as might be expected from the way the latter were constructed. The corresponding limiting power function of the standard F test under the normality assumption is also tabulated. For large n the distribution of F approximates to that of noncentral χ^2 with 2 degrees of freedom and parameter $\frac{1}{2}\pi\mu^2$.

7. Some generalizations. The m test can also be used for the parameters of the model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \tag{7.1}$$

under similar assumptions about the x's and the ϵ's. The lines in the (β_2, β_1) plane are

$$\beta_1 = (-x_{2i}/x_{1i})\beta_2 + y_i/x_{1i}. \tag{7.2}$$

If they are numbered in order of increasing x_{2i}/x_{1i}, the argument goes through exactly as before.

In principle the m test can be extended to the case of k parameters, though the distribution of m is not easily obtainable for $k > 2$. The model is then

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i. \tag{7.3}$$

The n hyperplanes

$$\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} = y_i \tag{7.4}$$

partition the k-dimensional parameter space into $1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{k}$ regions of which $2\{1 + \binom{n-1}{1} + \binom{n-1}{2} + \cdots + \binom{n-1}{k-1}\}$ are open and $\binom{n-1}{k}$ are closed. This is proved as follows.

Let $T_{n,k}$ be the number of regions into which a k dimensional Euclidean space is partitioned in general by n hyperplanes (i.e. $[k - 1]$ flats). Let $O_{n,k}$ of these be open and the remaining $C_{n,k}$ closed.

Consider the effect of adding one more hyperplane. It adds a new region for every existing region it intersects. The number of new regions added is therefore $T_{n,k-1}$ since the new hyperplane is itself partitioned into $T_{n,k-1}$ regions by the n existing hyperplanes. Hence $T_{n+1,k} = T_{n,k} + T_{n,k-1}$.

Clearly $T_{0,k} = 1$ for all k. Also $T_{n,1} = n + 1$, so that we can take $T_{n,0} = 1$. Let

$$G_n(z) = \sum_{k=0}^{\infty} T_{n,k} z^k.$$

Then $G_{n+1}(z) = (1 + z)G_n(z)$. Since $G_0(z) = 1/(1 - z)$, we get $G_n(z) = (1 + z)^n/(1 - z)$. Hence

$$T_{n,k} = 1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{k}.$$

The number of new closed regions added by the extra hyperplane is the same as the number of closed regions formed in itself by intersections with the n existing hyperplanes. Hence $C_{n+1,k} = C_{n,k} + C_{n,k-1}$. Since $C_{n,1} = n - 1$ we can take $C_{n,0} = 1$. Let

$$H_n(z) = \sum_{k=0}^{\infty} C_{n,k} z^k.$$

Then $H_{n+1}(z) = (1 + z)H_n(z)$. We have $C_{1,0} = 1$ and $C_{1,k} = 0$ for $k \geqq 1$ so that $H_1(z) = 1$ and $H_n(z) = (1 + z)^{n-1}$, and $C_{n,k} = \binom{n-1}{k}$. It follows easily that the number of open regions can be written as

$$O_{n,k} = 2\{1 + \binom{n-1}{1} + \binom{n-1}{2} + \cdots + \binom{n-1}{k-1}\}.$$

As before the sample signature is compared with the signatures of the open regions to find m. Note that the largest closed polyhedron is a confidence region with coefficient

$$1 - \frac{1}{2^{n-1}}\{1 + \binom{n-1}{1} + \binom{n-1}{2} + \cdots + \binom{n-1}{k-1}\} = \frac{1}{2^{n-1}}\{1 + \binom{n-1}{1} + \cdots + \binom{n-1}{n-k}\}$$

8. Repeated values of x. So far the possibility of repeated or "tied" values of x has been excluded. The presence of such tied x's introduces a difficulty similar to that found with the Brown and Mood test. We return to model (2.1) and consider the example for $n = 6$, illustrated in Fig. 3, where $x_3 = x_4 = x_5$.

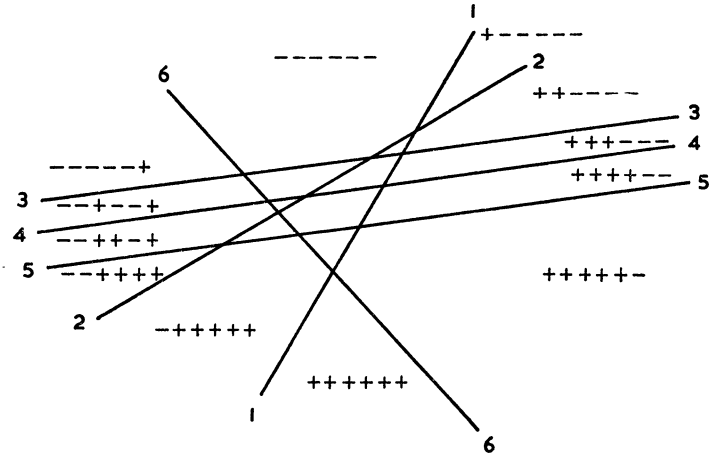

FIG. 3

The lines corresponding to the three tied x_i's are parallel, and the numbers 3, 4, 5 can be assigned at random to the three lines. With the ordering shown, the open regions of signature $---+++$, $----++$ have disappeared while the regions $--+---+$, $--+++-+$ previously regarded as closed are now open. With the ordering 543 replacing 345, the conjugate regions are the ones affected, while with any other ordering such as 435 one open region on either side is replaced by a region previously considered closed.

By assigning a random order to the tied x_i's the m test for (β_0, α_0) may still be applied as before, even when such ties are present. However, there will exist significance levels such that, for alternatives (β, α) lying in the direction of the tied lines, the power of the test never attains unity no matter how distant the point (β, α) is from (β_0, α_0), since some previously closed regions must extend to infinity. To avoid this difficulty the test has to be modified by relabelling the open regions.

Suppose the x_i's to occur in l tied groups with n_j in the jth group and $\sum_1^l n_j = n$. Thus

$$x_1 = x_2 = \cdots = x_{n_1} < x_{n_1+1} = x_{n_1+2} = \cdots x_{n_1+n_2} < \cdots < x_{n-n_1+1} = \cdots = x_n .$$

Fig. 4 illustrates the case $l = 3$, $n_1 = 3$, $n_2 = 4$, $n_3 = 2$. The 9 lines produce only 36 regions in the plane instead of the full 46, and in general $\frac{1}{2}n_j(n_j - 1)$ regions disappear for each tied group of n_j lines. Of course, some of the values of y in a tied group may also coincide, but in such cases we regard the corresponding regions as being present but of zero width. The test statistic is again the minimum number m of lines to be crossed in escaping to infinity from the point (β_0, α_0).

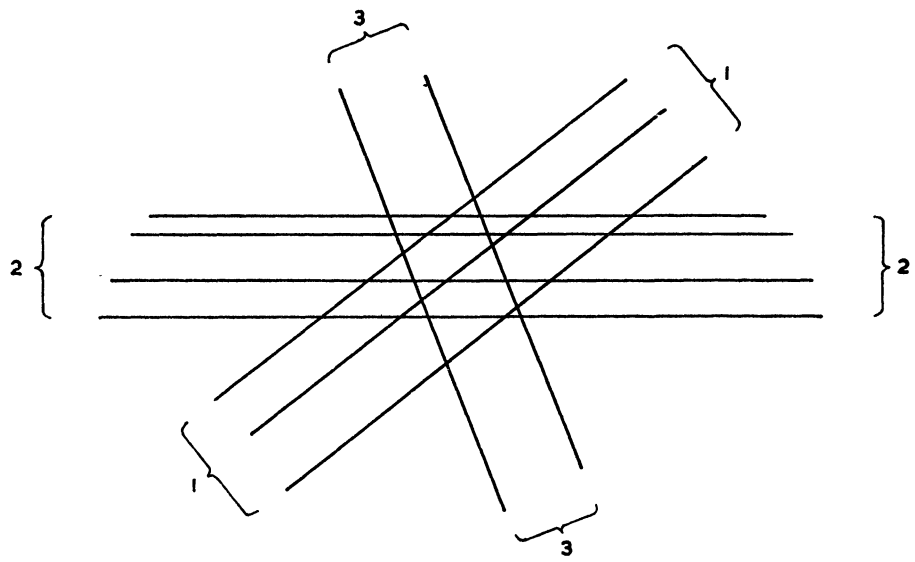

FIG. 4

The open regions are of two types, (i) those lying between two successive bands of tied lines and (ii) those formed between tied lines of the same group. The signatures of type (i) regions are invariant under sampling (for fixed x_i's) while those of type (ii) are not. In particular there is one type (i) region for which every ϵ_i is positive. We shall write its signature in the contracted form

$$(+)\ (+)\ (+)\ \cdots\ (+) \tag{8.1}$$

where the jth bracket $(+)$ stands for the n_j plus signs corresponding to the jth tied group. The open regions may then be characterized in the following way. Starting at the region just mentioned, we describe a clockwise circuit of the open regions. As we move through the first group of type (ii) regions the first $(+)$ changes successively into $n_1 - 1$ different sets of $+$'s and $-$'s, not all $-$, but the remaining brackets are unchanged. The particular combinations of signs for the $n_1 - 1$ regions will depend on the given sample, but in random sampling all $2^{n_1} - 2$ combinations of signs which are not all $+$ or all $-$ are possible for these regions. On passing beyond them into the next type (i) region every ϵ_i for $i = 1, 2, \cdots, n_1$ becomes negative and the signature of this region may be written in contracted notation as

$$(-)\ (+)\ (+)\ \cdots\ (+) \tag{8.2}$$

We introduce the symbol $(\sim)$ to denote any set of signs whatever for the ϵ_i's of a particular tied group of lines. Then the signature of any of the regions so far considered is included in the formula

$$(\sim)\ (+)\ (+)\ \cdots\ (+) \tag{8.3}$$

Therefore, this may be called the signature of the whole region lying between the last line of the lth tied group and the first line of the second tied group. Similarly

$$(8.4) \qquad (-)\ (\sim)\ (+)\ (+)\ \cdots\ (+)$$

characterizes the region extending from the last line of the first tied group to the first line of the third tied group. The two regions (8.3) and (8.4) are of course not disjoint, since both contain (8.2). Proceeding in this way we can cover all open regions by the following overlapping set of regions arranged in l conjugate pairs:

$$(8.5) \quad \begin{cases} 1 & (\sim)\ (+)\ (+)\ \cdots\ (+)\ (+) & (\sim)\ (-)\ (-)\ \cdots\ (-)\ (-) \\ 2 & (-)\ (\sim)\ (+)\ \cdots\ (+)\ (+) & (+)\ (\sim)\ (-)\ \cdots\ (-)\ (-) \\ 3 & (-)\ (-)\ (\sim)\ \cdots\ (+)\ (+) & (+)\ (+)\ (\sim)\ \cdots\ (-)\ (-) \\ \cdots & \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots & \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots\ \cdots \\ l-1 & (-)\ (-)\ (-)\ \cdots\ (\sim)\ (+) & (+)\ (+)\ (+)\ \cdots\ (\sim)\ (-) \\ l & (-)\ (-)\ (-)\ \cdots\ (-)\ (\sim) & (+)\ (+)\ (+)\ \cdots\ (+)\ (\sim) \end{cases}$$

9. The modified m distribution. Suppose the sample signature under (β_0, α_0) consists successively of r_1 +'s out of n_1, r_2 +'s out of n_2, $\cdots$ r_l +'s out of n_l. No ordering of the signs within tied groups is necessary. The numbers of disagreements of sign with the jth pair of conjugate signatures in (8.5) are d_j, d_j', where $d_j + d_j' = n - n_j$, and

$$(9.1) \qquad d_j = r_1 + r_2 + \cdots + r_{j-1} + (n_{j+1} - r_{j+1}) + (n_{j+2} - r_{j+2}) + \cdots + (n_l - r_l).$$

There are no disagreements of sign with the jth group, by definition of $(\sim)$. We require the distribution of

$$(9.2) \qquad m = \min_j \min\,(d_j, n - n_j - d_j).$$

It is perhaps unfortunate, when there are no tied x's, that d_j does not reduce to the previous t_j since the characterization of the open regions is different, but the present method seems most expeditious in the tied situation. On writing $s_j = 2r_j - n_j$, (9.1) becomes

$$(9.3) \qquad d_j = \tfrac{1}{2}(n - n_j) + \tfrac{1}{2}(s_1 + s_2 + \cdots + s_{j-1} - s_{j+1} - s_{j+2} - \cdots - s_l)$$

and

$$(9.4) \quad \begin{aligned} \Pr(m > m_0) &= \Pr(m_0 < d_j < n - n_j - m_0,\ j = 1, 2, \cdots, l) \\ &= \Pr\{\,|\, s_1 + s_2 + \cdots + s_{j-1} - s_{j+1} - s_{j+2} - \cdots - s_l \,| \\ &\qquad < n - n_j - 2m_0,\ j = 1, 2, \cdots l\}. \end{aligned}$$

The special case where $l = 2$ and the tied x's are respectively -1 and $+1$ is of interest. The model is

$$y_j = \begin{cases} \alpha - \beta + \epsilon_j, & j = 1, 2, \cdots, n_1, \\ \alpha + \beta + \epsilon_j, & j = n_1 + 1, n_1 + 2, \cdots, n_1 + n_2. \end{cases}$$

We are simultaneously testing whether the medians of two independently sampled populations are respectively $\alpha_0 - \beta_0$ and $\alpha_0 + \beta_0$. Then (9.4) reduces to

$$\Pr(m > m_0) = \Pr\{ |s_1| < n_1 - 2m_0, |s_2| < n_2 - 2m_0 \}$$

and we arrive at a particular kind of simultaneous sign test.

TABLE III

Values of $P_{\nu l}(m_0)$ for modified m test with l groups of ν tied x's

l \ m_0	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$\nu = 2$														
2	.750	1.0												
3	.281	.844	1.0											
4	.094	.469	.937	1.0										
5	.029	.205	.615	.967	1.0									
6	.009	.079	.316	.721	.984	1.0								
7	.003	.028	.141	.406	.797	.992	1.0							
8	.001	.010	.057	.300	.495	.853	.996	1.0						
9	0	.003	.022	.093	.321	.573	.893	.998	1.0					
10		0	.001	.008	.039	.134	.400	.639	.922	1.0				
$\nu = 3$														
2	.438	1.0												
3	.082	.355	.891	1.0										
4	.014	.113	.406	.824	1.0									
5	.002	.024	.123	.374	.736	.985	1.0							
6	0	.005	.030	.125	.340	.671	.948	1.0						
7		.001	.007	.033	.117	.305	.598	.885	.998	1.0				
8		0	.001	.008	.034	.109	.273	.534	.819	.984	1.0			
9			0	.003	.009	.033	.102	.240	.469	.745	.950	1.0	1.0	
10				0	.022	.009	.031	.082	.214	.414	.675	.901	.995	1.0

In general it is not possible to derive a convenient formula for (9.4), but particular cases can be evaluated by reformulating (9.4) in terms of a random walk with rather unusual boundary conditions and proceeding step by step. Write $w_j = s_1 + s_2 + \cdots + s_j$ and let d be the total number of negative signs in the sample signature. Then if $w_0 = 0$,

$$d_j = d - \tfrac{1}{2}n_j + \tfrac{1}{2}(w_{j-1} + w_j). \tag{9.5}$$

Also $w_l = n - 2d$, and for $m > m_0$, d can range from $m_0 + 1$ to $n - m_0 - 1$. Hence (9.4) becomes

$$\Pr(m > m_0) = \sum_{d=m_0+1}^{n-m_0-1} \Pr\{m_0 + \tfrac{1}{2}n_j < d + \tfrac{1}{2}(w_{j-1} + w_j) < n - m_0 - \tfrac{1}{2}n_j; \quad j = 1, 2, \cdots, l; \quad d + w_l = n - d\}. \tag{9.6}$$

We therefore consider the following random walk. Start at the point d and proceed in steps of s_j which can take values $2r_j - n_j$ with probabilities $2^{-n_j}\binom{n_j}{r_j}$. Absorption on boundary points at $m_0 + \frac{1}{2}n_j$, $n - m_0 - \frac{1}{2}n_j$ occurs when the *midpoint* of the step falls on or beyond these points. Thus it is possible for the path to overshoot the boundaries to some extent but not to stay outside for more than one step. The probability of arriving at the point $n - d$ after l steps is computed by enumerating the appropriate paths using a typical "binomial triangle" technique. Summation over d then gives $\Pr(m > m_0) = 1 - P_n(m_0)$. The distributions for equal groups of 2 and 3 given in Table III were computed in this way.

The case $m_0 = 0$ can be handled directly by observing that the open regions of type (i) account for $2l$ signatures while those of type (ii) can have $2\sum_1^l(2^{n_j} - 2)$ possible signatures; the largest closed polygon is therefore a confidence region with confidence coefficient

$$P_n(0) = 1 - \frac{1}{2^{n-1}}\left(\sum_{j=1}^{l} 2^{n_j} - l\right).$$

We finally remark that even when some values of x are not completely coincident, the corresponding lines in the (β, α) plane may be so nearly parallel that one would intuitively prefer to use the test which treats them as tied (though they would be kept distinct in calculating the ϵ's). But a rule for deciding between the two tests in such cases would have to depend on a comprehensive comparison of their power functions.

10. Acknowledgments. I am much indebted to W. H. Kruskal for many useful comments, and to W. Goldfarb and G. Chow for computational assistance.

REFERENCES

[1] G. W. Brown and A. M. Mood, "On median tests for linear hypotheses," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1950, p. 159–166.

[2] H. E. Daniels, "The theory of position finding," *J. Roy. Stat. Soc.* (B), Vol. 13, No. 2 (1951), pp. 186–207.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, 1950.

[4] A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1950.

[5] H. Theil, "A rank-invariant method of linear and polynomial regression analysis," *Indagationes Math.*, Vol. 12 (1950), pp. 85–91, 173–177, 467–482.