SOME QUICK SIGN TESTS FOR TREND IN LOCATION AND DISPERSION

By D. R. COX
Statistical Laboratory, University of Cambridge

AND A. STUART
Division of Research Techniques, London School of Economics

1. Introduction and summary

Many distribution-free tests have been devised to test the hypothesis of randomness of a series of N observations, i.e. the hypothesis that N independent random variables have the same continuous distribution function. Of these, the rank correlation tests are the most efficient tests against normal trend alternatives, but others are of some use in situations where speed and simplicity of computation are important.

In this paper, we discuss a class of simple sign tests, considered first as tests against trend in location. Optimum tests are found from the standpoint of asymptotic relative efficiency (a measure of local power in large samples), and it appears that the best of these tests may be preferred to the other simple tests considered in the literature, although they are, of course, less efficient than the rank correlation tests.

Similar tests are available for trend in dispersion, and the efficiency of these, in the normal situation. is investigated and compared with the test based on the maximum-likelihood estimator. Finally, we add a few remarks on sequential sign tests.

Readers not interested in the theory should look at §§ 10, 11 and 14, where there are brief statements of the tests and numerical examples.

2. The sign tests for trend in location

We consider a series of N independent observations from a standardized normal regression model with an upward trend, i.e.

$$H_1: \quad y_i = \alpha + \Delta i + \epsilon_i \quad (i = 1, 2, \ldots, N),$$

where $\Delta \geqslant 0$ and the ϵ_i are independent standardized normal variates. We wish to test the null hypothesis

$$H_0: \quad \Delta = 0,$$

using a distribution-free test statistic, so that our test will remain valid whatever the continuous distribution of the ϵ terms in the model, although naturally its efficiency will vary with the form of the distribution.

The most efficient known distribution-free tests of H_0 are those based on the rank correlation coefficients (Stuart, 1954), but our object here is specifically to find tests which are quick and simple to compute. We define for $i < j$ the score

$$h_{ij} = \begin{cases} +1 & \text{if} \quad y_i > y_j, \\ 0 & \text{if} \quad y_i < y_j. \end{cases}$$

h_{ij} is thus based on a comparison of the ith and jth in the series of observations. The distribution of the observations will be assumed continuous so that the possibility of ties can be ignored (see, however, § 16).

We confine ourselves throughout to comparisons of *independent* pairs of observations, i.e. no observation is compared with more than one other observation. (This is in contrast to the procedure used in calculating the rank correlation coefficients, where every observation is compared with every other observation in the series.) Since there are N observations, there can be no more than $\frac{1}{2}N$ such independent comparisons. We now assume N to be even and always take $i<j$. Our problem is to find the set of comparisons and the appropriate weights w_{ij} which will make the statistic

$$S = \Sigma w_{ij} h_{ij} \tag{1}$$

as efficient a test of H_0 as possible. The summation in (1) contains $\frac{1}{2}N$ terms, no suffix being repeated. All tests of the form (1) are distribution-free, since on the null hypothesis any h_{ij} is a $0-1$ variate with probabilities $(\frac{1}{2}, \frac{1}{2})$, whatever the distribution of the ϵ_i.

3. Asymptotic relative efficiency

We shall use as our criterion of efficiency the asymptotic relative efficiency (A.R.E.) of a test. If there are two consistent tests, s and t, of a hypothesis $H_0: \Delta = 0$, the A.R.E. is the reciprocal of the ratio of sample sizes required to attain the same power against the same alternative hypothesis H_1, taking the limit as the sample size N tends to infinity and as H_1 tends to H_0. (This second limiting process is necessary to keep the power of consistent tests bounded away from 1.) Pitman (1948) and Noether (1955) have shown that, if s and t both have normal limiting distributions on H_0 and H_1, the A.R.E. of s compared to t is given by

$$\text{A.R.E. } (s,t) = \lim_{N\to\infty} \left(\frac{R^2(s)}{R^2(t)}\right)^{1/r}, \tag{2}$$

where

$$R^2(X) = \left\{\left[\frac{\partial}{\partial\Delta} E(X \mid \Delta)\right]_{\Delta=0}\right\}^2 \Big/ D^2(X \mid \Delta = 0), \tag{3}$$

provided that r satisfies the equations

$$\lim_{N\to\infty} R^2(s)\, N^{-r} = R_1, \quad \lim_{N\to\infty} R^2(t)\, N^{-r} = R_2. \tag{4}$$

Here E and D^2 denote mean and variance as usual, while R_1 and R_2 are constants independent of N. The interpretation of the A.R.E. is discussed critically in § 9 below.

4. The best sign test

Since h_{ij} is a 0–1 variate,

$$E(h_{ij}) = \text{prob}\,(y_i > y_j),$$

and as $(y_j - y_i)$ is a normal variate with mean $(j-i)\,\Delta$ and variance 2 this is

$$E(h_{ij}) = G\left\{-\frac{(j-i)\,\Delta}{\sqrt{2}}\right\}, \tag{5}$$

where

$$G(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{(2\pi)}}\, e^{-\frac{1}{2}t^2} dt.$$

Now

$$\left[\frac{\partial}{\partial x} G(x)\right]_{x=0} = \frac{1}{\sqrt{(2\pi)}}, \tag{6}$$

so that, by (5) and (6),
$$E'(h_{ij}) = \left[\frac{\partial}{\partial\Delta} E(h_{ij})\right]_{\Delta=0} = \frac{(i-j)}{2\sqrt{\pi}}. \tag{7}$$

We now write $(j-i) = r_{ij}$. Using (7) in (1), we obtain
$$E'(S) = \Sigma w_{ij} E'(h_{ij}) = -\frac{1}{2\sqrt{\pi}} \Sigma w_{ij} r_{ij}. \tag{8}$$

We also have
$$D^2(S \mid \Delta = 0) = \Sigma w_{ij}^2 V(h_{ij} \mid H_0) = \tfrac{1}{4} \Sigma w_{ij}^2. \tag{9}$$

Equations (3), (8) and (9) give
$$R^2(S) = \frac{1}{\pi} \frac{(\Sigma w_{ij} r_{ij})^2}{\Sigma w_{ij}^2}, \tag{10}$$

and we now wish to maximize (10) to obtain the highest possible A.R.E. We do this in two stages. First we maximize (10) with respect to the w_{ij}, regarding the r_{ij} as fixed, and then we choose the supremum of these maxima for variations in the r_{ij}.

To maximize (10) for fixed r_{ij} and variation in the w_{ij}, we must maximize $\Sigma w_{ij} r_{ij}$ subject to Σw_{ij}^2 being held constant, i.e. we must unconditionally maximize
$$F = \Sigma w_{ij} r_{ij} - \lambda \Sigma w_{ij}^2.$$

It is clear from the conditions of the problem that each w_{ij} will be a function of the corresponding r_{ij}, so that on differentiating F for a stationary value we get
$$r_{ij} + w_{ij} \frac{\partial r_{ij}}{\partial w_{ij}} - 2\lambda w_{ij} = 0,$$

i.e.
$$\frac{r_{ij}}{w_{ij}} + \frac{\partial r_{ij}}{\partial w_{ij}} = 2\lambda.$$

This is satisfied by
$$w_{ij} = \lambda r_{ij}, \tag{11}$$

so that the required set of weights are proportional to the distances apart of the observations compared. The stationary value is a maximum. Substituting (11) into (10), we have
$$R^2(S) = \frac{1}{\pi} \Sigma r_{ij}^2. \tag{12}$$

This is the maximum value of $R^2(S)$ for a fixed set of r_{ij}. The r_{ij} are a set of $\frac{1}{2}N$ differences between pairs of integers chosen from the integers $1, 2, \ldots, N$. It is easily seen that Σr_{ij}^2 is largest when the pairs are $(1, N)$, $(2, N-1)$, $(3, N-2)$ and so on. In general
$$r_{ij} = (N-k+1) - k = N - 2k + 1 \quad (k = 1, 2, \ldots, \tfrac{1}{2}N) \tag{13}$$

so that
$$\Sigma r_{ij}^2 = \sum_{k=1}^{\frac{1}{2}N} (N-2k+1)^2 = \tfrac{1}{6} N(N^2-1), \tag{14}$$

and the supremum value of (12) is therefore
$$R^2(S_1) = \frac{N(N^2-1)}{6\pi} \sim \frac{N^3}{6\pi}. \tag{15}$$

We have denoted by S_1 the optimum S statistic
$$S_1 = \sum_{k=1}^{\frac{1}{2}N} (N-2k+1)\, h_{k, N-k+1},$$

for which

$$\left.\begin{aligned} E(S_1 \mid \Delta = 0) &= \tfrac{1}{2}\Sigma(N-2k+1) = \tfrac{1}{8}N^2, \\ D^2(S_1 \mid \Delta = 0) &= \tfrac{1}{4}\Sigma(N-2k+1)^2 = \tfrac{1}{24}N(N^2-1). \end{aligned}\right\} \tag{16}$$

The test based on S_1 is essentially a simplified version of Spearman's rank correlation test, which is in effect defined by

$$V = \sum_{i<j} (j-i)\, h_{ij}, \tag{17}$$

where the summation extends over all possible $\frac{1}{2}N(N-1)$ pairs of observations. Stuart (1954) has shown that

$$R^2(V) \sim \frac{N^3}{4\pi}, \tag{18}$$

so that using (15) and (18) in (2) and (4) with $r = 3$, we obtain for the A.R.E. of S_1 compared to V

$$\text{A.R.E.}\,(S_1, V) = (\tfrac{2}{3})^{\frac{1}{3}} = 0{\cdot}87. \tag{19}$$

The loss of A.R.E. involved in reducing the number of comparisons from $\frac{1}{2}N(N-1)$ to $\frac{1}{2}N$ is as little as 13%.

These values of the A.R.E. depend on the assumption of normality, but the calculation of the form of the optimum statistic, S_1, and also of the statistic, S_3, of § 5, does not. For (7) remains true, with a changed numerical factor, for general continuous distributions.

5. An unweighted sign test

The relatively high efficiency of S_1 compared to V leads us to construct, by analogy, a simplified version of Kendall's rank correlation test, which may be defined by

$$Q = \sum_{i<j} h_{ij}, \tag{20}$$

and gives equal weight to all $\frac{1}{2}N(N-1)$ comparisons. Q has the same A.R.E. as V (Stuart, 1954). The analogous sign test, based on $\frac{1}{2}N$ equally weighted independent comparisons, is

$$S_2 = \Sigma h_{ij}, \tag{21}$$

and using (10), we obtain, with all $w_{ij} = 1$,

$$R^2(S_2) = \frac{2}{N\pi}(\Sigma r_{ij})^2. \tag{22}$$

We now require to choose $\frac{1}{2}N$ pairs from the first N integers so that (22) or, equivalently, $\Sigma(j-i) = \Sigma r_{ij}$, takes its maximum value. This occurs whenever every i is chosen from the first $\frac{1}{2}N$ integers and every j from the last $\frac{1}{2}N$ integers. In particular, it occurs when every $r_{ij} = \frac{1}{2}N$ exactly, so that

$$S_2 = \sum_{k=1}^{\frac{1}{2}N} h_{k,\,\frac{1}{2}N+k}, \tag{23}$$

and (22) becomes

$$R^2(S_2) = \frac{N^3}{8\pi}. \tag{24}$$

Using (24) and (18), we obtain

$$\text{A.R.E.}\,(S_2, V) = (\tfrac{1}{2})^{\frac{1}{3}} = 0{\cdot}79, \tag{25}$$

while from (15)

$$\text{A.R.E.}\,(S_2, S_1) = (\tfrac{1}{2})^{\frac{1}{8}} = 0{\cdot}91. \tag{26}$$

Thus the simplified version of Kendall's rank correlation test is 21 % less efficient than Q or V, and 9 % less efficient than the simplified Spearman coefficient S_1. The use of S_2 is equivalent to a test considered by Theil (1950).

6. The best unweighted sign test

However, we can improve on the efficiency of S_2, and in fact get very nearly as high an efficiency as that of S_1, by 'throwing away' some of the $\frac{1}{2}N$ comparisons and retaining equal weights for the others. This was suggested by one of the present authors in the discussion of Foster & Stuart (1954); it leads to an increase in efficiency because, by comparing observations further apart, individual comparisons are made more sensitive.

In (1), let every w_{ij} be either 0 or 1, and let m $(\leqslant \frac{1}{2}N)$ be the number of non-zero w_{ij}. For our new statistic S_3 we have, as in (8),

$$E'(S_3) = -\frac{1}{2\sqrt{\pi}}\Sigma w_{ij} r_{ij} \quad (w_{ij} = 0 \text{ or } 1), \tag{27}$$

and from (9)

$$D^2(S_3 \mid \Delta = 0) = \tfrac{1}{4}m, \tag{28}$$

so that (3), (27) and (28) give

$$R^2(S_3) = \frac{1}{m\pi}(\Sigma w_{ij} r_{ij})^2 \quad (w_{ij} = 0 \text{ or } 1). \tag{29}$$

To maximize this by choice of m and r_{ij}, we again work in two stages. For fixed m, (29) will take its largest value when the comparisons given zero weights are based on the middle $(N-2m)$ observations, while every i is chosen from the first m observations and every j is chosen from the last m observations. In particular, this will be so when every $r_{ij} = (N-m)$ exactly, so that

$$S_3 = \sum_{k=1}^{m} h_{k,N-m+k} \tag{30}$$

and (29) becomes

$$R^2(S_3) = \frac{m(N-m)^2}{\pi}. \tag{31}$$

(31) is the largest possible value of $R^2(S_3)$ for fixed m. (S_2 is the special case of S_3 when $m = \frac{1}{2}N$.) We now choose m to maximize (31). Differentiating, we get

$$m = \tfrac{1}{3}N \tag{32}$$

for a maximum, so that finally

$$S_3 = \sum_{k=1}^{\frac{1}{3}N} h_{k,\frac{2}{3}N+k},$$

for which

$$\left.\begin{aligned} E(S_3) &= \tfrac{1}{6}N, \\ V(S_3) &= \tfrac{1}{12}N, \end{aligned}\right\} \tag{33}$$

and from (31),

$$R^2(S_3) = \frac{4N^3}{27\pi}. \tag{34}$$

From (34) and (18), we have

$$\text{A.R.E.}\,(S_3, V) = (\tfrac{16}{27})^{\frac{1}{3}} = 0{\cdot}84, \tag{35}$$

while from (15)

$$\text{A.R.E.}\,(S_3, S_1) = (\tfrac{8}{9})^{\frac{1}{3}} = 0{\cdot}96. \tag{36}$$

Compared with either V or S_1, S_3 has about 5 % higher efficiency than S_2, and in fact its efficiency is 96 % of that of S_1, so that for practical purposes it may be recommended instead of S_1 because it requires no weighting of the comparisons.

7. Comparison of the sign tests

In Table 1, the A.R.E. of the sign tests are tabulated, compared to each other, to the rank correlation tests, and to the best (parametric) test against normal regression, based on the sample regression coefficient b, which has a value of (3) given by

$$R^2(b) \sim \frac{N^3}{12}, \tag{37}$$

as follows immediately from the fact that b is an unbiased estimator of Δ with variance $12/\{N(N^2-1)\}$.

Table 1. *Asymptotic relative efficiencies of sign tests*

Test statistic	Asymptotic relative efficiency		
	Compared to S_1	Compared to rank correlation tests	Compared to best parametric test
$S_1 = \sum_{k=1}^{\frac{1}{3}N} (N-2k+1)\, h_{k,\, N-k+1}$	1·00	0·87	0·86
$S_2 = \sum_{k=1}^{\frac{1}{2}N} h_{k,\, \frac{1}{2}N+k}$	0·91	0·79	0·78
$S_3 = \sum_{k=1}^{\frac{1}{3}N} h_{k,\, \frac{2}{3}N+k}$	0·96	0·84	0·83

From (2), (18) and (37), it follows that the A.R.E. of either rank correlation coefficient compared to b is

$$\text{A.R.E.}\,(V, b) = \left(\frac{3}{\pi}\right)^{\frac{1}{3}} = 0{\cdot}98, \tag{38}$$

and not $3/\pi = 0{\cdot}95$ as given by Stuart (1954).

8. Comparison with A.R.E. of other tests

Apart from the two rank correlation tests already discussed, Stuart (1954) investigated the A.R.E. of three other distribution-free tests for trend in location. Two of these, the rank serial correlation test and the turning point test, were found to have zero values of R as defined by (3); the third, the difference-sign test, was found to have a value of r equal to 1 in (4), as against $r = 3$ for all the tests considered in this paper. It followed that the three tests mentioned all have A.R.E. zero compared to the rank correlation tests (and hence to all the tests discussed here). Noether (1955) gives general results which rigorize these conclusions.

A well-known and simple test which has not, as far as we know, previously been discussed from the point of view of A.R.E. is the median test, due to Brown & Mood (1951). The N (even) observations are divided into two sets of $\frac{1}{2}N$ consecutive observations. The test statistic is simply the number of observations in the first set which exceed the sample median y_m, and it is therefore defined by

$$B = \sum_{i=1}^{\frac{1}{2}N} b_{im}, \tag{39}$$

where

$$b_{im} = \begin{cases} 1 & \text{if} \quad y_i > y_m, \\ 0 & \text{if} \quad y_i < y_m. \end{cases}$$

The A.R.E. of B is easily obtained. We know that y_i is a normal variate with mean $(\alpha + i\Delta)$ and unit variance. It follows that the sample median y_m is asymptotically a normal variate with mean $(\alpha + \frac{1}{2}(N+1)\Delta)$ and variance of order N^{-1}. Since y_i and y_m are asymptotically independent, $(y_i - y_m)$ is asymptotically normal with mean $\Delta[i - \frac{1}{2}(N+1)]$ and unit variance, so that for $i < \frac{1}{2}(N+1)$

$$E(b_{im}) = \text{prob}\,(y_i > y_m) \sim 1 - G\{\Delta[\tfrac{1}{2}(N+1) - i]\}. \tag{40}$$

Using (6) in (40), we obtain

$$E'(b_{im}) \sim -\frac{1}{\sqrt{(2\pi)}}[\tfrac{1}{2}(N+1) - i], \tag{41}$$

so that from (39) and (41),

$$E'(B) = \sum_{i=1}^{\frac{1}{2}N} E'(b_{im}) \sim -\frac{1}{\sqrt{(2\pi)}} \sum_{i=1}^{\frac{1}{2}N} [\tfrac{1}{2}(N+1) - i] \sim -\frac{N^2}{8\sqrt{(2\pi)}}. \tag{42}$$

Also (Brown & Mood, 1951)

$$D^2(B \mid \Delta = 0) \sim \frac{N}{16}. \tag{43}$$

(42) and (43) give, in (3),

$$R^2(B) \sim \frac{N^3}{8\pi}. \tag{44}$$

Comparison of (44) with (24) shows that B has precisely the same A.R.E. as S_2, and is therefore slightly less efficient than S_1 and S_3. If the observations are available in serial order, S_3 is simpler to compute than B, which involves ranking all the observations to find the median, and then making $\frac{1}{2}N$ comparisons, as against $\frac{1}{3}N$ for S_3. There is therefore no reason to prefer B to S_3 in this case. If, however, the data were available graphically, B would be considerably easier to compute, and this would outweigh the slight loss of efficiency compared to S_3.

9. Comparison of the powers of tests

So far we have compared tests by the A.R.E. in the usual way. Before considering the power of the test S_3 in small samples it is convenient to examine the meaning of the A.R.E. more carefully. If the A.R.E. of a quick test relative to an efficient test is A, then asymptotically A^{-1} as many observations have to be made for the quick test to give the same local power as the efficient test. This is directly relevant if in designing an experiment a choice has to be made between, on the one hand, using an efficient method of analysis and on the other taking more observations and using a quick method of analysis. But it is not directly relevant to the choice of a method of analysis for a given body of data, because it depends in part on r, defined by (4), measuring the rate at which power increases with increasing N. For a given problem r is fixed and so the A.R.E. can be reinterpreted in terms of the power attained at a fixed sample size, but it seems preferable to compare tests directly in terms of power.

Consider a test based on a statistic S normally distributed with mean $E(S \mid \Delta)$ and standard deviation $D(S \mid \Delta)$, where the null hypothesis is $\Delta = 0$. The null hypothesis is rejected at the significance level α if

$$S > E(S \mid 0) + \lambda_\alpha D(S \mid 0), \tag{45}$$

where

$$G(-\lambda_\alpha) = \alpha. \tag{46}$$

The power of the test is $G[p(\Delta)]$, where

$$p(\Delta) = \frac{E(S \mid \Delta) - E(S \mid 0) - \lambda_\alpha D(S \mid 0)}{D(S \mid \Delta)}. \tag{47}$$

Now

$$p'(0) = \left(\frac{\partial p(\Delta)}{\partial \Delta}\right)_{\Delta=0} = \frac{E'(S \mid 0) + \lambda_\alpha D'(S \mid 0)}{D(S \mid 0)}. \tag{48}$$

In all the applications in this paper $D'(S \mid 0) = 0$, so that

$$p'(0) = \frac{E'(S \mid 0)}{D(S \mid 0)} = R(S). \tag{49}$$

Near $\Delta = 0$,

$$p(\Delta) = \Delta R(S) - \lambda_\alpha + O(\Delta^2), \tag{50}$$

and in applications the first two terms give, asymptotically in N, the whole of the power curve. Moreover, $R(S) \sim RN^{-\frac{1}{2}r}$ as $N \to \infty$ and comparable tests of a given hypothesis will have the same r; hence we usually need to consider just R. We call $p(\Delta)$ the *power deviate* and $p'(0)$ the *power derivative*. Asymptotically the graph of $p(\Delta)$ against Δ is linear, and tests at different significance levels are given by parallel straight lines. Or to put the same fact another way, the power curves are asymptotically linear when plotted on arithmetical probability paper.

Now consider the small sample theory with S possibly not normally distributed. Then if the power curves are plotted on probability paper they can be expected to form an approximately parallel set of curves approaching a set of parallel lines as the sample size increases and the distribution of S tends to normality. This is of course only a method of presenting the results of power calculations, but we shall find it very convenient both in assessing the small-sample behaviour and in comparing different tests.

Consider now two tests for which the asymptotic values of $R(S)$ are R_1, R_2. Then asymptotically in N the power curves for a given α are two lines on probability paper, the ratio of their slopes being R_1/R_2 independent of α; both lines intersect the probability axis at α.

A first consequence is that there is no simple general relation between the difference in the power of the two tests and the ratio R_1/R_2. If $R_1 \neq R_2$ we can, by taking α sufficiently small, make the difference in power between the two tests arbitrarily near unity. In practice we are probably only interested in $0{\cdot}20 \geqslant \alpha \geqslant 0{\cdot}001$, but the general conclusion remains that the difference in power between a quick test and an efficient test will be greatest for small α. Table 2 expresses this quantitatively; it shows for given R_1/R_2 the powers of the two tests at the point at which the difference in powers is greatest. The values in Table 2 are independent of N, but the values of Δ at which these powers are attained do depend on N. This is the restriction on the alternative hypothesis referred to in § 3. Thus if $R_1/R_2 = 0{\cdot}7$ and $\alpha = 0{\cdot}05$, the difference in power is greatest for the value of Δ at which the power of the efficient test is 77 % and of the quick test 51 %.

Now consider the power of S_3 in small samples. Two methods can be used. The first is to take the expansion (50) to higher powers of Δ and to introduce a correction for the non-normality of S based on an Edgeworth expansion. This may be shown to give good results even for very small N, and is a general method which could be used where direct numerical calculation is difficult. However, for S_3 it is much easier to calculate the power directly from the National Bureau of Standards tables of the binomial distribution (1950).

Table 2. *Asymptotic theory. Powers (per cent) of quick and efficient tests at points at which difference in power is greatest**

R_1/R_2 \ α	0·10	0·05	0·01	0·001
0·9	67, 73	63, 71	49, 60	54, 67
0·8	61, 74	56, 72	49, 71	43, 72
0·7	59, 80	51, 77	42, 77	39, 83
0·6	54, 84	47, 84	39, 86	29, 87
0·5	48, 88	41, 89	30, 90	20, 93
0·3	35, 96	27, 96	14, 97	7, 99

The power was computed in this way for $N = 15\,(15)\,135$, the significance level being the largest value $\leqslant 0{\cdot}05$. Under the null hypothesis the test statistic is distributed as $(\frac{1}{2}+\frac{1}{2})^{\frac{1}{3}N}$ and under the alternative hypothesis as $(p+q)^{\frac{1}{3}N}$, where

$$p = G\left(-\frac{\sqrt{2}\,N\Delta}{3}\right). \tag{51}$$

The power corresponding to given p, $\frac{1}{3}N$ can be read off directly from the tables and (51) solved for Δ. The results are given in Table 3. For comparative purposes the exact power of the parametric test based on the regression coefficient, b, has been computed for the same values of N and Δ. When the standard deviation about the regression line is known, the power is exactly $G[p(\Delta)]$, where

$$p(\Delta) = \{\tfrac{1}{12}N(N^2-1)\}^{\frac{1}{2}}\,\Delta - \lambda_\alpha. \tag{52}$$

To avoid rewriting the values of Δ in Table 4 the rows of both tables have been lettered, and each entry in Table 4 relates to the value of Δ shown above the corresponding entry in Table 3.

Asymptotically, the ratio of the R values of the two tests is, by (34) and (37), $4/(3\sqrt{\pi}) = 0{\cdot}75$; the interpretation of this in terms of power can be obtained from Table 2. The full curve in Fig. 1 and the full curves in Fig. 2 for $k = 0$ show the power curves for $N = 15$, 30, and the dotted lines are the corresponding asymptotic power curves. The small-sample power is lower than the value given by the asymptotic theory, the difference being quite appreciable in the region of 80–90 % power. The power curves of the most efficient test are exactly linear and differ from their asymptotic form only because of the very small difference between $\{N(N^2-1)\}^{\frac{1}{2}}$ and $N^{\frac{3}{2}}$. Hence the test S_3 is less efficient relative to b than the asymptotic

* These values were obtained graphically by drawing on probability paper lines whose ratio of slopes is R_1/R_2 and reading off the maximum difference in probability between them. The differences in power are determined accurately, but it is rather difficult to find the precise point of maximum difference. The values in Table 2 involve R_1, R_2 only through their ratio R_1/R_2.

Table 3. *Exact power of S_3 test against normal regression alternatives*

Values of the standardized regression coefficient, Δ, are given in parentheses, and the corresponding power appears immediately below.

Sample size (N) ...	15	30	45	60	75	90	105	120	135
Significance level α ...	0·031	0·011	0·018	0·021	0·022	0·049	0·045	0·040	0·036
a	(0·0035) 0·035	(0·0018) 0·013	(0·0012) 0·021	(0·0009) 0·026	(0·0007) 0·027	(0·0006) 0·062	(0·0005) 0·057	(0·0004) 0·053	(0·0004) 0·048
b	(0·0178) 0·050	(0·0089) 0·023	(0·0059) 0·042	(0·0044) 0·055	(0·0036) 0·064	(0·0030) 0·135	(0·0025) 0·134	(0·0022) 0·133	(0·0020) 0·130
c	(0·0358) 0·078	(0·0179) 0·046	(0·0119) 0·091	(0·0090) 0·126	(0·0072) 0·154	(0·0060) 0·291	(0·0051) 0·306	(0·0045) 0·317	(0·0040) 0·327
d	(0·0545) 0·116	(0·0272) 0·086	(0·0182) 0·173	(0·0136) 0·245	(0·0109) 0·306	(0·0091) 0·508	(0·0078) 0·542	(0·0068) 0·572	(0·0060) 0·598
e	(0·0742) 0·168	(0·0371) 0·149	(0·0247) 0·297	(0·0185) 0·416	(0·0148) 0·512	(0·0124) 0·730	(0·0106) 0·773	(0·0093) 0·807	(0·0082) 0·836
f	(0·0954) 0·237	(0·0477) 0·244	(0·0318) 0·461	(0·0238) 0·617	(0·0191) 0·727	(0·0159) 0·894	(0·0136) 0·924	(0·0119) 0·946	(0·0106) 0·961
g	(0·1190) 0·328	(0·0595) 0·376	(0·0397) 0·648	(0·0298) 0·804	(0·0238) 0·891	(0·0198) 0·974	(0·0170) 0·986	(0·0149) 0·992	(0·0132) 0·996
h	(0·1466) 0·444	(0·0733) 0·544	(0·0489) 0·823	(0·0366) 0·933	(0·0293) 0·975	(0·0244) 0·997	(0·0209) 0·999	(0·0183) 1·000	(0·0163) 1·000
i	(0·1812) 0·590	(0·0906) 0·736	(0·0604) 0·944	(0·0453) 0·989	(0·0362) 0·998	(0·0302) 1·000	(0·0259) 1·000	(0·0227) 1·000	(0·0201) 1·000
j	(0·2326) 0·774	(0·1163) 0·914	(0·0775) 0·995	(0·0582) 1·000	(0·0465) 1·000	(0·0388) 1·000	(0·0332) 1·000	(0·0291) 1·000	(0·0258) 1·000

Table 4. *Exact power of b test against normal regression alternatives*

Values of Δ are given in parentheses above the corresponding entry in Table 3.

Sample size (N) ...	15	30	45	60	75	90	105	120	135
Significance level α ...	0·031	0·011	0·018	0·021	0·022	0·049	0·045	0·040	0·036
a	0·036	0·013	0·023	0·027	0·030	0·066	0·062	0·057	0·054
b	0·059	0·030	0·056	0·074	0·088	0·179	0·182	0·183	0·183
c	0·103	0·074	0·143	0·201	0·249	0·429	0·457	0·481	0·502
d	0·171	0·157	0·300	0·416	0·509	0·722	0·764	0·799	0·827
e	0·267	0·294	0·519	0·673	0·775	0·919	0·945	0·962	0·973
f	0·395	0·485	0·746	0·877	0·941	0·988	0·994	0·997	0·999
g	0·551	0·699	0·911	0·975	0·993	0·999	1·000	1·000	1·000
h	0·772	0·880	0·984	0·998	1·000	1·000	1·000	1·000	1·000
i	0·879	0·977	0·999	1·000	1·000	1·000	1·000	1·000	1·000
j	0·979	0·999	1·000	1·000	1·000	1·000	1·000	1·000	1·000

theory suggests. The difference is greater for smaller α. The corresponding graphs to Figs. 1 and 2 for higher N show that for α in the range 0·01–0·05, the asymptotic theory applies well for $N \geqslant 60$.

The next thing is to investigate whether the form of S_3, involving the rejection of the middle third of the set of observations, can profitably be modified in small samples. Suppose that $(\frac{1}{3}N-2k)$ observations are rejected so that the number of comparisons is $(\frac{1}{3}N+k)$; the exact power function can be worked out from the binomial tables as before, but an immediate comparison is not possible because the significance levels for different values of α cannot be made equal, except by the artificial device of randomized tests. However, if the curves are plotted on probability paper they are almost parallel for different α and an

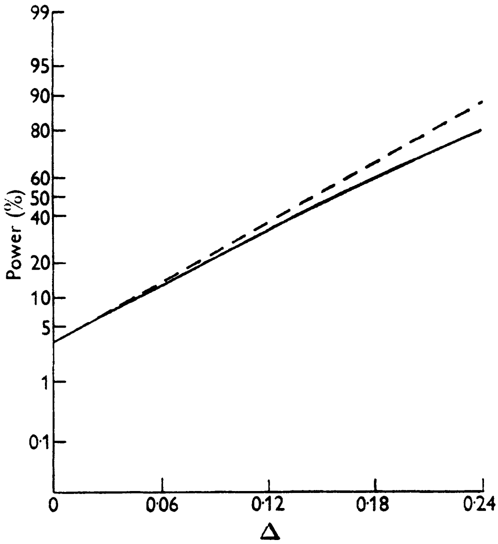

Fig. 1. Power of S_3 for $N=15$. —— Exact power. ---- Value from asymptotic theory.

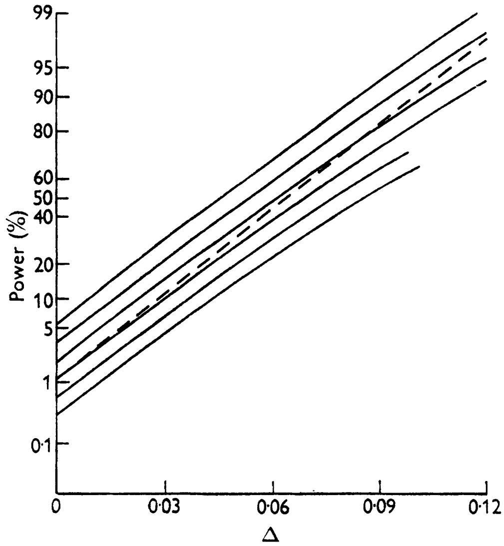

Fig. 2. Power of S_3 for $N=30$. —— Exact power. ---- Value from asymptotic theory. Full curves are in descending order $k=0, 1, 2, 0, 1, 2$, the second set having lower values of α than the first. Broken curve is $k=0$.

increase in power with change in k would be shown by decreasing curvature. As would be expected, a negative k leads to a loss of power. Fig. 2 shows for $N = 30$ the curves for $k = 0, 1, 2$. There is a tendency for the curvature to increase as α decreases, but there does not seem to be any systematic change with k. Therefore, although an increase in k increases the number of available significance levels, it does not appreciably increase power. Hence there seems to be little value in modifying the $\frac{1}{3}$ rule in small samples: similar calculations for $N = 15$ confirm this.

We have not made the corresponding investigations for S_1.

10. Examples of use of the sign tests against trend in location

To illustrate the S_1 and S_3 tests, we use the figures of annual rainfall at Oxford for the years 1858–1952, quoted by Foster & Stuart (1954, Table 9).

For S_1 we compare the kth observation with the $(N-k+1)$th, scoring 1 when the former is the larger and 0 when it is the smaller. The unit scores are then weighted by the distance apart of the observations compared, i.e. by $(N-2k+1)$. In this case $N = 95$ and is odd, so that we must ignore the middle observation and proceed with $N = 94$. The unit scores are those with weights as follows:

89, 83, 79, 77, 75, 71, 65, 59, 57, 55, 51, 49, 47, 45, 33, 31, 27, 15, 3.

The value of S_1 is the sum of these weights, 1011. From (16), with $N = 94$, we have

$$E(S_1) = 1104{\cdot}5, \quad D^2(S_1) = 34603{\cdot}75, \quad D(S_1) = 186{\cdot}0.$$

The observed value of S_1 thus represents a deviation from expectation of almost exactly one-half its standard error and is therefore in good agreement with the null hypothesis of zero trend.

If, alternatively, we were to use the simpler S_3 test we compare the kth observation with the $(\frac{2}{3}N+k)$th, scoring 1 or 0 as before, but no weighting is necessary. Since $N = 95$ and is not a multiple of three, we retain the extra observations (in accordance with our findings at the end of § 9 above) and compare each of the first thirty-two observations with the corresponding observation in the last thirty-two. For our sequence of scores we obtain

1 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 1 0 0 0 0,

the total score S_3 being 14. This clearly agrees well with the expected value of $\frac{1}{2} \times 32 = 16$. The standard error of S_3 is, from (33), $\sqrt{(\frac{1}{4} \times 32)} = 2{\cdot}83$, so that the deviation from expectation, corrected for continuity, is just over one-half of a standard error.

11. SIGN TESTS FOR TREND IN DISPERSION

We consider now tests for a trend not in location but in the dispersion about a fixed location. For example, in a regression problem we may want to test quickly whether the scatter about the regression curve increases as the independent variable increases.

Divide the series $x_1, \ldots, x_N$ into sets $x_1, \ldots, x_k; x_{k+1}, \ldots, x_{2k}; \ldots$, rejecting a few observations in the centre of the original series if N is not exactly divisible by k. The best choice of k is discussed below. For each set of k observations find the range, w, thus getting a series of ranges $w_1, \ldots, w_r$, where r is the integral part of N/k. The ranges are then tested for trend by one or other of the tests S_1 and S_3.

If the null hypothesis is that $x_1, \ldots, x_N$ are independently distributed with constant dispersion about a regression line, $w_1, \ldots, w_r$ are independent and identically distributed. If the regression is not linear the w's will be approximately identically distributed unless the trend within sets of observations varies appreciably.

In the next section, a valid test is obtained for any k, and the best value of k for detecting certain special forms of trend is found for large samples. The behaviour for small samples has not been investigated. The following provisional rules are suggested:

if	$N \geqslant 90$	take	$k = 5$,
if	$90 > N \geqslant 64$	take	$k = 4$,
if	$64 > N \geqslant 48$	take	$k = 3$,
if	$48 > N$	take	$k = 2$.

Except when N is very large it is probably advisable to use the weighted, rather than the unweighted, sign test.

12. Choice of k for dispersion tests

To investigate the theory of the test for trend in dispersion we take a special form for the null and alternative hypotheses. Suppose that $x_1, \ldots, x_N$ are independently normally distributed with constant mean and with standard deviations $\sigma(1), \ldots, \sigma(N)$, where $\sigma(n)$ varies at most slowly with n. Then the ranges $w_1, \ldots, w_r$ defined in § 11 have very nearly the distribution of ranges of k observations drawn from normal populations of standard deviations $\sigma_1 = \sigma(\frac{1}{2}k)$, $\sigma_2 = \sigma(\frac{3}{2}k)$, ... Patnaik (1950) has shown that a range of k observations can be represented to a close approximation as a multiple of a χ-variate with suitably chosen degrees of freedom, ν_k. Therefore $w_i^2\sigma_j^2/(w_j^2\sigma_i^2)$ is approximately an F variate with (ν_k, ν_k) degrees of freedom.

Hence
$$\text{prob}\,(w_i/w_j > 1) \simeq \int_0^{(\sigma_i/\sigma_j)^2} \frac{\Gamma(\nu_k)}{[\Gamma(\frac{1}{2}\nu_k)]^2} \frac{x^{\frac{1}{2}\nu_k - 1}}{(1+x)^{\nu_k}}\,dx$$
$$\simeq \frac{1}{2} + \left\{\left(\frac{\sigma_i}{\sigma_j}\right)^2 - 1\right\} A_k$$
where
$$A_k = \frac{\Gamma(\nu_k)}{[\Gamma(\frac{1}{2}\nu_k)]^2} (\tfrac{1}{2})^{\nu_k},$$
provided that $(\sigma_i/\sigma_j)^2 - 1$ is small.

If we assume that the trend in standard deviation is such that $\sigma(n) = \sigma_0 e^{\gamma n} \sim \sigma_0(1+\gamma n)$, where $\gamma N \ll 1$, so that γ is the fractional increase in standard deviation per observation, we have
$$(\sigma_i/\sigma_j)^2 - 1 \sim 2k\gamma(i-j)$$
and
$$\text{prob}\,(w_i/w_j > 1) \simeq \tfrac{1}{2} + 2k\gamma(i-j)\,A_k.$$

Consider first the application to the ranges of the unweighted sign test, S_3. From the $r \simeq N/k$ ranges we make approximately $\frac{1}{3}N/k$ independent comparisons in each of which $i - j \simeq \frac{2}{3}r$. Therefore if S is the total score, its mean and standard deviation are given by
$$E(S \mid \gamma) \simeq \frac{1}{3}\frac{N}{k}\, 4k\gamma\, \frac{2rA_k}{3} = \frac{8N^2 A_k \gamma}{9k},$$
$$D(S \mid \gamma) = \left(\frac{N}{3k}\right)^{\frac{1}{2}} + O(\gamma^2).$$
Therefore the power derivative, $p_3'(0)$, of the test is
$$p_3'(0) = \frac{E'(S \mid 0)}{D(S \mid 0)} \simeq \frac{8\sqrt{3}\,N^{\frac{3}{2}} A_k}{9\sqrt{k}}. \tag{53}$$

An exactly analogous calculation for the weighted sign test S_1 gives
$$p_1'(0) \simeq \frac{2\sqrt{6}\,N^{\frac{3}{2}} A_k}{3\sqrt{k}}. \tag{54}$$

Thus in both cases the asymptotically best value of k is the one that maximizes $A_k/\sqrt{k}$. From Patnaik's table of ν_k the values in Table 5 have been computed.

Now the number of ranges is N/k and the number of comparisons is one-half or one-third of this and is therefore small even when N is, by usual standards, quite large. Therefore it is advisable to use smaller values of k than the theoretical optimum in large samples. In the absence of an investigation of the small-sample properties of the test, the rule of § 11

Table 5. *Determination of efficiencies of different set sizes, k, for testing trend in dispersion*

k	$A_k/\sqrt{k}$	k	$A_k/\sqrt{k}$
2	0·112	6	0·167
3	0·141	7	0·169
4	0·158	8	0·170
5	0·164	9	0·169

is suggested. This is based on the considerations that there is little gain in taking $k > 5$ and that it is advisable, whenever possible, to have at least sixteen ranges.

If we substitute, in (53) and (54), the value $A_k/\sqrt{k} \simeq 0{\cdot}16$, we have

$$\left.\begin{aligned} p_3'(0) &\simeq 0{\cdot}246N^{\frac{3}{2}}, \\ p_1'(0) &\simeq 0{\cdot}261N^{\frac{3}{2}}. \end{aligned}\right\} \tag{55}$$

It remains to compare (55) with the corresponding quantity for the maximum-likelihood test of the corresponding parametric hypotheses.

13. A.R.E. of dispersion tests

For simplicity assume that $x_1, \ldots, x_N$ are normally and independently distributed with zero mean and that the standard deviation of x_n is $\sigma_0 e^{\gamma n}$, where γN is small. The log likelihood is

$$L = -\tfrac{1}{2}N\log(2\pi) - N\log\sigma_0 - \gamma\sum_1^N n - \frac{1}{2\sigma_n^2}\Sigma x_n^2 e^{-2\gamma n}.$$

If we differentiate and take expectations, retaining only the terms independent of γ, and letting N tend to infinity, we get

$$E\left(\frac{\partial^2 L}{\partial\sigma_0^2}\right) \sim -\frac{2N}{\sigma_0^2}, \quad E\left(\frac{\partial^2 L}{\partial\sigma_0\partial\gamma}\right) \sim -\frac{N^2}{\sigma_0}, \quad E\left(\frac{\partial^2 L}{\partial\gamma^2}\right) \sim -\tfrac{2}{3}N^3. \tag{56}$$

The large-sample variance of $\hat{\gamma}$, the maximum-likelihood estimate of γ, is given by inverting the Hessian matrix with elements (56). We get when γ is small and N is large

$$V(\hat{\gamma}) \sim 6/N^3. \tag{57}$$

Thus the power derivative of the test based on the maximum-likelihood estimate is

$$p_m'(0) = \frac{N^{\frac{3}{2}}}{\sqrt{6}} = 0{\cdot}408N^{\frac{3}{2}}. \tag{58}$$

(58) still applies if the mean in unknown or if a linear trend in mean has to be estimated.

From the formulae (55) and (58), and the fact that $p_x'(0) = R(x)$, it follows, on using (2) with $r = 3$, that the A.R.E.'s of the tests S_3, S_1 compared with the maximum-likelihood test are about 71 and 74 % respectively.

A test entirely analogous to the above tests can be found by calculating the variances within each set of k instead of the range. This is slightly more efficient in the parametric case but much of the simplicity of the test is lost and the increase in power may be shown to be trivial.

14. Examples of the use of sign tests against trend in dispersion

We again use for illustrative purposes the rainfall data quoted by Foster & Stuart (1954, Table 9). Using the provisional rule given in § 11 above, we take ranges of sets of five observations. Since $N = 95$, this gives us exactly nineteen sets, no rejection of observations being necessary. The nineteen ranges are:

9·64, 12·30, 12·01, 11·45, 5·43, 13·05, 9·86, 10·89, 6·95, 15·03,
11·34, 6·63, 12·19, 8·55, 4·80, 11·00, 7·76, 7·03, 10·98.

If we apply the test S_1 we drop the middle value and take the signs of 10·98–9·64, 7·03–12·30, ... down to 11·34–6·95, thus obtaining

score:	0	1	1	1	1	1	0	1	0
weight:	17	15	13	11	9	7	5	3	1

The total score is therefore 58, and from (16) with $N = 18$ the mean score is 40·5 and the variance is 242·25, so that the standard error is 15·6. The deviation from expectation is about 1·12 standard errors, and so the two-sided normal significance level is about 27 %. The exact significance level is, by enumeration, $73/256 \simeq 28\frac{1}{2}$ %.

If we use the test S_3 we reject the middle five of the nineteen ranges and take the signs of 12·19–9·64, etc. This gives scores

0 1 1 1 0 1 0

There is clearly good agreement with an equal probability for zeros and ones; significance would be tested in the binomial distribution $(\frac{1}{2}+\frac{1}{2})^7$. The test S_3 is not to be recommended in the present instance because with only seven comparisons the loss of sensitivity compared to the S_1 test would be considerable.

Thus although there is a slight indication that the dispersion decreases with time, both tests suggest that this could easily be a sampling fluctuation.

15. Sequential tests

Finally, we point out the possibility of constructing sequential tests for trend related to the tests considered above. While this paper was in preparation an abstract (Noether, 1954) appeared describing briefly a test rather similar to the one we had developed. Hence a full discussion will not be attempted here. However, some calculations in a special case suggest that the average sample size under the null and alternative hypotheses are, for the sequential sign tests, only a little greater than the corresponding parametric fixed sample size.

Sequential sign tests for trend are only likely to be of practical value under rather exceptional circumstances. For they require that observations are sufficiently easy to obtain for it to be worth while to use inefficient methods of analysis, and yet sufficiently difficult to obtain for the saving from the use of a sequential method to be important. A possible application is to the marking of a large number of examination scripts. If they are marked in alphabetical order it may be useful to test, as the marking proceeds, for a trend in the marks, which would indicate a changing standard of marking. A sequential method is appropriate and yet elaborate calculations would be out of place.

16. GENERAL COMMENTS

The calculation of the efficiency of the above tests and the determination of optimum weightings, etc., has been based on a particular type of alternative hypothesis. It is clear in a general way that the tests will remain effective for detecting monotone trends. Positive serial correlation among the observations would increase the chance of a significant answer even in the absence of a trend.

The occurrence of ties has been ignored in the above work. A small number of ties can be dealt with by counting one-half a comparison in each direction, i.e. if $y_i = y_j$ we calculate as if one-half a comparison has $y_i > y_j$ and one-half has $y_i < y_j$. If a substantial proportion of the comparisons are ties a special investigation is necessary or the comparisons should be randomized.

Estimates for the trend could be constructed from the test statistics S_1, S_3. It is very doubtful if such estimates would be of value; in any case, in much work with quick tests, if the trend is shown to be significant it can be estimated graphically.

REFERENCES

BROWN, G. W. & MOOD, A. M. (1951). On median tests for linear hypotheses. *Proceedings of the Second Berkeley Symposium*, pp. 159–66. University of California Press.

FOSTER, F. G. & STUART, A. (1954). Distribution-free tests in time-series based on the breaking of records. *J. R. Statist. Soc.* Series B (Methodological), **16**, 1–22.

NATIONAL BUREAU OF STANDARDS (1950). *Tables of the Binomial Probability Distribution*, Applied Mathematics Series, no. 6. Washington.

NOETHER, GOTTFRIED, E. (1954). Abstract of 'A sequential test of randomness against linear trend'. *Ann. Math. Statist.* **25**, 176.

NOETHER, GOTTFRIED E. (1955). 'The asymptotic relative efficiencies of tests of hypotheses'. *Ann. Math. Statist.* (to be published).

PATNAIK, P. B. (1950). The use of mean range as an estimator of variance in statistical tests. *Biometrika*, **37**, 78–87.

PITMAN, E. J. G. (1948). Lecture notes on 'Non-parametric statistical inference', University of North Carolina Institute of Statistics (mimeographed).

STUART, ALAN (1954). The asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. *J. Amer. Statist. Ass.* **49**, 147–57.

THEIL, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indag. math.* **12**, 85–91, 173–7.