The Statistical Sign Test
Author(s): W. J. Dixon and A. M. Mood
Reviewed work(s):
Source: *Journal of the American Statistical Association*, Vol. 41, No. 236 (Dec., 1946), pp. 557-566
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2280577
Accessed: 19/02/2013 17:39

THE STATISTICAL SIGN TEST*

W. J. DIXON
University of Oregon
A. M. MOOD
Iowa State College

This paper presents and illustrates a simple statistical test for judging whether one of two materials or treatments is better than the other. The data to which the test is applied consist of paired observations on the two materials or treatments. The test is based on the signs of the differences between the pairs of observations.

It is immaterial whether all the pairs of observations are comparable or not. However, when all the pairs are comparable, there are more efficient tests (the t test, for example) which take account of the magnitudes as well the signs of the differences. Even in this case, the simplicity of the sign test makes it a useful tool for a quick preliminary appraisal of the data.

In this paper the results of previously published work on the sign test have been included, together with a table of significance levels and illustrative examples.

INTRODUCTION

IN EXPERIMENTAL investigations, it is often desired to compare two materials or treatments under various sets of conditions. Pairs of observations (one observation for each of the two materials or treatments) are obtained for each of the separate sets of conditions. For example, in comparing the yield of two hybrid lines of corn, A and B, one might have a few results from each of several experiments carried out under widely varying conditions. The experiments may have been performed on different soil types, with different fertilizers, and in different years with consequent variations in seasonal effects such as rainfall, temperature, amount of sunshine, and so forth. It is supposed that both lines appeared equally often in each block of each experiment so that the observed yields occur in pairs (one yield for each line) produced under quite similar conditions.

The above example illustrates the circumstances under which the sign test is most useful:

(a) There are pairs of observations on two things being compared.

---

* This paper is an adaptation of a memorandum submitted to the Applied Mathematics Panel by the Statistical Research Group, Princeton University. The Statistical Research Group operated under a contract with the Office of Scientific Research and Development, and was directed by the Applied Mathematics Panel of the National Defense Research Committee.

(b) Each of the two observations of a given pair arose under similar conditions.

(c) The different pairs were observed under different conditions.

This last condition generally makes the t test invalid. If this were not the case (that is, if all the pairs of observations were comparable), the t test would ordinarily be employed unless there were other reasons, for example, obvious non-normality, for not using it.

Even when the t test is the appropriate technique many statisticians like to use the sign test because of its extreme simplicity. One merely counts the number of positive and negative differences and refers to a table of significance values. Frequently the question of significance may be settled at once by the sign test without any need for calculations.

It should be pointed out that, strictly speaking, the methods of this paper are applicable only to the case in which no ties in paired comparisons occur. In practice, however, even when ties would not occur if measurements were sufficiently precise, ties do occur because measurements are often made only to the nearest unit or tenth of a unit for example. Such ties should be included among the observations with half of them being counted as positive and half negative.

Finally, it is assumed that the differences between paired observations are independent, that is, that the outcome of one pair of observations is in no way influenced by the outcome of any other pair.

PROCEDURE

Let A and B represent two materials or treatments to be compared. Let x and y represent measurements made on A and B. Let the number of pairs of observations be n. The n pairs of observations and their differences may be denoted by:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

and

$$x_1 - y_1, x_2 - y_2, \cdots, x_n - y_n.$$

The sign test is based on the signs of these differences. The letter r will be used to denote the number of times the less frequent sign occurs. If some of the differences are zero, half of them will be given a plus sign and half a minus sign.

As an example of the type of data for which the sign test is appropriate, we may consider the following yields of two hybrid lines of corn obtained from several different experiments. In this example $n=28$ and $r=7$.

If there is no difference in the yielding ability of the two lines, the positive and negative signs should be distributed by the binomial distribution with $p=\frac{1}{2}$. The null hypothesis here is that each difference has a probability distribution (which need not be the same for all differences) with median equal to zero. This null hypothesis will obtain, for instance, if each difference is symmetrically distributed about a mean of zero, although such symmetry is not necessary. The null hypothesis will be rejected when the numbers of positive and negative signs differ significantly from equality.

YIELDS OF TWO HYBRID LINES OF CORN

Experiment Number	Yield of		Sign of $x-y$	Experiment Number	Yield of		Sign of $x-y$
	A	*B*			*A*	*B*	
1	47.8	46.1	+	4	40.8	41.3	−
	48.6	50.1	−		39.8	40.8	−
	47.6	48.2	−		42.2	42.0	+
	43.0	48.6	−		41.4	42.5	−
	42.1	43.4	−				
	41.0	42.9	−	5	38.9	39.1	−
					39.0	39.4	−
2	28.9	38.6	−		37.5	37.3	+
	29.0	31.1	−				
	27.4	28.0	−	6	36.8	37.5	−
	28.1	27.5	+		35.9	37.3	−
	28.0	28.7	−		33.6	34.0	−
	28.3	28.8	−				
	26.4	26.3	+	7	39.2	40.1	−
	26.8	26.1	+		39.1	42.6	−
3	33.3	32.4	+				
	30.6	31.7	−				

Table 1 gives the critical values of r for the 1, 5, 10, and 25 per cent levels of significance. A discussion of how these values are computed may be found in the appendix. A value of r less than or equal to that in the table is significant at the given per cent level.

Thus in the example above where $n=28$ and $r=7$, there is significance at the 5% level, as shown by Table 1. That is, the chances are only 1 in 20 of obtaining a value of r equal to or less than 8 when there is no real difference in the yields of the two lines of corn. It is concluded, therefore, at the *5% level of significance*, that the two lines have different yields.

In general, there are no values of r which correspond exactly to the levels of significance 1, 5, 10, 25 per cent. The values given are such that they result in a level of significance as close as possible to, but not exceeding 1, 5, 10, 25 per cent. Thus, the test is a little more strict,

TABLE 1

TABLE OF CRITICAL VALUES OF r FOR THE SIGN TEST

	Per Cent Level of Significance					Per Cent Level of Significance			
n	1	5	10	25	n	1	5	10	25
1	—	—	—	—	51	15	18	19	20
2	—	—	—	—	52	16	18	19	21
3	—	—	—	0	53	16	18	20	21
4	—	—	—	0	54	17	19	20	22
5	—	—	0	0	55	17	19	20	22
6	—	0	0	1	56	17	20	21	23
7	—	0	0	1	57	18	20	21	23
8	0	0	1	1	58	18	21	22	24
9	0	1	1	2	59	19	21	22	24
10	0	1	1	2	60	19	21	23	25
11	0	1	2	3	61	20	22	23	25
12	1	2	2	3	62	20	22	24	25
13	1	2	3	3	63	20	23	24	26
14	1	2	3	4	64	21	23	24	26
15	2	3	3	4	65	21	24	25	27
16	2	3	4	5	66	22	24	25	27
17	2	4	4	5	67	22	25	26	28
18	3	4	5	6	68	22	25	26	28
19	3	4	5	6	69	23	25	27	29
20	3	5	5	6	70	23	26	27	29
21	4	5	6	7	71	24	26	28	30
22	4	5	6	7	72	24	27	28	30
23	4	6	7	8	73	25	27	28	31
24	5	6	7	8	74	25	28	29	31
25	5	7	7	9	75	25	28	29	32
26	6	7	8	9	76	26	28	30	32
27	6	7	8	10	77	26	29	30	32
28	6	8	9	10	78	27	29	31	33
29	7	8	9	10	79	27	30	31	33
30	7	9	10	11	80	28	30	32	34
31	7	9	10	11	81	28	31	32	34
32	8	9	10	12	82	28	31	33	35
33	8	10	11	12	83	29	32	33	35
34	9	10	11	13	84	29	32	33	36
35	9	11	12	13	85	30	32	34	36
36	9	11	12	14	86	30	33	34	37
37	10	12	13	14	87	31	33	35	37
38	10	12	13	14	88	31	34	35	38
39	11	12	13	15	89	31	34	36	38
40	11	13	14	15	90	32	35	36	39
41	11	13	14	16	91	32	35	37	39
42	12	14	15	16	92	33	36	37	39
43	12	14	15	17	93	33	36	38	40
44	13	15	16	17	94	34	37	38	40
45	13	15	16	18	95	34	37	38	41
46	13	15	16	18	96	34	37	39	41
47	14	16	17	19	97	35	38	39	42
48	14	16	17	19	98	35	38	40	42
49	15	17	18	19	99	36	39	40	43
50	15	17	18	20	100	36	39	41	43

For $n > 100$, approximate values of r may be found by taking the nearest integer less than $\frac{1}{2}n - k\sqrt{n}$, where $k = 1.3$, 1, .82, .58 for the 1, 5, 10, 25 per cent values respectively. A closer approximation to the values of r is obtained from $\frac{1}{2}(n-1) - k\sqrt{n+1}$ and the more exact values of k, 1.2879, .9800, .8224, .5752.

on the average, than the level of significance which is indicated. For small samples the test is considerably more strict in some cases. For example, the value of r for $n=12$ for the 10 per cent level of significance actually corresponds to a per cent level less than 5.

The critical values of r in Table 1 for the various levels of significance were computed for the cases where either the $+$'s or $-$'s occur a significantly small number of times. Sometimes the interest may be in only one of the signs. For example, in testing two treatments, A and B, A may be identical with B except for certain additions which can only have the effect of improving B. In this case one would be interested only in whether the deficiency of minus signs (for differences in the direction A minus B) were significant or not. In cases of this kind the per cent levels of significance in Table 1 would be divided by two. Thus, 8 minus signs in a sample of 28 would correspond to the 2.5% level of significance.

SIZE OF SAMPLE

Even though there is no real difference, a sample of four or even five with all signs alike will occur by chance more than 5% of the time. Four signs alike will occur by chance 12.5% of the time and five signs alike will occur by chance 6.25% of the time. Therefore, at the 5% level of significance, it is necessary to have at least six pairs of observations even if all signs are alike before any decision can be made. As in most statistical work, more reliable results are obtained from a larger number of observations. One would not ordinarily use the sign test for samples as small as 10 or 15, except for rough or preliminary work.

The question may be raised as to the minimum sample size necessary to detect a given difference in two materials. Suppose that in an indefinitely large number of observations 30% $+$'s and 70% $-$'s are to be expected and that we wish the sample to be large enough to detect this difference at the 1% level of significance. Although no sample, however large, will make it absolutely certain that a significant difference will be found, the sample size can be chosen to make the probability of finding a significant result as near to certainty as is desired. In Table 2, this probability has been chosen as 95%; the minimum values of n (sample size) and the corresponding critical values of r to insure a decision 95% of the time are given for various actual percentages p_0 and levels of significance α.

The sign test merely measures the significance of departures from a 50–50 distribution. If the signs are actually distributed 45–55, then the departure from 50–50 is not likely to be significant unless the sample is quite large. Table 2 shows that if the signs are actually distributed 45–55, then one must take samples of 1,297 pairs in order to get a significant departure from a 50–50 distribution at the 5% level of significance. The number 1,297 is selected to give the desired significance 95% of the time; that is, if a large number of samples of 1,297 each were drawn from a 45–55 distribution, then 95% of those samples could be expected to indicate a significant departure (at the 5% level) from a 50–50 distribution.

TABLE 2

MINIMUM VALUES OF n NECESSARY TO FIND SIGNIFICANT DIFFERENCES 95% OF THE TIME FOR VARIOUS GIVEN PROPORTIONS

p_s	n				r			
	$\alpha=1\%$	5%	10%	25%	$\alpha=1\%$	5%	10%	25%
.45 (.55)	*1,777*	*1,297*	*1,080*	*780*	*833*	*612*	*512*	*373*
.40 (.60)	*442*	327	*267*	*193*	*193*	145	*119*	*87*
.35 (.65)	*193*	143	*118*	86	*78*	59	*49*	37
.30 (.70)	*106*	79	67	47	*39*	30	26	19
.25 (.75)	66	49	42	32	22	17	15	12
.20 (.80)	44	35	28	21	13	11	9	7
.15 (.85)	32	23	18	14	8	6	5	4
.10 (.90)	24	17	13	11	5	4	3	3
.05 (.95)	15	12	11	6	2	2	2	1

The italicized values are approximate. The maximum error is about 5 for the value of n, and 2 for the value of r. The values of n and r for 5% were taken from MacStewart (reference 1) who gives a table of values of n and r for a range of confidence coefficients (the above table uses only 95%) and a single value $\alpha=5\%$.

Of course, in practice one would not do any testing if he knew in advance the expected distribution of signs (that it was 45–55, for example). The practical significance of Table 2 is of the following nature: In comparing two materials one is interested in determining whether they are of about equal or of different value. Before the investigation is begun, a decision must be made as to how different the materials must be in order to be classed as different. Expressed in another way, how large a difference may be tolerated in the statement that "the two materials are of about equal value?" This decision, together with Table 2, determines the sample size. If one is interested in detecting a difference so small that the signs may be distributed 45–55, he must be prepared to take a very large sample. If, however, one is interested only in detecting larger differences, (for example, differences represented by a 70–30 distribution of signs), a smaller sample will suffice.

In many investigations, the sample size can be left undetermined, and only as much data accumulated as is needed to arrive at a decision. In such cases, the sign test could be used in conjunction with methods of sequential analysis. These methods provide a desired amount of information with the minimum amount of sampling on the average. A complete exposition of the theory and practice of sequential analysis may be found in references 3 and 4.

MODIFICATIONS OF THE SIGN TEST

When the data are homogeneous (measurements are comparable between pairs of observations), the sign test can be used to answer questions of the following kind:

1. Is material A better than B by P per cent?
2. Is material A better than B by Q units?

The first question would be tested by increasing the measurement on B by P per cent and comparing the results with the measurements on A. Thus, let

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \text{etc.}$$

be pairs of measurements on A and B, and suppose one wished to test the hypothesis that the measurements, x, on A were 5% higher than the measurements, y, on B. The sign test would simply be applied to the signs of the differences

$$x_1 - 1.05y_1, x_2 - 1.05y_2, x_3 - 1.05y_3, \text{etc.}$$

In the case of the second question the sign test would be applied to the differences

$$x_1 - (y_1 + Q), x_2 - (y_2 + Q), x_3 - (y_3 + Q), \text{etc.}$$

In either case, if the resulting distribution of signs is not significantly different from 50–50, the data are not inconsistent with a positive answer to the question. Usually there will be a range of values of P (or Q) which will produce a non-significant distribution of signs. If one determines such a range, using the 5% level of significance for example, then that range will be a 95% confidence interval for P (or Q).

Even when the data are not homogeneous, it may be possible to frame questions of the above kind, or it may be possible to change the scales of measurement so that such questions would be meaningful.

MATHEMATICAL APPENDIX

A. *Assumptions.* Let observations on two materials or treatments A and B be denoted by x and y, respectively. It is assumed that for any pair of observations (x_i, y_i) there is a probability $p (0<p<1)$ that

$x_i > y_i$ $(i=1, 2, \cdots, n)$; p is assumed to be unknown.[1] It is also assumed that the n pairs of observations (x_i, y_i), $(i=1, 2, \cdots, n)$ are independent; i.e., the outcome (+ or −) for (x_j, y_j) is independent of the outcome for (x_i, y_i) $(i \neq j)$.

B. *The Observations.* The purpose of obtaining observations (x_i, y_i) is to make an inference regarding p. The observed quantity upon which an inference is to be based is r, the number of +'s or −'s (whichever occur in fewer numbers) obtained from n paired observations (x_i, y_i). On the basis of the assumption above it follows that the probability of obtaining exactly r as the minimum number of +'s or −'s is:

$$\binom{n}{r}\left[p^r(1-p)^{n-r} + p^{n-r}(1-p)^r\right] \quad r = 0, 1, 2, \cdots, \frac{n-1}{2};\ n \text{ odd}$$

$$r = 0, 1, 2, \cdots, \frac{n-2}{2};\ n \text{ even}$$

$$\binom{n}{\frac{1}{2}n} p^{\frac{1}{2}n}(1-p)^{\frac{1}{2}n} \qquad r = \frac{n}{2};\ n \text{ even.}$$

C. *The Inference.* In the sign test the hypothesis being tested is that $p=\frac{1}{2}$; in other words that the distributions of the differences $x_i - y_i$ $(i=1, 2, \cdots, n)$ have zero medians. For the more general tests discussed in Section 5, the hypothesis is that the differences $x_i - f(y_i)$ $(i=1, 2, \cdots, n)$ have zero medians. The function $f(y)$ may be Py or $Q+y$ (where P and Q are the constants mentioned in Section 5) or any other function appropriate for comparison with x in the problem at hand.

The hypothesis that $p=\frac{1}{2}$ is tested by dividing the possible values of r into two classes and accepting or rejecting the hypothesis according as r falls in one or the other class. The classes are chosen so as to make small (say $\leqq \alpha$) the chance of rejecting the hypothesis when it is true and also to make small the chance of accepting the hypothesis when it is untrue. It can be shown that in a certain sense, the best set of rejection values for r is $0, 1, \cdots, R$, where R depends on α and n. R can be determined by solving for R = maximum i in the inequality:

$$\sum_{j=0}^{i}\binom{n}{j}\left(\frac{1}{2}\right)^n = I_{\frac{1}{2}}(n-i, i+1) \leqq \tfrac{1}{2}\alpha$$

where $I_x(a, b)$ is the incomplete beta function. Table 1 was computed in this way.

---

[1] An additional assumption is that the probability $A_i = B_i$ is 0; thus the probability $B_i > A_i$ is $(1-p)$.

D. *Sample Sizes.* When the sample size is small the sign test is likely to reject the hypothesis, $p=\frac{1}{2}$, only if p is near zero or one. If p is near, but not equal to $\frac{1}{2}$, the test is likely to reject the hypothesis, $p=\frac{1}{2}$, only when the sample is large.

The sample size required to reject the hypothesis $p=\frac{1}{2}$ at the α level of significance, $100\lambda\%$ of the time, may be determined by finding the largest i and smallest n which satisfy:

$$\sum_{j=0}^{i}\binom{n}{j}\left(\frac{1}{2}\right)^{n} \leqq \frac{1}{2}\alpha$$

and

$$\sum_{j=0}^{i}\binom{n}{j}p^{j}(1-p)^{n-j} \geqq \lambda \qquad\qquad p<\tfrac{1}{2}.$$

n and i are given in Table II for various values of p and α; λ was taken to be .95 in all cases. The tabular values for $1-p$ are the same as those for p because of the symmetry of the binomial distribution.

E. *Efficiency of the Sign Test.* Let $z=x-y$. Assume z is normally distributed with mean a and variance σ^2. The probabality of obtaining a + on a particular z_i is:

$$p=\frac{1}{\sqrt{2\pi}}\int_{-a/\sigma}^{\infty} e^{-\frac{1}{2}u^2}du.$$

An estimate of p involving only the signs of z_i $(i=1, 2, \cdots, n)$ yields an estimate of (a/σ). Cochran (reference 2) has shown that in large samples the variance of this estimate of (a/σ) is $2\pi pq\, e^{(a/\sigma)^2}/n$. We shall denote a/σ by c.

The efficiency of an estimate based on n independent observations is defined as the limit (as $n\rightarrow\infty$) of the ratio of the variance of an efficient estimate to that of the given estimate. An efficient estimate of c is:

$$\frac{\bar{z}}{\sqrt{\dfrac{\sum(z_i-\bar{z})^2}{n-1}}}=\frac{t}{\sqrt{n}}$$

where t is Student's t and $\bar{z}=\sum z_i/n$.

The variance of this estimate is $1/(n-2)$; thus the efficiency, E, of the sign test is $e^{-c^2}/2\pi pq$. If $c=0$, then $p=\frac{1}{2}$ and the efficiency is $2/\pi=63.7\%$.

The preceding discussion pertains to large values of n; for small values of n, the efficiency is a little better than 63.7%. Computations were made for several smaller values of n, namely, for $n=18$, 30, 44 pairs of observations at the 10% level of significance. It was found that the sign test using 18 pairs of observations is approximately equivalent to the t-test using 12 pairs of observations; for 30 pairs the equivalent t-test requires between 20 and 21 pairs; and for 44 pairs the equivalent t-test requires between 28 and 29 pairs. Cochran shows that the efficiency of r/n for estimating c decreases as $|c|$ increases.

REFERENCES

[1] W. MacStewart, "A note on the power of the sign test," *Annals of Mathematical Statistics*, Vol. 12 (1941), pp. 236–238.

[2] W. G. Cochran, "The efficiencies of the binomial series test of significance of a mean and of a correlation coefficient," *Journal Royal Statistical Society*, Vol. C, Part I (1937), pp. 69–73.

[3] A. Wald, "Sequential method of sampling for deciding between two courses of action," *Journal American Statistical Association*, Vol. 40 (1945), pp. 277–306.

[4] Statistical Research Group, Columbia University, *Sequential Analysis of Statistical Data: Applications* (1945), Columbia University Press, New York.