REFERENCES

[1] J. NEYMAN, "«Smooth Test» for goodness of fit," *Skandinavisk Aktuariedskrift*, (1937), pp. 149–199.

[2] G. A. BAKER, "The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 197–201.

[3] G. A. BAKER, "Transformations of bimodal distributions," *Annals of Math. Stat.*, Vol. 1 (1930), pp. 334–344.

[4] G. A. BAKER, "The relation between the means and variances, means squared and variance in samples from the combinations of normal populations," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 333–354.

[5] G. A. BAKER, "The significance of the product-moment coefficient of correlation with special reference to the character of the marginal distributions," *Jour. Am. Stat. Assoc.*, Vol. 25 (1930), pp. 387–396.

A NOTE ON THE POWER OF THE SIGN TEST

BY W. MAC STEWART

University of Wisconsin

1. Introduction. Let us consider a set of N non-zero differences, of which x are positive and $N - x$ are negative; and suppose that the hypothesis tested, H_0, implies, in independent sampling, that x will be distributed about an expected value of $N/2$ in accordance with the binomial $(\frac{1}{2} + \frac{1}{2})^N$. As a quick test of H_0, we may choose to test the hypothesis h_0 that x has the above probability distribution. Defining r to be the smaller of x and $N - x$, the test consists in rejecting h_0 and therefore H_0 whenever $r \leq r(\epsilon, N)$, where $r(\epsilon, N)$ is determined by N and the significance level ϵ.

2. Power of a test. In applying such a test it is of interest to know how frequently it will lead to a rejection of H_0 when H_0 is false and the situation H implies that the probability law of x is $(q + p)^N$, with $p \neq \frac{1}{2}$, thereby indicating an expectation of an unequal number of $+$ and $-$ differences. The probability of rejecting H_0 when H_1 implying $p = p_1$ is true, is termed the *power* of the test of H_0 relative to the alternative H_1.[1] Thus, from the point of view of experimental design the power (P) of the test of H_0 may be considered a function of the alternative hypothesis H_1, the significance level ϵ, and N. As such, the following observations may be noted:

1. The power P_2, for an assumed ϵ, N, and H_2 implying $p = p_2$ is greater than or equal to the power P_1 for ϵ, N and H_1 implying $p = p_1$ where $| p_2 - .50 | > | p_1 - .50 |$.

[1] For an extensive discussion of the power of a test, the reader is referred to J. Neyman and E. S. Pearson, *Statistical Research Memoirs*, Vol. 1 (1936), pp. 3–6.

2. The power P_2 for an assumed H_1, N, and ϵ_2, is greater than or equal to the power P_1 for H_1, N, and ϵ_1, where $\epsilon_2 > \epsilon_1$.

3. The power P_2 for an assumed H_1, ϵ, and N_2 is greater than or equal to the power P_1 for H_1, ϵ, and N_1 where $N_2 > N_1$.

Hence, to increase the power of the test of H_0 relative to a particular H_1, the methods implied in observations 2 and/or 3 may be employed. However, if any increase in an established ϵ is undesirable, the method implied in observation 3 is the alternative.

3. Explanation of table. In the interests of efficiency and economy, two questions then arise: (1) What is the minimum value of N, which, at the significance level ϵ, will give the test of H_0 a power $P > \beta$, relative to a particular alternative hypothesis H_1? (2) For this minimum value of N corresponding to ϵ, what is the maximum value of r? Stated in another manner, the questions are these: "What is the smallest number (min N) of paired samples that must be employed in conjunction with the Sign Test in order that the test of H_0, at the significance level ϵ, shall have a power $P > \beta$ relative to an alternative hypothesis H_1?" (2) If x of these paired samples give rise to a positive difference, and (min $N - x$) a negative difference, and if r be defined as the smaller of x and (min $N - x$); then, what is the maximum value that r may attain and still have the results, at the level ϵ, judged significant?

Table I provides the answers to these questions for the significance level $\epsilon \leqq .05$; and (1) for H_1 implies $p = p_1$ for values of p_1 from .60 to .95 (and thereby from .40 to .05) at intervals of .05; (2) for values of β from .05 to .95 at intervals of .05, and also for $\beta > .99$. For example, assume that a power $P > .80$ relative to the alternative hypothesis H_3 ($p_1 = .70$) is desired. In Table I, the entry appearing in the column headed H_3 ($p_1 = .70$), and in the row $P > .80$ is 49,17—indicating that 49 paired samples are required, of which 17 or less must be of one sign (+ or −) and hence 32 or more must be of the opposite sign in order that the results be significant at the .05 level.

Because of the discreteness of the binomial distribution, it is impossible to maintain the level of significance at .05 or even arbitrarily close to that figure and still hold to the criterion that N shall be at a minimum. For that reason, particularly when min N is small, results significant at .05 according to Table I may be significant at a level ϵ' where ϵ' is considerably less than .05. In general, however, and in particular when min N is large (greater than 50) both the quantities $(.05 - \epsilon')$ and $(P - \beta)$ are small.

4. Illustrative example. Goulden[2] describes a simple experiment in identifying varieties of wheat. In this experiment, a wheat "expert" is presented paired grain samples of two particular varieties of wheat. The object of the

[2] C. H. Goulden, *Methods of Statistical Analysis*, John Wiley and Sons, New York, 1939, p. 2.

experiment is to test the ability of the expert to differentiate between the two varieties by arranging the pairs so that samples of one variety are on the left, say, and samples of the other variety are on the right.

In a problem of this type, it is desirable to have a sufficiently large number, N, of paired samples in order that the following conditions be fulfilled: (1) The probability that a person possessing no discriminating ability pass the test

TABLE I

Minimum number of paired samples and maximum values of related r

$H_0 \sim p_0 = .50$

(5% level of significance, i.e., $\epsilon \leq .05$)

(min N, max r)

Power	H_8 $p_1=.95$	H_7 $p_1=.90$	H_6 $p_1=.85$	H_5 $p_1=.80$	H_4 $p_1=.75$	H_3 $p_1=.70$	H_2 $p_1=.65$	H_1 $p_1=.60$
$0 < P \leq .05$	—	—	—	—	—	—	7,0	6,0
$P > .05$	—	—	—	—	—	7,0	6,0	9,1
$P > .10$	—	—	—	—	7,0	6,0	9,1	17,4
$P > .15$	—	—	—	8,0	6,0	9,1	12,2	25,7
$P > .20$	—	—	—	7,0	10,1	13,2	17,4	37,12
$P > .25$	—	—	8,0	6,0	14,2	12,2	23,6	44,15
$P > .30$	—	—	7,0	11,1	9,1	18,4	25,7	56,20
$P > .35$	—	—	6,0	10,1	12,2	17,4	30,9	65,24
$P > .40$	—	8,0	—	9,1	16,3	20,5	35,11	74,28
$P > .45$	—	7,0	11,1	—	15,3	26,7	42,14	89,35
$P > .50$	—	6,0	10,1	13,2	18,4	25,7	44,15	101,40
$P > .55$	—	—	9,1	12,2	17,4	30,9	51,18	112,45
$P > .60$	—	—	14,2	15,3	20,5	36,11	56,20	125,51
$P > .65$	7,0	11,1	13,2	19,4	23,6	35,11	63,23	143,59
$P > .70$	6,0	10,1	12,2	18,4	25,7	40,13	67,25	158,66
$P > .75$	—	9,1	16,3	17,4	28,8	44,15	79,30	175,74
$P > .80$	—	14,2	15,3	20,5	30,9	49,17	90,35	199,85
$P > .85$	11,1	12,2	18,4	25,7	35,11	56,20	101,40	227,98
$P > .90$	9,1	15,3	17,4	28,8	42,14	65,24	114,46	263,115
$P > .95$	12,2	17,4	23,6	35,11	49,17	79,30	143,59	327,145
$P > .99$	15,3	23,6	30,9	44,15	67,25	110,44	199,85	453,205

through sheer guesswork be less than ϵ; and (2) if past experience has proven that an expert *does* possess the ability to discriminate between the varieties to the extent of placing a proportion, p_1, of the pairs correctly in the long run, then the probability that he will pass the test be P.

Under these conditions, how large an N is required, and for that N, what is the maximum number of pairs that may be incorrectly placed without failing

the test? For alternative hypothesis H_4 ($p_1 = .75$), and for $P > .90$, referring to Table I, it is seen that 42 paired samples must be employed and not more than 14 may be placed incorrectly. Under the same alternative hypothesis, if it be required merely that $P > .50$ (i.e., an expert with an ability of .75 have better than an even chance of passing), then only 18 paired samples are necessary and not more than 4 may be arranged incorrectly.

Thus, before conducting an experiment in which the Sign Test is to be employed, if the experimenter first decides what power the test must have relative to a certain alternative hypothesis; then from the accompanying table he may learn the minimum number of paired samples that are necessary; and the related maximum value of r.

If this procedure is not followed, and an experimenter employs, say 6 paired samples, he may (as can be seen from the table) discover, to his dismay, that "experts" of ability .75 will be unrecognized more than 80% of the time.

MOMENTS OF THE RATIO OF THE MEAN SQUARE SUCCESSIVE DIFFERENCE TO THE MEAN SQUARE DIFFERENCE IN SAMPLES FROM A NORMAL UNIVERSE

By J. D. Williams

Phoenix, Arizona

The following result may have considerable application to trend analysis. The specific problem was proposed to me by R. H. Kent.

Consider a sample $0_n : X_1, X_2, \cdots, X_n$ from a normal population with zero mean and variance σ^2, the variates being arranged in temporal order. We seek the moments of the ratio of δ^2 to S^2, where

$$(n-1)\delta^2 = \sum_{j=1}^{n-1} (X_j - X_{j+1})^2 \tag{1}$$

and

$$nS^2 = \sum_{j=1}^{n} (X_j - \bar{X})^2. \tag{2}$$

Here $\bar{X}$ is the mean of the X_j. In order to simplify the algebra, we will work with quantities A and B defined by

$$\begin{aligned} 2\sigma^2 A &= (n-1)\delta^2, \\ 2\sigma^2 B &= nS^2. \end{aligned} \tag{3}$$

The characteristic function for the joint distribution of A and B is

$$\begin{aligned} \varphi(t_1, t_2) &= E(e^{At_1 + Bt_2}) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \int_{-\infty}^{\infty}\int\cdots\int \exp\left(At_1 + Bt_2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n} X_j^2\right) \prod_{j=1}^{n} dX_j, \end{aligned} \tag{4}$$