Psychological Review
Vol. 67, No. 3, 1960

DIRECTIONAL STATISTICAL DECISIONS

HENRY F. KAISER

University of Illinois

This paper has two purposes. First, we shall point out a seemingly common logical error in the statistical interpretation given results of two-sided tests of statistical hypotheses. A correct interpretation of the traditional two-sided test would appear to make this classic procedure of essentially negligible interest. Second, we shall outline an appropriate treatment of the problem with which two-sided statistical tests seem concerned and contrast this procedure with the one-sided test. Throughout the paper, we shall indicate the relationship of our discussion to the prolonged controversy on one-sided tests versus two-sided tests (Burke, 1953, 1954; Goldfried, 1959; Hick, 1952; Jones, 1952, 1954; Kimmel, 1957; Marks, 1951, 1953).

The arguments developed in this paper are based on logical considerations in statistical inference. (We do not, of course, suggest that statistical inference is the only basis for scientific inference.) Our statistical interpretation and development stem primarily from the decision-theoretic position of Wald (1939, 1950).

The Nondirectional Two-Sided Test

Consider the traditional two-sided test. For example, we wish to test the null hypothesis:

$$H_2: \mu_X - \mu_Y = 0$$

against the obvious two-sided alternative:

$$H_{13}: \mu_X - \mu_Y \neq 0$$

where μ_X and μ_Y are the population means of the normally distributed random variables X and Y, and where $\sigma_X = \sigma_Y = \sigma$ is unknown.[1] This, of course, is an example of the classic *t* test.

The error in statistical interpretation of perhaps the majority of those who have used this test lies in the decision or statement made if the null hypothesis is rejected. When this occurs obviously we accept the alternative hypothesis that the population means are different. While this is correct, *we cannot logically make a directional statistical decision or statement when the null hypothesis is rejected on the basis of the direction of the difference in the observed sample means.* Our a priori alternative hypothesis merely states a nondirectional difference; logically, then, we may only state or decide upon a nondirectional difference if this alternative is accepted.

It seems difficult to imagine a problem for which this traditional test could give results of interest. To find a difference or a "significant" effect and not be able to decide in which direction this difference or effect lies, seems a sterile way to do business. One escape would be to conduct the traditional nondirectional two-sided test, and then if the alternative hypothesis is accepted, to gather new data and attempt to decide upon the direction provoked by the original nondirectional two-sided affair with the appropriate

[1] Our designation of the null hypothesis as H_2 rather than H_0 is unconventional; however, the exposition seems logically clearer if we use the subscripts 1, 2, and 3 to refer to negative, zero, and positive differences, respectively.

one-sided test. This two-stage procedure, while correct, obviously wastes data. A more efficient, single-stage procedure is described in the section after next.

THE DIRECTIONAL ONE-SIDED TEST

Consider the one-sided test. We wish, for example, to test the null hypothesis:

$$H_{12}\colon \mu_X - \mu_Y \leq 0$$

against the one-sided alternative:

$$H_3\colon \mu_X - \mu_Y > 0$$

One point of confusion concerning the above statement of the null hypothesis sometimes occurs because traditionally the definition of a null hypothesis has been restricted to the hypothesis of no difference—e.g., our H_2 of the previous section. Under this latter interpretation, the one-sided test would be for deciding between the null hypothesis H_2 and the alternative H_3—leaving the left flank unguarded. Statistically, this restriction is not necessary; a statistical hypothesis simply is a statement about the probability distribution(s) of observable random variable(s) (Neyman, 1950, p. 250). Any such statement, such as our H_{12} above, if it is the hypothesis being tested (in the sense that falsely rejecting it may occur with maximum probability given by the level of significance) is then a null hypothesis (Neyman, 1950, p. 259). The well-entrenched adjective "null" is probably misleading for it implies an unnecessary restriction on statements of hypotheses to be tested.

On the other hand, Burke (1953, 1954) has argued not unconvincingly that stating the null hypothesis in a one-sided test as a nonpositive difference may often be scientifically naive; the difference between the scientific hypotheses corresponding to H_1 and H_2 may be such that it would not be wise (extrastatistically) to toss them into the same null pot, where they remain indistinguishable.

Of course, with the one-sided test we are in the much more palatable position than with the traditional two-sided test of being able to make a directional statistical decision if the alternative hypothesis is accepted.

THE DIRECTIONAL TWO-SIDED TEST

Let us say we are interested in making a directional decision if we attain "statistical significance" and yet wish to guard against differences in both directions. This section outlines a solution of this problem for the example considered here, statistical decisions about differences between population means of normally distributed random variables with equal variance.

To do this we consider briefly the notion of a statistical decision function (Wald, 1950). A statistical decision function prescribes a correspondence between one of k possible decisions as a function of n possible observational outcomes (Neyman, 1950, p. 10). In applied statistics n is usually infinite; for example, in our problem the possible values of t are the $n = \infty$ possible outcomes. On the other hand, in conventional applied statistics k is usually either two or infinite; when k is two, we have hypothesis testing using two-valued statistical decision functions, and when k is infinite we have the problem of estimation (deciding along a continuum of points or intervals). Either of the t tests considered in the two previous sections uses a two-valued statistical decision function or, less solemnly, is a two-decision procedure, because in each case there are two, and only two, possible decisions contemplated: a decision to accept (not reject) the null hypothesis or a decision to accept the alternative. The two two-decision procedures are different, of course, both because of the nature of the hypotheses tested and because of the different correspondence established between the possible outcomes and the two decisions; i.e., the critical regions or tail(s) for the two tests are different for rejecting the null hypothesis.

Wald's (1939, 1950) contribution of the notion of a statistical decision function integrates into a single general theory what prior to 1939 were thought of as two more or less distinct branches of statistics, hypothesis testing and estimation. In this most general framework, conventional hypothesis testing is represented by two-valued statistical decision functions while estimation involves statistical decision functions of infinitely many values.

However, there is no reason why we should not consider the zone in-between: *k*-decision procedures, $2 < k < \infty$. And this is precisely what we shall do to give a correct single-stage solution to the directional two-sided decision problem. For this problem requires a *three*-valued statistical decision function (Lehmann, 1950); we wish to decide among

$$H_1\colon \mu_X - \mu_Y < 0$$

and

$$H_2\colon \mu_X - \mu_Y = 0$$

and

$$H_3\colon \mu_X - \mu_Y > 0$$

The difference between the traditional nondirectional two-sided test and the directional two-sided test proposed in this section may be seen by considering the possible errors which may occur in making a wrong decision. For the classic nondirectional test, only two errors are possible: (*a*) the error of deciding that there is a difference, when, in fact, the null hypothesis is true—an error of the first kind (α error), or (*b*) the error of not detecting that the null hypothesis is false, i.e., deciding that there is no difference when in fact there is—an error of the second kind (β error). The four possible situations may be represented conveniently in the following four-fold table:

		Nature	
		H_2	H_{13}
Decision about Nature	H_2	correct decision	β error
	H_{13}	α error	correct decision

For the directional test of this section, there are six possible errors. They may readily be seen as the off-diagonal cells in the following nine-fold table:

		Nature		
		H_1	H_2	H_3
Decision about Nature	H_1	correct decision	α_{12} error	γ_{13} error
	H_2	β_{21} error	correct decision	β_{23} error
	H_3	γ_{31} error	α_{32} error	correct decision

Any one of the possible errors in the above table is symbolized uniquely by the subscripts used: the first subscript indicates which hypothesis is decided upon, while the second indicates the hypothesis which obtains in Nature. We have added the unnecessary α, β, and γ, to provide comparability with the notation used in classical hypothesis testing. Thus, the α_{12} and α_{32} errors are similar to the α error in the usual two-decision

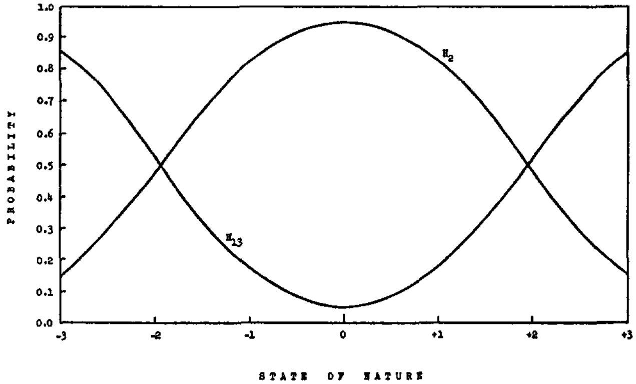

Fig. 1. Performance characteristic of the nondirectional two-sided test with $\alpha = .05$. State of Nature in units of the standard error of the difference between the means. Rule of inductive behavior (for large samples): decide upon H_2 when $-1.960 \leq t \leq +1.960$, decide upon H_{13} when $|t| > +1.960$.

problem; either involves making a false decision of difference when there is none. The β_{21} and β_{23} errors are similar to the β error; either involves not detecting a difference. The particularly repugnant γ_{13} and γ_{31} errors —"errors of the third kind"—have no parallel in classical hypothesis testing, as these "gamma errors" involve deciding upon a difference in the wrong direction.[2]

The difference between the directional and nondirectional two-sided tests may be illustrated quantitatively if we contrast their *performance characteristics.* The performance characteristic of a k decision procedure is the system of k functions, each of which gives the probability, as a function of the model describing the state of Nature, of accepting one of the k decisions contemplated (Neyman, 1950, p. 11). Figure 1 shows the performance characteristic for the classic nondirectional two-sided (equal tails) t test with level of significance α (= .05). It consists of two functions, each giving the probability of deciding upon the hypothesis (H_2 or H_{13}) indicated.[3] Note that the two functions are redundant; the curve giving the probability of accepting H_2, Wald's operating characteristic, is complementary to the curve giving the probability of accepting H_{13}, the Neyman-Pearson power function. More generally, of course, any performance characteristic has this sort of redundancy: since the k possible decisions are mutually exclusive and jointly exhaustive, the probability of making any $k - 1$ of them is sufficient to give the desired information.

Figure 2 shows the three functions of the performance characteristic of our three-decision procedure, the directional two-sided test. For this illustration the probability of making each of the α_{12} and α_{32} errors has been set at one-half the level of significance used for Fig. 1; this makes it convenient to compare the directional and nondirectional two-sided tests when the directional test is carried out under the guise of the traditional nondirectional test with level of significance α. Several comparisons of this three-decision procedure with the traditional test seem worth mentioning. When $\alpha_{12} = \alpha_{32} = \frac{1}{2}\alpha$, comparing Figs. 1 and 2:

1. The probability of accepting the null hypothesis H_2 is the same for either test for all states of Nature.

2. The probability of correctly ac-

[2] Mosteller (1948) seems to have coined the expression "errors of the third kind."

[3] Dixon and Massey (1957, Ch. 14) and Walker and Lev (1953, pp. 161–167) give excellent elementary discussions of how these functions may be computed.

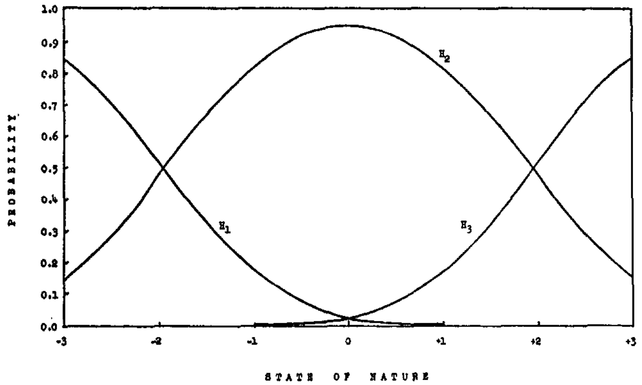

FIG. 2. Performance characteristic of the directional two-sided test with $\alpha_{12} = \alpha_{32} = .025$. State of Nature in units of the standard error of the difference between the means. Rule of inductive behavior (for large samples): decide upon H_1 when $t < -1.960$, decide upon H_2 when $-1.960 \leq t \leq +1.960$, decide upon H_3 when $t > +1.960$.

cepting either H_1 or H_3 in the directional test is less than the probability of correctly accepting H_{13} in the nondirectional test, and, for a given state of Nature, this loss of power is equal to the probability of making the nasty error of the third kind.

3. The probability of making a gamma error is always less than $\frac{1}{2}\alpha$.

An alternative treatment for the directional two-sided problem would be to make the probabilities of each of the α_{12} and α_{32} errors traditional values, like .05 or .01, rather than the .025 or .005 generated by an incorrect interpretation of the nondirectional test.

DISCUSSION

It seems obvious that the traditional two-sided test should almost never be used. If, as is typical, not rejecting the null hypothesis is a result of little scientific concern, then this test may be said never to give results of direct scientific interest because accepting the nondirectional alternative H_{13} is merely a generator of directional alternative hypotheses.[4]

Since we are proposing that almost without exception the directional two-sided test should replace the traditional nondirectional test, it seems appropriate to contrast the one-sided test with this three-decision procedure. The performance characteristic of the one-sided test is given in Fig. 3. The level of significance ($\alpha = .05$) in this illustration is the same as in Fig. 1. In comparing Figs. 2 and 3, then, we compare the one-sided test with the three-decision procedure where $\alpha_{12} = \alpha_{32} = \frac{1}{2}\alpha$. The loss of performance engendered by guarding both sides with the directional two-sided test is readily seen. Only from the traditional point of view of correctly accepting the null hypothesis, or controlling errors of the first kind, is the two-sided test as good; the two-sided test is markedly

[4] When the alternative hypothesis is so nonspecific as the nondirectional H_{13}, a compelling argument (Jones, 1955) may be made not to test hypotheses at all—that a more appropriate statistical procedure is estimation, e.g., for our problem, first find a point estimate for $\mu_X - \mu_Y$ and then determine a confidence interval about this point. Indeed, it might be argued that even the directional alternatives H_1 and H_3 are too nonspecific relative to the null hypothesis H_2, because their dimensionality in the parameter space is greater than that of H_2. A completely "balanced" or "symmetric" theory of testing hypotheses would seem logically to require that all k hypotheses under consideration have the same dimensionality in the parameter space. Such symmetry for our procedure could occur if H_1 and H_3 were chosen as specific negative and positive differences rather than any positive and negative differences.

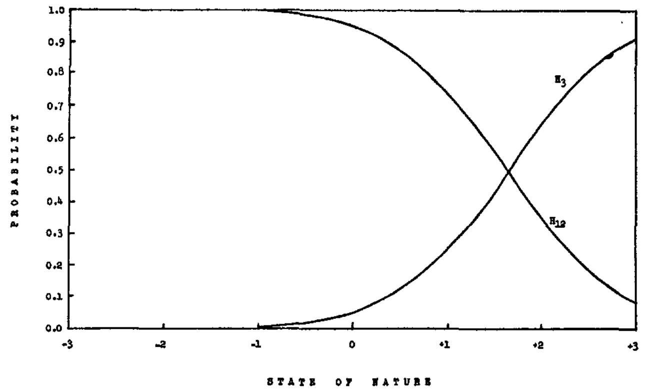

Fig. 3. Performance characteristic of the directional one-sided test with $\alpha = .05$. State of Nature in units of the standard error of the difference between the means. Rule of inductive behavior (for large samples): decide upon H_{12} when $t \leq +1.645$, decide upon H_3 when $t > +1.645$.

less powerful (in the Neyman-Pearson sense) than the one-sided test for correctly rejecting the null hypothesis in the "right" direction, from the viewpoint of the one-sided test. Also, with the two-sided test, there is always the possibility of the repulsive gamma errors.

To equate the power of the directional two-sided test with that of the one-sided test with level of significance α, it is sufficient to use this three-decision procedure with $\alpha_{12} = \alpha_{32} = \alpha$. Compare Figs. 3 and 4.

A nice feature of this comparison is that there is no difference in the critical values to t in the tail corresponding to the alternative hypothesis for the one-sided test. See Fig. 5. Thus the traditional and delicate problem of changing the number of sides in midstream and/or fudging with a posteriori alpha values cannot arise. The distinction between these two tests lies in whether differences in the "wrong" direction, from the viewpoint of the one-sided test, can lead to a decision in this direction. For the three-decision procedure proposed in this paper, this may happen; for the traditional one-sided test, it may not, as the null hypothesis there includes all nonpositive differences. At first glance, then, it might seem that one would always prefer the three-decision procedure because it guards against differences in both directions —differences which of course may be decided upon directionally. It is suggested that this argument is not to be taken lightly; consult Burke (1953, 1954) for an extended and convincing

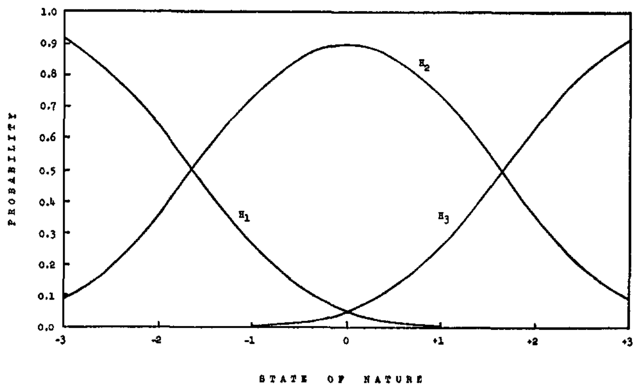

Fig. 4. Performance characteristic of the directional two-sided test with $\alpha_{12} = \alpha_{32} = .05$. State of Nature in units of the standard error of the difference between the means. Rule of inductive behavior (for large samples): decide upon H_1 when $t < -1.645$, decide upon H_2 when $-1.645 \leq t \leq +1.645$, decide upon H_3 when $t > +1.645$.

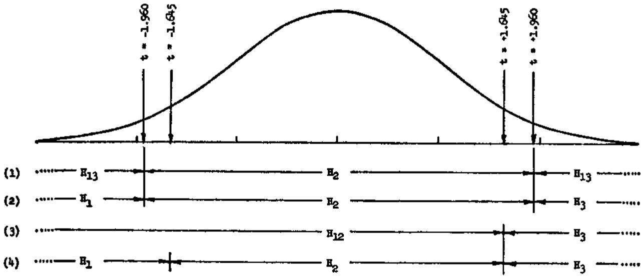

FIG. 5. Ranges of values of t (for large samples) leading to the decision indicated for four statistical tests. The numbers in parentheses at the left are the same as the numbers of the above figures giving the performance characteristics of these tests: (1) nondirectional two-sided test with $\alpha = .05$, (2) directional two-sided test with $\alpha_{12} = \alpha_{32} = .025$, (3) directional one-sided test with $\alpha = .05$, (4) directional two-sided test with $\alpha_{12} = \alpha_{32} = .05$.

defense of the scientific desirability of procedures which will detect differences in both directions.

However, the choice is not completely clear cut. Consider again Fig. 4, the performance characteristic of the three-decision procedure where $\alpha_{12} = \alpha_{32} = \alpha$. In comparing Figs. 3 and 4, one traditionally serious disadvantage of the directional two-sided test obtains: in the unlikely event that the null hypothesis H_2 is true, i.e., the population mean difference is exactly equal to zero, then the probability of accepting this null hypothesis is only $1 - \alpha_{12} - \alpha_{32} = 1 - 2\alpha$ for the two-sided test as compared with probability $1 - \alpha$ of accepting H_{12} when the one-sided test is used.

We have not attempted to settle the scientific issue of one-sided versus two-sided tests. However, it is hoped that the problem has been recast so as to eliminate confusion arising from failing to distinguish directional from nondirectional two-sided tests. As for the scientific issues briefly outlined in this section, a more detailed and perhaps more compelling defense of either test may be found in the papers referred to in the first paragraph of this paper, if it is remembered that these writers are almost certainly referring to our three-decision procedure when speaking of two-sided tests.

NOTES

The directional two-sided test proposed in this paper need not necessarily be developed explicitly as a three-decision procedure. An alternative approach would be simultaneously to make two one-sided tests (Hodges & Lehmann, 1954): H_{12} against H_3 and H_{23} against H_1. If both these two-decision procedures are carried out simultaneously and with the same data, at level of significance α, it is readily seen that we have exactly the equivalent of the three-decision procedure illustrated and described in Fig. 4.

It is perhaps worth pointing out that it is surely not necessary and may not always be best to set $\alpha_{12} = \alpha_{32}$ in our directional two-sided test; after all, the α_{12} and α_{32} errors may entail very different consequences. Indeed, we may envision a continuum of possible partitions of $\alpha_{12} + \alpha_{32}$ from a left tail critical one-sided test through the equal tails three-decision procedure described in this paper to a right tail critical one-sided test.

It has been convenient to discuss problems about differences between means. The rationale and application of the three-decision procedure outlined above may easily be extended to other problems involving other parameters where the traditional alternative hypothesis lies on both sides of the null hypothesis.

Finally, it might also be noted that the statistical notion of the number of sides or "tails" bears no necessary relation to the scientific notion of whether the test is directional or nondirectional. For example, the one-sided t test is directional scientifically and one-sided statistically, while most traditional F and chi square tests are nondirectional scientifically and one-sided statistically. As such, directional decisions cannot properly be made with such F and chi square tests and they are to be thought of merely as hypothesis generators for scientifically more explicit statistical decision procedures.

REFERENCES

BURKE, C. J. A brief note on one-tailed tests. *Psychol. Bull.*, 1953, **50**, 384–387.

BURKE, C. J. Further remarks on one-tailed tests. *Psychol. Bull.*, 1954, **51**, 587–590.

DIXON, W. J., & MASSEY, F. J. *Introduction to statistical analysis.* (2nd ed.) New York: McGraw-Hill, 1957.

GOLDFRIED, M. R. One-tailed tests and "unexpected" results. *Psychol. Rev.*, 1959, **66**, 79–80.

HICK, W. E. A note on one-tailed and two-tailed tests. *Psychol. Rev.*, 1952, **59**, 316–318.

HODGES, J. L., & LEHMANN, E. L. Testing the approximate validity of statistical hypotheses. *J. Roy. statist. Soc., Lond., Ser. B (Methodological)*, 1954, **16**, 261–268.

JONES, L. V. Tests of hypotheses: one-sided vs. two-sided alternatives. *Psychol. Bull.*, 1952, **49**, 43–46.

JONES, L. V. A rejoinder on one-tailed tests. *Psychol. Bull.*, 1954, **51**, 585–586.

JONES, L. V. Statistical theory and research design. *Annu. Rev. Psychol.*, 1955, **6**, 405–430.

KIMMEL, H. D. Three criteria for the use of one-tailed tests. *Psychol. Bull.*, 1957, **54**, 351–353.

LEHMANN, E. L. Some principles of the theory of testing hypotheses. *Ann. math. Statist.*, 1950, **21**, 1–26.

MARKS, M. R. Two kinds of experiments distinguished in terms of statistical operations. *Psychol. Rev.*, 1951, **58**, 179–184.

MARKS, M. R. One- and two-tailed tests. *Psychol. Rev.*, 1953, **60**, 207–208.

MOSTELLER, F. A k-sample slippage test for an extreme population. *Ann. math. Statist.*, 1948, **19**, 58–65.

NEYMAN, J. *First course in probability and statistics.* New York: Holt, 1950.

WALD, A. Contributions to the theory of statistical estimation and testing hypotheses. *Ann. math. Statist.*, 1939, **10**, 299–326.

WALD, A. *Statistical decision functions.* New York: Wiley, 1950.

WALKER, H. M., & LEV, J. *Statistical inference.* New York: Holt, 1953.

(Received June 3, 1959)