

REPRINTS AND REFLECTIONS

Tests of significance considered as evidence*

By Joseph Berkson, MD

Division of Biometry and Medical Statistics, Mayo Clinic

‘After all, the higher statistics are only common sense reduced to numerical appreciation.’—Karl Pearson

There was a time when we did not talk about tests of significance; we simply did them. We tested whether certain quantities were significant in the light of their standard errors, without inquiring as to just what was involved in the procedure, or attempting to generalize it. In recent years tests of significance have been more broadly conceived as tests of hypotheses, and they have been generalized as *t* tests, *F* tests and certain amplifications of these, such as analysis of variance or of covariance. It is hardly an exaggeration to say that statistics, as it is taught at present in the dominant school, consists almost entirely of tests of significance, though not always presented as such, some comparatively simple and forthright, others elaborate and abstruse. Behind this is a doctrine of analysis that consists of setting up what is called a ‘null hypothesis’ and testing it. Indeed, in this conception not only does this procedure characterize the method of statistics, but it is considered to be the very essence of all experimental science. In his well known book, *The Design of Experiments*, R.A. Fisher wrote, ‘Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.’¹

What is this null hypothesis procedure? I quote from a recent text.²

We have just set up the hypothesis that our sample of 900, which has a mean of 15 071 miles, is a random sample drawn from the population having a known mean of 15 200 miles ... Such a hypothesis is called a *null* hypothesis since our computations undertake to nullify it. The procedure may be summarized into three steps: (1) Set up the hypothesis that the true difference is zero. (2) Upon the basis of this hypothesis determine the probability that such a difference as the one observed might occur because of sampling variations. (3) Draw a conclusion concerning the hypothesis. If such observed difference could hardly have occurred by chance, we have cast much doubt upon the hypothesis. We therefore abandon the hypothesis and conclude that the observed difference is significant.

This I believe is a fair if abbreviated statement of the essential procedure as it is generally understood. If the experience at hand would occur only very infrequently in a given hypothesis, the hypothesis is considered disproved.

The argument has an apparent plausibility and for many years I adhered to it. However, set against experience with actual problems, reflection has led me to the conclusion that it is erroneous, and that a re-evaluation will lead to clearer comprehension in the application of tests of significance and also serve as a corrective of some of its misuses.

In the first place, the argument seems to be basically illogical. Consider it in symbolic form. It says ‘If *A* is true, *B* will happen sometimes; therefore if *B* has been found to happen, *A* can be considered disproved.’ There is no logical warrant for considering an event known to occur in a given hypothesis, even if infrequently, as disproving the hypothesis.

More to the present point, the argument does not seem to accord with what would be the mode of reasoning in ordinary rational discourse, nor with the rationale of usual procedures as they are observed in the scientific laboratory. Suppose I said, ‘Albinos are very rare in human populations, only one in fifty thousand. Therefore, if you have taken a random sample of 100 from a population and found in it an albino, the population is not human.’ This is a similar argument but if it were given, I believe the rational retort would be, ‘If the population is not human, what is it?’ A question would be asked that demands an *affirmative* answer. In the null hypothesis schema we are trying only to nullify something: ‘The null hypothesis is never proved or established but is possibly disproved in the course of experimentation.’ But ordinarily evidence does not take this form. With the corpus delicti in front of you, you do not say, ‘Here is evidence against the hypothesis that no one is dead.’ You say, ‘Evidently someone has been murdered.’

Nor do you find experimentalists typically engaged in disproving things. They are looking for appropriate evidence for affirmative conclusions. Even if the mediate purpose is the disestablishment of some current idea, the immediate objective of a working scientist is likely to be to gain affirmative evidence in favor of something that will refute the allegation which is under attack.

Does this mean that the application of tests of significance is in basic discord with rational scientific procedure? I am not sure. I think that there is a possibility of using them soundly, but the rule of inference on which they are supposed to rest has been misconceived, and this has led to certain fallacious uses.

Consider the objective of testing whether a distribution is normal. One could validly say, ‘If the distribution is normal and, the skewness of the sample, g_1 , having been calculated, if a die

* A paper presented at the 103rd annual Meeting of the American Statistical Association, New York, December 29, 1941.

Reprinted with permission from *The Journal of the American Statistical Association*. Copyright 1942 by the American Statistical Association. All rights reserved. Tests of significance considered as evidence. Joseph Berkson. *J Am Statist Assoc* 1942;37:325–35.

of 100 faces, five which are black, is thrown at random, a black face will occur only five times in 100.' No one would suggest that the finding of a black face on a die following such a calculation is any reason for rejecting the null hypothesis that the distribution is normal. But when one says, 'If the distribution is normal, a value of $g_1/S_{g_1} \geq 1.96$ will occur only five times in 100,' the finding of such a value of g_1/S_{g_1} is taken as reason for rejecting the null hypothesis. What is the essential difference between the two situations? Following the procedures which were outlined for dealing with a null hypothesis, one should reject the hypothesis that the distribution is normal on the finding of a black face, for it is surely an event rare in the circumstance of the distribution being normal. The difference appears to be that we recognize that if distribution actually were *abnormal* (skew), the occurrence of a black face still would not be expected, but a large value of g_1/S_{g_1} would be expected. The latter constitutes evidence *in favor* of skewness. We may discern, as operating in the realm of tests of significance, a principle that I suggest is generally operative in scientific inquiry; it is this. The finding of an event which is *frequent* under a hypothesis H_1 can be taken as evidence *in favor* of H_1 . If H_0 is a contradictory alternative to H_1 for which the event would not be frequent, then per corollary the finding of the event is, in so far, evidence in disfavor of H_0 .

At this point I can imagine the question rising, 'What difference does it make whether you say that you *reject* H_0 because for it the event is not frequent, or because you are *accepting* the alternative H_1 for which it is frequent?' To this the first answer must be that it would seem to be a sound idea to get one's head clear as to what are the principles on which one is really acting. If an event has occurred, the definitive question is not, 'Is this an event which would be rare if H_0 is true?' but 'Is there an alternative hypothesis under which the event would be relatively frequent?' If there is no plausible alternative at all, the rarity is quite irrelevant to a decision, and if there is such an alternative, the decisive question is, 'Would the event be relatively frequent?' Secondly, the pursuit of a false principle for testing the null hypothesis will lead to false conclusions that will be avoided if one is consciously guided by the principle suggested here as being the correct one. I shall cite an example.

As an illustration of a test of linearity under the caption, 'Test of straightness of regression line,' R.A. Fisher utilizes data relating the temperature to the number of eye facets of *Drosophila melanogaster*, the facet number being measured in factorial units. An analysis of variance procedure is utilized for the test and, the calculations having been made, Fisher says,

The deviations from linear regression are evidently larger than would be expected, if the regression were really linear, from the variations within the arrays. For the value of z we have 1.2434 while the 1 per cent point is about .488. There can therefore be no question of the statistical significance of the deviations from the straight line ... the departure from linearity was markedly significant.³

I have plotted the data of mean facet number in relation to temperature together with the least square line and they are shown in Chart I.

It was found by the significance test as applied that this regression was not straight, but on inspection it appears as

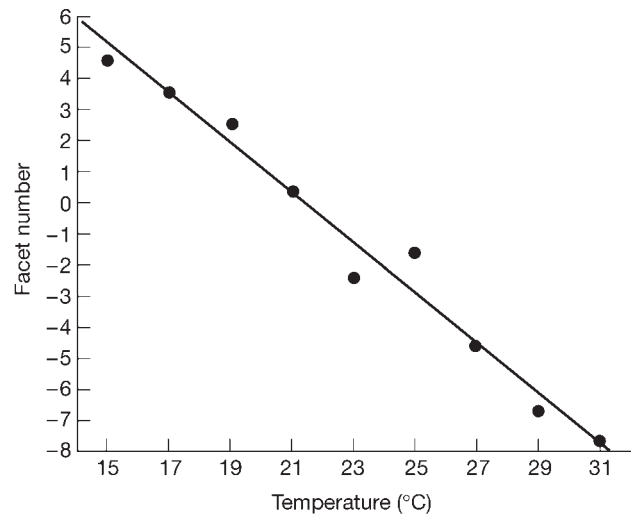


Chart I Mean number of eye facets of *Drosophila melanogaster* raised at different temperatures and best fitting straight line by method of least squares

Source: Data from R.A. Fisher, *Statistical Methods for Research Workers*. London, Oliver and Boyd, 1938, p. 260.

straight a line as one can expect to find in biological material. What has betrayed the author is a faithful adherence to an unsound principle: to wit, reject the null hypothesis tested, in this case that the regression is linear, if the P of the test is small.

Let us consider the problem according to the principle advanced here. The event which has been found to have happened, in this case the small P , is to be considered as evidence in favor of any hypothesis under which it would be a frequent occurrence. Under what hypothesis would the P , considering its mode of calculation, be a frequent occurrence? If the regression were curvilinear, a small P is to be expected relatively frequently. In so far as this is so, a small P is evidence *in favor of curvilinearity* and because of this and *primarily because of this*, a small P can be considered evidence in disfavor of its alternative, linearity. But also a small P is to be expected relatively frequently if the regression is linear and the variability heteroscedastic; hence a small P is also evidence in favor of linearity plus heteroscedasticity. Or again a small P is to be expected frequently if the regression is linear and a value of the abscissal variate, in this case the temperature, is not constant but subject to fluctuation. And there may be other conditions which, with linearity, would produce a small P relatively frequently. The small P is favorable evidence for any or several of these. Which of these shall be taken to have been demonstrated by the evidence of the small P will have to be determined by other evidence, possibly other statistical tests. In this case my own judgment would be, not that the regression is nonlinear, but that the temperature has varied during each or some of the experiments. At least that would explain the small P .

According to what is advocated here, we cannot lay down any pat axiomatic rules such as 'A very small P disproves the hypothesis tested,' or 'Equally, a very high P disproves the hypothesis,' for it is not primarily the infrequency of the P which gives the finding its meaning. Each test will have to

be examined and the circumstances in which it is applied will have to be examined, to find out, as best we can, whether any particular regions of P will occur relatively frequently in the case of an alternative to the tested hypothesis. There are situations in which a very large P will be frequent in an alternative, and in these circumstances, but *only in these circumstances*, a very high P can be said to disfavor the null hypothesis. I cite an example.

If with $(n + 1)$ observations from a frequency distribution of a variate x the quantity ns^2/\bar{x} is calculated, where $\bar{x} = \Sigma x/(n + 1)$ and $s^2 = \Sigma(x - \bar{x})^2/n$, it is known that the quantity is distributed in random samples as χ^2 for n degrees of freedom, if the distribution is Poisson.

Small values of P , say $P \leq 0.05$, will occur with the small frequency of five times in 100. If, however, the distribution is what has been called supernormal, a distribution that is known to characterize certain physical situations, the variance σ^2 is greater than the mean μ , and in random samples large values of the quantity χ^2 , and correspondingly low values of $P \leq 0.05$ will be more frequent than five in 100. The finding of a $P \leq 0.05$ therefore can be taken as preponderant favorable evidence for the super-Poisson, and hence as unfavorable to the null hypothesis tested that the distribution is Poisson. Similarly, if the distribution is Poisson, large values of P , say $P \geq 0.95$, will occur with the small frequency, five times in 100. If, however, the distribution is Bernoullian-binomial or sub-Poisson, the variance σ^2 will be less than the mean μ , and small values of the quantity χ^2 and correspondingly large values of $P \geq 0.95$ will be more frequent than five times in 100. The finding of a $P \geq 0.95$ therefore can be taken as preponderant favorable evidence for the Bernoullian or sub-Poisson, and hence as unfavorable to the null hypothesis tested that the distribution is Poisson. Here then is a case in which either a very low value of P or a very high value can be considered as warrant for rejecting the null hypothesis. There are other such cases, but the rule is not general.

So much for the meaning of P 's which are relatively frequent in the case of an alternative, and in so far, are evidence in disfavor of the null hypothesis tested. In the cases in which a very low P or very high P is evidence in favor of an alternative, what can we say of the finding of a middle value of P , say a P in the region 0.3 to 0.7? Statistical authors are not very clear about this. For the most part they merely confine themselves to statements that a low P disproves and one which is not low does not disprove. In some cases they say explicitly that a low P *disproves* but one which is not low does not *prove* the null hypothesis. What such a P should mean according to the principle advanced here is unequivocally clear. Since by definition such P 's will occur frequently in the case in which the null hypothesis is true, the finding of one is to be taken as *prima facie* evidence *in favor* of the *null hypothesis*. That is in fact the way the statistician uses them, in contradistinction to the way he says they should be used when he describes the testing of the null hypothesis.

This was somewhat amusingly illustrated at one of our meetings. One of our most eminent members gave a paper presenting the application of the lambda test and used for illustration data designed to test a certain Mendelian hypothesis. The data having been examined and the test applied, a P of about 0.6 was found. 'We can say therefore,' he remarked, 'that the results substantiate the hypothesis.' He applied the

test illustratively to several other sets of data successively and getting a P of considerable size, each time he said, 'The results therefore substantiate the hypothesis.' When he was finished, an equally eminent mathematical colleague rose to object and said, 'You cannot say that the results of the test support the hypothesis; all you are able to say is that they have not in these data disproved it.' The most interesting part of the colloquy is that the first mathematician accepted the correction!

This I find is rather typical. In the abstract the mathematical statistician insists that the middle value of P only fails to refute the hypothesis; but if he is dealing with real data and gets interested in the physical problem in hand, he forgets his statistical principles and relapses to the rules of inference applied generally in such problems.

That statisticians with real problems in hand do interpret a middle P as positive support for the null hypothesis can be readily illustrated by innumerable examples to be found in the literature. I shall cite one that is in a field in which I once did some work. 'Student,' in his classic paper on the error of count with a hemocytometer,⁴ used a series of data to examine whether the actual distribution in the hemocytometer followed the Poisson distribution, as it should on certain physical assumptions. He applied the Pearson chi-square test to a number of series and finding the P 's taken together fairly large, he concluded that the distribution was sensibly Poisson, and that therefore the variability could be taken as the square root of the average count. If this positive conclusion in favor of the null hypothesis tested was not obtained from the relatively high P 's, then his statistical work was entirely irrelevant. Other examples of the use by statisticians of relatively high P 's for demonstration of the null hypothesis are easily found if one keeps a weather eye open for them.

When I say that a middle value of P is to be considered valid evidence in favor of the null hypothesis, I have by no means resolved all the pertinent questions that may be asked regarding it. I do not say anything has been 'proved' or 'disproved.' I leave to others the use of these words, which I think are quite inadmissible as applying to anything that can be accomplished by statistics. All I say is that what we have is in the nature of positive supporting evidence. Whether the evidence is of sufficient weight to be convincing is another matter.

The development of what should be taken to affect the weight of the evidence is beyond anything I wish to undertake but a few pertinent remarks I do wish to make. Whereas it can be said that the evidence provided by a small P correctly evaluated is broadly independent of the number in the sample from which it has been calculated, this is not true for such evidence as is provided by a P in the middle region, say 0.3 to 0.7. Consider Table I depicting the hypothetical results of a physician's judgments based on a serological test, designed to ascertain the sex of a fetus *in utero*. Examine experience 1, divest yourself of formal rules, and consider what would be your reaction. I think I can fairly guess that it would be something like this: 'We cannot say anything from this experience: it certainly does not present any convincing evidence that the physician can discriminate between the sexes. But I should not want to say either that he cannot discriminate. The experience is too small for any conclusion.' With experience 2 I think you would say, or at any rate I should: 'There is no question in my mind; quite evidently the physician does not possess any ability

Table 1 Hypothetical results: determination of sex

Category	Experience 1			Experience 2		
	Total	Judgement of sex		Total	Judgement of sex	
		Correct	Incorrect		Correct	Incorrect
Expected by chance	10	5	5	1000	500	500
Physician's judgment	10	6	4	1000	505	495
<i>P</i>		0.38			0.38	

to discriminate by this serological test between the sexes. The experiment is quite large enough, and if he could discriminate to any significant degree we should see it in the results, which we do not.'

Now for both experiences, the *P*, which is the probability of obtaining by chance as good a result as the one obtained, on the null hypothesis that the probability of either sex is a half, is the same, namely, 0.38. But the experience 2, being based on large numbers, is convincing positive evidence of the truth of the null hypothesis within practical limits. I do not intend to attempt to analyze what is the justification for the added conviction provided when the numbers are large, beyond suggesting that it has the same basis as what has been argued here is the general principle of inference which is operative throughout. When the numbers are small, a middle *P* will occur with considerable frequency if the null hypothesis is true or if an alternative is; with large numbers such a *P* will occur frequently in the case of the null hypothesis but not in the case of a practical alternative. Hence with large numbers, a middle *P* provides probative evidence in favor of the null hypothesis.

Here we have disclosed one fundamental weakness in the position of those who contend that small samples can be effectively utilized in statistical investigations if the calculations of the *P*'s are correctly made. If it were a fact that conclusions are drawn only when the *P* is very small and the null hypothesis disproved, then so far as concerns the main considerations here developed, there would be a certain validity to this view, for small *P*'s are more or less independent, in the weight of the evidence they afford, of the numbers in the sample. But if actually it is the fact that conclusions will be drawn from *P*'s which are not small, then only very considerable numbers in the sample are reliable.

If a test for the difference between means has yielded a large or middle *P*, it does not merely fail to disprove the null hypothesis that the true means are equal; it furnishes *affirmative evidence* that the means are substantially equal. If the numbers on which the test is based are large, the evidence will have convincing weight; otherwise not. Contrariwise a low *P* points affirmatively towards the alternative that the means are unequal. It is the merit of some kinds of tests that they indicate unequivocally the specific alternative toward which they point.⁵ Such are tests for the difference between means or the difference between variances or tests for skewness. Other tests such as the frequency χ^2 or some applications of the analysis of variance do not have this characteristic. In Table II is presented an experience of mortalities following certain operations with and without the use of a vaccine for the prevention of peritonitis. Four tests are given for the 'null hypothesis' that the true mortality rates are identical for patients with and without vaccine: (1) the probability of getting as many differences in the favorable

direction as found; (2) the appropriate *P* for the χ^2 test of the four-fold table constituted by the totals; (3) the Fisher test of combining the value of $\chi^2 = -2 \ln Px^2$; (4) the summation of the χ^2 and degrees of freedom for the separate operations. The resulting *P*'s are considerably different. In terms of the usual rationalization, each of these tests is equally valid for testing the null hypothesis. If the null hypothesis were true, that is, if the vaccine were ineffective and the mortality for any operation were the same whether the vaccine were used or not, the appropriate limiting value of each test function would occur only infrequently—one just as infrequently as the other. But the tests are differently sensitive to the presence of different *alternatives*. In terms of the Neyman-Pearson formulation they have different powers for any particular alternative, and hence are likely to give different results in any particular case. How blind is the procedure of doing some test of significance, when there is no knowledge at hand as to whether it is likely to show a significant result or not show one, no matter how importantly different the facts may be from the hypothesis tested. The importance of this consideration is underscored when we realize that in practical applications the failure to show significant result will be taken to corroborate the null hypothesis. It is an important but neglected task of mathematical statistics to investigate what alternatives are particularly pointed to by specified findings with different tests.

I should like to see the development of investigation of the finding of middle *P*'s. I am not ready to say what this should be or just what it would lead to. But this is an example of what I mean. With the development that we now have, which emphasizes the low *P*'s, we find such statements as the following in the literature, and it is typical of the essential procedure in many fields in which statistical tests are applied. A standard curve for estimating dosage from mortality has been established with its confidence zones, from a first set of data. A set of data for another drug is to be used for estimating the potency of a second drug. But realizing the possibility that the standard curve may not be applicable any more, the author counsels the use of some controls to see whether the standard curve still applies for the first drug. He says, 'When the controls have been shown to agree with the standard of the regression line by the appropriate χ^2 or *t* test, the first curve can be used.' Now what is meant by this is that if the test does not show a low *P*, the curve can be used, which is to say that if the test shows a middle *P*, the curve *will* be used. It should be clear on consideration that if there is a real discrepancy of a given size between the present conditions and the curve, a *P* which is not low will result with small numbers, while with the same discrepancy a low *P* will result if the numbers are large. The use of the suggested rule could easily be disastrous if drugs were standardized on the basis of it and small numbers were used. Investigation should be made which

Table II Mortality rates for operations with and without use of vaccine: tests of significance of differences

Type of operation	Vaccine			No vaccine			Mortality difference per cent
	Operations	Hospital deaths		Operations	Hospital deaths		
		Number	Per cent		Number	Per cent	
A	107	2	1.9	142	4	2.8	-0.9
B	28	3	10.7	60	9	15.0	-4.3
C	21	3	14.3	34	5	14.7	-0.4
D	21	4	19.0	34	8	23.5	-4.5
E	47	3	6.4	45	4	8.9	-2.5
F	21	1	4.8	26	2	7.7	-2.9
Total	245	16	6.53	341	32	9.38	-2.85
Test	<i>P</i>						
1. Signs	0.016						
2. Total difference mortality	0.11						
3. Combination of <i>P</i> 's—Fisher	0.91						
4. Summation of χ^2 and D.F.	0.98						

could result in a rule not such as just given, but rather of the following kind: 'If the control is tested with data including so and so degrees of freedom and if the test results in a *P* of this amount or higher, the curve may be accepted as stable.'

References

- ¹ R.A. Fisher, *The Design of Experiments*. Ed 2, London, Oliver and Boyd, Ltd., 1937, p. 19.
- ² F.E. Croxton and DJ Cowden. *Applied General Statistics*. New York, Prentice-Hall, Inc., 1940, p. 310.

³ R.A. Fisher, *Statistical Methods for Research Workers*. Ed 7, London, Oliver and Boyd, Ltd., 1938, pp. 259–265.

⁴ Student, 'On the Error of Counting With a Hemocytometer,' *Biometrika* **5**: 351–360, 1906–1907.

⁵ Elsewhere I have suggested that those tests are ones which in principle can be stated alternatively and equivalently in terms of an estimate and its confidence limits. Joseph Berkson, 'Comments on Dr. Madow's "Note on Tests of Departure from Normality" with Some Remarks Concerning Tests of Significance.' This JOURNAL **46**: 539–541, December 1941.