SOME DIFFICULTIES OF INTERPRETATION ENCOUNTERED IN THE APPLICATION OF THE CHI-SQUARE TEST*

By Joseph Berkson, M.D.
Division of Biometry and Medical Statistics,
The Mayo Clinic, Rochester, Minnesota

The remarks that I have to make are not derived from any considerations of the mathematics of the chi-square test.[1] I have a considerable interest in mathematical statistics, but very little competency in it. You will not hear anything about cards or black and white balls from me. I shall speak as a practitioner who has frequently applied the test to real observations, made seriously for the solution of concrete scientific problems. I have used the chi-square test to help make decisions as to the character of experimental data in situations in which I had every reason to think it was appropriate. I have used it in the same spirit in which we doctors use, say, the Wassermann test, to help make decisions in situations where we think a patient may have syphilis. In the course of these experiences I have encountered numerous situations in which the test did not adequately perform the function for which I thought I could use it, and I shall present a few examples *seriatim:*

I. I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small. Having observed this, and on reflection, I make the following dogmatic statement, referring for illustration to the normal curve: "If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large—for instance, on the order of 200,000—the chi-square P will be small beyond any usual limit of significance."

This dogmatic statement is made on the basis of an extrapolation of the observation referred to and can also be defended as a prediction from *a priori* considerations. For we may assume that it is practically certain that any series of real observations does not actually follow a normal curve *with absolute exactitude* in all respects, and no matter how

* A paper presented at the Ninety-ninth Annual Meeting of the American Statistical Association, Atlantic City, New Jersey, December 27, 1937.

[1] In this discussion I mean, by the chi-square test, the comparison of two sets of frequencies in which chi-square is the sum of the terms $(o-t)^2/t$ calculated from the observed and theoretical frequencies, not other tests using the chi-square distribution, such as the testing of the significance of the difference of an observed and theoretical variance.

small the discrepancy between the normal curve and the true curve of observations, the chi-square P will be small if the sample has a sufficiently large number of observations in it.

If this be so, then we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all![2]

II. In Table 1 and Chart I are shown four series of observations of basal metabolism for humans, each representing a different situation. To each series a normal curve has been fitted and a chi-square test for goodness of fit made. Judging the results in the routine way, and using $P=0.05$ as the limit of significance, we would say that the first fit ($P=0.62$) is good, *i.e.*, not rejected; the second and third ($P=0.02$) would be rejected; about the fourth ($P=0.996$), there is a question as to what should be done. Fisher says: (*Statistical Methods*, 4th ed., p. 83) "Values over 0.999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. . . . In these cases the hypothesis considered is as definitely disproved as if P had been 0.001." I can only interpret this to mean, applied here, that the hypothesis that the distribution comes from a normal universe is to be rejected just as definitely as it would have been if the P had been 0.004.

Now, in these four instances, when I considered them from my own personal viewpoint, I actually made the following decisions: The first I considered a good fit, *i.e.*, I accepted the conclusion that on the evidence at hand these observations follow the normal curve. (I will not stop for the hair-splitting question as to whether I accept or merely do not reject. This question is operationally meaningless, for I had to tell my readers whether I thought the distribution was sensibly normal or not, on the evidence at hand. Of course I knew I might be wrong, just as I would if I made a positive diagnosis of syphilis by the Wassermann test.) In the second case I rejected the hypothesis that these observations follow the normal curve. In these first two cases,

[2] Lest this be interpreted as a comment upon all tests of significance, I should like to note, without attempting here to adequately amplify the point, that there is an important distinction between the physical connotation of a test for, say, the significance of a difference between means or variances and a chi-square difference. We conceive a *true* difference of means, or a *true* difference of variances, which corresponds to the *true* distributions. These can be operationally defined. The tests are, so to speak, comments upon our estimates of these *true* differences. But there is nothing that corresponds to a *true* chi-square difference between the *true* distributions. The chi-square corresponds to no definable specific character of the *true* distribution. It is not a descriptive parameter like the standard deviation.

then, I agreed with the routine conclusion. In the third case I did *not* reject the hypothesis that the observations follow the normal curve, and in the fourth, I accepted the hypothesis of normality with confidence. In the last two cases, therefore, there is a difference between the decision made on the routine test and what I actually did in practice.

TABLE 1
BASAL METABOLISM OBSERVATIONS

Deviation from mean in class units* (mid-value)	Series 1			Series 2			Series 3			Series 4		
	Obs.	Th.	$O-T$	Obs.	Th.	$O-T$	Obs.	Th.	$O-T$	Obs.	Th.	$O-T$
below −4	1						1			10	9.0	+1.0
−3.5	3	25.3	+1.7	2	21.1	−8.1	1	14.8	−1.8	12	10.7	+1.3
−2.5	23			11			11			17	18.4	−1.4
−1.5	53	45.3	+7.7	48	32.9	+15.1	47	39.8	+7.2	27	26.0	+1.0
−0.5	68	69.9	−1.9	53	49.5	+3.5	67	72.4	−5.4	31	30.9	+0.1
+0.5	71	69.9	+1.1	49	49.5	−0.5	80	72.4	+7.6	30	30.9	−0.9
+1.5	36	45.3	−9.3	26	32.9	−6.9	26	39.8	−13.8	26	26.0	0
+2.5	19	19.1	−0.1	12	15.3	−3.3	17	14.8	+6.2	17	18.4	−1.4
+3.5	5	6.2	+0.8	1	5.8	+0.2	4			10	10.7	−0.7
over +4	2			5						10	9.0	+1.0
Total	281	281		207	207		254	254		190	190	
Median†	−0.11	0		−0.17	0		0	0		−0.07	0	
S.D.†	1.49			1.57			1.27			2.39		
Skewness	+0.08±0.05			+0.32±0.06			0			+0.09±0.06		
χ^2	3.5			12.4			10.1			0.7		
D.F.	5‡			4			3			7		
P	0.62			0.02			0.02			0.996		

* For experiment 1 the class unit includes 2 calories per square meter per hour; for experiments 2, 3, and 4 it includes 1 calorie per square meter per hour.

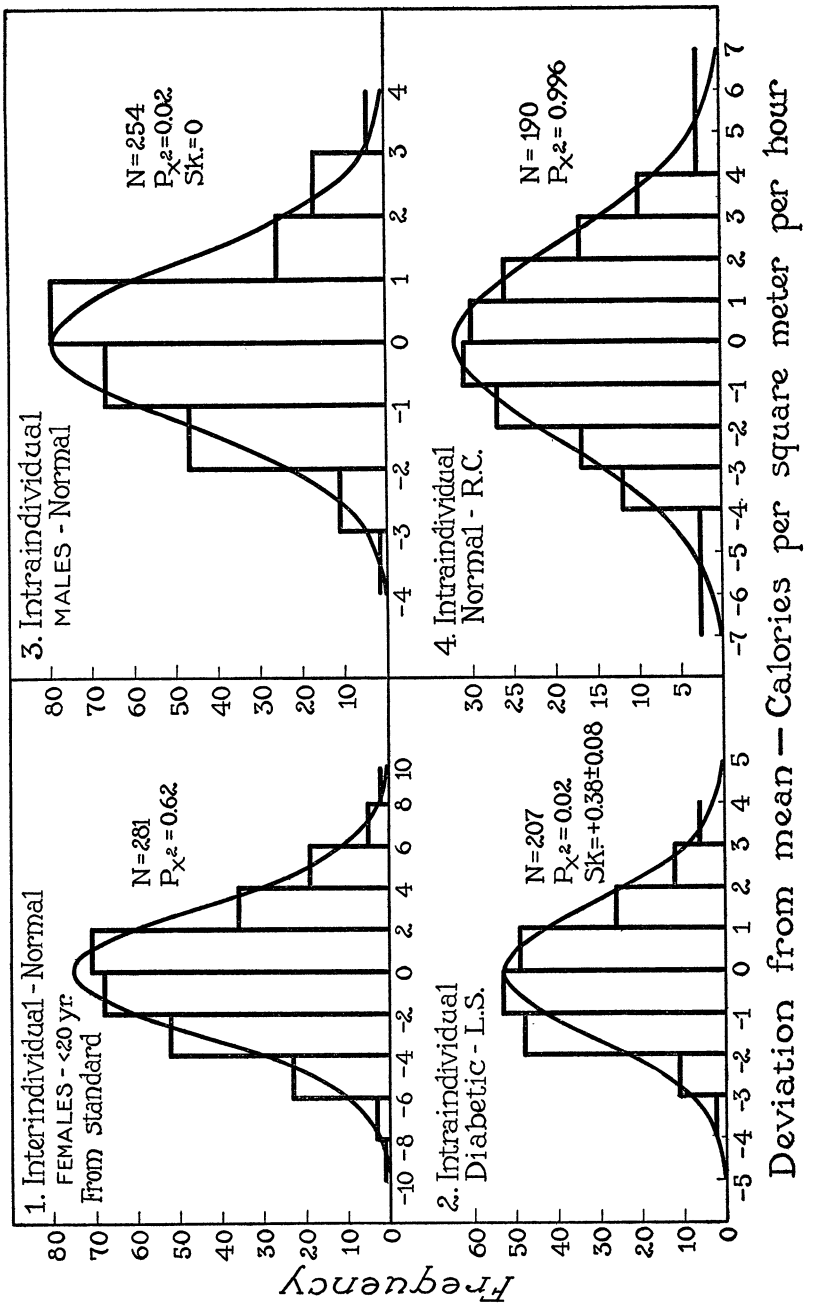
† Calculated from original data measured to 0.1 calorie, ungrouped.

‡ Deviations in this experiment are measured not from the observed but from a standard mean, and the normal curve fitted around the theoretical mean of zero, using the observed total frequency and S.D. Hence the D.F. equals classes minus 2.

If I differ with the conclusion of the test, I may inquire on what basis I made my decision and what explains my difference. For this purpose I can outline my own view as to what is the valid logical basis of the decision in any case, even the routine one. My statement of the reasoning involved will differ in certain respects, I believe, from that which would be given, say, by Fisher. Take the first two cases, in which the decisions agree. I believe the viewpoint as represented by Fisher, say, would, briefly, be something as follows:

I set up for the first case the null hypothesis that there is no difference between this distribution and a normal one. I make a test to see

CHART I


1. Interindividual - Normal
FEMALES - <20 yr.
From standard
N=281
P χ² = 0.62
2. Intraindividual
Diabetic - L.S.
N=207
P χ² = 0.02
Sk = +0.38±0.08
3. Intraindividual
MALES - Normal
N=254
P χ² = 0.02
Sk.=0
4. Intraindividual
Normal - R.C.
N=190
P χ² = 0.996
Frequency
Deviation from mean — Calories per square meter per hour

how frequently such an experience as I have at hand would appear on this null hypothesis and, using the chi-square test for this purpose, I arrive at the conclusion that an experience with as large a chi-square difference would occur six times out of ten. There is, using the arbitrary limit of $P = 0.05$, no disproof of the null hypothesis. In the second case we go through the same reasoning and reach the conclusion that, on the null hypothesis that there is no difference, so large a chi-square would occur only two times in a hundred, and the null hypothesis is therefore rejected on account of the rarity of such an experience on this hypothesis.

I believe this reasoning is fallacious and that, logically, there is no ground for rejection, whatever the size of the P, if the consideration is limited to that P alone. However, since I came to the same conclusion and did, for instance, reject, in the second case, what do I think is the valid reasoning?

I would say: I have a set of observations at hand which I think may be normally distributed. (I think so because I have seen observations of a similar character that I was satisfied followed sensibly the normal curve.) I also think they may not follow the normal curve but some regular non-normal curve. (I think this may be because I have seen bodies of data like this which do not follow the normal curve but which were, for instance, skew, etc.)[3] I then make an inquiry along the following lines: If the observations come from a normal distribution, how frequently would such a chi-square as I got occur? The conclusion is, "Quite rarely—only two times in a hundred." I then make an inquiry, not stated and not calculated, but I believe absolutely necessary for the completion of a valid argument, as follows: If the distribution is non-normal, this experience, judged by a chi-square difference, would occur quite frequently. (All I have to do is imagine that the non-normal curve has the observed skew character of the distribution.) I therefore reject the normal hypothesis on the principle that I accept that one of alternative considered hypotheses on which the experienced event would be more frequent. I say the rejection of the null hypothesis is valid only on the willingness to accept an alternative (this alternative not necessarily defined precisely in all respects).

Now the line of reasoning that I have described, as contrasted with what I have described as the more usual, would explain why my decision differs from the routine one in the third and fourth cases.

[3] The importance of the consideration of a set of alternative hypotheses in statistical reasoning has been set forth very entertainingly by Fry, 1933 ("A Mathematical Theory of Rational Inference," published in *Scripta Mathematica*, II (1934), 204–221.) The argument developed here under II may be considered an application of the general viewpoint advanced by Fry to the specific question of the interpretation of the chi-square test.

With regard to the third case, after I have tried the chi-square test, I have reached the conclusion, that on the hypothesis of no difference from normality, a distribution with so large a chi-square would occur rarely. So far we are in exactly the same position as we were at this point in the second case. But now let me examine the probability that this experience would occur if the original supply were a regular non-normal one. Would this experience occur more frequently? There is no reason to say so. The distribution is perfectly symmetrical, i.e., the skewness is zero (there were exactly 50 per cent of the cases on each side of the mean), and a cursory examination of the differences from expected frequencies in the different classes shows they are not systematic, i.e., the plus deviations and minus deviations alternate in random order. Such a distribution is not to be expected frequently from any plausible non-normal curve. We therefore have *no reason at hand for rejection of the normal curve.*

My view is that *there is never any valid reason for rejection of the null hypothesis except on the willingness to embrace an alternative one.*[4] No matter how rare an experience is under a null hypothesis, this does not warrant logically, and in practice we do not allow it, to reject the null hypothesis if, for any reasons, no alternative hypothesis is credible. The fact that statisticians talk in terms of the null hypothesis and disproving it, is due to the circumstance that the numerical calculations can usually be made only with this part of the problem. The fact that the alternative hypotheses cannot be dealt with *numerically* should not lead to the fallacious conclusion that they do not form an integral part of the necessary logical structure by which the null hypothesis is rejected.

It is easy to see by the reasoning that I have given why, in my view, there is absolutely no reason for rejecting the normal fit for the fourth case. The event at hand has been proved, by the finding of a $P=0.996$, to be of a very rare kind under the normal hypothesis, surely, and on the basis of the principle that such a rare event is itself a warrant for rejection, we should reject it here, and Fisher seems to say we should.

[4] It is not necessary, for the purposes in hand, to define the alternatives at the outset. Indeed it may be economical not to do so. As the argument is advanced here, rejection of the null hypothesis depends on the acceptance of an alternative according to which the observed event would be impressively more frequent. It is understood, of course, that the alternative need not be defined completely or precisely. We will be interested, therefore, in the end, only in such other hypotheses as agree with the observations, and we may wisely wait till the analysis is made to select for consideration only such other hypotheses as are agreeable. The procedures may therefore be outlined in logical order as follows: (1) The chi-square P is evaluated; if it is not unusually low, the null hypothesis may be accepted so far as the evidence in hand is concerned. (2) If the P is unusually low, other hypotheses, agreeing with the data (i.e., which would yield a P like the one found) are to be considered. For any such a one we will then wish to know: (a) is it alternative, i.e., different from the null hypothesis in a respect pertinent to the problem in hand?; (b) does it have *a priori* plausibility? If both these conditions are fulfilled, the null hypothesis is rejected in favor of the alternative.

But in my formulation we must now ask whether the experience would be comparatively *frequent* under a non-normal distribution. Of course, the experience would be comparatively *infrequent*. If the fit seems miraculous on the hypothesis of normality, it would be even more so on a hypothesis of non-normality. Were an alternative hypothesis, to the effect that the data had been falsified to render them normal, pertinent and tenable in the circumstances, there would be ground for rejection. But even then what would be rejected is not the hypothesis but the data.

TABLE 2

DATA FROM A STUDY TO DETERMINE THE EFFECTIVENESS OF A CERTAIN VACCINE IN THE PREVENTION OF THE COMMON COLD

	Total Group			Males			Females		
	Number	Affected		Number	Affected		Number	Affected	
		No.	Per cent		No.	Per cent		No.	Per cent
Experimental	143	121	84.6	70	57	81.4	73	64	87.7
Control	157	145	92.4	80	72	90.0	77	73	94.8
Total	300	266	88.7	150	129	86.0	150	137	91.3
Difference, Control-Experimental			+7.8			+8.6			+7.1
P; χ^2 for 1 D.F.	0.02			0.07			0.06		
P; Normal distribution Difference/S.E. difference	0.02			0.07			0.06		
P; Normal distribution Mean difference/S.E. mean difence				0.02					
P; Sum of χ^2 for 2 D.F.				0.05					

III. In Table 2 are given the results from an experiment that was performed to determine the effectiveness of a certain vaccine in the prevention of the common cold. If we consider the total group, we see that the experimental group had 7.8 per cent fewer affected than did the control group. If this difference is examined by the chi-square test, we get a P of 0.02, which is significant, we will say. (Half the P of the chi-square tables is used since only positive differences are relevant here.) The same question can, of course, be answered by examining this difference with its S.E., the latter evaluated as $\sqrt{PQ(1/n_e+1/n_c)}$ where P is the rate for the total group; these two answers are identical, as we know, for the four-fold table. Now let us look at the subdivision of the data into males and females. Using the same procedures in each

four-fold table, we get for the males a $P = 0.07$, not significant, and for the females $P = 0.06$, not significant. We may combine the results for the two sexes by using the mean of the two differences. Examining this in the light of its S.E., we again get a $P = 0.02$, the same value as before for the combined group. Looked at this way, we substantiate the previous conclusion that the significance of the difference for the experience taken as a whole is measured by a P of 0.02. Now suppose we combine the experience of the males and females by using the theorem that for independent tables we may use the sum of the chi-squares with the sum of the degrees of freedom to obtain a P. We get $P = 0.05$, a value that may appear reasonable considering the P's for the males and females separately. But the P obtained this way, 0.05, is not significant, while the value previously obtained, 0.02, is. In the problem here cited there are methods available for combining the males and females to perform a chi-square test other than that of summing the chi-squares and the degrees of freedom, and I am only citing this example to demonstrate that the P obtained in the latter way may not be a good basis for judgment. I think that where the chi-square test, using the sum of the chi-squares and the degrees of freedom, fails in this example is in not being sensitive to the similar directional character of the difference. Since the chi-square function squares the differences from expectation, it destroys the value, if there be one, of knowing the sign. One can, of course, make additional different tests depending on the expectation of the distribution of signs. But my point here is that the chi-square test routinely used is not doing this and, except where you haul out an example, as I have done, you would not know it because the general direction of the results when the P is obtained by summing of the chi-squares and the degrees of freedom is reasonable. The matter becomes of practical importance when the separate tables cannot be validly combined into a single table, and this as well as another point I shall attempt to illustrate next.

IV. In Table 3 is a resume of five series of observations on the number of blood cells counted in a hemocytometer chamber. The cells in each of 400 squares of the entire hemocytometer chamber were enumerated. Since it had been domonstrated previously by "Student" (*Biometrika*, 5 (1906–1907), 351–360) that this distribution is theoretically Poisson, and this being an important matter for me to know about, I, like "Student" with his series, compared the Poisson distribution by the chi-square test for each of the five experiments. The test was made for each series in the usual way. That is, each observed mean determined a Poisson, which was used to calculate the theoretical frequency for each number of cells. The chi-square test was performed,

TABLE 3
CHI-SQUARE TEST FOR "GOODNESS OF FIT," POISSON DISTRIBUTION

Erythrocytes	Experiment I		II		III		IV		V	
	Th.	Ob.	Th.	Ob.	Th.	Ob.	Th.	Ob.	Th.	Ob.
0	↑	0	↑	0	↑	0	↑	1	↑	2
	6.92				6.77				12.45	
1	↓	5	17.31	2	↓	6	14.33	1	↓	10
2	17.81	19	↓	11	17.50	17	↓	10	27.83	21
3	35.63	33	27.64	20	35.17	38	24.05	21	49.28	52
4	53.48	49	44.86	49	53.01	52	40.56	32	65.44	63
5	64.21	59	58.25	54	63.91	65	54.72	55	69.53	77
6	64.25	78	63.03	75	64.23	69	61.53	79	61.55	62
7	55.10	57	58.46	70	55.33	46	59.32	64	46.72	46
8	41.35	38	47.44	42	41.70	40	50.03	45	31.03	41
9	27.58	29	34.22	34	27.95	28	37.51	35	18.32	11
10	16.56	19	22.22	24	16.86	22	25.31	28	9.73	8
11	9.04	9	13.12	9	9.24	10	15.53	16	↑	5
12	↑	5	↑	7	↑	5	↑	10	↓	2
13	↓	0	↓	1	↓	1	↓	3	8.12	0
	8.07		13.45		8.33		17.11			
14	↓	0	↓	1	↓	1	↓	0	↓	0
15	↓	0	↓	1	↓	0	↓	0	↓	0
m	6.04		6.49		6.03		6.75		5.31	
χ^2	6.48		11.35		4.20		10.01		9.33	
D.F.	10		9		10		9		9	
P	0.77		0.25		0.94		0.35		0.41	
s	2.46	2.33	2.55	2.33	2.46	2.45	2.60	2.39	2.30	2.18
ns^2/m	357.53		332.51		396.29		338.30		357.80	
P^*	0.14		0.01		0.94		0.03		0.14	

Total $\chi^2 = 41.37$; $P = 0.71$ for 47 degrees of freedom.
Total $ns^2/m = 1782.43$; $P = 0.0006$.

* The P here is the probability of getting in random samples from a Poisson distribution so large a discrepancy between s^2 and m as that observed. The appropriate P is therefore the one corresponding to a difference from the mean chi-square as large in either direction as that observed, i.e., the P for a discrepancy in one direction is doubled. The observed chi-square is given by ns^2/m where n is the number of degrees of freedom, which for each experiment is 399, and for the total chi-square is 1995.

comparing theoretical and observed frequencies and using as degrees of freedom 2 less than the number of classes, since the theoretical distribution and the observed data agree with respect to the mean and total number. Table 3 gives the results. It is seen that in no single instance would the Poisson be rejected by the routine chi-square test, and considering the entire series together, by adding the chi-squares and the degrees of freedom we get a P of 0.71, no value for rejection. I, then, as a practitioner would say, as "Student" did say with a similar experience, that the cells followed the Poisson distribution. Having, therefore, decided to accept on the basis of the chi-square test that the Poisson distribution applies, I now go forward with my

experiment on the assumption that I can say that the standard deviation is equal to the square root of the mean, which is true for the Poisson distribution. In fact, it was in order to be able to calculate the variability from the mean that I made the test of the Poisson in the first place. It was natural for me, then, to set down for each experiment a comparison of the variance and the mean. When this was done, it was found that for each experiment the observed value of the variance was *less* than the expected value for the Poisson! A glance at the table should convince one that the S.D. is really less than $\sqrt{m}$, for in each experiment it is less, and it is very improbable that five random discrepancies would be simultaneously in the same direction by chance. But if one wishes to make statistical tests for this, they can be made in a number of ways. One can test the ratio between the mean of the differences and its standard error in the classic way, or by the *t*-test, and the mean difference is found to be exceedingly significant. Whatever way the test is made, in fact, the difference from the hypothesis that the variance is equal to the mean is in the order of 4 sigmas. Interestingly enough, one of the ways of testing whether the variance is equal to the mean is to use the chi-square function (not the chi-square test as it has been discussed up to now). For the Poisson distribution ns^2/m is equal to chi-square for n degrees of freedom, where n is the number of degrees of freedom used in calculating s. Table 3 shows the P value obtained by using this test for each of the experiments; and for all the experiments taken together it is 0.0006. We have now the surprising result that, considering the experiment as a whole, the chi-square test for goodness of fit of the Poisson shows no reason for rejection, whereas any test for the principal characteristic of the Poisson—namely, that the standard deviation is equal to the square root of the mean—shows indubitably that this is not true.[5]

I should attribute this discrepancy in conclusions, according as to whether they are drawn from the application of the chi-square test for goodness of fit or from a direct test for the agreement of the variance and mean, to two defects of the chi-square test considered as a test to be applied to a situation such as described here. The first is the nonspecific character of the chi-square test. The test is frequently referred to as a test for the "goodness of fit," but it is such a test only in the tautologic sense that it tests whether chi-square fits. A test can be applied only as respects a certain measurement. We recognize regularly that finding a significant difference between the mean of an

[5] "Student's" observations using yeast cells do not agree with mine in regard to the relation of the standard deviation to the mean, and, therefore, with "Student's" observations I should have drawn the same conclusions that he did. What I mean is that he did not find it necessary to supplement the chi-square test.

experiment and a hypothetical curve does not warrant rejection of the curve in other respects, say, the standard deviation. These two statistics are functions, and it is because these functions are related to certain specific physical characteristics that a significant difference in respect of them has great meaning. So, too, a significant difference tested by another function reflects a significant difference in kurtosis or skewness, etc. Every function is only a variable mathematically, and an independent investigation is required to reveal what this variable represents physically. If I say that a significant difference found by testing the probability of an experienced value of a certain variable divulges a difference only as respects the character represented by that variable, I may ask, "*What characteristic does the chi-square variable represent?*" I don't think there *is* any specific characteristic, and I believe that is one of the chief deficiencies of the chi-square test so far as its value for practice is concerned.

The second defect has to do with why, though we may be sure there is a small regular difference between the distribution of cells in the hemocytometer chamber and the Poisson, even the combined experience embracing a frequency of 2,000 hemocytometer divisions and about 13,000 erythrocytes did not divulge this as a significant difference when the various experiences were combined by summing the chi-squares and the degrees of freedom. This, I think, is another exemplification of the point I was trying to make under III. The differences were all in one direction, but the chi-square test for different independent samples combined, effected by adding the chi-squares and the degrees of freedom, was insensitive to this fact.

There is no room here for further elaboration of these views, or even for an adequate summary of the points already made. I may, however, record my impression that in practice the chi-square test is being relied on too much and too uncritically. As an exploratory tool for preliminary survey it may have some usefulness. But for any more searching analysis—as, for instance, if one wishes to base some theoretical development on the frequency function—one should seek first to ascertain functions that refer specifically to the questions at hand, and apply statistical tests that are sensitive to variations in those specific functions.