THE CLASSICAL PROBLEM OF INFERENCE— GOODNESS OF FIT

OSCAR KEMPTHORNE
IOWA STATE UNIVERSITY

1. General introduction

As a preface to my lecture, I find it necessary to discuss in general terms the status of the statistical art and what we should mean by the term "inference." It seems to me that over the whole history of human thought there have been two basic underlying ideas of inference:

(a) what may best be expressed, perhaps, by the colloquialism "making sense of data";
(b) the choice of an action in a prespecified class of possible actions on the basis of data, costs, risks, and opinions.

Of course, to attempt to characterize the whole of statistics in some such way as the preceding is rather like attempting to characterize mathematics by a few brief common sense statements, and this is obviously foredoomed to failure. But the attempt has been made by others, who with a zeal approaching that of religious fanatics attempt to convince the world that there is one true religion, the one they are preaching. We should feel a considerable debt to van Dantzig [10], [11] for calling attention to the phenomenon of "Statistical Priesthood" with which our profession is now plagued. He gave us just two examples and pointed out the moral. It is curious that even in its activities unrelated to ethics, humanity searches for a religion. At the present time, the religion being "pushed" the hardest is Bayesianism. A few years ago it was decision theory. The actions of the proponents are like those of the religious evangelist. It is characteristic of new religions that they are intolerant of the old ones. It seems obvious that the only religion we should uphold is that there is no true religion. I find myself quite intolerant of the several cults.

My own preference is to say that the bulk of the activities of statisticians is encompassed by one or other of the two basic ideas expressed above, and I like to denote them by (a) Statistical Inference and (b) the Theory of Decision-making. It is remarkable that over the years we have had many papers and even books which take the view that statistical inference is a part of decision-making.

Some aspects of the reported work were supported by NSF Grant G-14237. Journal Paper No. J-5461 of the Iowa Agricultural and Home Economics Station, Ames, Iowa. Project 890.

Such people presumably take the view that the writing down of conclusions can be regarded as an act in a specified class of actions, to which costs and risks can be applied. The problem here is partially one of semantics: what does one mean by an action, for instance? My main criticism of the point of view which takes (a) to be a part of (b) is that it has in my opinion proved to be rather sterile with regard to the general problem of making sense of data, the problem of condensation of data, and so on. Before 1940 we had in the total history of mankind perhaps 100 man years of intellectual effort in the direction of statistics; in the Forties perhaps another 100; in the Fifties perhaps 200, and in the Sixties so far perhaps 400. But the advances in the art of making sense of data have not been at all commensurate. There have, however, been great advances in the theory of decision-making. Our knowledge of possible rules for terminal decisions has expanded very rapidly. It seems, however, to have escaped attention that there is a vast difference between what one is entitled to think on the basis of the data and what action one should take on the basis of the data.

What is the main problem of data interpretation? I believe it to be the development of a condensation of the data which in some imprecise way does not throw away any of the information in the data. It may be said that I am using vague expressions which cannot be given precise meaning and I would agree. But I would then say that the history of Science is full of examples of the dangers of narrowly prescribed specifications and frameworks of thought.

It is a consequence of the above view that one of the basic problems of data interpretation is the problem of model specification. I am highly amused by a statement of Sir Ronald Fisher ([4], p. 314): "As regards problems of specification, these are entirely a matter for the practical statistician." But this curious remark is balanced by a later one on the same page: "The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration." It is probably fairly generally agreed that the beginning of what we call statistical inference in the sense (a) was the development of the χ^2 goodness-of-fit test. This was really a most remarkable piece of work and stands out in my opinion as one of the great ideas of human thought. It is true that Pearson made some errors with regard to the concept of degrees of freedom, and it was necessary for Fisher to clear them up. But this should not detract from the magnitude of Pearson's step.

To emphasize the problem of model specification let me give a common example. An experimenter has compared 6 treatments in a randomized block design with regard to their effect on the growth of mice. The statistician says, "Ah, yes! Randomized Blocks, so *the* model is $y_{ij} = \mu + b_i + t_j + e_{ij}$ with the errors normally and independently distributed around zero with common variance σ^2." I have difficulty imagining a more blatant travesty of common sense, let alone scientific method. It is really appalling. But this is taught extensively. Books say, "the appropriate model is. . . ." Fortunately, in recent years there has been some deeply considered attack on the problem, particularly by Anscombe and Tukey.

My interest in the problem of goodness of fit was stimulated by consideration of the problem of transformations, primarily in connection with the analysis of comparative experiments. I was then led to the simplest situation, the case of a single sample. The problem of whether observations should be analyzed on an arithmetic scale or a logarithmic scale is surely one of the most elementary problems, but surprisingly little has been done on it. The application of some goodness-of-fit procedures seemed appropriate, and the χ^2 goodness-of-fit test seemed a reasonable candidate.

I hope that a little of what I have to say is new. But if not, a representation of old material may still be of value. A paper by Slakter [9] has numerical results closely related to some of the results I shall present.

2. A brief historical review

The literature on goodness-of-fit tests is vast. Shapiro and Wilk [8] for instance, give a list of about 70 papers on the subject. I found the review by David [3] very helpful. Even the literature on the χ^2 goodness-of-fit test is very extensive. The procedure is, of course, to divide the distribution into mutually exclusive cells and to form the criterion

$$\sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i} \tag{2.1}$$

where O_i is the observed frequency, and E_i the expected frequency in the i-th cell. In the case of a discrete data, the discrete classification of the distribution provides a partition into classes for application of the χ^2 procedure. It is customary, however, to invoke some rule such as that the expectations of cells should be greater than 5, or in some cases greater than 2. For example, Cramér [1] says:

"When the χ^2 test is applied in practice, and all the expected frequencies np_i are ≥ 10, the limiting χ^2-distribution tabulated in table III gives as a rule the value χ_p^2 corresponding to a given $P = p/100$ with an approximation sufficient for ordinary purposes. If some of the np_i are <10, it is usually advisable to pool the smaller groups, so that every group contains at least 10 expected observations, before the test is applied. When the observations are so few that this cannot be done, the χ^2 tables should not be used, but some information may still be drawn from the values of $E(\chi^2)$ and $D(\chi^2)$ calculated according to (30.1.1)."

In fact, this matter is quite obscure and there is considerable arbitrariness in the application of the procedure to discrete data, as anyone who has looked at data discovers. One can vary the "answer" that is obtained by the choice of grouping the possible classes. Individual judgment is always called into play. In the continuous distribution case, which should always be characterized by some phrase such as the "so-called" continuous case because we can never observe a continuous random variable except with a grouping error, the situation is much worse with regard to applying the χ^2 goodness-of-fit test. "How

is one to make up the cells?", "Where are they centered?", "How many cells should one use?" are questions on which the personal arbitrariness of the tester seems to enter. Again, anyone who has used the procedure has met these questions, and answered them in his own somewhat arbitrary way. Even if one has the rule that every cell should have an expectation greater than or equal to 5, there is the matter of placement of the classes. Also, the imposition of such rules has apparently led to the view that the χ^2 goodness-of-fit test is not consistent. One attack on this question was by Mann and Wald [6] who were led to the rule that asymptotic maximum power in a certain sense of the χ^2 test was achieved when the number of cells is proportional to $2(N-1)^{1/5}$, where N is the number of observations.

The purpose of the present paper is to describe and evaluate partially a completely objective simple rule, namely divide the distribution, fitted on the basis of N observations, into N equal parts each with probability $1/N$. This gives N cells each with expectation equal to unity. Count the number x_i in each cell. Then the χ^2 criterion

$$K = \sum_i \frac{(x_i - e_i)^2}{e_i} \tag{2.2}$$

becomes

$$\begin{aligned} K &= \sum_i (x_i - 1)^2 = \sum_i x_i^2 - 2\sum x_i + N \\ &= \sum_i x_i^2 - N \\ &= \sum_i x_i(x_i - 1). \end{aligned} \tag{2.3}$$

The evaluation of the criterion is then made by reference to the χ^2 distribution with degrees of freedom equal to $(N - 1 - p)$, where p is the number of parameters fitted. In evaluating the criterion, however, it is to be noted that K can take only even integral values, so that one obtains the probability of exceeding $(K - 1)$ for the mathematical χ^2 distribution. It is to be noted that in the case of continuous data this rule is objective, and there is no room for personal choice on number and location of cells. It is assumed that the resulting classes are still wide relative to the grouping interval of observations, though clearly this will not be true with very large samples. In very large samples the grouping error of observations would have to be considered.

The present paper gives some preliminary results on the above procedure. It contains a discussion of the distribution of the Pearsonian criterion with k equally likely classes in the case when no parameters are estimated, this discussion being relevant and appropriate to the case of any continuous distribution. Then some Monte Carlo results are given on the distribution of K for the case of the normal distribution, in which the mean and variance are estimated. Finally a few power comparisons are made, and some discussion on the relevance of power is presented.

3. The distribution of K_k with a completely specified distribution

We consider the partition of a continuous distribution into equiprobable classes each with probability $1/k$, and denote the χ^2 criterion by K_k. In the case of N observations with k equal to N, the criterion is the K given above. If we denote the observed numbers in a total sample of N, which are in the classes, $i = 1, 2, \cdots, k$, by x_i, the probability of $x_1, x_2, \cdots, x_k$ is

$$\frac{N!}{x_i!}\left(\frac{1}{k}\right)^N. \tag{3.1}$$

The χ^2 quantity is equal to

$$K_k = \sum_i \frac{(x_i - (N/k)^2}{(N/k)} \tag{3.2}$$

so that

$$\begin{aligned} \frac{N}{k} K_k &= \sum_i \left(x_i - \frac{N}{k}\right)^2 \\ &= \sum x_i^2 - 2\frac{N}{k}\sum x_i + \frac{N^2}{k} \\ &= \sum x_i^2 - \frac{N^2}{k}, \end{aligned} \tag{3.3}$$

or

$$\frac{N}{k}(K_k + N) = \sum x_i^2 = S \text{ (say)}. \tag{3.4}$$

Hence, to get the moments of K_k we first obtain the moments of S. The first two moments of K_k are well known, but we include them for completeness.

First moment of K_k. One has

$$S = \sum x_i^2 = \sum x_i(x_i - 1) + \sum x_i \tag{3.5}$$

and

$$Ex_i(x_i - 1) = \frac{N_2}{k^2} \tag{3.6}$$

where $N_s = N(N-1)(N-2)\cdots(N-s+1)$. Hence $E(S) = N((N-1)/k) + N$ and

$$\begin{aligned} E(K_k) &= (N-1) + k - N \\ &= (k-1). \end{aligned} \tag{3.7}$$

Variance of K_k. The term S^2 is equal to

$$S^2 = \sum_i x_i^4 + \sum_{i,i'}{}^{\neq} x_i^2 x_{i'}^2, \tag{3.8}$$

but $x^4 = x_4 + 6x_3 + 7x_2 + x_1$ where, as before,

$$x_s = x(x-1)(x-2)\cdots(x-s+1). \tag{3.9}$$

Such relationships are verified easily by writing, for example,

$$x^6 = x_6 + \alpha_5 x_5 + \alpha_4 x_4 + \alpha_3 x_3 + \alpha_2 x_2 + \alpha_1 x_1 + \alpha_0 \tag{3.10}$$

and successively placing x equal to 0, 1, 2, 3, 4, and 5. This device is helpful throughout. The reason for using this mode of expression is that $E(x_s) = (N_s/k^s)$, and $Ex_{i(s)}x_{i'(s')} = (N_{s+s'}/k^{s+s'})$, and so on. Hence,

$$E(S^2) = k\left\{\frac{N_4}{k^4} + 6\frac{N_3}{k^3} + 7\frac{N_2}{k^2} + \frac{N_1}{k}\right\} + k(k-1)\left\{\frac{N_4}{k^4} + 2\frac{N_3}{k^3} + \frac{N_2}{k^2}\right\} \tag{3.11}$$

so that $V(S) = N(N-1)2((k-1)/k^2)$ and

$$V(K_k) = 2(k-1)(N-1)/N. \tag{3.12}$$

Third moment of K_k. One can write

$$S^3 = \sum_i x_i^6 + 3\sum_{i,i'}^{\neq} x_i^4 x_{i'}^2 + \sum_{i,i',i''}^{\neq} x_i^2 x_{i'}^2 x_{i''}^2, \tag{3.13}$$

but

$$x^6 = x_6 + 15x_5 + 65x_4 + 90x_3 + 31x_2 + x_1, \tag{3.14}$$

$$x^4 = x_4 + 6x_3 + 7x_2 + x_1, \tag{3.15}$$

and $x^2 = x_2 + x_1$, so that

(3.16)

$$E(S^3) = k\left\{\frac{N_6}{k^6} + 15\frac{N_5}{k^5} + 65\frac{N_4}{k^4} + 90\frac{N_3}{k^3} + 31\frac{N_2}{k^2} + \frac{N_1}{k}\right\} + 3k(k-1)\left\{\frac{N_6}{k^6} + 6\frac{N_5}{k^5} + 7\frac{N_4}{k^4} + \frac{N_3}{k^3} + \frac{N_5}{k^5} + 6\frac{N_4}{k^4} + 7\frac{N_3}{k^3} + \frac{N_2}{k^2}\right\} + k(k-1)(k-2)\left\{\frac{N_6}{k^6} + 3\frac{N_5}{k^5} + 3\frac{N_4}{k^4} + \frac{N_3}{k^3}\right\},$$

and the third moment of S is $\mu_3(S) = E(S^3) - 3V(S)E(S) - E^3(S)$ so that, after some tedious algebra,

$$\mu_3(S) = 8N_3\frac{(k-1)}{k^3} + 4N_2\frac{(k-1)(k-2)}{k^3} = \frac{8N(N-1)(N-2)(k-1)}{k^3} + \frac{4N(N-1)(k-1)(k-2)}{k^3}. \tag{3.17}$$

Hence,

$$\mu_3(K_k) = 8(k-1)\frac{(N-1)(N-2)}{N^2} + \frac{4(k-1)(k-2)(N-1)}{N^2}. \tag{3.18}$$

Fourth moment of K_k. Obtaining the fourth moment was very tedious. Using relationships such as

$$x^8 = x_8 + 28x_7 + 266x_6 + 1050x_5 + 1701x_4 + 966x_3 + 127x_2 + x_1 \tag{3.19}$$

and expressing polynomials in N as linear functions of N_s, I obtained

$$\mu_4(S) = 12(k-1)(k+3)\frac{N_4}{k^4} + \left[144k - 384 + \frac{240}{k}\right]\frac{N_3}{k^3} + \left[8k - 32 + \frac{48}{k} - \frac{24}{k^2}\right]\frac{N_2}{k^2}. \tag{3.20}$$

Hence,

$$\mu_4(K_k) = 12(k-1)(k+3)\frac{N_4}{N^4} + \left[144k - 384 + \frac{240}{k}\right]k\frac{N_3}{N^4} + \left[8k - 32 + \frac{48}{k} - \frac{24}{k^2}\right]k^2\frac{N_2}{N^4}. \tag{3.21}$$

If N is large and k fixed, the moments are very close to the moments of the theoretical χ^2 distribution with $(k-1)$ degrees of freedom, which is part of the basis for the use of the theoretical χ^2 table. The inconsistency of the k-group χ^2 test arises because the test detects only deviations from the multinomial obtained by grouping the continuous distribution which is being examined. This inconsistency can, however, be removed by letting k increase with N, and the purpose of the tedious calculation of the moments of K_k was to examine this matter. The obvious candidate mentioned above is to let k equal N. For this case the second moment about the mean is $2((N-1)^2/N)$ which is equal to $2[N - 2 + (1/N)]$, the third moment is $12((N-1)^2(N-2)/N^2)$, which is $12[N - 4 + (5/N) - (2/N^2)]$, and the fourth moment is equal to $12N^2 + 120N + 0(1)$.

The ratios of these moments to the moments of χ^2 with $(N-1)$ degrees of freedom are

Moment	*Ratio*
First	1
Second	$1 - \frac{1}{N}$
Third	$\frac{3}{2}\left(1 - \frac{1}{N}\right)$
Fourth	$\left(1 + \frac{8}{N}\right)$.

Obviously with k equal to N, the first, second, and fourth moments go to those of the χ^2 distribution quite rapidly. The ratio of third moments, however, tends to $\frac{3}{2}$. With both N and k large, the third moment is essentially

$$8(k-1) + 4\frac{(k-1)(k-2)}{N}, \tag{3.22}$$

so the ratio to the theoretical moment is $1 + \frac{1}{2}((k-2)/N)$, or if k is equal to rN and both large, the ratio is $1 + (r/2)$. It might appear, therefore, that unless r is small, the distribution would not tend to the theoretical one. However, another aspect "saves the day," namely that as the degrees of freedom become

large, the χ^2 distribution tends to the normal distribution. The skewness of the distribution of K_k tends to

$$\frac{\sqrt{2}(2 + r)}{\sqrt{k}} \tag{3.23}$$

which with increasing k, and fixed r tends to zero. It therefore appears that if the number of equiprobable classes k used in the goodness-of-fit test is equal to rN, a reasonable approximation for r sizeable relative to unity is to suppose that K_k is normally distributed with a mean of $(k - 1)$ and a variance $2(k - 1)[1 - (1/N)]$.

The results given above seem to tell us that for testing the goodness of fit to continuous distributions, the variety of rules in the literature, stating that cell expectations should be greater than 10, or greater than 5, or greater than 2, seem to be quite irrelevant. In fact, the distribution of K_k is not disturbed appreciably, apparently, if the number of classes is of the order of the number of observations. If k is equal to N, the distribution of K_k is asymptotically normal. It is clear that if k is of greater order, peculiar results obtain. If, for instance, k equals N^2, then the third moment is approximately $4N^3 + 8N^2$, and the skewness would be

$$\frac{(4N^3 + 8N^2)}{2N^2\sqrt{2N}} = \sqrt{2} \tag{3.24}$$

which does not go to zero with increasing N. Similarly, the fourth moment would be approximately $12N^2 + 144N^3 + 8N^4$ with kurtosis of approximately 2 for indefinitely large N.

The whole question of choice of the number of classes for the goodness-of-fit test seems therefore to be still quite an open one. My initial view was that having the number of classes equal to N, the number of observations would be a good choice. Of course this would be modified as soon as the inevitable grouping error of observations from a continuous distribution is met. It seems clear, however, that the larger the number of cells, the greater is the sensitivity of the test to deviations in the tails. It is extremely unlikely that any particular choice can be shown to be best for all circumstances.

The following sections give a few empirical results on the case k equal to N, for a null composite hypothesis of normality with data arising from a normal distribution and from two distributions for which a generating program could be written very quickly. Obviously much more computation needs to be done as well as some theoretical work on the whole matter.

4. Monte Carlo results on the distribution of K

The mathematical results above hold for the case of a completely specified distribution, which we may note has no conditions on its dimensionality. In the case of a multivariate distribution, one merely splits up the distribution into N equiprobable regions, and an intuitively reasonable way of doing this

is to base the regions on the equiprobability contours. The grouping error of observations will, however, cause problems.

I have not yet obtained any mathematical results on the effect on the distribution of K when parameters of the distribution are estimated. I imagine that exactly the same type of result as was obtained by Fisher will hold, but the mathematics are not so easy because in the present case the cell expectations remain constant. In the cases considered by Pearson and Fisher, the cells were fixed, and asymptotically the cell frequencies increase and have a multivariate normal distribution, whose exponent is distributed as χ^2.

In envisaging the test described, I felt that it was primarily to be regarded as a test of distribution shape and not of location or scale, though obviously it has power asymptotically with regard to *any* alternatives. I therefore had computations performed for the case when parameters are estimated. Also my initial interest was in tests of normality, so I had computations made on samples from a normal distribution. Obviously, however, the test can be applied to *any* continuous distribution, and there is no reason to surmise that the distribution of K depends on the nature of the true continuous distribution from which the samples originate. Of course, it will be necessary that the method of "estimation" of parameters must be in some sense efficient or else the value of χ^2 will be too large (Fisher [5]).

Monte Carlo computations of the distribution of K were done by drawing sample (y) of size 10, 20, and 50 from normal distributions $N(\mu, \sigma^2)$ estimating the mean and variance of the normal distribution by

$$\begin{aligned} \hat{\mu} &= \text{ave } y, \\ \hat{\sigma}^2 &= \frac{1}{N-1} \sum (y - \text{ave } y)^2, \end{aligned} \tag{4.1}$$

and then comparing the actual sample with the fitted distribution.

Case I. Samples of size 10. As stated above, the possible values of K are even integers, so the obvious continuity correction was made. In view of the projected use of the test as a tail area test, the most appropriate comparison is to compare tail frequencies of the empirical distribution with those of the χ^2 distribution with 7 degrees of freedom. Of course, a test of goodness of fit would use the cell frequencies and not the tail frequencies. The reduction from 9 to 7 degrees of freedom was made because two parameters are estimated. Theoretical frequencies were taken from table 7 of Pearson and Hartley [7]. I have not bothered to make a goodness-of-fit test of the Monte Carlo results because the agreement and lack of agreement is quite obvious. In the case of samples of size 10, the frequency of the class "16 or over" observed was .0258, whereas the theoretical value is .0360; the expected number is therefore 180 and the observed number was 129, which is clearly discrepant. However, from the viewpoint of use, the reporting of a significance level as .0360, when it is close to .0258, cannot be regarded as a serious defect.

It is worth noting that the mean value of K was found to be 7.43, and the

variance of K was 12.3. The above theory and the usual rule of subtracting unity for each parameter estimated suggests that the mean would be 7.0 and the variance would be 12, so the agreement is really rather good. Actually the mean is significantly greater than 7, but the tail areas do not seem to have been disturbed seriously by the change in mean and variance from the theoretical values for the χ^2 distribution with 7 degrees of freedom.

TABLE I

MONTE CARLO DISTRIBUTION OF K FOR 5000 SAMPLES OF SIZE 10 FROM A NORMAL POPULATION

	Proportion $\geq K$	
Value of K	Observed	Expected
0	1.000	
2	.995	
4	.941	
6	.732	
8	.499	
10	.284	.253
12	.133	.139
14	.080	.072
16	.0258	.0360
18	.0122	.0174
20	.0078	.0082
22	.0040	.0038
24+	.0004	.0017

Case II. Samples of size 20. One thousand samples of size 20 from a normal distribution were generated, and the distribution of K was estimated. The comparison is given in table II. Actually, the agreement of tail areas for the empirical distribution and the χ^2 distribution with 17 degrees of freedom seems quite remarkable. The mean of the distribution of K was estimated to be 17.4, as opposed to the theoretical value 17, and the variance as 30.90, which is to be compared with an expected value from the theory presented above of 32 and the value 34 for the theoretical χ^2 distribution for 17 degrees of freedom. Apparently, the discrepancy in mean and variance do not affect the tail areas appreciably.

Case III. Samples of size 50. Five hundred samples of size 50 were used, and the comparison with the χ^2 distribution for 47 degrees of freedom is given in table III. Because Pearson and Hartley [7] give tail areas for 46 and for 48 degrees of freedom and for even valued abscissa, simple linear interpolation was used. Actually, better interpolation could be done, but it was not deemed necessary with the sample size considered. The agreement is really quite remarkable. The discrepancy in mean was 47.37 compared to a theoretical value of 47, and in variance 87.6 compared to 94 (or approximately 92, suggested by the theory given above). It is again quite curious that discrepancy in mean

TABLE II

MONTE CARLO DISTRIBUTION OF K FOR 1000 SAMPLES OF SIZE 20 FROM A NORMAL POPULATION

	Proportion $\geq K$	
Value of K	Observed	Expected
6	1.000	
8	.996	
10	.972	
12	.905	
14	.788	
16	.627	
18	.460	
20	.331	.329
22	.217	.226
24	.153	.149
26	.092	.095
28	.061	.058
30	.039	.034
32	.025	.020
34	.014	.011
36	.008	.006
38	.005	.003
40	.001	.001

and variance do not affect the upper tail areas appreciably. There may be a small discrepancy with the extreme upper tail, and a larger sample could be examined to check on this point. The moral is, however, quite obvious, namely

TABLE III

MONTE CARLO DISTRIBUTION OF K FOR 500 SAMPLES OF SIZE 50 FROM A NORMAL POPULATION

Proportion $\geq K$			Proportion $\geq K$		
Value of K	Observed	Expected	Value of K	Observed	Expected
26	1.000		54	.232	.26
28	.998		56	.188	.20
30	.994		58	.130	.15
32	.982		60	.102	.11
34	.962		62	.084	.08
36	.936		64	.066	.06
38	.884		66	.048	.04
40	.818		68	.034	.03
42	.742		70	.026	.02
44	.674		72	.018	.014
46	.574	.55	74	.018	.009
48	.452	.45	76	.012	.006
50	.380	.39	78	.010	.004
52	.310	.32	80	.010	.003

that the use of the theoretical χ^2 distribution is quite unlikely to be even slightly misleading.

5. Power of the K test

As stated above, it is obvious that the K test has some power with regard to any alternative. The test is obviously consistent. A detailed examination of power has not yet been made. Obviously there is the possibility of theoretical development. Shapiro and Wilk [8] report their W-test for normality and some comparison of power they made with the χ^2-test (it is not clear how this was defined), $\sqrt{b_1}$, b_2, Kolmogorov-Smirnov, Cramér-Von Mises, and a weighted Cramér-Von Mises test. Their table 10 suggests that the W_n-test is greatly superior to the others. A comparison of the K tests with all these is planned but has not been done because of lack of funds. Two cases are particularly easy to program; when the parent distribution is the triangular distribution and when it is χ^2 with 1 degree of freedom. In most cases in Shapiro and Wilk [8], the b_2 test was fairly good relative to most of the others, except the W_n-test. In order to get evidence cheaply on power of the K test, I have therefore applied the K test to 1000 samples from each of the two distributions named above.

(a) *Triangular distribution.* Five hundred samples were drawn from a triangular distribution and tested for normality. We take size of test from table II and obtain the results in table IV.

TABLE IV

SENSITIVITY OF K TEST FOR NORMALITY WITH SAMPLES FROM A TRIANGULAR DISTRIBUTION

Value of K	Size of test	Estimated power
24	.153	.956
26	.092	.900
28	.061	.814
30	.039	.712
34	.014	.458
38	.005	.290
40	.001	.208

(b) *Chi-square distribution with one d.f.* Results were obtained similarly for testing of normality, when the data originate from the χ^2 distribution with 1 degree of freedom, and are given in table V. The preliminary results show, by using the results given by Shapiro and Wilk [8] that the K test merits further examination and consideration.

6. Concluding remarks

I now return to the matters discussed at the beginning of my lecture.

In a repetitive situation, like acceptance sampling, in which decisions are

TABLE V

SENSITIVITY OF K TEST OF NORMALITY WITH SAMPLES FROM χ^2 (ONE D.F.) DISTRIBUTION

Value of K	Size of test	Estimated power
24	.153	.946
26	.092	.922
28	.061	.866
30	.039	.818
34	.014	.726
38	.005	.616
40	.001	.572

terminal, such as accepting or rejecting the lot, it is obvious that one has to map the sample space onto the decision space and that one has to consider the properties in repetitions of the mapping rule. Clearly, if a most powerful decision rule exists, it should be used. Also the costs of observation and risks of erroneous decisions have to be included in the formulation. Also it seems quite obvious that the decision-maker's personal opinions about the class of repetitive situations he will meet are relevant and should be included in the whole formulation. Even in a so-called repetitive situation goodness of fit of model is relevant. It is not sufficient merely to assume that a particular statistical model fits the situation.

It seems clear to me, however, that the accumulation of knowledge is not a repetitive process but rather an evolutionary one. No new situation is exactly like a previous situation except with regard to some parameter values. An essential part of the application of any model to data is the application of goodness-of-fit evaluations. It might be hoped that there would be one way of evaluating goodness of fit which is superior to all others, but obviously such a hope is foredoomed to failure. The literature on goodness of fit suggests that particular goodness-of-fit procedures are in some sense best with regard to the lack of fit they are designed to detect. It appears, however, that optimality in one direction is always accomplished at the expense of optimality in other directions. We are studying numerically the joint behavior of several goodness-of-fit tests, but I do not have any results yet. Such results seem to be essential for an overall intelligent approach to the problem.

The goodness-of-fit tests so far proposed seem to fall into essentially four main categories:

(a) those based on occupancy of cells determined by the hypothesized or fitted distribution, of which the Pearson chi-square test is the classic case;
(b) those based on the comparison of the cumulative sample frequency and the cumulative population frequence, like the Kolmogorov-Smirnov test, or the tests based on the differences of population cumulative frequencies between successive sample points;

(c) the comparison of the ordered statistic with the expected value of the ordered statistic, as is done in one way by Shapiro and Wilk [8];
(d) comparison of functions of moment statistics with theoretical values.

The comparison of tests is not easy. Asymptotic theory has been developed for some cases, and hopefully it gives a reliable indication of the sensitivity of the tests in various "directions." I have, however, an uneasy feeling that much asymptotic theory is based on the premise that observations of unlimited accuracy are possible. Of course, this is not the case, and with largish samples and a "reasonable" grouping error, ties will occur. This seems to affect all the tests but not with the same force. Many test criteria "blow up" when there are ties. From one point of view this is not unreasonable. If observations have unlimited accuracy, the hypothesis that they arise from a continuous distribution is untenable as soon as ties occur. The occurrence of "ties" appears to be an unavoidable embarrassment to the person who develops theory for so-called continuous observations. I find the discussion in the literature on this quite unsatisfying. The "answer" one obtains by some tests depends critically on how "ties" are broken. To suggest using some extraneous device to break ties, amounts to basing one's opinions with regard to a situation on an independent source of noise and this seems totally repugnant. This is quite unrelated to the use of a coin-toss to make up one's mind when one is unable to do so in any other way.

Of the classes of tests outlined above, many of the tests of classes (b) and (c) encounter the problems arising from grouping of observations in a violent form. The tests of class (d) are relatively not bothered by this, because we have strong intuitions that a moment-like function of a grouped sample is very close in behavior to that of an ungrouped sample. The tests based on occupancy of cells will encounter difficulties from grouping error, but it would appear that these difficulties are mild, and that reasonable smoothing devices will not disturb the distribution of the test criterion very much.

I am inclined to the view that with small samples, when ties will be very infrequent with a reasonable grouping error, tests based on order statistics will prove to be reasonably sensitive in many diverse directions. With intermediate and large samples, I am inclined to think that occupancy type tests give the data analyzer generally satisfactory answers. The question is not "Do the data come from such and such a distribution?", because one can be sure they never do, but "Is such and such a distribution a reasonable model for the description of the data?". It may well prove to be the case that a good overall procedure will combine an occupancy test and some sort of extreme value test. [See David [2] in connection with such a possibility.]

I wish to make one final point on inference. When faced with a sample the statistician makes a goodness-of-fit test for normality, and then constructs some limits on the parameters. The probability stated to be associated with these limits is always stated to be that conditional on normality, say. If, instead, one nominates a test of goodness of fit, and then delimits the parameter

values for which there is a fit within a specified significance level, one obtains distributions which are consonant with the whole data. Such limits will be different, and may be wider and might therefore be thought to be not as good as the conditional ones. But they seem to me at least to give an answer to the informational question. This seems to be an example of how the notion of "most powerful," and even the notion of sufficiency have led our profession astray.

ACKNOWLEDGMENT

I am indebted to H. T. David for several informative discussions, to J. D. Atkinson for help in computation, and to L. Jordan for checking the algebra.

REFERENCES

[1] HARALD CRAMÉR, *Mathematical Methods of Statistics*, Princeton, Princeton University Press, 1946.

[2] H. T. DAVID, "Order statistics and statistics of structure (*d*)," *Ann. Math. Statist.*, Vol. 36 (1965), pp. 897–906.

[3] ———, "Goodness of fit," to be published in *International Encyclopedia of the Social Sciences*, 1966.

[4] R. A. FISHER, "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. London Ser. A*, Vol. 222 (1922), pp. 309–368.

[5] ———, "The conditions under which χ^2 measures the discrepancy between observations and hypothesis," *J. Roy. Statist. Soc.*, Vol. 87 (1924), pp. 442–450.

[6] H. J. MANN and A. WALD, "On the choice of the number of class intervals in the application of the chi-square test," *Ann. Math. Statist.*, Vol. 13 (1942), pp. 306–317.

[7] E. S. PEARSON and H. O. HARTLEY, *Biometrika Tables for Statisticians*, Vol. 1, Cambridge, Cambridge University Press, 1945.

[8] S. SHAPIRO and M. B. WILK, "An analysis of variance test for normality (complete samples)," *Biometrika*, Vol. 52 (1965), pp. 591–611.

[9] M. J. SLAKTER, "A comparison of the Pearson chi-square and Kolmogorov goodness-of-fit tests with respect to validity," *J. Amer. Statist. Assoc.*, Vol. 60 (1965), pp. 854–858.

[10] D. VAN DANTZIG, "Statistical Priesthood" (Savage on Personal Probabilities 1), *Statistica Neerlandica*, Vol. 11, Nr. 1 (1957), pp. 1–16.

[11] ———, "Statistical Priesthood II" (Sir Ronald on Scientific Inference), *Statistica Neerlandica*, Vol. 11, Nr. 4 (1957), pp. 185–200.