RECENT DEVELOPMENTS IN CHI-SQUARE TESTS FOR GOODNESS OF FIT

David S. Moore
Purdue University*

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #459

July 1976

*Preparation of this paper was supported by the Air Force Office of Scientific Research under Grant No. AFOSR-72-2350. This is the text of an invited lecture presented at the American Statistical Association Annual Meeting, Boston, Mass., August 1976.

(v) Size and Sparsity. A matrix is sparse if, say, 90% of its elements are zero. The two important classes are:

Dense matrices (usually small, can be stored in central memory as arrays, $2 \leq n \leq 100$).

Sparse matrices (usually large, not storeable as arrays, $500 \leq n \leq 10^6$).

Small, sparse matrices do occur. In fact, dense matrices are often reduced to these forms as a preliminary step.

(vi) Class of Matrix. Some matrix properties are exploitable in computation. Here are some common types: real, symmetric, positive definite, band structured ($a_{ij} = 0$ if $|i-j| > w$).

D. Relation to Other Areas

The boundary between Numerical Linear Algebra, Optimization, Mathematical Software, and Differential Equations is fuzzy.

II. CHRONOLOGICAL DEVELOPMENT SINCE 1945

A. Direct Methods

(i) The Fall and Rise of Gaussian Elimination

In high school one learns how to solve a system of linear equations by elimination of variables and backsubstitution. In the late 1940's this technique was implemented on the new digital computers. The inevitable roundoff errors are the only source of error but analysis, and a few unfortunate experiences, suggested that these tiny errors could cascade and ruin the calculation. Gaussian elimination was rejected for systems with more than 10 unknowns as being unstable (1946). Other methods were explored. Pessimism reigned. It seemed incredibly difficult to give a rigorous analysis of the simplest matrix methods when executed in noisy arithmetic. For example, addition is not associative

$$A + (B+C) \neq (A+B) + C$$

and one seems forced to deal with pseudo-addition and pseudo-multiplication at every turn.

The breakthrough, or rather the enlightenment, came between 1955 and 1960. Through a systematic use of backward error analysis Wilkinson showed that floating point computations were rather easy to understand. This backward analysis (intuitively, ask not for the error but for the problem you actually solved) had been used previously but, for various reasons, its implications were not appreciated. It permitted escape from the pseudo-operations. Not only was Gaussian elimination vindicated but almost all matrix computations were illuminated and understood by an ever increasing band of users.

RECENT DEVELOPMENTS IN CHI-SQUARE TESTS FOR GOODNESS OF FIT

David S. Moore*
Purdue University

1. INTRODUCTION

A chi-square statistic for testing fit is a quadratic form having as its arguments the numbers of observations falling in each of a finite set of cells. It will soon appear that when the freedom to choose the cells and the quadratic form in various ways is exercised, this class includes a wide variety of statistics. Chi-square tests are the oldest, and still one of the most common, class of tests of fit. Surveys still of value are given by Cochran (1952) and Watson (1959). These tests are usually less powerful than such other general tests of fit as the various EDF statistics, and are also less powerful than tests designed to test specific hypotheses, such as the Shapiro-Wilk test for univariate normality. The choice of material in this paper is guided by this fact, and by several theses which follow from it.

The first thesis is that chi-square tests of fit must compete for use on the basis of adaptibility and ease of use. EDF tests are somewhat discomfited when (as is usual) we wish to test fit to a parametric family of distributions rather than to a single specified distribution. What is worse,

---

*Preparation of this paper was supported by the Air Force Office of Scientific Research under Grant No. AFOSR-72-2350. This is the text of an invited lecture presented at the American Statistical Association Annual Meeting, Boston, Mass., August 1976.

these tests fall into disarray when the data are discrete or multivariate. The classical Pearson chi-square test is not affected (except for changing degrees of freedom) by these complications. The Pearson test does require that unknown parameters be estimated by the minimum chi-square (or asymptotically equivalent) method. Recent work has provided alternative chi-square statistics which retain the advantages of the Pearson statistic while allowing alternative methods of estimating parameters, data-dependent cells, and other features which increase the flexibility of chi-square methods.

A second thesis: chi-square statistics actually having a (limiting) chi-square null distribution have a much stronger claim to practical usefulness. Ease of use requires the ability to obtain (1) the observed value of the test statistic, and (2) critical points for the test statistic. The calculations required for (1) in chi-square statistics are typically iterative solutions of nonlinear equations and evaluation of quadratic forms, perhaps with matrix expressed as the inverse of a given symmetric pd matrix. These are not serious barriers to practical use, given the current availability of computer library routines. Critical points often require much more effort, so that (2) is the main determinant of ease of use.

The large sample null distribution of a chi-square statistic is typically that of a linear combination of independent chi-square random variables. Theorem 4.2 of Moore and Spruill (1975) is a quite general result of this form. If infinite linear combinations are allowed, EDF statistics also have distributions of this form. See Section 3 of Stephens (1976) for a summary and references. There are effective methods for computing critical points for such distributions (Moore (1971), Section 4; Dahiya and Gurland (1972); Stephens (1976).) But a user must actually write a program to do this, and separate tables must be constructed for each hypothesized family. This is a worthwhile endeavor when fit to a particular family will be tested repeatedly. But since the work required for a chi-square test is similar to that for a

more powerful EDF or specialized test, use of a chi-square test is not justified. The argument on behalf of the second thesis is now clear. There are very general methods for constructing statistics having a chi-square distribution in large samples. Indeed, the Pearson statistic is one such method. These are the chi-square tests which can most effectively compete for the attention of users.

It is already clear that I would not recommend use of a chi-square statistic in such standard situations as testing whether a random sample comes from a univariate normal population. I hope my support for a third thesis is also clear. In many situations, especially when data are discrete, multivariate, or censored, or when parameters must be estimated in an uncommon model, chi-square tests of fit are superior in practice to their competitors.

Section 2 surveys recent work on chi-square tests of fit in a quite selective way, guided by the first two theses. Section 3 presents several examples of the use of these statistics, and attempts to illustrate the third thesis. The emphasis in this paper is entirely on construction of chi-square statistics having tabled large-sample distributions. Although this emphasis is, I believe, justified both by the discussion above and by the volume of recent work in this area, some aspects of great practical importance are not covered here. Chief among these is the study of exact distributions and of the accuracy of the large-sample approximations. Readers interested in these aspects might begin with Good, Gover and Mithcell (1970). I have also omitted any systematic discussion of power or efficiency. The performance of several of the statistics reviewed in Section 2 has not yet been adequately studied, so that no survey of this aspect is yet possible.

2. CONSTRUCTION OF CHI-SQUARE STATISTICS

I will divide the chi-square statistics to be reviewed into two classes. The first I call "standard," as it contains all statistics whose large sample theory is similar to that of the classical Pearson statistic. A general account of this theory appears in Moore and Spruill (1975). The second, or "non-standard" class contains some interesting statistics not falling within the range of this general theory. For reasons stated in the introduction, I have restricted attention to statistics having tabled large-sample distributions, usually chi-square distributions.

2.1 STANDARD STATISTICS

Suppose that $X_1,\ldots,X_n$ are independent and identically distributed r.v.'s taking values in R^p and having unknown df G. Suppose also that $F(\cdot|\theta)$ is a family of df's indexed by an m-dimensional parameter θ taking values in an open set Ω in R^m. We wish to test the hypothesis

$$H_0\colon G(\cdot) = F(\cdot|\theta) \qquad \text{for some } \theta \text{ in } \Omega.$$

If $E_1,\ldots,E_M$ are cells which partition R^p, the cell probability for E_σ under H_0 with θ true is

$$p_\sigma(\theta) = \int_{E_\sigma} dF(x|\theta) \ .$$

Since θ is unknown, it is estimated by some estimator $\theta_n = \theta_n(X_1,\ldots,X_n)$, and the estimated cell probabilities are $p_\sigma(\theta_n)$. The number of $X_1,\ldots,X_n$ falling in E_σ will be denoted by $N_{n\sigma}$. Denote by $V_n(\theta)$ the M-vector of standardized cell frequencies having σth component

$$[N_{n\sigma} - np_\sigma(\theta)]/(np_\sigma(\theta))^{\frac{1}{2}} \ .$$

If $Q_n = Q_n(X_1,\ldots,X_n)$ is a possibly data-dependent M x M symmetric nnd matrix, the general form of a standard chi-square statistic is

$$V_n(\theta_n)'Q_nV_n(\theta_n) \ . \tag{2.1}$$

That is, we are concerned with arbitrary nnd quadratic forms in the standardized cell frequencies.

There is an additional feature which greatly increases the flexibility of these statistics, fortunately without increasing their complexity in practice. This is the use of data-dependent cells. Suppose then that each cell $E_{n\sigma} = E_{n\sigma}(X_1,\ldots,X_n)$ is a p-dimensional rectangle with vertices converging in probability to the vertices of a fixed cell $E_\sigma(\theta_0)$ when H_0 holds and θ_0 is the true parameter value. A common example is the use of cell boundaries $\bar{X}_n + a_\sigma s_n$ ($\bar{X}_n$ and s_n are the sample mean and standard deviation) in testing fit to the univariate normal family. Since the cell frequencies $N_{n\sigma}$ no longer have a multinomial distribution, the theoretical study of these statistics becomes more complex. But in practice, the use of random cells has no effect: the limiting distribution of the statistics (2.1) with random cells $E_{n\sigma}$ is exactly the same (under H_0 and θ_0 true) as if the limiting cells $E_\sigma(\theta_0)$ were used. This is true even when the estimator θ_n is the minimum chi-square estimator computed from the random cells. These results hold under mild regularity conditions. Details may be found in Section 4 of Moore and Spruill (1975).

1. The Pearson Statistic. Here $Q_n \equiv I_M$, the M x M identity matrix, and is estimated by $\bar{\theta}_n$, the minimum chi-square estimator which is the solution of

$$\sum_{\sigma=1}^{M} \left(\frac{N_{n\sigma}}{p_\sigma(\theta)}\right)^2 \frac{\partial p_\sigma(\theta)}{\partial \theta_j} = 0 \qquad j=1,\ldots,m.$$

It is asymptotically equivalent, and computationally simpler, to take $\bar{\theta}_n$ to be the maximum likelihood estimator from the multinomial $N_{n\sigma}$ found as the solution of

$$\sum_{\sigma=1}^{M} \frac{N_{n\sigma}}{p_\sigma(\theta)} \frac{\partial p_\sigma(\theta)}{\partial \theta_j} = 0 \qquad j=1,\ldots,m. \tag{2.2}$$

The Pearson statistic is

$$P_n(\bar{\theta}_n) = \sum_{\sigma=1}^{M} \frac{[N_{n\sigma} - np_\sigma(\bar{\theta}_n)]^2}{np_\sigma(\bar{\theta}_n)} = V_n(\bar{\theta}_n)'V_n(\bar{\theta}_n) \quad .$$

2. The Rao-Robson Statistic. Here we estimate θ by the MLE $\hat{\theta}_n$ from $X_1,\ldots,X_n$. The Pearson statistic $P_n(\hat{\theta}_n)$ does not have a chi-square distribution. As Chernoff and Lehmann (1954) discuss in detail, the distribution of $P_n(\hat{\theta}_n)$ in general depends on the unknown θ and has critical points known only to fall between those of the $\chi^2(M-1)$ and $\chi^2(M-m-1)$ distributions. When the number of cells M is large, these are often useful bounds. As A. R. Roy and G. S. Watson noticed (see Watson (1959) for discussion and references), the use of data-dependent cells can render the distribution of $P_n(\hat{\theta}_n)$ θ-free in location-scale cases. But this distribution is not F-free, and is not χ^2. In accordance with the second thesis, I will therefore ignore this option in favor of one offered by Rao and Robson (1974). Their idea is to ask: What quadratic form in $V_n(\hat{\theta}_n)$ has the $\chi^2(M-1)$ limiting law?

To answer this question, let $J(\theta)$ be the m x m information matrix for $F(x|\theta)$ and define the M x m matrix $B(\theta)$ with (i,j)th entry

$$p_i(\theta)^{-\frac{1}{2}} \frac{\partial p_i(\theta)}{\partial \theta_j} \quad .$$

Then nB'B is the information about θ in the cell frequencies $N_{n\sigma}$. If J-B'B, which is always nnd, is pd, we can write

$$S(\theta) = I_M + B(\theta)[J(\theta) - B(\theta)'B(\theta)]^{-1}B(\theta)'$$

The Rao-Robson statistic is

$$R_n = V_n(\hat{\theta}_n)'S(\hat{\theta}_n)V_n(\hat{\theta}_n)$$

and has the $\chi^2(M-1)$ limiting null distribution. This form simplifies considerably since $\Sigma_1^M \partial p_\sigma/\partial\theta_j = 0$ implies that

(2.3) $$V_n'B = n^{-\frac{1}{2}} \left(\sum_{\sigma=1}^{M} \frac{N_{n\sigma}}{p_\sigma} \frac{\partial p_\sigma}{\partial \theta_1}, \ldots, \sum_{\sigma=1}^{M} \frac{N_{n\sigma}}{p_\sigma} \frac{\partial p_\sigma}{\partial \theta_j} \right)$$

Further simplification can be achieved in location-scale cases by the use of random cells for which $p_\sigma(\hat{\theta}_n) \equiv 1/M$. When m=1, the Rao-Robson statistic is, using (2.3),

(2.4) $$R_n = \sum_{\sigma=1}^{M} \frac{(N_\sigma - np_\sigma)^2}{np_\sigma} + \frac{1}{nD} \left(\sum_{\sigma=1}^{M} \frac{N_\sigma}{p_\sigma} \frac{dp_\sigma}{d\theta} \right)^2$$

where

$$D = J - \sum_{\sigma=1}^{M} \frac{1}{p_\sigma} \left(\frac{dp_\sigma}{d\theta} \right)^2$$

and J, p_σ, $dp_\sigma/d\theta$ are all evaluated at $\theta = \hat{\theta}_n$. Rao and Robson (1974) give several examples of the computation of this statistic, using data-dependent cells in some examples. Their simulations of power and some theoretical work by Moore and Spruill, (1975), Section 7, and Spruill (1976) suggest strongly that the Rao-Robson statistic generally has high power relative to other chi-square tests of fit.

3. The Wald's Method Statistic. For general estimators θ_n of θ one can ask what quadratic form in $V_n(\theta_n)$ has the chi-square limiting null distribution with the largest possible number of degrees of freedom. Moore (1976) has shown that a general answer has the following form. Suppose that when H_0 and θ_0 are true, $V_n(\theta_n)$ has a limiting M-variate normal distribution $N_M(0,\Sigma(\theta_0))$. If $\Sigma(\theta)^-$ is any generalized inverse of the limiting covariance matrix $\Sigma(\theta)$, then

$$W_n(\theta_n) = V_n(\theta_n)'\Sigma(\theta_n)^- V_n(\theta_n)$$

has the $\chi^2(k)$ limiting null distribution, k being the rank of $\Sigma(\theta_0)$ and the largest possible number of degrees of freedom. This rather abstract recipe is made more usable by the fact that typically W_n is invariant under choice of Σ^- and can be computed as the quadratic form corresponding to the

inverse of a nonsingular matrix depending on the method of estimation used. Details appear in Moore (1976).

The statistic $W_n(\bar{\theta}_n)$ is the Pearson statistic, and $W_n(\hat{\theta}_n)$ is the Rao-Robson statistic. In other cases $W_n(\theta_n)$ is much more complicated to compute.

4. The Dzhaparidze-Nikulin Statistic. Faced with the complexity of the Wald's method statistic for general θ_n, we might willingly sacrifice degrees of freedom (and presumably power) for ease of computation. Dzhaparidze and Nikulin (1974) offer the statistic

$$D_n(\theta_n) = V_n'(I_M - B(B'B)^{-1}B')V_n$$

where V_n and B are evaluated at $\theta = \theta_n$. They claim that $D_n(\theta_n)$ has the $\chi^2(M-m-1)$ limiting null distribution whenever θ_n approaches the true θ_0 at the usual $n^{\frac{1}{2}}$ rate. (Their proof of this claim appears to be defective. I can verify that $D_n(\theta_n)$ has a $\chi^2(k)$ distribution for $k \leq M-m-1$, and in most cases that $k = M-m-1$, but I cannot yet prove this last result in the generality claimed.) Computation of D_n is again simplified by (2.3). Dzhaparidze and Nikulin give no examples of the use of their statistic.

2.2 NONSTANDARD STATISTICS

1. The Kempthorne Statistic. If the number of cells M is allowed to increase with the sample size n at a rate faster than $o(n^{\frac{1}{2}})$, the nature of the large-sample theory of chi-square statistics changes radically. Kempthorne (1968) proposes such a test: given n observations, use the Pearson sum of squares for n cells each having probability 1/n under H_0. That is, for cells $N_{n1},\ldots,N_{nn}$ chosen in this manner, the Kempthorne statistic is

$$K_n = \sum_{\sigma=1}^{n} (N_{n\sigma} - 1)^2 = \sum_{\sigma=1}^{M} N_{n\sigma}(N_{n\sigma} - 1) \quad .$$

For the case of testing fit to a completely specified distribution, the $N_{n\sigma}$ are multinomial and it follows from, e.g., Morris (1975) that K_n has a normal limiting null distribution. One expects that this result will be

unaffected by estimating unknown parameters, but to my knowledge this has not been proved. The theory of general chi-square statistics when M increases more rapidly than $o(n^{\frac{1}{2}})$ is largely unexplored. Some preliminary simulations suggest that K_n is superior in power to standard chi-square tests only for very short-tailed alternatives, and may be quite inferior in other cases.

2. The O'Reilly-Quesenberry Statistic. O'Reilly and Quesenberry (1973) propose a transformation approach to testing fit. Specifically, they give a "conditional probability integral transformation" (the form of which depends on the hypothesized family $F(\cdot|\theta)$) of $X_1,\ldots,X_n$ into $U_1,\ldots,U_{np-m}$, where under H_0 the U_i are independent uniform r.v.'s on the unit interval. Quesenberry (1975) has extended the class of families $F(\cdot|\theta)$ for which such transformations are available. One can now test the U_i for uniformity by any available test, thereby obtaining a test of fit to the parametric family $F(x|\theta)$ having a null distribution which is both θ-free and F-free. Perhaps the tests suggested by Weiss (1974), (1976) deserve consideration for use in this last step. They represent an interesting theoretical development, but are not discussed in this survey since they test fit only to a specified distribution, which is taken to be uniform.

O'Reilly and Quesenberry, by applying a chi-square statistic to the U_i, obtain particular members of the following class of nonstandard chi-square tests. Rather than base cell frequencies on cells E_σ (fixed) or $E_{\sigma n}(X_1,\ldots,X_n)$ (data-dependent) into which all of $X_1,\ldots,X_n$ are classified, the cells used to classify each succesive X_i are functions $E_{\sigma i}(X_1,\ldots,X_i)$ of $X_1,\ldots,X_i$ only. Thus additional observations do not require reclassification of earlier observations, as in the usual data-dependent cell case. O'Reilly and Quesenberry show from their conditional transformation approach the existence of functions $E_{\sigma i}(X_1,\ldots,X_i)$ such that the cell frequencies $N_{n1},\ldots,N_{nM}$ computed as above have the multinomial distribution with any specified set

$\{p_1,\ldots,p_M\}$ of cell probabilities. The O'Reilly-Quesenberry statistic is then the Pearson statistic

$$O_n = \sum_{\sigma=1}^{M} \frac{(N_{n\sigma} - np_\sigma)^2}{np_\sigma}$$

for these $N_{n\sigma}$, and has the $\chi^2(M-1)$ limiting null distribution.

The authors show by example that in common cases the boundaries of the appropriate cells $E_{\sigma i}(X_1,\ldots,X_i)$ can be obtained fairly simply. In terms of distribution theory, these statistics are direct competitors of the Rao-Robson statistics, which also achieve a $\chi^2(M-1)$ limiting null distribution. I know of no work on power comparisons which might aid a choice between the two methods. Obtaining the cells for the O'Reilly-Quesenberry statistic for a specific family $F(\cdot|\theta)$ requires the computation of the minimum variance unbiased estimate of $F(\cdot|\theta)$, so that a practitioner wishing to test fit to a relatively uncommon family (I have argued that this is the case in which chi-square tests are most defensible) will generally find the Rao-Robson statistic easier to obtain. It would be valuable to have available the specific recipes for O_n for testing fit to common multivariate families, where competition from EDF tests falls off. A first effort in this direction (multivariate normal with known covariance matrix) appears in Section 5 of O'Reilly and Quesenberry (1973). Finally, it should be noted that the very ingenious conditional probability integral transformation approach constructs and obtains the limiting null distribution of only specific members of the general class of "sequential-cell" chi-square statistics. The theory of this class of statistics does not fall within the scope of Moore and Spruill (1975), and offers much room for future work.

3. Easterling's Test. Easterling (1976) provides a very interesting approach to parameter estimation based on tests of fit. Roughly speaking, he advocates replacing the usual confidence intervals for θ in $F(\cdot|\theta)$ based

on the acceptance regions of a test of

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

with intervals based on the acceptance regions of tests of fit to completely specified distributions,

$$H_0^*: G(\cdot) = F(\cdot|\theta_0)$$

$$H_1^*: G(\cdot) \neq F(\cdot|\theta_0) \quad .$$

(I say "roughly speaking" to avoid discussion of the philosophy of inference aspect of Easterling's paper, which is not relevant here.) In the course of his discussion, Easterling suggests rejecting the family $\{F(x|\theta): \theta \text{ in } \Omega\}$ as a model for the data if the (say) 50% confidence interval for θ based on acceptance regions for H_0^* is empty. It is this "implicit test of fit" that I wish to comment on, using the chi-square case to make some observations which apply as well when other tests of H_0^* are employed.

Taking then the standard chi-square statistic for H_0^* when $Y_n = (X_1,\ldots,X_n)$ is observed,

$$T_n(Y_n,\theta_0) = \sum_{\sigma=1}^{M} \frac{[N_{n\sigma} - np_\sigma(\theta_0)]^2}{np_\sigma(\theta_0)} \quad ,$$

and denoting by $\chi_\alpha^2(M-1)$ the upper α-point of the $\chi^2(M-1)$ distribution, the $(1-\alpha)$-confidence interval is empty if and only if

$$T_n(Y_n,\theta) > \chi_\alpha^2(M-1) \qquad \text{for all } \theta \text{ in } \Omega. \tag{2.5}$$

But if $\bar{\theta}_n$ is the minimum chi-square estimator, (2.5) holds if and only if

$$T_n(Y_n,\bar{\theta}_n) > \chi_\alpha^2(M-1) \quad . \tag{2.6}$$

When any $F(x|\theta)$ is true, $T_n(Y_n,\bar{\theta}_n)$ has the $\chi^2(M-m-1)$ distribution, and the probability of the event (2.6) can be explicitly computed. It is less than α, but close to α when M is large.

Thus Easterling's suggestion essentially reduces to the use of standard tests of fit with parameters estimated by the minimum distance method

corresponding to the test statistic employed. He is a bit surprised that as many as 16% of simulated exponential observations failed this test for exponentiality (using the Anderson-Darling statistic) with $\alpha = .50$. I am surprised that so low a percentage failed.

3. EXAMPLES

The examples below are chosen to illustrate the versatility of chi-square methods and yet be sufficiently simple for compact presentation. Consequently, the hypothesized families are all univariate and for the most part allow explicit computation. Where computer work is needed, the examples are incomplete. I hope to provide more details in the final version of this paper. The examples are restricted to standard chi-square tests, since (some of) these are easier to adapt to a new problem than is the O'Reilly-Quesenberry statistic.

Example 1. We wish to test fit to the double exponential family having density function

$$f(x|\theta) = \frac{1}{2\theta_2} e^{-|x-\theta_1|/\theta_2} \qquad -\infty < x < \infty$$

$$\Omega = \{(\theta_1,\theta_2): -\infty < \theta_1 < \infty,\ 0 < \theta_2 < \infty\} .$$

The MLE $\hat{\theta}_n = (\hat{\theta}_{1n},\hat{\theta}_{2n})$ from a random sample $X_1,\ldots,X_n$ is

$$\hat{\theta}_{1n} = \text{median } (X_1,\ldots,X_n)$$

$$\hat{\theta}_{2n} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{\theta}_{1n}| \quad .$$

Since $f(x|\theta)$ is a location-scale family, we follow the usual practice of using data-dependent cells to achieve equal estimated cell probabilities. If the successive cell boundaries are $\hat{\theta}_{1n} + a_\sigma \hat{\theta}_{2n}$, then

(3.1)
$$p_\sigma(\theta) = \int_{\hat\theta_{1n} + a_{\sigma-1}\hat\theta_{2n}}^{\hat\theta_{1n} + a_\sigma \hat\theta_{2n}} \frac{1}{2\theta_2} e^{-|x-\theta_1|/\theta_2}\, dx$$

and

$$p_\sigma(\hat\theta_n) = \int_{a_{\sigma-1}}^{a_\sigma} \frac{1}{2} e^{-|t|} dt$$

which depends only on the a_σ. Using an even number of cells, say $M = 2\nu$ and choosing the a_σ symmetrically about 0 as $á_{\nu+i} = -a_{\nu-i} = c_i$, where

$$c_i = -\log\left(1 - \frac{i}{\nu}\right) \qquad i=0,\ldots,\nu$$

(in particular, $a_0 = -\infty$, $a_\nu = 0$, $a_M = \infty$) gives $p_\sigma(\hat\theta_n) \equiv 1/M$.

To obtain the Rao-Robson statistic R_n, first compute $\partial p_\sigma/\partial\theta_j$ by interchanging differentiation and integration in (3.1). Substituting $\hat\theta_n$ for θ gives

$$\frac{\partial p_\sigma}{\partial \theta_1}(\hat\theta_n) = \begin{cases} -1/M\hat\theta_{2n} & \sigma=1,\ldots,\nu \\ +1/M\hat\theta_{2n} & \sigma=\nu+1,\ldots,M \end{cases}$$

(3.2)
$$\frac{\partial p_\sigma}{\partial \theta_2}(\hat\theta_n) = \frac{1}{2\hat\theta_{2n}}\left(c_{i-1}e^{-c_{i-1}} - c_i e^{-c_i}\right) \qquad \begin{cases} \sigma = \nu+i,\ \nu-i+1 \\ i = 1,\ldots,\nu \end{cases}$$

Set $d_i = c_{i-1}e^{-c_{i-1}} - c_i e^{-c_i}$. Then

$$B(\hat\theta_n)'B(\hat\theta_n) = \hat\theta_{2n}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & \nu\Sigma_1^\nu d_i^2 \end{pmatrix}$$

Since the information matrix for $f(x|\theta)$ is $\theta_2^{-1} I_2$, the matrix $J(\hat\theta_n) - B(\hat\theta_n)'B(\hat\theta_n)$ has rank 1 and the Rao-Robson statistic is not defined. This is an unusual situation.

We can fall back on either the Pearson statistic or the Dzhaparidze-Nikulin statistic $D_n(\hat\theta_n)$. The latter is more natural when $\hat\theta_n$-dependent cells

are used, and much easier to compute. Since these cells have the practical advantage of adjusting to the location and spread of the data, I prefer to use them and therefore recommend the use of D_n. Computation of D_n is almost trivial because (i) here

$$\sum_{\sigma=1}^{M} N_{n\sigma} \frac{\partial p_\sigma}{\partial \theta_1}(\hat{\theta}_n) = 0 \tag{3.3}$$

by (3.2) and the definition of the median; (ii) B'B is diagonal; and (iii) the relation (2.3) applies. The resulting statistic is

$$D_n(\hat{\theta}_n) = \frac{M}{n} \sum_{\sigma=1}^{M} (N_\sigma - \frac{n}{M})^2 - \frac{M}{n} \frac{1}{2\ \Sigma_1^\nu d_i^{\,2}} [\sum_{i=1}^{\nu} d_i (N_{\nu+i} + N_{\nu-i+1})]^2$$

and has the $\chi^2(M-3)$ limiting null distribution.

[Comments. The first term in D_n is the Pearson statistic for this choice of cells. It has the distribution of $\chi^2(M-3) + \lambda_1\chi^2(1)$ where the χ^2's denote independent chi-square r.v.'s with the indicated degrees of freedom and $0 < \lambda_1 < 1$. The number λ_1 is θ-free because of our use of data-dependent cells. If J-B'B were nonsingular, another term of the form $\lambda_2\chi^2(1)$ would appear in this representation. The second term in D_n can be thought of as cancelling the $\lambda_1\chi^2(1)$. If θ_2 were known, we would have by (3.3) (compare (2.2), recalling that $p_\sigma(\hat{\theta}_n) = 1/M$) the unusual situation in which the MLE $\hat{\theta}_n(X_1,\ldots,X_n)$ is the same as the grouped data MLE $\overline{\theta}_n(N_{n1},\ldots,N_{nM})$. So for the location-parameter double exponential family, we can choose cells as here and then use the Pearson statistic

$$P_n(\hat{\theta}_n) = \frac{M}{n} \sum_{\sigma=1}^{M} (N_{n\sigma} - \frac{n}{M})^2$$

with the $\chi^2(M-2)$ distribution. The use of data-dependent cells is essential to this result.]

Example 2. We wish to test fit to the negative exponential family having density function

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta} \qquad 0 < x < \infty$$

$$\Omega = \{\theta: \ 0 < \theta < \infty\} \ .$$

We have not a full random sample, but rather Type II censored data. That is, we observe the order statistics up to the sample α-quantile,

$$X_{(1)} < X_{(2)} < \ldots < X_{([n\alpha])} \ ,$$

where $[n\alpha]$ is the greatest integer in $n\alpha$ and $0 < \alpha < 1$. Such data are common in life testing situations. It is natural to make use of random cells with sample quantiles $\xi_i = X_{([n\delta_i])}$ as cell boundaries. Here $\xi_0 = 0$, $\xi_1 = \infty$ and

$$0 = \delta_0 < \delta_1 < \ldots < \delta_{M-1} = \alpha < \delta_M = 1 \ ,$$

so that the $n - [n\alpha]$ unobserved X_i fall in the rightmost cell. Although the cell frequencies $N_{n\alpha}$ are now fixed, the general theory of Moore and Spruill (1975) applies to this choice of cells. The use of order statistics as cell boundaries was considered by Witting (1959) and Bofinger (1973), but this application to censored data seems new. I will discuss several chi-square tests of fit based on this choice of cells. For references to previous literature on tests of fit for censored data, see Lurie, Hartley, and Stroud (1974). This example can be taken as a response to their claim that "the chi-square criterion is not generally applicable to testing the fit of Type II censored samples."

The Pearson Statistic. Estimate θ by the grouped data MLE found as the solution of (2.2). That equation becomes in this case

$$\sum_{\sigma=1}^{M} N_{n\sigma} \frac{\xi_{\sigma-1} e^{-\xi_{\sigma-1}/\theta} - \xi_{\sigma} e^{-\xi_{\sigma}/\theta}}{e^{-\xi_{\sigma-1}/\theta} - e^{-\xi_{\sigma}/\theta}} = 0 \tag{3.4}$$

which is easily solved iteratively to obtain $\overline{\theta}_n = \overline{\theta}_n(\xi_1, \ldots, \xi_{M-1})$. The test statistic is

$$P_n(\bar{\theta}_n) = \sum_{\sigma=1}^{M} \frac{[N_{n\sigma} - np_\sigma(\bar{\theta}_n)]^2}{np_\sigma(\bar{\theta}_n)}$$

where

$$N_{n\sigma} = [n\delta_\sigma] - [n\delta_{\sigma-1}] \qquad \text{(nonrandom)}$$

$$p_\sigma(\theta) = e^{-\xi_{\sigma-1}/\theta} - e^{-\xi_\sigma/\theta} \qquad \text{(random)} .$$

The limiting null distribution is $\chi^2(M-2)$.

The Dzhaparidze-Nikulin statistic. It is tempting to use a more efficient estimator of θ than $\bar{\theta}_n$. The MLE (also the MVUE) of θ based on the observed order statistics is given by Epstein and Sobel (1953) as

$$\tilde{\theta}_n = \frac{1}{[n\alpha]} \{ \sum_{i=1}^{[n\alpha]} X_{(i)} + (n - [n\alpha])X_{([n\alpha])} \} .$$

Since $\tilde{\theta}_n$ has a simple explicit form, we have a double incentive to use it. When m=1, it is very easy to calculate $D_n(\tilde{\theta}_n)$ as

$$D_n(\tilde{\theta}_n) = \sum_{\sigma=1}^{M} \frac{(N_{n\sigma} - np_\sigma)^2}{np_\sigma} - \frac{1}{n \sum_{1}^{M} \frac{1}{p_\sigma} \left(\frac{dp_\sigma}{d\theta} \right)^2} \left(\sum_{\sigma=1}^{M} \frac{N_{n\sigma}}{p_\sigma} \frac{dp_\sigma}{d\theta} \right)^2$$

where $p_\sigma(\theta)$ is given above,

$$\frac{dp_\sigma}{d\theta} = \theta^{-2}(\xi_{\sigma-1} e^{-\xi_{\sigma-1}/\theta} - \xi_\sigma e^{-\xi_\sigma/\theta})$$

and both are evaluated at $\theta = \tilde{\theta}_n$. Although $D_n(\tilde{\theta}_n)$ has an explicit form, while $P_n(\bar{\theta}_n)$ requires numerical solution of (3.4), the practical advantage of D_n in ease of use is not great. Both statistics have the $\chi^2(M-2)$ limiting null distribution; their relative efficiency has not been studied.

The Wald's Method statistic. It is quite probable that a more powerful chi-square test than either P_n or D_n can be obtained by computing the

"natural" quadratic form in the $N_{n\sigma} - np_\sigma(\tilde{\theta}_n)$ given by Wald's method. The first step is to show that $V_n(\tilde{\theta}_n)$ has a limiting multivariate normal $N(0,\Sigma(\theta))$ distribution when $f(x|\theta)$ is true. This follows from Theorem 6.1 and Section 7 of Shorack (1969) after writing each component of $V_n(\tilde{\theta}_n)$ as a linear combination of order statistics plus an asymptotically negligible remainder. The resulting $\Sigma(\theta)$ is quite involved. I hope to investigate whether $\Sigma(\tilde{\theta}_n)^-$ can be obtained in usable form.

Example 3. We suspect that a population is described by a member of the univariate normal family, but the data come to us rounded to the nearest integer. We must therefore test fit to the discrete family with probability function

$$f(x|\theta) = \Phi\left(\frac{x+\frac{1}{2}-\theta_1}{\theta_2}\right) - \Phi\left(\frac{x-\frac{1}{2}-\theta_1}{\theta_2}\right) \qquad x=0, \pm1, \pm2,\ldots$$

$$\Omega = \{(\theta_1,\theta_2): \ -\infty < \theta_1 < \infty, \ 0 < \theta_2 < \infty \}$$

where Φ is the standard normal df. This problem is not uncommon in survey data when respondents give numerical replies. An example appears in Carlson (1975), where the data are predictions of future price index levels by business economists.

The unknown parameters θ_1 and θ_2 are not the mean and standard deviation of $f(x|\theta)$, and $\overline{X}$,s are not consistent estimators of θ_1,θ_2. The MLE $\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n})$ can be found as the solution of the log-likelihood equations

$$\sum_{i=1}^{n} \frac{\varphi\left(\frac{x_i + \frac{1}{2} - \theta_1}{\theta_2}\right) - \varphi\left(\frac{x_i - \frac{1}{2} - \theta_1}{\theta_2}\right)}{\Phi\left(\frac{x_i + \frac{1}{2} - \theta_1}{\theta_2}\right) - \Phi\left(\frac{x_i - \frac{1}{2} - \theta_1}{\theta_2}\right)} = 0$$

(3.4)

$$\sum_{i=1}^{n} \frac{(x_i + \frac{1}{2} - \theta_1)\,\varphi\left(\frac{x_i + \frac{1}{2} - \theta_1}{\theta_2}\right) - (x_i - \frac{1}{2} - \theta_1)\,\varphi\left(\frac{x_i - \frac{1}{2} - \theta_1}{\theta_2}\right)}{\Phi\left(\frac{x_i + \frac{1}{2} - \theta_1}{\theta_2}\right) - \Phi\left(\frac{x_i - \frac{1}{2} - \theta_1}{\theta_2}\right)} = 0$$

where φ,Φ are the standard normal pdf and df. Solution of these equations, computation of $J(\hat{\theta}_n)$, computation of $B(\hat{\theta}_n)$ for fixed cells having boundaries at half-integer points, and finally computation of the Rao-Robson statistic R_n is an exercise in computer usage, made easier by library routines for evaluation of Φ. (There are some preliminary theoretical problems involving the equations (3.4). It appears that they may have multiple roots for small n. I do not know if practical problems will arise from this.)

The Pearson statistic offers considerable computational advantages here. Though the equations (2.2) here differ in computational complexity from (3.4) only in having M rather than n terms, that and the relative simplicity of the Pearson sum of squares are clear advantages. The Pearson statistic is still a defensible choice in many problems. Turning to Carlson's data, we do encounter a difficulty (in addition to probable loss of power relative to R_n). In 25 of his 140 samples, the data led him to use only 3 cells. While R_n remains applicable with the $\chi^2(2)$ distribution, P_n cannot be used. Carlson used a very rough approximation, estimating the parameters by $\bar{X}$ and s and using critical points for the Watson-Roy random cell version of $P_n(\hat{\theta}_n)$. It would be interesting to reassess his data using R_n or (leaving out 25 samples) even P_n.

REFERENCES

[1] BOFINGER, E. (1973). Goodness-of-fit using sample quantiles. J. Roy. Statist. Soc. Ser. B 35 277-284.

[2] CARLSON,J. A. (1975). Are price expectations normally distributed? J. Amer. Statist. Assoc. 70 749-754.

[3] CHERNOFF, H. and LEHMANN, E. L. (1954). The use of maximum-likelihood estimates in χ^2 test for goodness of fit. Ann. Math. Statist. 25 579-586.

[4] COCHRAN, W. G. (1952). The χ^2 goodness-of-fit test. Ann. Math. Statist. 23 315-345.

[5] DAHIYA, R. C. and GURLAND, J. (1973). Pearson chi-square test of fit with random intervals. Biometrika 59 147-153.

[6] DZHAPARIDZE, K. O. and NIKULIN, M. S. (1974). On a modification of the standard statistics of Pearson. Theor. Probability Appl. 19 851-853.

[7] EASTERLING, R. G. (1976). Goodness of fit and parameter estimation. Technometrics 18 1-9.

[8] EPSTEIN, B. and SOBEL, M. (1953). Life testing. J. Amer. Statist. Assoc. 48 486-502.

[9] GOOD, I. J., GOVER, T. N. and MITCHELL, G. L. (1970). Exact distributions for χ^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. J. Amer. Statist. Assoc. 65 267-283.

[]0] KEMPTHORNE, O. (1968). The classical problem of inference-goodness of fit. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1 235-249.

[11] LURIE, D., HARTLEY, H. O. and STROUD, M. R. (1974). A goodness of fit test for censored data. Comm. Statist. 3 745-753.

[12] MOORE, D. S. (1971). A chi-square statistic with random cell boundaries. Ann. Math. Statist. 42 147-156.

[13] MOORE, D. S. (1976). Generalized inverses, Wald's method and the construction of chi-squared tests of fit. J. Amer. Statist. Assoc. to appear.

[14] MOORE, D. S. and SPRUILL, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. Ann. Statist.3 599-616.

[15] MORRIS, C. (1975). Central limit theorems for multinomial sums. Ann. Statist. 3 165-188.

[16] O'REILLY, F. J. and QUESENBERRY, C. P. (1973). The conditional probability integral transformation and applications to obtain composite chi-square goodness-of-fit tests. Ann. Statist. 1 74-83.

[17] QUESENBERRY, C. P. (1975). Transforming samples from truncation parameter distribution to normality. Comm. Statist. 4 1149-1155.

[18] RAO, K. C. and ROBSON, D. S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. Comm. Statist. 3 1139-1153.

[19] SHORACK, G. R. (1969). Asymptotic normality of linear combinations of functions of order statistics. Ann. Math. Statist. 40 2041-2050.

[20] SPRUILL, M. C. (1976). A comparison of chi-square goodness-of-fit tests based on approximate Bahadur slope. Ann. Statist. 4 409-412.

[21] STEPHANS, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann. Statist. 4 357-369.

[22] WATSON, G. S. (1959). Some recent results in chi-square goodness-of-fit tests. Biometrics 15 440-468.

[23] WEISS, L. (1974). The asymptotic sufficiency of a relatively small number of order statistics in tests of fit. Ann. Statist. 2 795-802.

[24] WEISS, L. (1976). Asymptotic properties of Bayes tests of nonparametric hypotheses. In S. S. Gupta and D. S. Moore (Eds.) Statistical Decision Theory and Related Topics II. Academic Press, to appear.

[25] WITTING, H. (1959). Über einen χ^2 - Test, dessen klassen durch geordnete Stichprobenfunktionen festgelegt werden. Ark. Mat. 10 468-479.