MEASURES OF LACK OF FIT FROM TESTS OF CHI-SQUARED TYPE*

by

David S. Moore
Purdue University

Technical Report #83-12

April 1983

Purdue University
West Lafayette, IN

*Research supported by the National Science Foundation under Grant MCS 81-21948.

MEASURES OF LACK OF FIT FROM TESTS OF CHI-SQUARED TYPE*

DAVID S. MOORE

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Abstract: If X^2 is the Pearson chi-squared statistic for testing fit, then X^2/n has long been considered an associated measure of the degree of lack of fit. Here we consider two classes of statistics of chi-squared type, each having X^2 as a member. The first is a class of directed divergence statistics discussed by Cressie and Read, the second consists of nonnegative definite quadratic forms in the standardized cell frequencies. We investigate the large sample behavior of T/n, where T is any of these statistics. A number of auxiliary results on the Cressie-Read statistics are also obtained. The measures are illustrated by application to data from classical physics compiled by Stigler.

AMS 1970 Subject Classification: 62G10.

Key words: Goodness of fit, minimum discrepancy estimators, Pearson statistic.

Running title: Measures of Lack of Fit

*Research supported by the National Science Foundation under Grant MCS 81-21948.

1. INTRODUCTION

It is a commonplace of inference that the magnitude of an effect as well as its statistical significance should be reported, and that these two concepts are not identical. In particular, significance is strongly influenced by the sample size n, while the magnitude or degree of effect present ought not to depend on n. We consider here measures of the degree of lack of fit of data to a parametric family of distributions that are naturally associated with test statistics based on discrete or grouped data. The paradigm such statistic for testing fit is the Pearson chi-squared statistic, X^2. X^2 itself measures the significance of the lack of fit. The associated measure of the degree of lack of fit is X^2/n.

This measure (and its square root) have long been employed in a variety of contexts. They remain popular in psychometrics, but are not recommended as measures of association in contingency tables. Fleiss (1981), Section 5.2, gives a discussion and references on these points. See also Bishop, Fienberg and Holland (1975), Chapter 11. The strongest arguments against X^2-based measures of association are: (1) the availability of more easily interpreted measures such as the odds ratio; (2) the fact that the value of X^2/n depends on whether a two-way table is studied prospectively, retrospectively, or naturalistically; and (3) the fact that the value of X^2/n depends on the cutting points used when continuous distributions underlie the two-way table. Here, however, we are concerned with testing fit to parametric families of distributions, e.g. with tests of normality. Only the third argument retains its force in this setting. The dependence of the statistic on the choice of cells in continuous cases is indeed a drawback of chi-squared-like methods. Yet it is this discretizing of the data that allows the use of these tests with standard critical points when parameters must be estimated from the data and when the data are multivariate. The flexibility of X^2 and related statistics is responsible for their continued use, and this in turn warrants a systematic study of the associated measures of lack of fit.

Cressie and Read (1983) have systematized the theory of tests of fit based on the multinomial distributions of cell counts by pointing out that a family of measures of discrepancy between finite probability distributions gives rise to the chi-squared, Neyman modified chi-squared, log likelihood ratio, Freeman-Tukey, and many other test statistics as special cases. These tests, and the minimum-discrepancy estimators based on them, have identical asymptotic properties under the null hypothesis and local alternatives, but not under distant alternatives. Read (1982b) has provided some guidance as to the sensitivity of various tests in this class to different types of deviations from the null hypothesis, and hence to the choice of discrepancy measure in practice.

Suppose that $X_1,\ldots,X_n$ (which may be multidimensional) are iid with common cdf G, and are to be tested for fit to a family of cdf's $\mathcal{F} = \{F(\cdot|\theta)\colon \theta \text{ in } \Omega\}$. The parameter space Ω is an open set in Euclidean m-space. Partition the range of X_j into k cells, whose probabilities are $p_i(\theta)$ under $F(\cdot|\theta)$, $i=1,\ldots,k$. If N_i are the observed cell frequencies and θ_n an estimator of θ based on $X_1,\ldots,X_n$, the Cressie-Read statistics are

$$R^{\lambda}(\theta_n) = 2nI^{\lambda}(N/n : p(\theta_n))$$

where N and $p(\theta)$ are the vectors of N_i and $p_i(\theta)$ and I^{λ} is a directed divergence between probability distributions on k outcomes. The real number λ indexes the class of divergences employed, with $R^1 = X^2$.

Another class of statistics for testing fit, studied in detail by Moore and Spruill (1975), consists of nonnegative definite quadratic forms in the standardized cell frequencies $[N_i - np_i(\theta_n)]/[np_i(\theta_n)]^{1/2}$. The Pearson statistic is the sum of squares, and hence the simplest member of this class. These statistics can also be considered as based on a measure of the discrepancy of N/n from $p(\theta_n)$. Unlike the Cressie-Read statistics, the Moore-Spruill statistics have differing asymptotic behavior under the null hypothesis.

If T_n is a member of either of these classes, then in regular cases T_n has a nondegenerate limiting distribution under $H_0 : G$ in $\mathfrak{F}$. Thus T_n measures the significance of lack of fit. But T_n/n will be seen to measure directly the discrepancy between the empiric distribution of $X_1,\ldots,X_n$ and $F(\cdot|\theta_n)$, obtained in the grouped data setting as a discrepancy between N/n and $p(\theta_n)$. Under H_0, $T_n/n \to 0$ a.s., but we will show that under G not in $\mathfrak{F}$, T_n/n has an a.s. limit that is a corresponding measure of the discrepancy between G and $F(\cdot|\theta_0)$, where $\theta_n \to \theta_0$ a.s. (G). When θ_n is the minimum-T_n estimator, $F(\cdot|\theta_n)$ is the "closest" member of $\mathfrak{F}$ to the empiric distribution of the observations, and we will see that $F(\cdot|\theta_0)$ is then the "closest" member of $\mathfrak{F}$ to G. Thus T_n/n estimates the "distance" of the true G from $\mathfrak{F}$, where the particular "distance" can be chosen for sensitivity to specific types of alternatives. The purpose of this paper is to study the large sample behavior of the measures T_n/n.

Section 2 summarizes our results in the case of the Pearson statistic, and so forms an introduction to the more general study. Section 3 introduces the Cressie-Read and Moore-Spruill statistics. Data-dependent cells are often employed in practice, and were allowed in Moore and Spruill (1975). We point out in Section 3 that if the random cells converge to fixed cells as $n \to \infty$, all of our results (and those of Cressie and Read) for statistics based on the limiting set of fixed cells extend to the random cell case.

Section 4 discusses the behavior of estimators θ_n under G not in the hypothesized family $\mathfrak{F}$. Convergence $\theta_n \to \theta_0$ a.s., and identification of the limit θ_0, are needed for our study of measures of lack of fit. Such results are available for many classes of estimators, but are not given by Read (1983) in his study of minimum-R^λ estimators. We give a very general result of this kind. Section 5 contains the main results for measures of lack of fit based on both Cressie-Read and Moore-Spruill statistics. We discuss pointwise convergence, the relation to approximate Bahadur slope, and asymptotic expansions that can lead to asymptotic normality.

Finally, Section 6 presents an example using the data compiled by Stigler (1977) from 18th and 19th century measurements of physical constants. Since series of repeated measurements "ought to" be approximately normal, we measure degree of nonnormality. The example affords an opportunity to discuss several practical matters, such as the choice of cells.

It is apparent from the outline above that this paper contains some complements to the work of Cressie and Read on general statistics for testing fit based on multinomial data. Nonetheless, the primary purpose is to increase understanding of certain measures of lack of fit by a thorough study of the large-sample properties of these measures.

2. The Pearson Statistic

The Pearson statistic for testing the fit of $X_1,\ldots,X_n$ to the family $\mathfrak{F}$ when θ is estimated by $\theta_n(X_1,\ldots,X_n)$ is

$$X^2(\theta_n) = \sum_{i=1}^{k} \frac{[N_i - np_i(\theta_n)]^2}{np_i(\theta_n)}$$

$$= n \sum_{i=1}^{k} \frac{[N_i/n - p_i(\theta_n)]^2}{p_i(\theta_n)} \quad .$$

The second expression makes it clear that $X^2(\theta_n)/n$ is a measure of the discrepancy between the empirical probabilities N_i/n and the probabilities $p_i(\theta_n)$ estimated under H_0: G in $\mathfrak{F}$. If G is in $\mathfrak{F}$, then $X^2(\theta_n)/n \to 0$ a.s. in regular cases. Our concern is the behavior of $X^2(\theta_n)/n$ when G is not in $\mathfrak{F}$.

Estimators. Common classes of estimators θ_n have the property that $\theta_n \to \theta_0$ a.s. under G, where θ_0 depends on G as well as on the estimation procedure employed. For example, if θ_n is the MLE of θ in $\mathfrak{F}$ based on $X_1,\ldots,X_n$ then Huber (1967) and Perlman (1972) give quite general conditions for a.s. convergence. In this case, θ_0 is the point in Ω at which $E_G[-\log f(X|\theta)]$ is minimized, where f is the density function corresponding to F. Similar results for minimum contrast estimators are given by Pfanzagl (1969).

Specializing these general results to the case in which θ_n is the grouped data MLE of θ in $\mathfrak{F}$ based on the cell frequencies $N_1,\ldots,N_k$ we obtain that in regular cases θ_n is the solution of the equations

$$(2.1) \qquad \sum_{i=1}^{k} \frac{N_i/n}{p_i(\theta)} \frac{\partial p_i}{\partial \theta_j} = 0 \qquad j = 1,\ldots,m$$

and that $\theta_n \to \theta_0$ a.s. (G), where θ_0 satisfies

$$(2.2) \qquad \sum_{i=1}^{k} \frac{\pi_i}{p_i(\theta)} \frac{\partial p_i}{\partial \theta_j} = 0 \qquad j = 1,\ldots,m$$

where $\pi = (\pi_1,\ldots,\pi_k)$ is the vector of cell probabilities under G.

A natural choice of θ_n is the minimum chi-squared estimator, the value of θ in Ω minimizing $X^2(\theta)$. This is asymptotically equivalent to the grouped data MLE under G in $\mathfrak{F}$ and under contiguous alternatives, but not under G not in $\mathfrak{F}$. In regular cases, this θ_n satisfies

$$(2.3) \qquad \sum_{i=1}^{k} \left(\frac{N_i/n}{p_i(\theta)}\right)^2 \frac{\partial p_i}{\partial \theta_j} = 0 \qquad j = 1,\ldots,m$$

and under G, $\theta_n \to \theta_0$ a.s. where θ_0 satisfies

$$(2.4) \qquad \sum_{i=1}^{k} \left(\frac{\pi_i}{p_i(\theta)}\right)^2 \frac{\partial p_i}{\partial \theta_j} = 0 \qquad j = 1,\ldots,m \quad .$$

A general theorem implying these results will be given in Section 4. Note that if $G(\cdot) = F(\cdot|\theta_0)$ is in $\mathfrak{F}$, then $\theta_n \to \theta_0$ is just a.s. consistency.

<u>Pointwise convergence.</u> We have seen that common estimators θ_n

will satisfy $\theta_n \to \theta_0$ a.s. (G). It is then easy to see that

$$(2.5) \qquad \frac{X^2(\theta_n)}{n} \to \sum_{i=1}^{k} \frac{[\pi_i - p_i]^2}{p_i} \quad \text{a.s. (G)}$$

where $p_i = p_i(\theta_0)$. The measure $X^2(\theta_n)/n$ is thus a consistent estimator of the measure $d = \Sigma_1^k[\pi_i - p_i]^2/p_i$ of discrepancy between the true cell probabilities π_i and p_i. When θ_n is the minimum-X^2 estimator, $p(\theta_0)$ is the closest point to π among all $p(\theta)$ for θ in Ω, in the sense that θ_0 satisfying (2.4) has smallest discrepancy d among all θ in Ω. Thus $X^2(\theta_n)/n$ consistently estimates a measure of the "distance" of the true G from $\mathfrak{F}$. This measure depends on the choice of cells, but if G, $F(\cdot|\theta_0)$ have pdf's g,f with respect to Lebesque measure, then as the partition of the range of X_j into cells is refined, d approaches $\int(g-f)^2/f$, an integral discrepancy measure.

Another interpretation of $X^2(\theta_n)/n$ is offered by the fact that the limit d is the approximate Bahadur slope of the Pearson statistic $X^2(\theta_n)$ at the alternative G. Since the slope d determines (asymptotically) the sample size n required for $X^2(\theta_n)$ to reach a stated P-value (computed from the limiting null distribution, which is chi-squared if θ_n is the minimum-X^2 estimator) against G, this fact suggests an easy-to-grasp restatement of the measure $X^2(\theta_n)/n$ of lack of fit. If $X^2(\theta_n)/n = c$ is observed, and c_α is the level α critical point of the limiting null distribution of $X^2(\theta_n)$ (i.e. $P_{H_0}[X^2(\theta_n) \geq c_\alpha] \to \alpha$), then $N_\alpha = c_\alpha/c$ is the number of observations required for an effect of size c to reach the level of significance α. N_α is thus an alternative to c as a measure of lack of fit. For example, if for a sample of size n = 100, $X^2 = 28.1$ with the $\chi^2(9)$ limiting null distribution, then the P-value is 0.0009 and $X^2/n = 0.281$ is the estimated discrepancy. An effect of this size would require $N_{.05} = 61$ observations to be found significant at level $\alpha = 0.05$. Note that while N_α calls attention to the fact that an effect of any fixed magnitude will be significant for n sufficiently large, it does not in itself report which of many possible measures of effect magnitude was employed.

A final interpretation of X^2/n, though one so far afield that we will not discuss it, is given by the concept of resistance to rejection proposed by Ylvisaker (1977). Ylvisaker shows that in the no-estimation case, the resistance to rejection of the critical region $X^2 \geq c$ is proportional to $(c/n)^{\frac{1}{2}}$.

<u>Asymptotic normality.</u> When $p_i(\theta)$ are continuous and (as happens in regular cases) $n^{\frac{1}{2}}(p_i(\theta_n) - p_i) = O_p(1)$ under G, expansion of $X^2(\theta_n)$ in Taylor series shows that

$$(2.6) \qquad n^{\frac{1}{2}}(X^2(\theta_n)/n-d) = 2\sum_{i=1}^{k} \frac{\pi_i}{p_i}\, n^{\frac{1}{2}}(N_i/n - \pi_i)$$

$$- \sum_{i=1}^{k} \left(\frac{\pi_i}{p_i}\right)^2 n^{\frac{1}{2}}(p_i(\theta_n) - p_i) + o_p(1).$$

When $p_i(\theta)$ are differentiable and an appropriate expansion of $n^{\frac{1}{2}}(\theta_n - \theta_0)$ exists (see Huber (1967) for such expansions in the MLE case), then (2.5) will imply asymptotic normality of $X^2(\theta_n)/n$. We can then use $X^2(\theta_n)/n$ to obtain approximate confidence intervals for d, as well as for point estimation.

In general, the variance of the normal limiting law is so complex as to defeat use. But if θ_n is the minimum chi-square estimator and

$p_i(\theta)$ are continuously differentiable at θ_0, then θ_0 satisfies (2.4) and expansion of $p_i(\theta)$ about θ_0 shows that the second term on the right in (2.6) is $o_p(1)$. It follows at once from asymptotic normality of the N_i that under G,

$$
\begin{aligned}
& n^{\frac{1}{2}}(X^2(\theta_n)/n - d) \xrightarrow{\mathcal{D}} N(0,\tau^2) \\
(2.7) \qquad & \\
& \tau^2 = 4\{\sum_{i=1}^{k} \pi_i^3/p_i^2 - (\sum_{i=1}^{k} \pi_i^2/p_i)^2\}.
\end{aligned}
$$

In the no-estimation case of testing fit to $F(\cdot|\theta_0)$ with known θ_0, (2.7) was obtained in another context by Broffitt and Randles (1977). Since π_i and p_i can be estimated by N_i/n and $p_i(\theta_n)$, τ^2 in (2.7) can easily be estimated to obtain approximate confidence intervals for d. Note that (2.7) does not hold for θ_n other than the minimum chi-squared estimator, even for θ_n (such as the grouped-data MLE) asymptotically equivalent under H_0.

3. General Statistics of Chi-Squared Type

We will consider two general classes of statistics for testing fit from the cell frequencies $N_1,\ldots,N_k$. The Pearson statistic is the only common member of these classes. The first class was introduced by Cressie and Read (1983), with full detail given by Read (1982a). If $p = (p_1,\ldots,p_k)$ and $q = (q_1,\ldots,q_k)$ are probability distributions on k points, define for $-\infty < \lambda < \infty$

$$I^{\lambda}(p:q) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} p_i[(p_i/q_i)^{\lambda} - 1]$$

to be the directed divergence of p from q of order λ. For $\lambda = -1, 0$ the divergence is defined by continuity in λ. Cressie and Read discuss the properties of I^{λ} and its relation to other divergences. Only for $\lambda = -1/2$ is I^{λ} a metric, the Hellinger or Matusita distance.

The Cressie-Read statistics for testing the fit of $X_1, \ldots, X_n$ to the family $\mathfrak{F}$ of distributions are

$$R^{\lambda}(\theta_n) = 2nI^{\lambda}(N/n : p(\theta_n))$$

where $N = (N_1, \ldots, N_k)$ is the vector of cell frequencies, $p(\theta) = (p_1(\theta), \ldots, p_k(\theta))$ the vector of cell probabilities under $F(\cdot|\theta)$, and $\theta_n = \theta_n(X_1, \ldots, X_n)$ an estimator of θ. This family includes the Pearson ($\lambda = 1$), Neyman modified chi-squared ($\lambda = -2$), log likelihood ratio ($\lambda = 0$), modified log likelihood ratio ($\lambda = -1$), and Freeman-Tukey ($\lambda = -1/2$) statistics. A useful choice of θ_n is the minimum discrepancy estimator, for which $R^{\lambda}(\theta_n) = \inf R^{\lambda}(\theta)$ over θ in Ω. Cressie and Read show that when $F(\cdot|\theta_0)$ is true and the regularity conditions of Birch (1964) hold:

(A) If θ_n is any estimator satisfying
$n^{\frac{1}{2}}(\theta_n - \theta_0) = O_p(1)$, then $R^{\lambda}(\theta_n) = X^2(\theta_n) + o_p(1)$.

(B) The minimum R^{λ} estimators for all λ share a common asymptotic expansion and are BAN estimators of θ.

(C) For θ_n any BAN estimator of θ,
$R^{\lambda}(\theta_n) \xrightarrow{\mathcal{L}} \chi^2_{k-m-1}$.

Cressie and Read establish many other results as well, but (A), (B), (C) illustrate the principle that the statistics $R^{\lambda}(\theta_n)$ share the behavior of the Pearson statistic $R^1(\theta_n) = X^2(\theta_n)$ under H_0: G in $\mathfrak{F}$.

This is also true under contiguous alternatives, but not under fixed G outside $\mathfrak{F}$.

It is often convenient in practice to employ data-dependent cells rather than fixed cells in tests of fit of chi-square type. This is done in the examples of Section 6. In this case, the cell frequencies N_i are no longer multinomial. But when the random cell boundaries (which are functions of the data $X_1,\ldots,X_n$) converge in probability as $n \to \infty$ to a set of nonrandom cell boundaries (which are generally functions of G, the cdf of the X_j), it can be shown in general that the asymptotic behavior of $X^2(\theta_n)$ based on the random cells is the same as if the limiting nonrandom cells had been employed. This is done by Moore and Spruill (1975) for rectangular cells (in particular for intervals when X_j are real), and by Pollard (1979) for cells of quite general shape. By combining Lemma 4.1 of Moore and Spruill (1975) with the work of Cressie and Read the following result can be obtained. <u>If conditions</u> (A1) - (A3) <u>of Moore and Spruill</u> (1975) <u>and the conditions of Birch</u> (1964) <u>hold, then</u> (A), (B), (C) <u>above remain true when</u> $R^\lambda(\theta_n)$ <u>is based on data-dependent cells.</u> The details of the proof are similar to arguments in Moore and Spruill, and will not be given. Asymptotic results under contiguous alternatives can be similarly extended.

Our concern in Sections 4 and 5 is with the behavior of θ_n and $R^\lambda(\theta_n)$ when G is not in $\mathfrak{F}$. It is easy to see that when (a) the cells are rectangles E_{in} whose vertices converge a.s. as $n \to \infty$ to the vertices of nonrandom cells E_i, (b) G is continuous at the vertices of the E_i, (c) the cell probabilities $p_{in}(\theta) = \int_{E_{in}} dF(x|\theta)$ are continuous in θ and the vertices of E_{in}, then Theorem 5.1 on pointwise convergence of $R^{\lambda}(\theta_n)/n$ remains true, the limit being the same as if cells E_i had been employed for all n. A similar statement holds for convergence in probability. Having made these remarks, we will not explicitly consider random cells in Sections 4 and 5.

The second class of generalizations of the Pearson statistic that we discuss is studied in detail by Moore and Spruill (1975). Let $V_n(\theta)$ be the k-vector of the standardized cell frequencies $[N_i - np_i(\theta)]/[np_i(\theta)]^{\frac{1}{2}}$, and let M_n be a (possibly random) sequence of k×k nonnegative definite matrices. The statistics are the quadratic forms

$$T^M(\theta_n) = V_n'(\theta_n)M_nV_n(\theta_n),$$

where θ_n is again an estimator of θ. The Pearson statistic is the sum of squares $X^2(\theta_n) = T^I(\theta_n)$ obtained when $M_n = I$, the k×k identity matrix. Statistics of form T^M are also measures of the discrepancy of N/n from $p(\theta_n)$; similar distance measures are widely used in statistics, e.g. in assessing influential cases in linear models, Beckman and Cook (1983), p. 140.

The statistics T^M are of interest because for a given estimation method θ_n, one can usually find matrices M_n such that $T^M(\theta_n) \xrightarrow{L} \chi^2_{k-1}$ under H_0. The proper M_n was obtained by Rao and Robson (1974) when θ_n is the MLE of θ from $X_1,\ldots,X_n$. Moore (1977) showed in general how to choose M_n so that the asymptotic null distribution of T^M is chi-squared with maximal degrees of freedom. The general method includes the Rao-Robson result, and when θ_n is the minimum chi-squared estimator gives $M_n = I$ with k-m-1 as the largest available degrees of freedom. LeCam. Mahan and Singh (1983) show that this choice of M_n for given θ_n has some asymptotic optimality properties.

Thus statistics T^M are usually chosen to obtain χ^2 asymptotic critical points, and also an asymptotic optimality property, after the estimator θ_n has been selected. In this case, θ_n is not the minimum-T^M estimator, and the minimum-T^M estimator is of little interest. The statistics $R^\lambda(\theta_n)$, on the other hand, do not even have θ-free limiting null distributions in general for estimators θ_n other than the minimum-R^λ estimators (all of which are asymptotically equivalent under H_0).

4. Asymptotic Behavior of Estimators

In order to discuss the behavior of measures of fit to parametric families, we must establish the behavior of estimators θ_n of θ in $F(\cdot|\theta)$ under G not in $\mathfrak{F}$. As was mentioned in Section 2, convergence $\theta_n \to \theta_0$ a.s. (G) is known for common classes of estimators, such as MLE's, that might be employed in the statistics $T^M(\theta_n)$, and the limit θ_0 can be identified. We require a similar result for θ_n the minimum-R^λ estimator. Because of the specific form of R^λ, a.s. convergence can be proved without the hard-to-verify compactness assumptions employed in the literature for MLE's. Parallel results for convergence in probability hold, both here and in Section 5, but will not be stated separately.

Suppose that $N = (N_1,\ldots,N_k)$ has the multinomial (n,π) distribution for $\pi = (\pi_1,\ldots,\pi_k)$. Define $Q_n(\theta) = I^\lambda(N/n\colon p(\theta))$ and $Q(\theta) = I^\lambda(\pi\colon p(\theta))$. In general, θ_n minimizing Q_n converges a.s. to θ_0 minimizing Q. Here is one such result.

(A1) There is a θ_0 in Ω such that $Q(\theta) > Q(\theta_0)$ for all $\theta \neq \theta_0$ in Ω.

(A2) For any $\delta > 0$, there is an $\varepsilon > 0$ such that

$$\inf_{|\theta-\theta_0|>\delta} Q(\theta) \geq Q(\theta_0) + \varepsilon.$$

Assumption A2 is implied by the common assumption (e.g. Birch) that the map $\theta \to p(\theta)$ is continuous and has a continuous inverse at θ_0.

Theorem 4.1. If all $\pi_i > 0$ and (A1), (A2) hold, then $\theta_n \to \theta_0$ a.s. for any sequence θ_n satisfying

$$(4.1) \qquad Q_n(\theta_n) - \inf_\theta Q_n(\theta) \to 0 \text{ a.s.}$$

Proof. The details of the proof differ slightly for various λ. We give the proof for $\lambda > 0$, the more difficult case. Note first that (A1) and $\pi_i > 0$ imply $p_i(\theta_0) > 0$. Next, if $(N_i/n)^{\lambda+1}/p_i(\theta)^\lambda \geq M$ for any i and θ, then $Q_n(\theta) \geq (M-1)/\lambda(\lambda+1)$. This with a.s. convergence of N_i/n to $\pi_i > 0$ implies that there is a $c > 0$ such that a.s. $p_i(\theta_n) \geq c$ eventually, for all i. Let

$$\Omega_c = \{\theta \text{ in } \Omega\text{: } p_i(\theta) \geq c \text{ for all i}\}.$$

We can assume θ_n in Ω_c. Now $Q_n(\theta_0) \geq \inf_\theta Q_n(\theta)$ with (4.1) implies that a.s.

$$Q(\theta_0) = \lim Q_n(\theta_0) \geq \limsup Q_n(\theta_n).$$

But

$$\sup_{\Omega_c} |Q_n(\theta) - Q(\theta)| \to 0 \text{ a.s.}$$

so that $\lim\sup Q_n(\theta_n) = \lim\sup Q(\theta_n)$. So, since $Q(\theta_0) \leq Q(\theta_n)$,

$$Q(\theta_0) \leq \lim\inf Q(\theta_n) \leq \lim\sup Q(\theta_n) \leq Q(\theta_0)$$

and therefore $Q(\theta_n) \to Q(\theta_0)$ a.s. A2 then implies that θ_n must eventually stay in the neighborhood $|\theta-\theta_0| < \delta$ for any $\delta > 0$, i.e., that $\theta_n \to \theta_0$ a.s.

Note that if $\pi = p(\theta_0)$ for some θ_0 in Ω, then $Q(\theta_0) = 0$ and we have proved a.s. consistency of θ_n satisfying (4.1) under (A_2). This is a much stronger consistency result than appears in Cressie and Read (1983) or Read (1983), where the emphasis is on asymptotic normality. In regular cases, $\inf_\theta Q_n(\theta)$ is actually attained at a point θ_n satisfying (for $\lambda \neq -1$)

$$(4.2) \qquad \sum_{i=1}^{k} \left(\frac{N_i/n}{p_i(\theta)}\right)^{\lambda+1} \frac{\partial p_i}{\partial \theta_j} = 0 \qquad j = 1,\ldots,m$$

and θ_0 satisfies

$$(4.3) \qquad \sum_{i=1}^{k} \left(\frac{\pi_i}{p_i(\theta)}\right)^{\lambda+1} \frac{\partial p_i}{\partial \theta_j} = 0 \qquad j = 1,\ldots,m.$$

The familiar equations (2.1), (2.3) are cases of (4.2), and (2.2), (2.4) are cases of (4.3). Under conditions stronger than those of Theorem 4.1, one can establish asymptotic normality of $n^{\frac{1}{2}}(\theta_n-\theta_0)$ under G not in $\mathcal{F}$ by following Huber's (1967) treatment of the MLE case.

5. Measuring Degree of Lack of Fit

Associated with statistics $R^{\lambda}(\theta_n)$ or $T^{M}(\theta_n)$ for testing the significance of lack of fit are natural measures $R^{\lambda}(\theta_n)/n$ and $T^{M}(\theta_n)/n$ for estimating the degree of lack of fit. Once pointwise convergence of θ_n under G is established, pointwise convergence of these measures follows at once.

Let $\pi = (\pi_1,\ldots,\pi_k)$ be the cell probabilities under G, $p(\theta) = (p_1(\theta),\ldots,p_k(\theta))$ the cell probabilities under $F(\cdot|\theta)$, and $p = p(\theta_0)$, where θ_0 is the limit under G of θ_n. Here are our assumptions.

(B1) $\theta_n \to \theta_0$ a.s. (G), $p(\theta)$ is continuous at θ_0, and $\pi_i > 0$, $p_i > 0$ for $i = 1,\ldots,k$.

For statistics T^M, we also require that $M_n \to M_0$ a.s. (G) for a nonrandom matrix M_0. Convergence of M_n under $F(\cdot|\theta)$ is required by the large sample theory of Moore and Spruill (1975), and convergence under general G is true in all practical examples, such as the Rao-Robson (1974) statistic. Finally, let $b = (b_1,\ldots,b_k)'$ with $b_i = (\pi_i-p_i)/p_i^{\frac{1}{2}}$, so that if d is as in Section 2, $d = b'b = 2I^1(\pi:p)$.

Theorem 5.1. If (B1) holds, then

$$\frac{R^{\lambda}(\theta_n)}{n} \to 2I^{\lambda}(\pi:p) \text{ a.s. (G)}$$

If in addition $M_n \to M_0$ a.s. (G), then

$$\frac{T^M(\theta_n)}{n} \to b'M_0b \text{ a.s. } (G).$$

<u>Proof.</u> The result for R^λ is immediate from continuity of I^λ in its arguments. That for T^M follows from the fact that $n^{-\frac{1}{2}} V_n(\theta_n) \to b$ a.s. (G).

The result (2.5) for the Pearson statistic is a case ($\lambda = 1$, $M_n = I$) of both parts of Theorem 5.1. When θ_n is the minimum-R^λ estimator and Theorem 4.1 applies, then

$$2I^\lambda(\pi:p) = \inf_\theta 2I^\lambda(\pi:p(\theta))$$

so that $R^\lambda(\theta_n)/n$ is a consistent estimator of the discrepancy of G from $\mathfrak{F}$. In the case of T^M, θ_n is typically not the minimum-T^M estimator. But $F(\cdot|\theta_0)$ is often the closest member of $\mathfrak{F}$ to G by some other measure, usually one based on the raw data rather than on grouped data. For example, if θ_n is the raw-data MLE, then its a.s. limit θ_0 satisfies

$$E_G[-\log f(X|\theta_0)] = \inf_\theta E_G[-\log f(X|\theta)]$$

for f a density function corresponding to F. Thus $F(\cdot|\theta_0)$ is the closest point in $\mathfrak{F}$ to G in the sense of entropy, and $T^M(\theta_n)/n$ estimates a grouped-data distance of G from $F(\cdot|\theta_0)$.

Standard arguments identify the limits in Theorem 5.1 as the approximate Bahadur slopes of the respective test statistics. Spruill (1976), who considers certain $T^M(\theta_n)$ statistics, gives as Lemma 1 a result that implies the following.

Theorem 5.2. *When statistics* $R^\lambda(\theta_n)$ *and* $T^M(\theta_n)$ *have* θ*-free limiting distributions under* H_0, *and the conclusions of Theorem* 5.1 *hold, the approximate Bahadur slopes at* G *are* $2I^\lambda(\pi:p)$ *and* $b'M_0b$, *respectively.*

Statistics in practical use usually have θ-free limiting null distributions (data-dependent cells can facilitate this, as in the example of Section 6). When this is not so, the limits in Theorem 5.1 are the approximate slopes of the statistics when H_0: G in $\mathcal{F}$ is replaced by the simple hypothesis $G(\cdot) = F(\cdot|\theta_0)$.

If it is desired to obtain the asymptotic distribution of R^λ/n and T^M/n, the following result can be employed. The proof is straightforward.

Theorem 5.3. *If* (B1) *holds and* $n^{\frac{1}{2}}(p(\theta_n)-p) = O_p(1)$, *then under* G

$$n^{\frac{1}{2}}\left[\frac{R^\lambda(\theta_n)}{n} - 2I^\lambda(\pi:p)\right] = \left[\frac{2}{\lambda(\lambda+1)} + \frac{2}{\lambda+1}\right]\sum_{i=1}^{k}\left(\frac{\pi_i}{p_i}\right) n^{\frac{1}{2}}\left(\frac{N_i}{n} - \pi_i\right)$$

$$- \frac{2}{\lambda+1}\sum_{i=1}^{k}\left(\frac{\pi_i}{p_i}\right)^{\lambda+1} n^{\frac{1}{2}}\left(p_i(\theta_n) - p_i\right) + o_p(1)$$

If in addition $M_n \to M_0(P_G)$, *then*

$$n^{\frac{1}{2}}\left[\frac{T^M(\theta_n)}{n} - b'M_0b\right] = 2b'M_0\{p_i^{-\frac{1}{2}}\, n^{\frac{1}{2}}\left(\frac{N_i}{n} - \pi_i\right)\}$$

$$- b'M_0\left\{\frac{\pi_i+p_i}{p_i^{3/2}}\, n^{\frac{1}{2}}\left(p_i(\theta_n) - p_i\right)\right\} + o_p(1)$$

In Theorem 5.3 only, $\{a_i\}$ denotes the k-vector with components a_i. In regular cases,

$$n^{\frac{1}{2}}(p(\theta_n) - p) = D_p(\theta_0)n^{\frac{1}{2}}(\theta_n-\theta_0) + o_p(1)$$

where D_p is the matrix of derivatives $\partial p_i/\partial\theta_j$, and $n^{\frac{1}{2}}(\theta_n-\theta_0)$ is asymptotically a sum of n iid r.v.'s with zero mean. Since $n^{\frac{1}{2}}(N/n - \pi)$ has a similar form, the central limit theorem with Theorem 5.3 establishes asymptotic normality of R^λ/n and T^M/n under G. When θ_n is the minimum-R^λ estimator and $D_p(\theta)$ is continuous at θ_0, then (4.3) shows that in regular cases the second term in the expansion of $R^\lambda(\theta_n)/n$ in Theorem 5.3 is $o_p(1)$. Results (2.6) and (2.7) for the Pearson statistic are cases of Theorem 5.3.

6. Example

In order to illustrate the use and limitations of measures of lack of fit, we must first choose a statistic from the broad classes considered. In this section we use the familiar Pearson statistic. Since the data sets are repeated measurements of physical constants, $\mathfrak{F}$ is taken to be the family of univariate normal distributions. For assessing the significance of lack of fit in this case, both simulations by Rao and Robson (1974) and asymptotic theory by LeCam et. al. (1983) give reason to prefer the Rao-Robson statistic. The Rao-Robson divergence measure $b'M_0b$ (see Theorem 5.1) is sufficiently more complex than the Pearson measure

$$(6.1) \qquad d(\pi:p) = \sum_{i=1}^{k} \frac{(\pi_i - p_i)^2}{p_i}$$

that we prefer the Pearson statistic for illustrative purposes.

It remains to choose the cells, including the number of cells k, and the estimator θ_n of the parameters μ and σ^2 of the normal family. In testing fit to a single distribution using the Pearson X^2, there are compelling reasons to use cells equiprobable under H_0: (a) the test is then unbiased and has an optimality property within this class of tests (Cohen and Sackrowtiz (1975)); (b) Mann and Wald (1942) establish a minimax-type optimality property; (c) computational work, e.g. Roscoe and Byars (1971) and Larntz (1978), shows that the chi-squared distribution is a more accurate approximation in equiprobable cases. Data-dependent cells having boundaries of the form $\bar{X}+c_i s$ ($\bar{X}$, s are the sample mean and standard deviation) allow cells equiprobable under the estimated normal distribution $N(\bar{X},s^2)$ to be used in testing fit to $\mathcal{F}$. The asymptotic properties of the statistic under H_0 are then identical to those obtained by employing cells equiprobable under $N(\mu_G,\sigma_G^2)$, where (μ_G,σ_G^2) are the mean and variance of the true cdf G.

Mann and Wald (1942) also found approximately optimal k in terms of the sample size n and desired significance level α. The optimum is very broad, and more accurate approximations by Schorr (1974) confirm that about half the Mann-Wald value is preferable. We recommend using k about half the Mann-Wald value for α = 0.05, that is, approximately $k = 2n^{2/5}$. (This is not an endorsement of α = 0.05, or any other fixed α, in tests of fit. The Mann-Wald k decreases with α, but overstates the optimum k, so a small α is appropriate in our guideline.) Now this choice of k is defensible for assessing significance, but the discrepancy measure $d(\pi:p)$ of (6.1) depends on k, so that d's for different k's are not comparable. In discussing the data sets, it is therefore convenient to also use a common k, k = 7 in our case.

Finally, we will estimate $\theta = (\mu,\sigma^2)$ by $\theta_n = (\bar{X},s^2)$, the raw-data MLE's. With the random cells as above, $\theta_n \to \theta_0 = (\mu_G,\sigma_G^2)$ a.s. (G), whether or not G is in $\mathfrak{F}$, and under H_0: G in $\mathfrak{F}$, the limiting null distribution of $X^2(\theta_n)$ does not depend on θ_0. This distribution is not chi-squared, but is that of a r.v. $\sum_{i=1}^{k-3} Z_i^2 + \lambda_1 Z_{k-2}^2 + \lambda_2 Z_{k-1}^2$, where the Z_i are iid (N(0,1) r.v.'s. The characteristic roots λ_i have simple expressions given on p. 345 of Watson (1957). This distribution is easily computable; the P-values below were obtained by the method of Section 4 of Moore (1971). Note that $p_i = p_i(\theta_0) = 1/k$ whether or not G is in $\mathfrak{F}$, so that $X^2(\theta_n)/n$ converges to

$$d(\pi:p) = k \sum_{i=1}^{k} (\pi_i - 1/k)^2$$

where π_i are the probabilities under G of k cells equiprobable under $N(\mu_G,\sigma_G^2)$.

Stigler (1977) presents several data sets from historically significant experiments in classical physics. We will consider the following:

Data Set 1: Cavendish's 1789 measurements of the mean density of the earth (n = 29).

Data Set 2: Michelson's 1879 measurements of the velocity of light (n = 100).

Data Set 3: Newcomb's 1882 measurements of the velocity of light ($n = 66$).

Data Set 4: Newcomb's data less one egregious outlier, which Newcomb himself eliminated ($n = 65$).

Data Set 5: Michelson's 1891 supplementary measurements of the velocity of light ($n = 23$).

Stigler also presents several sets of data from Short's 1763 determinations of the parallax of the sun. These data are not strictly speaking iid, and show puzzling variations in degree of nonnormality. We will not consider them here.

Stigler discusses these fascinating data in some detail. For his purposes, he breaks the larger data sets into groups of about $n = 20$. Under the heading "Are real data normal?" he examines 20 such groups collectively as potentially a sample of normal data sets. The emphasis here, on the other hand, is on comparing the degree of nonnormality of the individual data sets.

Table 1 displays the analysis. First, Data Sets 1-5 were analyzed using the number k of cells recommended for the sample sizes n. Comparison of sets 3 and 4 shows the effect of the outlier in Newcomb's data, both on the significance and the degree of nonnormality. Data Set 3 is not considered further. For direct comparison, Data Sets 1, 2, 4, 5 were next analyzed with $k = 7$, the number of cells appropriate for the smallest sets. In addition, for each of the sample sizes $n = 23, 29, 65, 100$ (matching Data Sets 1, 2, 4, 5), 1000 random samples

Table 1

Data Set	n	k	X^2	P-value	X^2/n
1	29	7	1.655	0.867	0.057
2	100	13	26.620	0.003	0.266
3	66	11	43.333	<0.0001	0.657
4	65	11	15.723	0.052	0.242
5	23	7	7.130	0.159	0.310
1	29	7	1.655	0.867	0.057
2	100	7	7.520	0.137	0.075
4	65	7	7.908	0.118	0.122
5	23	7	7.130	0.159	0.310
IMSL	100	7	4.336	0.429	0.043
IMSL	65	7	4.495	0.407	0.069
IMSL	29	7	4.341	0.428	0.150
IMSL	23	7	4.531	0.402	0.197

were generated using the normal random variable routine GGNML from the IMSL library. The X^2 column contains the mean value of the Pearson statistics from these samples.

Consider first the IMSL results. The limiting null distribution of $X^2(\theta_n)$ has expected value 4.469, which is closely matched by the sample means displayed in the X^2 column. Other sample moments also closely match those expected under the hypothesis of normality. The P-values given are for the mean X^2. Since the mean of the theoretical null distribution of X^2 exceeds its median, the P-value of the sample mean will be <0.5. The X^2/n entries for the IMSL samples show the convergence to zero with increasing n that is expected under normality. The IMSL samples are quite closely normal, and provide a standard of comparison for the sets of real data.

Turning to the real data sets, note first the considerable effect of the choice of k on both the significance and the degree of nonnormality for Data Sets 2 and 4. The strength of this effect on significance is sometimes overlooked by users of chi-squared tests; it argues for an "objective" choice of k and p_i such as that discussed above and employed in the first portion of Table 1. For comparing degree of nonnormality, a common k is needed. The Cavendish data fit the normal family very well, better than even the average IMSL sample of the same size. Stigler observes that Cavendish's measurements with a torsion balance are considered "an ideal example of scientific experimentation." It is not surprising that his data are closer to normality than those of Michelson and Newcomb, who reflected light between a rotating mirror and a fixed mirror 600 to several thousand meters distant.

The Michelson and Newcomb data sets of sizes 23, 65 and 100 show a pattern similar to that of the IMSL samples: X^2 and its P-value are quite stable, and therefore X^2/n decreases with increasing n. The discrepancy X^2/n is in each case between 1.5 and 2 times that of the corresponding IMSL mean result. Some of this discrepancy may be due to positive dependence among the observations, which tends to inflate X^2. Michelson's Data Set 2 in particular shows long runs of similar values. Although the velocity of light experiments used similar apparatus, they of course differed in many details, and in fact Michelson's data are velocities while Newcomb's are passage times. The stability of X^2 for Data Sets 2, 4, 5 is remarkable. (This is not an artifact of the choice $k = 7$. For example, with $k = 11$ Data Set 2 yields $X^2 = 13.520$, again close to Data Set 4 for this k.)

Of the data sets considered, only Data Set 3 is distinctly non-normal. Data Set 1 appears very close to normality. The velocity of light data, while not as close to the normal family as the supposedly normal IMSL samples, show collectively behavior that suggests convergence of X^2/n to a small value. We had a priori expected these data to show significant nonnormality, resulting in P-values decreasing with n, and perhaps stable X^2/n. This pattern appears in simulations with nonnormal distributions. As Stigler observes, the measurements of Newcomb and Michelson are pioneering work with novel apparatus, and might be expected to be less regular than more routine series of laboratory measurements. We have found no evidence against the common assumption that series of careful measurements are normally distributed.

Acknowledgment. I am grateful to Regina Becker for assistance with the computing required for Section 6.

References

Beckman, R. J. and R. D. Cook (1983). Outlier........s. Technometrics 25, 119-149.

Birch, M.W. (1964). A new proof of the Pearson-Fisher Theorem. Ann. Math. Statist. 35, 817-824.

Bishop, Y.M.M., S.E. Fienberg and P.W. Holland (1975). Discrete Multivariate Analysis. MIT Press, Cambridge.

Broffit, J.D. and R.H. Randles (1977). A power approximation for the chi-square goodness of fit test: simple hypothesis case. J. Amer. Statist. Assoc. 72, 604-607.

Cohen, A. and H.B. Sackrowitz (1975). Unbiasedness of the chi-square, likelihood ratio and other goodness of fit tests for the equal cell case. Ann. Statist. 3, 959-964.

Cressie, N. and T.R.C. Read (1983). Multinomial goodness-of-fit tests. J. Royal Statist. Soc. Ser. B, to appear.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, 2nd Ed. Wiley, New York.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, 221-233.

Larntz, K. (1978). Small sample comparisons of exact levels for chi-squared goodness of fit statistics. J. Amer. Statist. Assoc. 73, 253-263.

LeCam, L., C. Mahan and A. Singh (1983). An extension of a theorem of H. Chernoff and E.L. Lehmann. In M.H. Rizvi, J.S. Rustagi and D. Siegmund, Eds., Recent Advances in Statistics: Papers in Honor of Herman Chernoff, Academic Press, New York, 303-337.

Mann, H. and A. Wald (1942). On the choice of the number of class intervals in the application of the chi-square test. Ann. Math. Statist. 13, 306-317.

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

Moore, D.S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. J. Amer. Statist. Assoc. 72, 131-137.

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

Perlman, M.D. (1972). On the strong consistency of approximate maximum likelihood estimators. Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1, 263-281.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimators. Metrika 14, 249-272.

Pollard, D. (1979). General chi-square goodness-of-fit tests with data-dependent cells. Z. Wahr. verw. Geb. 50, 317-331.

Rao, K. C. and D. R. Robson (1974). A chi-square statistic for goodness-of-fit within the exponential family. Comm. Statist. 3, 1139-1153.

Read, T.R.C. (1982a). On choosing a goodness-of-fit test. Ph.D. thesis, Flinders University of South Australia.

Read, T.R.C. (1982b). Small sample comparisons for the power divergence goodness-of-fit statistics. J. Amer. Statist. Assoc., submitted.

Read, T.R.C. (1983). Minimum distance parameter estimation for the multinomial model. Canadian J. Statist., submitted.

Roscoe, J.T. and J.A. Byars (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. J. Amer. Statist. Assoc. 66, 755-759.

Schorr, B. (1974). On the choice of the class intervals in the application of the chi-square test of goodness of fit. Math. Operations Forsch. u. Statist. 5, 357-377.

Spruill, M.C. (1976). Cell selection in the Chernoff-Lehmann chi-square statistic. Ann. Statist. 4, 375-383.

Stigler, S.M. (1977). Do robust estimators work with real data? Ann. Statist. 5, 1055-1078.

Ylvisaker, D. (1977). Test resistance. J. Amer. Statist. Assoc. 72, 551-556.