Z. Wahrscheinlichkeitstheorie verw. Gebiete
50, 317 – 331 (1979)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

General Chi-square Goodness-of-fit Tests with Data-dependent Cells

David Pollard*

Dept. of Statistics, Yale University, Box 2179 Yale Station, NewHaven, CT06520, USA

Dedicated to Leopold Schmetterer, on his sixtieth birthday

Summary. The goodness-of-fit of a parametric model for non-categorical data can be tested using the χ^2 statistic calculated after grouping the data into a finite number of disjoint cells. Work of Watson, Čebyšev, Moore and others shows that the classical limit distributions still hold even for certain methods of grouping that depend on the data themselves. These results are generalised to cover a much wider class of methods of grouping; the parameters can be estimated from either the grouped or the ungrouped data. The proofs use a Central Limit Theorem for Empirical Measures due to Dudley. The grouping cells are allowed to come from any Donsker class for the underlying sampling distribution.

§ 1. Introduction

Although the χ^2 goodness-of-fit test is formally a method for testing a hypothesis specifying the cell probabilities of a multinomial distribution, it can also be applied to non-categorical data. For example, to test whether a sample of independent observations could have come from some specified distribution P on the real line, a partition of the line into disjoint sets $C_1, \dots, C_k$ can be used to group the data into categories. If these sets were determined independently of the data, and if they were held fixed throughout the sampling, the problem would be reduced to testing the fit of a multinomial distribution with cell probabilities $P(C_1), \dots, P(C_k)$. If the specification of P involved unknown parameters, which were estimated *from the grouped data*, the classical theory for the χ^2 test would be applicable.

Watson (1959) argued, however, that in practice the C_i's are not fixed: their choice is to some extent dependent on the data. For certain types of data-dependence, he outlined a proof showing that, when s parameters are estimated from the grouped data, the test statistic still has a limiting χ^2_{k-s-1} distribution.

* Most of the work for this paper was carried out with the support of a fellowship of the Alexander von Humboldt Foundation at the Ruhr-Universität Bochum

Another problem aries with the estimation of the unknown parameters from the grouped data. It is usually simpler to construct estimates from the ungrouped data. Chernoff and Lehmann (1954) proved that, for fixed cells, these estimators lead to a limiting distribution of the form $\chi^2_{k-s-1}+\lambda_1\chi^2_1+\lambda_2\chi^2_1+\dots+\lambda_s\chi^2_1$: a linear combination of independent χ^2 variates, with the weights λ_i lying between zero and one. In general, these λ_i's will depend on the true value of the unknown parameter. Watson (1957, 1958, 1959) and Roy (unpublished work described by Watson (1959)) considered the same problem with data-dependent cells: the limit distribution is the same. This extension is significant if the unknown parameters are those of scale and location, for then the data-dependent cells can be chosen to ensure that the λ_i's depend only on the shape of the family (for example, normal or exponential), and not on the values of the unknown parameters. For a rigorous derivation of these results - using the theory of weak convergence - see the papers by Čebyšev (1971), Moore (1971), and Moore and Spruill (1975).

The cells treated by Čebyšev were random intervals of the real line; Moore (and Moore and Spruill) allowed rectangles in $\mathbb{R}^m$ with edges parallel to the coordinate axes for use with multivariate data. Recently Moore and Stubblebine (1978) have proposed using data-dependent cells whose boundaries are hyperellipses in a test for multivariate normality; but their cited references do not seem to provide the necessary theory for such cells. Indeed all of the results to date appear tied to the use of rectangular cells, or cells that might be transformed into rectangles - the available theory for the weak convergence of empirical distribution functions can be applied for rectangular cells. With a very general form of Central Limit Theorem for empirical measures due to Dudley (1978) these limitations on the cell shape can be removed.

Dudley's results will be applied in this paper in considering χ^2 tests with data-dependent cells that are not necessarily rectangular. The analogues of the results just discussed, where estimates are based on the ungrouped data, carry over readily (Sect. 5). The related problem where the estimators are calculated from the grouped data is also solved. To avoid overburdening the proof with details that might obscure the main idea, I consider first (Sect. 2) the case where the cells are fixed. The method is a refinement of an approach due to Birch (1964). The proof for data-dependent cells is spread between two sections: precise formulation of assumptions and results in Sect. 3, and details of the proof in Sect. 4.

It is assumed that the data are grouped into (possibly data dependent) cells $\Gamma_{N1},\dots,\Gamma_{Nk}$ that converge in an appropriate sense to fixed cells $\Gamma_1,\Gamma_2,\dots,\Gamma_k$ - the assumption usually made in the literature. The proof follows the usual plan of showing that calculation of the χ^2 statistics with the cells Γ_{Ni} is asymptotically equivalent to using the cells Γ_i, but this is done more for simplicity of exposition, rather than of necessity: as outlined in Sect. 6, the methods in this paper could be used to prove a more general result where no convergence of the cells Γ_{Ni} need be assumed. As the extension is straightforward, I have contented myself in this paper with the simpler result proved under unnecessarily restrictive conditions.

§ 2. The Goodness-of-fit Test with Fixed Cells

The χ^2 test has a long history. It was first proposed by Pearson (1900); but he incorrectly specified the limit distribution for the case where parameters must be estimated. The correct result was established by Fisher in a series of papers beginning with Fisher (1922a). Cramér (1946) gave a proof under precisely stated regularity conditions; but when the likelihood equation has multiple roots, his proof runs into difficulties with selection of a consistent root. Completely rigorous proofs are due to Birch (1964) and Rao (1965, § 5e). The argument to be developed in this section is adapted from the work of Birch and Rao, and from refinements of Birch's method due to Dudley (1976); the notation has been chosen to facilitate the generalisation to data dependent cells.

Independent observations are classified according to which of k disjoint cells $C_1, C_2, \ldots, C_k$ they fall into: write $p_N(i)$ for the proportion of the first N observations that lands in C_i. A model specifies the underlying cell probabilities $p(i, \boldsymbol{\theta})$ up to an unknown parameter $\boldsymbol{\theta}$, which ranges over a subset Θ of $\mathbb{R}^s$. For each $\boldsymbol{\theta}$, these cell probabilities sum to one. The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained by maximising $\sum N p_N(i) \log p(i, \boldsymbol{\theta})$. Equivalently we may choose an estimator $\tilde{\boldsymbol{\theta}}_N$ to minimise

$$L_N(\boldsymbol{\theta}) = 2N \sum_{i=1}^{k} p_N(i) \log [p_N(i)/p(i, \boldsymbol{\theta})].$$

Evaluating the function

$$X_N^2(\boldsymbol{\theta}) = N \sum_{i=1}^{k} [p_N(i) - p(i, \boldsymbol{\theta})]^2 / p(i, \boldsymbol{\theta})$$

at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_N$ gives us the usual χ^2 goodness-of-fit statistic. (As we shall see later, minimising L_N is asymptotically equivalent to minimising X_N^2.)

For ease of notation, form the cell frequencies $p_N(i)$ into a $k \times 1$ column vector $\mathbf{p}_N$, and the probabilities $p(i, \boldsymbol{\theta})$ into a vector $\mathbf{p}(\boldsymbol{\theta})$. Since $N\mathbf{p}_N$ has a multinomial distribution based on cell probabilities $\mathbf{p}(\boldsymbol{\theta}_0)$ – the true unknown value of $\boldsymbol{\theta}$ is denoted by $\boldsymbol{\theta}_0$ – the multivariate version of the Central Limit Theorem (Breiman 1968, p. 238) ensures asymptotic normality of $\sqrt{N}[\mathbf{p}_N - \mathbf{p}(\boldsymbol{\theta}_0)]$ as N tends to infinity. For a more precise statement, write $\mathbf{r}$ for the $k \times 1$ column vector with elements $\sqrt{p(i, \boldsymbol{\theta}_0)}$, and Λ for the $k \times k$ diagonal matrix $\operatorname{diag}[\sqrt{p(1, \boldsymbol{\theta}_0)}, \ldots, \sqrt{p(k, \boldsymbol{\theta}_0)}]$. Then

$$\text{MCLT:} \quad \mathbf{v}_N = \sqrt{N}(\mathbf{p}_N - \mathbf{p}(\boldsymbol{\theta}_0)) \xrightarrow{\mathscr{D}} N(\mathbf{0}, V),$$

where $V = \Lambda^2 - \mathbf{p}(\boldsymbol{\theta}_0)\,\mathbf{p}(\boldsymbol{\theta}_0)' = \Lambda(I_k - \mathbf{r}\mathbf{r}')\Lambda$. Writing the covariance matrix in this form emphasises the role to be played by the matrix $I_k - \mathbf{r}\mathbf{r}'$, which represents the projection onto the orthogonal complement of the one dimensional space spanned by the unit vector $\mathbf{r}$.

Birch's (1964) assumptions were:

A1: the true value $\boldsymbol{\theta}_0$ is an interior point of Θ;
A2: for every neighbourhood U of $\boldsymbol{\theta}_0$,

$$\inf\{\|\mathbf{p}(\boldsymbol{\theta})-\mathbf{p}(\boldsymbol{\theta}_0)\|: \boldsymbol{\theta}\notin U\}>0;$$

A3: each component of $\mathbf{p}(\boldsymbol{\theta}_0)$ is strictly positive;
A4: the map $\boldsymbol{\theta}\mapsto\mathbf{p}(\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0$, i.e. there exists a $k\times s$ matrix D for which

$$\mathbf{p}(\boldsymbol{\theta})=\mathbf{p}(\boldsymbol{\theta}_0)+D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|) \text{ near } \boldsymbol{\theta}_0;$$

A5: the matrix D has full rank, i.e. rank $(D)=s$.

To avoid minor measurability problems, let us add to this list:
A6: the map $\boldsymbol{\theta}\mapsto\mathbf{p}(\boldsymbol{\theta})$ is continuous.
In view of A4, this hardly reduces the generality of the theorem. The O_p, o_p notation of Mann and Wald will be used; the advantages of this notation have been detailed by Chernoff (1956).

Theorem 1. *Let $\tilde{\boldsymbol{\theta}}_N$ be any sequence of estimators satisfying $L_N(\tilde{\boldsymbol{\theta}}_N)=\inf\{L_N(\boldsymbol{\theta}): \boldsymbol{\theta}\in\Theta\}+o_p(1)$. Then, if assumptions* A1 *to* A6 *hold,*

$$X_N^2(\tilde{\boldsymbol{\theta}}_N)\xrightarrow{\mathscr{D}}\chi^2_{k-s-1}.$$

The first step of the proof will be to prove that $\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)-\mathbf{p}(\boldsymbol{\theta}_0)$ is of order $O_p(1/\sqrt{N})$: this will be deduced from MCLT and some inequalities (Lemma 1) undoubtedly well known in Information Theory. Assumption A2 will then imply that $\tilde{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0$ is of order $o_p(1)$. Inequalities derived using A4 and A5 (Lemma 2) will strengthen this to $O_p(1/\sqrt{N})$ so that only values of $\boldsymbol{\theta}$ in an $O_p(1/\sqrt{N})$ neighbourhood of $\boldsymbol{\theta}_0$ need be considered. Over such a neighbourhood, both $L_N(\cdot)$ and $X_N^2(\cdot)$ will be approximated by a simple quadratic $Q_N(\cdot)$ in $\boldsymbol{\theta}-\boldsymbol{\theta}_0$, whose minimum can be found explicitly. The Continuous Mapping Theorem will give the limiting distribution of this minimum.

The first of the lemmas gives bounds on, and approximations to, what is sometimes called the I-divergence (cf. Csiszár (1975)) between two distributions. If $\mathbf{x}$ and $\mathbf{y}$ are two $k\times 1$ vectors of non-negative real numbers summing to one, define $I(\mathbf{x},\mathbf{y})=\sum_{i=1}^{k} x_i\log(x_i/y_i)$, with $0\log(0/y)$ defined to be zero for every y. Then by consideration of the Taylor expansion of $x_i\log x_i$ about y_i, the inequality $\log(1+z)\leqq z$, and the Taylor expansion of $(1+\delta_i)\log(1+\delta_i)$ for $\delta_i=(x_i-y_i)/y_i$, one obtains the three results of the first lemma (cf. Birch (1964) or Rao (1965, §5e)).

Lemma 1.

(i) $\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\leqq I(\mathbf{x},\mathbf{y})$

(ii) $I(\mathbf{x},\mathbf{y})\leqq\sum(x_i-y_i)^2/y_i$

(iii) *for each positive ε, there exists a constant C (depending on ε and k) such that*

$$|I(\mathbf{x},\mathbf{y})-\tfrac{1}{2}\textstyle\sum(x_i-y_i)^2/y_i|\leqq C\,\|\mathbf{x}-\mathbf{y}\|^3$$

whenever $y_i\geqq\varepsilon$ *for* $i=1,2,\dots,k$.

Lemma 2. *There exist positive real numbers* δ, m *and* M *such that, whenever* $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta$:

(i) $m\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leqq\|\mathbf{p}(\boldsymbol{\theta})-\mathbf{p}(\boldsymbol{\theta}_0)\|$;

(ii) $\|\mathbf{p}(\boldsymbol{\theta})-\mathbf{p}(\boldsymbol{\theta}_0)\|\leqq M\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|$.

Proof. For (i) take $2m=\inf\{\|D\mathbf{t}\|:\|\mathbf{t}\|=1\}$, where D is the derivative matrix of A4. This is strictly positive because of A5. When $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|$ is small enough, $D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ dominates the error term in A4. The proof of (ii) is similar. □

Proof of Theorem 1. Apply Lemma 1 (i) with $\mathbf{x}=\mathbf{p}_N$ and $\mathbf{y}=\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)$ to get $N\|\mathbf{p}_N-\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)\|^2\leqq L_N(\tilde{\boldsymbol{\theta}}_N)$; then apply Lemma 1 (ii) with $\mathbf{x}=\mathbf{p}_N$ and $\mathbf{y}=\mathbf{p}(\boldsymbol{\theta}_0)$ to get $L_N(\boldsymbol{\theta}_0)\leqq 2X_N^2(\boldsymbol{\theta}_0)$. By MCLT, the difference $\mathbf{p}_N-\mathbf{p}(\boldsymbol{\theta}_0)$ is of order $O_p(1/\sqrt{N})$; together with A3 this guarantees that $X_N^2(\boldsymbol{\theta}_0)$ is of order $O_p(1)$. Thus $N\|\mathbf{p}_N-\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)\|^2\leqq L_N(\tilde{\boldsymbol{\theta}}_N)\leqq L_N(\tilde{\boldsymbol{\theta}}_0)+o_p(1)=O_p(1)$, which implies that $\mathbf{p}_N-\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)$ is of order $O_p(1/\sqrt{N})$. It follows that $\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)-\mathbf{p}(\boldsymbol{\theta}_0)=O_p(1/\sqrt{N})$. This is more than enough to prove that $\mathbb{P}\{\tilde{\boldsymbol{\theta}}_N\in U\}\to 1$ for any neighbourhood of $\boldsymbol{\theta}_0$. Apply this to the open ball with radius equal to the δ of Lemma 2. With probability tending to one, therefore, the inequality $\|\tilde{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0\|\leqq m^{-1}\|\mathbf{p}(\tilde{\boldsymbol{\theta}}_N)-\mathbf{p}(\boldsymbol{\theta}_0)\|$ will be satisfied. Consequently $\tilde{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0$ must be of order $O_p(1/\sqrt{N})$.

The next step in the argument involves showing that $X_N^2(\boldsymbol{\theta})$ and $L_N(\boldsymbol{\theta})$ are close for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$. This is a very old idea introduced by Fisher (1922b). Similarly, replacing $\mathbf{p}(\boldsymbol{\theta})$ by $\mathbf{p}(\boldsymbol{\theta}_0)+D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ in the numerator, and by $\mathbf{p}(\boldsymbol{\theta}_0)$ in the denominator, we obtain a quadratic function $Q_N(\cdot)$ of $\boldsymbol{\theta}-\boldsymbol{\theta}_0$ approximating $X_N^2(\cdot)$ near $\boldsymbol{\theta}_0$. To be able to deduce from these approximations that $X_N^2(\tilde{\boldsymbol{\theta}}_N)$, $L_N(\tilde{\boldsymbol{\theta}}_N)$ and the minimum of $Q_N(\cdot)$ are close, we need the approximation to hold uniformly well in a region containing $\tilde{\boldsymbol{\theta}}_N$ and the point where $Q_N(\cdot)$ is minimised. This requires closer attention to the remainder terms in the approximations.

Consider the random neighbourhood $B_N=\{\boldsymbol{\theta}\in\Theta\colon\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leqq\rho_N\}$ of $\boldsymbol{\theta}_0$. The radius ρ_N is a random variable of order $O_p(1/\sqrt{N})$ to be specified more precisely presently; for the moment we need only that $\rho_N\geqq\|\tilde{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0\|$.

From Lemma 2 (ii), $\sup\{\|\mathbf{p}(\boldsymbol{\theta})-\mathbf{p}(\boldsymbol{\theta}_0)\|:\boldsymbol{\theta}\in B_N\}=O_p(1/\sqrt{N})$. Thus, with probability tending to one, $p(i,\boldsymbol{\theta})$ will be uniformly bounded away from 0 throughout B_N; Lemma 1 (iii) with $x_i=p_N(i)$ and $y_i=p(i,\boldsymbol{\theta})$ therefore provides the bound

$$\sup\{|X_N^2(\boldsymbol{\theta})-L_N(\boldsymbol{\theta})|\colon\boldsymbol{\theta}\in B_N\}=O_p(1/\sqrt{N})=o_p(1).$$

Similarly, if the i^{th} component of the vector $D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ is written as $[D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]_i$, then

$$\sup\{|X_N^2(\boldsymbol{\theta})-N\sum_i(p_N(i)-p(i,\boldsymbol{\theta}_0)-[D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]_i)^2/p(i,\boldsymbol{\theta}_0)|\colon\boldsymbol{\theta}\in B_N\}=o_p(1).$$

This is easier to comprehend in matrix notation. Set

$$\begin{aligned} Q_N(\boldsymbol{\theta}) &= N[\mathbf{p}_N-\mathbf{p}(\boldsymbol{\theta}_0)-D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]'\Lambda^{-2}[\mathbf{p}_N-\mathbf{p}(\boldsymbol{\theta}_0)-D(\boldsymbol{\theta}-\boldsymbol{\theta}_0)] \\ &= \|\Lambda^{-1}\mathbf{v}_N-\Lambda^{-1}D\sqrt{N}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2. \end{aligned}$$

Regard $Q_N(\cdot)$ as a function whose domain is not just Θ, but the whole of $\mathbb{R}^s$. Then

$$\sup\{|X_N^2(\boldsymbol{\theta})-Q_N(\boldsymbol{\theta})|:\ \boldsymbol{\theta}\in B_N\}=o_p(1).$$

Now ρ_N can be specified. Let $\boldsymbol{\theta}_N^*$ be the value of $\boldsymbol{\theta}$ at which $Q_N(\cdot)$ achieves its global minimum $\|\Pi\Lambda^{-1}\mathbf{v}_N\|^2$, where Π denotes the projection onto the orthogonal complement of the column space of $\Lambda^{-1}D$. Because $\sqrt{N}(\boldsymbol{\theta}_N^*-\boldsymbol{\theta}_0)$ is a linear function of $\mathbf{v}_N$, it must be of order $O_p(1)$. Choose ρ_N as any $O_p(1/\sqrt{N})$ quantity bigger than both $\|\tilde{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0\|$ and $\|\boldsymbol{\theta}_N^*-\boldsymbol{\theta}_0\|$. As long as $\boldsymbol{\theta}_N^*\in\Theta$ - which will happen with probability tending to one, since $\boldsymbol{\theta}_0$ is an interior point of Θ - such a choice for ρ_N ensures that $Q_N(\cdot)$ achieves its global minimum at a point of B_N. Hence

$$\inf\{Q_N(\boldsymbol{\theta}):\ \boldsymbol{\theta}\in B_N\}=\|\Pi\Lambda^{-1}\mathbf{v}_N\|^2+o_p(1).$$

Over the random neighbourhood B_N, the functions $L_N(\cdot)$ *and* $Q_N(\cdot)$ lie within a band of uniform $o_p(1)$ width around $X_N^2(\cdot)$. Thus

$$\begin{aligned} X_N^2(\tilde{\boldsymbol{\theta}}_N) &= L_N(\tilde{\boldsymbol{\theta}}_N)+o_p(1) \\ &= \inf\{L_N(\boldsymbol{\theta}):\ \boldsymbol{\theta}\in B_N\}+o_p(1) \\ &= \inf\{Q_N(\boldsymbol{\theta}):\ \boldsymbol{\theta}\in B_N\}+o_p(1) \\ &= \|\Pi\Lambda^{-1}\mathbf{v}_N\|^2+o_p(1). \end{aligned}$$

The Continuous Mapping Theorem and MCLT give

$$X_N^2(\tilde{\boldsymbol{\theta}}_N)\xrightarrow{\mathscr{D}}\|\Pi\Lambda^{-1}\mathbf{v}\|^2$$

where $\Lambda^{-1}\mathbf{v}\sim N(\mathbf{0},\ I_k-\mathbf{r}\mathbf{r}')$. This limit distribution is the same as that of $\|\Pi(I_k-\mathbf{r}\mathbf{r}')\mathbf{Z}\|^2$, with $\mathbf{Z}$ a $N(\mathbf{0},I_k)$ variable. Since $\mathbf{r}'\Lambda^{-1}D=\mathbf{1}'D=\mathbf{0}'$, the matrix $\Pi(I_k-\mathbf{r}\mathbf{r}')$ represents the orthogonal projection onto the $k-s-1$ dimensional space orthogonal to the space spanned by $\mathbf{r}$ and the columns of $\Lambda^{-1}D$. It follows that $X_N^2(\tilde{\boldsymbol{\theta}}_N)$ has the desired χ^2_{k-s-1} limit distribution.

§ 3. Data-Dependent Cells: Notation and Statement of Theorem

The model prescribes a family $\{P(\cdot,\boldsymbol{\theta}):\ \boldsymbol{\theta}\in\Theta\}$ of probability measures on a measure space $(\mathscr{X},\mathscr{F})$; once again, Θ is a subset of $\mathbb{R}^s$. Independent observations are taken on the distribution $P(\cdot,\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is fixed but unknown. It is assumed that each $P(\cdot,\boldsymbol{\theta})$ is absolutely continuous with respect to a fixed σ-finite measure μ on $\mathscr{F}$. It will prove convenient to work with $\xi(\cdot,\boldsymbol{\theta})$, the square root of the density function, because these functions are elements of the Hilbert space $L^2(\mu)$. For example, to avoid measurability problems it will suffice to assume that the map $\boldsymbol{\theta} \to \xi(\cdot, \boldsymbol{\theta})$ from Θ into $L^2(\mu)$ is continuous.

As the analogue to A4, it will be assumed that ξ is differentiable in $L^2(\mu)$ mean at $\boldsymbol{\theta}_0$. That is, assume that there exists an $s \times 1$ column vector $\dot{\boldsymbol{\xi}}$ of functions in $L^2(\mu)$ such that the $L^2(\mu)$ norm of $\xi(\cdot, \boldsymbol{\theta}) - \xi(\cdot, \boldsymbol{\theta}_0) - \dot{\boldsymbol{\xi}}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is of order $o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|)$ near $\boldsymbol{\theta}_0$. By several applications of the Schwarz inequality, it can be shown (Example 2.3 of Pollard 1980) that the vector measure $\Delta(\cdot)$, defined by

$$\Delta(F) = 2 \int_F \dot{\boldsymbol{\xi}}(x)\, \xi(x, \boldsymbol{\theta}_0)\, \mu(dx),$$

plays the role of a derivative $\frac{\partial}{\partial \boldsymbol{\theta}} P(\cdot, \boldsymbol{\theta})$; that is

$$\sup\{|P(F, \boldsymbol{\theta}) - P(F, \boldsymbol{\theta}_0) - \Delta(F)'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)| : F \in \mathscr{F}\} = o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) \text{ near } \boldsymbol{\theta}_0.$$

The set $\mathscr{X}$ is to be partitioned into k disjoint cells - k is fixed throughout the paper - and a χ^2 test performed on the number of observations falling into these cells. The cells are to be chosen from a class $\mathscr{C} \subseteq \mathscr{F}$. Equip $\mathscr{C}$ with the topology generated by the $L^2(P(\cdot, \boldsymbol{\theta}_0))$ norm, and with the corresponding Borel structure. The partitions of $\mathscr{X}$ into $\mathscr{C}$ cells correspond to elements of the class $\mathscr{G} = \{\gamma \in \mathscr{C}^k : \gamma_1, \ldots, \gamma_k \text{ disjoint and } \bigcup_1^k \gamma_i = \mathscr{X}\}$; here $\gamma_1, \ldots, \gamma_k$ denote the components of γ. Equip $\mathscr{G}$ with its product topology and Borel structure. A partition of $\mathscr{X}$ into data-dependent cells $\Gamma_{N1}, \ldots, \Gamma_{Nk}$ determines a map $\boldsymbol{\Gamma}_N$ from the underlying probability space into $\mathscr{G}$. Call $\boldsymbol{\Gamma}_N$ a *random element* of $\mathscr{G}$ if it is a measurable map.

From the sample of size N on the distribution $P(\cdot, \boldsymbol{\theta}_0)$, an empirical measure $P_N(\cdot)$ can be constructed. For each fixed $\gamma \in \mathscr{G}$ there is a vector $P_N(\gamma)$ of cell frequencies; to the $\mathbf{p}_N$ of Sect. 2 corresponds the vector $P_N(\boldsymbol{\Gamma}_N)$. Similarly, as analogues of $X_N^2(\boldsymbol{\theta})$ and $L_N(\boldsymbol{\theta})$ we have

$$X_N^2(\gamma, \boldsymbol{\theta}) = N \sum_{i=1}^{k} [P_N(\gamma_i) - P(\gamma_i, \boldsymbol{\theta})]^2 / P(\gamma_i, \boldsymbol{\theta}),$$

$$L_N(\gamma, \boldsymbol{\theta}) = 2N \sum_{i=1}^{k} P_N(\gamma_i) \log [P_N(\gamma_i)/P(\gamma_i, \boldsymbol{\theta})].$$

The estimate $\tilde{\boldsymbol{\theta}}_N$ should be chosen to minimise $L_N(\boldsymbol{\Gamma}_N, \cdot)$; the desired asymptotic distribution for $X_N^2(\boldsymbol{\Gamma}_N, \tilde{\boldsymbol{\theta}}_N)$ is then χ^2_{k-s-1}. Precise conditions under which this holds will be given in Theorem 2 below.

Corresponding to the matrix D of A4 is the $k \times s$ matrix $D(\gamma)$, defined for each $\gamma \in \mathscr{G}$, having rows $\Delta(\gamma_1)', \ldots, \Delta(\gamma_k)'$. Norm differentiability of $\xi(\cdot, \boldsymbol{\theta})$ ensures the existence of a non-negative function $\alpha(\cdot)$, of order $o(1)$ near zero, for which

$$\sup\{\|P(\gamma, \boldsymbol{\theta}) - P(\gamma, \boldsymbol{\theta}_0) - D(\gamma)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| : \gamma \in \mathscr{G}\} \leqq \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \cdot \alpha(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|).$$

Without loss of generality, we may assume α to be continuous and strictly increasing; this ensures that it has a continuous inverse $\alpha^{-1}(\cdot)$. Since the random cells $\boldsymbol{\Gamma}_N$ will be assumed to converge in probability (in the sense of the topology on $\mathscr{G}$) to a set of fixed cells $\boldsymbol{\Gamma}\in\mathscr{G}$, that is $P(\Gamma_{Ni}\backslash\Gamma_i\cup\Gamma_i\backslash\Gamma_{Ni},\boldsymbol{\theta}_0)\xrightarrow{\mathbb{P}}0$, only the behaviour of $D(\cdot)$ near $\boldsymbol{\Gamma}$ will be of interest: by analogy with A5, we shall need $D(\boldsymbol{\Gamma})$ to be of full rank.

Finally, to replace MCLT, it will be necessary to assume that $v_N(\cdot)=\sqrt{N}(P_N(\cdot)-P(\cdot,\boldsymbol{\theta}_0))$, regarded as a random element of the function space $D_0(\mathscr{C},P(\cdot,\boldsymbol{\theta}_0))$ defined by Dudley (1978), converges in distribution to a Gaussian process $v(\cdot)$ in the sense of Dudley's Central Limit Theorem for Empirical Measures: it will be assumed that $\mathscr{C}$ is a Donsker class for $P(\cdot,\boldsymbol{\theta}_0)$. Theorems 5.1 and 5.7, and Proposition 7.12 of Dudley (1978) help to identify many Donsker classes relevant to χ^2 tests with random cells when $\mathscr{X}$ is an Euclidean space. For example, for any fixed m, the class of all sets expressible as intersections of at most m open or closed half-spaces is a Donsker class for any probability measure on any $\mathbb{R}^p$; regions generated as differences of hyperellipsoids have the same property – this is needed to complete the arguments of Moore and Stubblebine (1978). The Donsker class property will ensure that $v_N(\boldsymbol{\Gamma}_N)$ convergences in distribution to $v(\boldsymbol{\Gamma})$, and that $\sup\{\|v_N(\gamma)\|: \gamma\in\mathscr{G}\}=O_p(1)$.

Theorem 2. *Suppose that $\tilde{\boldsymbol{\theta}}_N$ is a consistent estimate for $\boldsymbol{\theta}$ satisfying*

$$L_N(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)=\inf\{L_N(\boldsymbol{\Gamma}_N,\boldsymbol{\theta}):\ \boldsymbol{\theta}\in\Theta\}+o_p(1)$$

where:

(a) *$\{\boldsymbol{\Gamma}_N\}$ is a sequence of random elements of $\mathscr{G}$ converging in probability to a fixed $\boldsymbol{\Gamma}\in\mathscr{G}$;*

(b) *the true value $\boldsymbol{\theta}_0$ is an interior point of Θ;*

(c) *each component of $P(\boldsymbol{\Gamma},\boldsymbol{\theta}_0)$ is positive;*

(d) *$\xi(\cdot,\boldsymbol{\theta})$, the square root of the density, is diffeeentiable in $L^2(\mu)$ mean at $\boldsymbol{\theta}_0$, with derived vector $\dot{\boldsymbol{\xi}}$;*

(e) *the $k\times s$ matrix $D(\boldsymbol{\Gamma})=2\int\xi(x,\boldsymbol{\theta}_0)\boldsymbol{\Gamma}\dot{\boldsymbol{\xi}}(x)'\mu(dx)$ has rank s;*

(f) *the class $\mathscr{C}$ from which the random cells are chosen is a Donsker class for $P(\cdot,\boldsymbol{\theta}_0)$.*

Then $X_N^2(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)\xrightarrow{\mathscr{D}}\chi^2_{k-s-1}$.

The integrand in (e) is the $k\times s$ matrix whose (i,j)th element is the product of the $L^2(\mu)$ *function* $\xi(\cdot,\boldsymbol{\theta}_0)$, the indicator function of Γ_i, and the $L^2(\mu)$ function that is the jth component of $\dot{\boldsymbol{\xi}}$; the elements of this matrix are therefore integrable with respect to μ. The differentiability condition (d) is borrowed from LeCam (1970), who showed that it is weaker than the differentiability conditions usually encountered in proofs of asymptotic normality of maximum likelihood estimators. It is easy to check that the convergence condition (a) is weaker than the assumptions of Watson (1959), or Moore and Spruill (1975). Consistency of $\tilde{\boldsymbol{\theta}}_N$ is assumed, in preference to imposing a messy condition analogous to A2. As the discussion in Sect. 5 will show, such consistency is an unimportant consideration in the application of the theorem; $\boldsymbol{\theta}_N$ would not be obtained by direct minimisation, anyway.

The proof of Theorem 2 follows closely on the pattern established for Theorem 1. In essence, the idea is that a maximum likelihood estimate $\tilde{\theta}_N(\gamma)$ could be calculated for each fixed $\gamma \in \mathscr{G}$ by minimising $L_N(\gamma, \cdot)$. Theorem 1 would show that $X_N^2(\gamma, \tilde{\theta}_N(\gamma)) \to \chi_{k-s-1}^2$ for each such γ satisfying the requirements of that theorem. Use of the estimate $\tilde{\theta}_N = \tilde{\theta}_N(\Gamma_N)$ should therefore be equivalent, in the limit, to mixing over a family of χ_{k-s-1}^2 distributions, provided Γ_N is asymptotically independent of $X_N^2(\Gamma_N, \theta_0)$ in some sense; convergence in probability of Γ_N is just one way of achieving this asymptotic independence (see Sect. 6). The details in the proof of Theorem 2 are aimed at showing that each of the approximation arguments of Sect. 2 can be made uniform with respect to γ in some shrinking neighbourhood of Γ.

The theorem could be extended to include asymptotic results under sequences of alternatives - results along the lines of those described by Moore and Spruill (1975) - by following the procedure sketched in Sect. 6 of Pollard (1980).

§ 4. Proof of Theorem 2

To simplify the proof slightly, we can assume that Γ_N takes values in

$$\mathscr{G}_0 = \{\gamma \in \mathscr{G}: \operatorname{rank} D(\gamma) = s \text{ and } P(\gamma_i, \theta_0) > 0 \text{ for each } i\}.$$

This is an open subset of $\mathscr{G}$ which, because of (c) and (e), contains Γ. Condition (a) would ensure that Γ_N lies in $\mathscr{G}_0$ with probability tending to one in any case.

The $\sqrt{N}$-consistency of $\tilde{\theta}_N$ is obtained from a uniform analogue of Lemma 2.

Lemma 3. *There exist continuous, positive functions $m(\cdot)$, $M(\cdot)$ and $\delta(\cdot)$ on $\mathscr{G}_0$ for which:*

$$\textit{if} \quad \|\theta - \theta_0\| < \delta(\gamma) \quad \textit{then}$$
$$m(\gamma)\|\theta - \theta_0\| \leqq \|P(\gamma, \theta) - P(\gamma, \theta_0)\| \leqq M(\gamma)\|\theta - \theta_0\|.$$

Proof. The matrix $D(\gamma)$ is a continuous function of γ, since the vector measure $\Delta(F) = 2\int_F \xi(x, \theta_0)\dot{\xi}(x)\,\mu(dx)$ is absolutely continuous with respect to $P(F, \theta_0) = \int_F \xi(x, \theta_0)^2\,\mu(dx)$. Thus the function

$$2m(\gamma) = \inf\{\|D(\gamma)\mathbf{t}\| : \|\mathbf{t}\| = 1\}$$

is continuous and positive on $\mathscr{G}_0$. Define $\delta(\gamma)$ to be $\alpha^{-1}(m(\gamma))$, with α as defined in Sect. 3.

If $\|\theta - \theta_0\| < \delta(\gamma)$ then $\alpha(\|\theta - \theta_0\|) < \mathrm{m}(\gamma)$, and so

$$\begin{aligned}\|P(\gamma, \theta) - P(\gamma, \theta_0)\| &\geqq \|D(\gamma)(\theta - \theta_0)\| - \|\theta - \theta_0\|\,\alpha(\|\theta - \theta_0\|)\\ &\geqq m(\gamma)\|\theta - \theta_0\|.\end{aligned}$$

The function $M(\cdot)$ can be defined by $M(\gamma) = \sup\{\|D(\gamma)\mathbf{t}\| : \|\mathbf{t}\| = 1\} + m(\gamma)$. □

As with Theorem 1, the proof begins with two applications of Lemma 1 to obtain the inequalities

$$\begin{aligned} N\|P_N(\boldsymbol{\Gamma}_N)-P(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)\|^2 &\leqq L_N(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)\\ &\leqq L_N(\boldsymbol{\Gamma}_N,\boldsymbol{\theta}_0)+o_p(1)\\ &\leqq 2X_N^2(\boldsymbol{\Gamma}_N,\boldsymbol{\theta}_0)+o_p(1). \end{aligned}$$

By conditions (a) and (c), $P(\Gamma_{Ni},\boldsymbol{\theta}_0)\xrightarrow{\mathbb{P}}P(\Gamma_i,\boldsymbol{\theta}_0)>0$; thus $P(\Gamma_{Ni},\boldsymbol{\theta}_0)^{-1}$ is $O_p(1)$ for each i. Each of the terms in the numerators of the summands is bounded by

$$\begin{aligned} &2\|v_N(\boldsymbol{\Gamma}_N)\|^2\\ &\leqq 2\sup\{\|v_N(\gamma)\|^2\colon \gamma\in\mathscr{G}_0\}\\ &\xrightarrow{\mathscr{D}} 2\sup\{\|v(\gamma)\|^2\colon \gamma\in\mathscr{G}_0\} \end{aligned}$$

by the Continuous Mapping Theorem. It follows that $\|\sqrt{N}[P_N(\boldsymbol{\Gamma}_N)-P(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)]\|^2$ is bounded by a quantity of order $O_p(1)$. Since $\sqrt{N}[P_N(\boldsymbol{\Gamma}_N)-P(\boldsymbol{\Gamma}_N,\boldsymbol{\theta}_0)]=v_N(\boldsymbol{\Gamma}_N)=O_p(1)$, we can conclude that $P(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)-P(\boldsymbol{\Gamma}_N,\boldsymbol{\theta}_0)=O_p(1/\sqrt{N})$. Continuity of $m(\cdot)$ and condition (a) imply that $m(\boldsymbol{\Gamma}_N)^{-1}=O_p(1)$. Lemma 3 completes the proof that $\tilde{\boldsymbol{\theta}}_N-\boldsymbol{\theta}_0=O_p(1/\sqrt{N})$.

Define the random neighbourhood B_N of $\boldsymbol{\theta}_0$ as in the proof of Theorem 1. The steps leading to the bound

$$\sup\{|X_N^2(\boldsymbol{\Gamma}_N,\boldsymbol{\theta})-L_N(\boldsymbol{\Gamma}_N,\boldsymbol{\theta})|\colon \boldsymbol{\theta}\in B_N\}=o_p(1)$$

then parallel those followed before.

For $\gamma\in\mathscr{G}_0$ define the $k\times k$ matrix $\Lambda(\gamma)=\operatorname{diag}[\sqrt{P(\gamma_1,\boldsymbol{\theta}_0)},\dots,\sqrt{P(\gamma_k,\boldsymbol{\theta}_0)}]$, and the quadratic $Q_N(\gamma,\boldsymbol{\theta})=\|\Lambda(\gamma)^{-1}v_N(\gamma)-\sqrt{N}\Lambda(\gamma)^{-1}D(\gamma)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|^2$. It is easy to show that $\Lambda(\cdot)$, and hence $\Lambda(\cdot)^{-1}D(\cdot)$, is continuous on $\mathscr{G}_0$. The matrix $\Pi(\gamma)$ for projection onto the orthogonal complement of the column space of $\Lambda(\gamma)^{-1}D(\gamma)$ is therefore continuous on $\mathscr{G}_0$ – use the full rank assumption built into the definition of $\mathscr{G}_0$. Once again the global minimum of $Q_N^2(\boldsymbol{\Gamma}_N,\cdot)$ occurs within $O_p(1/\sqrt{N})$ of $\boldsymbol{\theta}_0$; the radius of B_N can be chosen to take this into account. No essentially new features enter the argument from now until we reach the conclusion that

$$X_N^2(\boldsymbol{\Gamma}_N,\tilde{\boldsymbol{\theta}}_N)=\|\Pi(\boldsymbol{\Gamma}_N)\Lambda(\boldsymbol{\Gamma}_N)^{-1}v_N(\boldsymbol{\Gamma}_N)\|^2+o_p(1).$$

At that stage we need the full force of condition (f) to conclude that the right hand side of this last expression converges in distribution to $\|\Pi(\boldsymbol{\Gamma})\Lambda(\boldsymbol{\Gamma})^{-1}v(\boldsymbol{\Gamma})\|^2$. This follows by an application of the Continuous Mapping Theorem using the functional $h(z,\gamma)=\|\Pi(\gamma)\Lambda(\gamma)^{-1}z(\gamma)\|^2$, defined on $D_0(\mathscr{C},P(\cdot,\boldsymbol{\theta}_0))\otimes\mathscr{G}_0$, applied to the random elements $(v_N,\boldsymbol{\Gamma}_N)$ – the measurability difficulties associated with non-separability can be overcome as in the argument in Theorem 5.6 of Pollard (1980), in modifying Theorem 4.4 of Billingsley (1968) to prove that $(v_N,\boldsymbol{\Gamma}_N)\xrightarrow{\mathscr{D}}(v,\boldsymbol{\Gamma})$. The remainder of the proof then follows that for Theorem 1; we are essentially back in the situation of working with fixed cells $\boldsymbol{\Gamma}$.

§ 5. Calculation of $\tilde{\theta}_N$ – Analogues of the Chernoff-Lehmann Result

Direct calculation of the $\tilde{\boldsymbol{\theta}}_N$ to minimise $L_N(\boldsymbol{\Gamma}_N, \cdot)$ would usually prove troublesome – it would involve calculating $P(\boldsymbol{\Gamma}_N, \boldsymbol{\theta})$ for many different values of $\boldsymbol{\theta}$. One way of avoiding such problems is the Method of Scoring (Rao 1965, §5G) – an iterative procedure for calculating $\tilde{\boldsymbol{\theta}}_N$ starting with a preliminary $\sqrt{N}$-consistent estimate $\boldsymbol{\theta}_N^*$, such as the maximum likelihood estimator based on the ungrouped data. A similar method was suggested by Watson (1959, p. 451); it corresponds to the modified χ^2 statistics of Dzhaparidze and Nikulin (1974) and Dudley (1976). An equivalent geometrical way of viewing these prodedures, when applied to finding a $\tilde{\boldsymbol{\theta}}_N$ for general random cells, is described in this section. The theoretical justification for the assertions to be made is similar in detail to the proofs of Sect. 4; to avoid tedious repetitions, I leave these details to the reader.

The idea is to find a quadratic approximation to $X_N^2(\boldsymbol{\Gamma}_N, \tilde{\boldsymbol{\theta}}_N)$ by making Taylor expansions about $\boldsymbol{\theta}_N^*$, instead of about $\boldsymbol{\theta}_0$. This can be justified if $\xi(\boldsymbol{\theta})$ is continuously differentiable in $L^2(\mu)$ norm in a neighbourhood of $\boldsymbol{\theta}_0$, with the $L^2(\mu)$ norm of $\xi(\boldsymbol{\theta}+\mathbf{t})-\xi(\boldsymbol{\theta})-\dot{\boldsymbol{\xi}}(\boldsymbol{\theta})'\mathbf{t}$ uniformly of order $o(\mathbf{t})$ as $\mathbf{t}$ tends to zero. By analogy with Sect. 3, define matrices $D(\gamma, \boldsymbol{\theta})=2\int \xi(x, \boldsymbol{\theta})\gamma\dot{\boldsymbol{\xi}}(x, \boldsymbol{\theta})'\mu(dx)$ and $\Lambda(\gamma, \boldsymbol{\theta}) = \operatorname{diag}[\sqrt{P(\gamma_1, \boldsymbol{\theta})}, \ldots, \sqrt{P(\gamma_k, \boldsymbol{\theta})}]$. Notice that $D(\gamma, \boldsymbol{\theta}_0)=D(\gamma)$ and $\Lambda(\gamma, \boldsymbol{\theta})=\Lambda(\gamma)$. Then for values of $\boldsymbol{\theta}$ in any random neighbourhood of radius $O_p(1/\sqrt{N})$ about $\boldsymbol{\theta}_0$, both $X_N^2(\boldsymbol{\Gamma}_N, \boldsymbol{\theta})$ and $L_N(\boldsymbol{\Gamma}_N, \boldsymbol{\theta})$ can be uniformly approximated within $o_p(1)$ by

$$Q_N^*(\boldsymbol{\theta})=N\,\|\Lambda(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)^{-1}[P_N(\boldsymbol{\Gamma}_N)-P(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)-D(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)(\boldsymbol{\theta}-\boldsymbol{\theta}_N^*)\|^2.$$

The radius should be chosen so that, with probability tending to one, the random neighbourhood covers the region where $L_N(\boldsymbol{\Gamma}_N, \cdot)$, $X_N^2(\boldsymbol{\Gamma}_N, \cdot)$ and $Q_N^*(\cdot)$ come close to their infima. The test statistic is to be approximated by the global minimum of $Q_N^*(\cdot)$, which takes the form $Z_N^2(\boldsymbol{\theta}_N^*)=\|\Pi_N^*\Lambda(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)^{-1}\sqrt{N}[P_N(\boldsymbol{\Gamma}_N) - P(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)]\|^2$ where Π_N^* denotes the projection onto the orthogonal complement of the column space of $\Lambda(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)^{-1}D(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)$. This corresponds to a single iteration of the Method of Scoring to find a $\tilde{\boldsymbol{\theta}}_N$ to minimise $X_N^2(\boldsymbol{\Gamma}_N, \cdot)$, starting at $\boldsymbol{\theta}_N^*$ and using $Q_N^*(\cdot)$ as the approximation to $X_N^2(\boldsymbol{\Gamma}_N, \cdot)$. Since $Z_N^2(\boldsymbol{\theta}_N^*)$ can be calculated directly from the data, it is a convenient goodness-of-fit statistic. What is more, the methods of Sect. 4 show that $Z_N^2(\boldsymbol{\theta}_N^*)-\|\Pi(\boldsymbol{\Gamma}_N)\Lambda(\boldsymbol{\Gamma}_N)^{-1}\nu_N(\boldsymbol{\Gamma}_N)\|^2 = o_p(1)$; the statistic $Z_N^2(\boldsymbol{\theta}_N^*)$ does indeed have the desired χ_{k-s-1}^2 limit distribution.

The alternative measure of fit, $X_N^2(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)$, investigated by Watson (1958, 1959) and others (see Sect. 1) has the same form as $Z_N^2(\boldsymbol{\theta}_N^*)$, except that the projection matrix Π_N^* is omitted. The resulting asymptotic distribution need no longer be χ_{k-s-1}^2 – indeed, extra assumptions have to be made about $\boldsymbol{\theta}_N^*$ before an asymptotic distribution can be shown to exist. It is usually assumed (see, for example, Moore (1971)) that

$$\sqrt{N}(\boldsymbol{\theta}_N^*-\boldsymbol{\theta}_0)=J^{-1}(1/\sqrt{N})\sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}\log f(x_i, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}+o_p(1);$$

generally J is the information matrix. With such a choice for J, this asymptotic form may be written more compactly in our notation as

$$\sqrt{N}(\boldsymbol{\theta}_N^* - \boldsymbol{\theta}_0) = J^{-1} v_N(2\dot{\xi}/\xi(\boldsymbol{\theta}_0)) + o_p(1);$$

with J being the variance matrix of $2\dot{\xi}/\xi(\boldsymbol{\theta}_0)$ calculated under the distribution $P(\cdot, \boldsymbol{\theta}_0)$. That is

$$\begin{aligned} J &= \int 4\dot{\xi}\dot{\xi}'/\xi(\boldsymbol{\theta}_0)^2 \, dP(\cdot, \boldsymbol{\theta}_0) \\ &= 4 \int \dot{\xi}\dot{\xi}' \, d\mu, \end{aligned}$$

since

$$\begin{aligned} &\int 2\dot{\xi}/\xi(\boldsymbol{\theta}_0) \, dP(\cdot, \boldsymbol{\theta}_0) \\ &\quad = \int 2\dot{\xi}\xi(\boldsymbol{\theta}_0) \, d\mu \\ &\quad = 0. \end{aligned}$$

[As an element of $L^2(\mu)$, the function $\xi(\boldsymbol{\theta})$ is of constant length one; thus it is orthogonal to its derivative.]

We can therefore approximate

$$\begin{aligned} &\sqrt{N}[P_N(\boldsymbol{\Gamma}_N) - P(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)] \\ &\quad = v_N(\boldsymbol{\Gamma}_N) + \sqrt{N}[P(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_0) - P(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)] \end{aligned}$$

by

$$\begin{aligned} &v_N(\boldsymbol{\Gamma}_N) - D(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)\sqrt{N}(\boldsymbol{\theta}_N^* - \boldsymbol{\theta}_0) \\ &\quad = v_N(\boldsymbol{\Gamma}_N) - D(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*) J^{-1} v_N(2\dot{\xi}/\xi(\boldsymbol{\theta}_0)) + o_p(1). \end{aligned}$$

This gives the asymptotic form

$$X_N^2(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*) = \|\mathbf{h}_N\|^2 + o_p(1),$$

where

$$\mathbf{h}_N = A(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)^{-1}[v_N(\boldsymbol{\Gamma}_N) - D(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*) J^{-1} v_N(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))].$$

As Π_N^* projects orthogonal to the columns of $A(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)^{-1} D(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)$, we can decompose $X_N^2(\boldsymbol{\Gamma}_N, \boldsymbol{\theta}_N^*)$ as a sum

$$\begin{aligned} &\|\Pi_N^* \mathbf{h}_N\|^2 + \|(I_k - \Pi_N^*)\mathbf{h}_N\|^2 + o_p(1) \\ &\quad = Z_N^2(\boldsymbol{\theta}_N^*) + \|(I_k - \Pi_N^*)\mathbf{h}_N\|^2 + o_p(1). \end{aligned}$$

It is the middle term here that contributes the additional terms, which were identified by Chernoff and Lehmann (1954), in the limit distribution. The limiting joint behaviour of these components is obtained by the same method as that used at the end of Sect. 4.

The results of Dudley (1978) can be modified to show that

$$[v_N(\cdot), v_N(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))] \xrightarrow{\mathscr{D}} [v(\cdot), v(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))]$$

- this amounts to proving that $\mathscr{C}$ augmented by the components of $2\dot{\xi}/\xi(\boldsymbol{\theta}_0)$, which are all square integrable with respect to $P(\cdot, \boldsymbol{\theta}_0)$, forms a Donsker class (of functions). Since Π_N^* converges in probability to $\Pi(\Gamma)$, the projection orthogonal to the columns of $\Lambda(\Gamma)^{-1}D(\Gamma)$, a Continuous Mapping Theorem argument shows that

$$[\Pi_N^* \mathbf{h}_N, (I_k - \Pi_N^*)\mathbf{h}_N] \xrightarrow{\mathscr{D}} [\Pi(\Gamma)\mathbf{h}, (I_k - \Pi(\Gamma))\mathbf{h}]$$

where

$$\mathbf{h} = \Lambda(\Gamma)^{-1}[v(\Gamma) - D(\Gamma)J^{-1}v(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))].$$

As $v(\cdot)$ is a zero mean Gaussian process with covariance kernel $\operatorname{cov}[v(f_1), v(f_2)] = P(f_1 f_2, \boldsymbol{\theta}_0) - P(f_1, \boldsymbol{\theta}_0)P(f_2, \boldsymbol{\theta}_0)$ (see p. 900 of Dudley (1978)), the covariance structure of $\mathbf{h}$ can be determined explicitly. For this it is convenient to introduce the vector $\mathbf{g} = \Lambda(\Gamma)^{-1}D(\Gamma)J^{-1}v(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))$, because

$$\begin{aligned}\operatorname{cov}[\mathbf{h}, \mathbf{g}] &= \Lambda(\Gamma)^{-1}\operatorname{cov}[v(\Gamma), v(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))]J^{-1}D(\Gamma)'\Lambda(\Gamma)^{-1}\\ &\quad - \Lambda(\Gamma)^{-1}D(\Gamma)J^{-1}\operatorname{var}[v(2\dot{\xi}/\xi(\boldsymbol{\theta}_0))]J^{-1}D(\Gamma)'\Lambda(\Gamma)^{-1}\\ &= \Lambda(\Gamma)^{-1}D(\Gamma)J^{-1}D(\Gamma)'\Lambda(\Gamma)^{-1}\\ &\quad - \Lambda(\Gamma)^{-1}D(\Gamma)J^{-1}JJ^{-1}D(\Gamma)'\Lambda(\Gamma)^{-1}\\ &= 0.\end{aligned}$$

This means that $\mathbf{h}$ and $\mathbf{g}$ are independent, and

$$\begin{aligned}\operatorname{var}[\mathbf{h}] &= \operatorname{var}[\mathbf{h} + \mathbf{g}] - \operatorname{var}[\mathbf{g}]\\ &= I_k - \mathbf{r}\mathbf{r}' - \Lambda(\Gamma)^{-1}D(\Gamma)J^{-1}D(\Gamma)'\Lambda(\Gamma)^{-1}\end{aligned}$$

with $\mathbf{r}' = (\sqrt{P(\Gamma_1, \theta_0)}, \ldots, \sqrt{P(\Gamma_k, \theta_0)})$ as before. Knowing $\operatorname{var}[\mathbf{h}]$, we can deduce that $\operatorname{cov}[\Pi(\Gamma)\mathbf{h}, (I_k - \Pi(\Gamma))\mathbf{h}] = 0$, because $\Pi(\Gamma)$ commutes with $(I_k - \mathbf{r}\mathbf{r})$, and $\Pi(\Gamma)\Lambda(\Gamma)^{-1}D(\Gamma) = 0$. The limit distribution of $X_N^2(\Gamma_N, \theta_N^*)$ is therefore the sum of two independent components, one of which is χ^2_{k-s-1} distributed, the other the quadratic form $\|(I_k - \Pi(\Gamma))\mathbf{h}\|^2$ in the zero-mean normal vector $\mathbf{h}$. By a suitable rotation, this second component can be cast into the form $\sum \lambda_i \chi_1^2$, a linear combination of independent χ_1^2 variates. The λ_i's are the eigenvalues of the variance matrix of $(I_k - \Pi(\Gamma))\mathbf{h}$ which, using the expression for $\operatorname{var}[\mathbf{h}]$ just obtained, reduces to $I_k - \Pi(\Gamma) - \Lambda(\Gamma)^{-1}D(\Gamma)J^{-1}D(\Gamma)'\Lambda(\Gamma)^{-1}$. The eigenvalues are, therefore, all non-negative and less than the corresponding eigenvalues of $I_k - \Pi(\Gamma)$. As this last matrix represents a projection onto an s dimensional space, we have found a geometric interpretation for the result of Chernoff and Lehmann (1954): all except s of the eigenvalues λ_i are zero, and the non-zero eigenvalues all lie between zero and one.

§ 6. Convergence of the Random Cells is Unnecessary

The convergence in probability of Γ_N to Γ played two distinct roles. On the one hand, it ensured that quantities such as $m(\Gamma_N)^{-1}$ and $\Lambda(\Gamma_N)^{-1}$ were of order

$O_p(1)$, and that $Q_N(\boldsymbol{\Gamma}_N, \cdot)$ achieved its global minimum within some $O_p(1/\sqrt{N})$ neighbourhood of $\boldsymbol{\theta}_0$. This enabled the arguments of Sect. 2 to be carried over to show that $X_N^2(\boldsymbol{\Gamma}_N, \tilde{\boldsymbol{\theta}}_N) = \|\Pi(\boldsymbol{\Gamma}_N)\,\Lambda(\boldsymbol{\Gamma}_N)^{-1}\,v_N(\boldsymbol{\Gamma}_N)\|^2 + o_p(1)$. On the other hand, convergence of $\boldsymbol{\Gamma}_N$ was used to prove that $(v_N(\cdot), \boldsymbol{\Gamma}_N) \to (v(\cdot), \boldsymbol{\Gamma})$, in order that a Continuous Mapping Theorem would give the limiting distribution for $\Pi(\boldsymbol{\Gamma}_N)\,\Lambda(\boldsymbol{\Gamma}_N)^{-1}\,v_N(\boldsymbol{\Gamma}_N)$.

In its first role, convergence of $\boldsymbol{\Gamma}_N$ could have been replaced by a uniform tightness condition: for each $\varepsilon > 0$ there exists a compact subset $\mathscr{K}_\varepsilon$ of $\mathscr{G}_0$ such that $\mathbb{P}\{\boldsymbol{\Gamma}_N \in \mathscr{K}_\varepsilon\} > 1 - \varepsilon$ for all N large enough. Continuity on $\mathscr{G}_0$ of maps such as $\gamma \mapsto m(\gamma)^{-1}$ and $\gamma \mapsto \Pi(\gamma)$ would then have secured the required boundedness conditions.

In its second role, convergence of $\boldsymbol{\Gamma}_N$ to a fixed, non-random $\boldsymbol{\Gamma}$ imposed a form of asymptotic independence between $v_N(\cdot)$ and $\boldsymbol{\Gamma}_N$. Clearly, if $\boldsymbol{\Gamma}_N$ were actually independent of $v_N(\cdot)$, the conditional distribution $\Pi(\boldsymbol{\Gamma}_N)\,\Lambda(\boldsymbol{\Gamma}_N)^{-1}\,v_N(\boldsymbol{\Gamma}_N)|\boldsymbol{\Gamma}_N = \gamma$ could be analysed using the results of Sect. 2 for $\Pi(\gamma)\Lambda(\gamma)^{-1}\,v_N(\gamma)$; the limit distribution would come out as a mixture of χ^2_{k-s-1} distributions – which is again χ^2_{k-s-1}. (This can be proved by applying a result of Wichura (1970) to find a version $\tilde{v}_N(\cdot)$ of the empirical process that converges uniformly with respect to $\mathscr{C}$. The processes $\Pi(\gamma)\,\Lambda(\gamma)^{-1}\,\tilde{v}_N(\gamma)$ converge uniformly on compact subsets of $\mathscr{G}_0$. If the distributions of the $\boldsymbol{\Gamma}_N$'s are uniformly tight, Theorem 5.5 of Billingsley (1968) then justifies integrating out with respect to these distributions.) Asymptotic independence between $v_N(\cdot)$ and $\boldsymbol{\Gamma}_N$ has the same effect, at least when asymptotic independence is interpreted in the sense that $(v_N(\cdot), \boldsymbol{\Gamma}_N)$ behave asymptotically like the pair $(v(\cdot), \boldsymbol{\Gamma})$. Such an independence concept still makes sense even when $\boldsymbol{\Gamma}$ is not constant. Theorem 2 could be strengthened by requiring only that $\boldsymbol{\Gamma}_N$ converges in probability to a (possibly random) element of $\mathscr{G}_0$. The convergence in Dudley's (1978) Central Limit Theorem for Empirical Measures can be strengthened to convergence in the mixing sense of Rênyi (cf. the argument in Theorem 16.3 of Billingsley (1968)), from which it again follows (cf. Theorem 4.5 of Billingsley (1968)) that $(v_N(\cdot), \boldsymbol{\Gamma}_N) \xrightarrow{\mathscr{D}} (v(\cdot), \boldsymbol{\Gamma}^*)$, where $v(\cdot)$ and $\boldsymbol{\Gamma}^*$ are independent and $\boldsymbol{\Gamma}^*$ has the same distribution as $\boldsymbol{\Gamma}$. It would be more interesting though to find an asymptotic independence condition that put no convergence requirements on $\boldsymbol{\Gamma}_N$, but under which Theorem 2 continued to hold.

References

Billingsley, P.: Convergence of Probability Measures. New York: Wiley 1968

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

Breiman, L.: Probability. Reading, Mass.: Addison-Wesley 1968

Chernoff, H., Lehmann, E.L.: The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit. Ann. Math. Statist. **25**, 579–586 (1954)

Chernoff, H.: Large sample theory: parametric case. Ann. Math. Statist. **27**, 1–22 (1956)

Čebyšev, D.M.: Certain chi-square type tests for continuous distributions. Theor. Probability Appl. **16**, 1–22 (1971)

Cramér, H.: Mathematical Methods of Statistics. Princeton: Princeton Univ. Press 1946

Csiszár, I.: *I*-divergence geometry of probability distributions and minimization problems. Ann. Probability **3**, 146–158 (1975)

Dudley, R.M.: On χ^2 tests of composite hypotheses. (Unpublished manuscript) 1976

Dudley, R.M.: Central limit theorems for empirical measures. Ann. Probability **6**, 899–929 (1978)

Dzhaparidze, K.O., Nikulin, M.S.: On a modification of the standard statistics of Pearson. Theor. Probability Appl. **19**, 851–863 (1974)

Fisher, R.A.: On the interpretation of χ^2 from contingency tables, and the calculation of *P*. J. Roy. Statist. Soc. **85**, 87–94 (1922a)

Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc. London A **222**, 309–368 (1922b)

LeCam, L.: On the assumptions used to prove asymptotic normality of maximum likelihood estimates. Ann. Math. Statist. **41**, 802–828 (1970)

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

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

Moore, D.S., Stubblebine, J.B.: Chi-square tests for multivariate normality with application to common stock prices. Preprint: Purdue Univ. (1978)

Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. **50**, 157–175 (1900)

Pollard, D.: The minimum distance method of testing. Metrika (1980)

Rao, C.R.: Linear Statistical Inference and Its Applications. New York: Wiley 1965

Watson, G.S.: The χ^2 goodness-of-fit test for normal distributions. Biometrika **44**, 336–348 (1957)

Watson, G.S.: On chi-square goodness-of-fit tests for continuous distributions. J. Roy. Statist. Soc., Ser. B **20**, 44–61 (1958)

Watson, G.S.: Some recent results in chi-square goodness-of-fit tests. Biometrics **15**, 440–468 (1959)

Wichura, M.J.: On the construction of almost uniformly convergent random variables with given weakly convergent image laws. Ann. Math. Statist. **41**, 284–291 (1970)

Received March 14, 1979; In revised form August 7, 1979