[6] R. Curtain, *Stochastic evolution equations with general white noise disturbance*, Control Theory Centre Report 1975.
[7] I. I. Gihman, A. V. Skorohod, *Stochastic differential equations*, Berlin 1972.
[8] E. Hille, R. S. Philips, *Functional analysis and semigroups*, A. M. S., Providence, R. I. 1957.
[9] [R. Lipcer, A. Širyayev] Р. Липцер, А. Ширяев, *Статистика случайных процессов*, Москва 1974.
[10] J. L. Lions, *Quelques methodes de résolution des problémes aux limites non lineaires*, Dunod-Gauthier Villars, Paris 1969.
[11] H. McKean, *Stochastic integrals*, Academic Press, New York 1969.
[12] M. Metivier, G. Pistone, *Une formule d'isométrie pour l'intégrale stochastique hilbertienne et equations d'evolution lineaires stochastique*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 33 (1975).
[13] P. A. Meyer, *Probability and potentials*, 1966.
[14] E. Pardoux, *Non linear stochastic partial differential equations*, C. R. Acad. Sci. Paris, 1972.
[15] B. Rozowskii, *Stochastic differential equations in infinite dimensional spaces and filtering problems*, Proc. Druskininkai, Nov. 1974.
[16] R. B. Vinter, *A representation of solutions to stochastic delay equations*, Imperial College Report, 1975.
[17] J. Zabczyk, *On stability of infinite-dimensional linear stochastic systems*, this volume, pp. 273–281.

Presented to the Semester
Probability Theory
February 11–June 11, 1976

PROBABILITY THEORY
BANACH CENTER PUBLICATIONS, VOLUME 5
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1979

ON χ^2 TESTS OF COMPOSITE HYPOTHESES

R. M. DUDLEY

Department of Mathematics, Massachusetts Institute of Technology, Cambridge Mass., USA

For a χ^2 test with m cells and a composite hypothesis in an s-dimensional submanifold V, Birch [4] showed that simple differentiability of V suffices to give a limiting χ^2_{m-s-1} distribution. Dzhaparidze and Nikulin [14] introduced modified X^2 statistics for composite hypotheses, which are easier to compute than the classical ones. Here, these results are proved for topologically non-trivial manifolds (such as circles and spheres).

1. Introduction

For background on the χ^2 test, we refer to Cramér [11], Lancaster [18] and C. R. Rao [25]. Let $\mathscr{L}(X)$ denote the probability distribution or *law* of a random variable X. Let χ_d^2 denote a χ^2 variable with d degrees of freedom, i.e. $\mathscr{L}(\chi_d^2) = \mathscr{L}(G_1^2+\ldots+G_d^2)$ where G_i are independent standard normal variables. Let $N(m, C)$ denote a Gaussian (normal) distribution on $\boldsymbol{R}^d$ with mean vector m and covariance matrix C. The characteristic function of a χ^2 variable is given by

$$E\exp(it\chi_d^2) = (1-2it)^{-d/2} = f(t)^{-d}, \tag{1.1}$$

where $f(t) = (1-2it)^{1/2}$, using the continuous branch of the square root with positive real part.

Let S be a finite set with m elements, say $S = \{1, 2, \ldots, m\}$. In the applications, S often results from decomposing a more general space into m cells. Let p and q be probability measures on S, $p\{j\} = p_j$, $q\{j\} = q_j$, $j = 1, \ldots, m$, with $p_j > 0$ for all j.

Let $Y_1, Y_2, \ldots$ be i.i.d. (independent and identically distributed) with distribution q. Given n, let $n_j = n_j(\omega, n)$ be the number of values of $i \leqslant n$ such that $Y_i = j$. Let

$$X^2 := \sum_{1\leqslant j\leqslant m} (n_j-np_j)^2/np_j.$$

If $q = p$, the central limit theorem in $\boldsymbol{R}^m$ implies that $\mathscr{L}(X^2) \to \mathscr{L}(\chi_{m-1}^2)$ as $n \to \infty$ for the usual convergence of laws. Thus if q is unknown but Y_j can be observed, the hypothesis $q = p$ can be tested by the χ^2 test using the X^2 statistic.

K. Pearson [24] proposed the test, approximating $\Pr(X^2 \geqslant M)$ by $\Pr(\chi^2_{m-1} \geqslant M)$. Of most interest are values of M which give statistical "significance" at conventional levels, such as $\Pr(\chi^2_{m-1} \geqslant M) = .05$, .01 or .001.

The approximation of X^2 by χ^2 is considered adequate if $np_i \geqslant 5$ for all i (Cochran, [10]; Roscoe and Byars, [27]; Yarnold, [29]).

2. Some evidence on the X^2 approximation

We report here on one Monte Carlo experiment for the X^2 statistic with $n = 80$, $m = 16$, and $p_j = 1/16$ for $j = 1, \ldots, 16$. In 10,000 iterations, using a total of 800,000 pseudo-random numbers from 1 to 16, the results were as follows, where $N(X^2 \geqslant M)$ denotes the number of cases in which $X^2 \geqslant M$.

M	$10^4\Pr(\chi^2_{15} \geqslant M)$	$N(X^2 \geqslant M)$
22.3	1,000	966
25.0	500	458
27.5	250	245
30.6	100	102
32.8	50	56
37.7	10	11

The agreement is excellent. The statistic $\max_j n_j$, whose distribution in cases of significance is known (Doornbos and Prins, [13]) was used as a check on the randomness of the pseudo-random number generator, also with excellent results.

The agreement of the laws of X^2 and χ^2 seems much better than would be expected from known "Berry–Esséen" results on speeds of convergence in the central limit theorem. It is thus a challenge to probabilists to explore this approximation further, also in the more complicated case of composite hypotheses (Sections 4–6 below).

3. Some differential geometry

We will review some basic definitions and constructions.

Let X and Y be two Euclidean spaces. The usual norm on such a space will be written $|\cdot|$. A function f from an open set $U \subset X$ into Y is called *differentiable at* p with *derivative* $f'(p)$ iff $f'(p)$ is a linear map from X into Y such that

$$|f(x)-f(p)-f'(p)(x-p)|/|x-p| \to 0 \quad \text{as} \quad x \to p.$$

This $f'(p)$ is also called a *total* or *Fréchet* derivative and may be written $(Df)(p)$.

The chain rule holds: *if f maps U into an open set $V \subset Y$, g maps Y into another Euclidean space Z, $p \in U$, and $f'(p)$ and $g'\big(f(p)\big)$ exist, then*

$$(g \circ f)'(p) = g'\big(f(p)\big) \circ \big(f'(p)\big).$$

Let $L(X, Y)$ denote the set of all linear transformations from X into Y. Then $L(X, Y)$ is also a finite-dimensional real vector space, on which all norms are equivalent. When the derivatives exist, $f'(p) \in L(X, Y)$, $f''(p) \in L\big(X, L(X, Y)\big)$, etc. We say f is k *times continuously differentiable*, or C^k, iff $f^{(k)}$ exists and is continuous on U.

Given a topological space S, a *chart* or *local coordinate system* (U, f) is a homeomorphism f of a connected open set $U \subset S$ onto an open set in $\boldsymbol{R}^m$ for some m. Two charts (U, f) and (W, g) are said to be C^k-*related* iff $f(U)$ and $g(W)$ have the same dimension and on $g(U \cap W)$, $f \circ g^{-1}$ is C^k with $D(f \circ g^{-1})$ nonsingular. (If $U \cap W = \emptyset$, the condition is vacuously satisfied.)

A C^k *atlas* is a collection $\mathscr{A}$ of charts on S such that

(a) $\forall p \in S\ \exists (U, f) \in \mathscr{A}\colon\ p \in U$,

(b) any two charts in $\mathscr{A}$ are C^k related.

An atlas $\mathscr{A}$ will be called *complete* iff for every chart (U, f) which is C^k-related to each chart in $\mathscr{A}$, we have $(U, f) \in \mathscr{A}$.

A C^k *manifold* is a pair $(S, \mathscr{A})$ where S is a connected Hausdorff space and $\mathscr{A}$ is a complete atlas on S. We will assume $k \geqslant 1$ when writing C^k.

The *dimension* $\dim S$ of a manifold S is defined as the common dimension of the ranges of the charts.

Every atlas $\mathscr{A}$ is a subset of a unique complete atlas, namely the set of all charts C^k related to all charts in $\mathscr{A}$. On $\boldsymbol{R}^n$, the identity map I is a chart and $\{I\}$ is an atlas, so for each k, $(\boldsymbol{R}^n, \mathscr{A}_k)$ is a C^k manifold for some atlas $\mathscr{A}_k \supset \{I\}$. Any open subset of a C^k manifold V becomes a C^k manifold, using those charts whose domains are included in the subset. A C^k manifold is also a C^r manifold for any $r = 0, \ldots, k-1$, since a C^k atlas is also a C^r atlas (no longer complete, but we can complete it).

Given two C^k manifolds $(M, \mathscr{A})$ and $(N, \mathscr{B})$, and some $r = 0, 1, \ldots,$ or k, a mapping f from M into N is called C^r iff for any $p \in M$, chart $(U, g) \in \mathscr{A}$ with $p \in U$, and chart $(V, h) \in \mathscr{B}$ with $f(p) \in V$, $h \circ f \circ g^{-1}$ is C^r on its domain.

A *parametrized curve* is a C^1 function from an open interval in $\boldsymbol{R}$ into a manifold.

Given a C^k manifold V, $p \in V$, and two parametrized curves f and g with $f(0) = g(0) = p$, we say that "$f'(0) = g'(0)$" iff for some chart (U, h) with $p \in U$, $(h \circ f)'(0) = (h \circ g)'(0)$. Then if (V, j) is another chart with $p \in V$, $D(j \circ h^{-1})\big(h(p)\big)$ exists and by the chain rule, $(j \circ f)(0) = (j \circ g)'(0)$.

Thus the relation "$f'(0) = g'(0)$" is an equivalence relation. The equivalence class of f for this relation will be called $f'(0)$. The set of all such $f'(0)$ will be called the *tangent space* V_p to S at p. If $\dim S = s$, V_p can be made into an s-dimensional real vector space, since for any chart h, the set of all $(h \circ f)'(0)$ is such a vector space, and the structure is preserved by the linear isomorphisms $D(j \circ h^{-1})$.

For C^∞ manifolds, V_p can be defined as the set of all linear maps L from C^∞ functions on V into $\boldsymbol{R}$ such that $L(fg) = L(f)g(p)+f(p)L(g)$. Such a linear map L is called a *derivation at* p. For C^k functions with finite k, however, the linear space of derivations at p becomes infinite-dimensional (Newns and Walker, [22]) and so much larger than V_p.

Given a C^∞ map F from a manifold M into another one N, and a parametrized curve g in M, $F \circ g$ is a parametrized curve in N. This transformation of curves preserves the equivalence relation $f'(0) = g'(0)$, by the chain rule. Thus it defines a linear map dF from M_p into $N_{F(p)}$ for each p.

If $N = \mathbf{R}^1$, then each tangent space N_q can be identified with $\mathbf{R}^1$ in a natural way. Let $L(M_p, \mathbf{R}) := M_p^*$, the dual space of M_p, called the *cotangent space* at p. For $F\colon M \to \mathbf{R}$ and $p \in M$, $dF(p) \in M_p^*$. This differential dF is not "infinitesimal", although many of the classical differential formulas hold for it.

Let V be a manifold of dimension s. Let $(e_j)_i = \delta_{ji}$, $e_j \in \mathbf{R}^s$. Then for $p \in V$ and a chart (U, x) with $x(p) = 0$, and $j = 1, \ldots, s$, $t \to x^{-1}(te_j)$ is a C^1 curve in V. Its tangent vector at 0 is called $\partial/\partial x_j|_p$. We can write $g = G \circ x$ where G is a C^1 function on a neighborhood of 0 in $\mathbf{R}^s$. Then $\partial g/\partial x_j|_p = G_j(0)$, the usual partial derivative with respect to the jth coordinate. Also, if V is an open set in $\mathbf{R}^s$ with chart given by the identity (the usual coordinates), then $\partial/\partial x_j|_p$ has its usual meaning.

The tangent vectors $\partial/\partial x_j|_p$ form a basis of V_p. Since $dx_i(\partial/\partial x_j|_p) = (\partial x_i/\partial x_j)(p) = \delta_{ij}$, the differentials dx_i at p form a basis of V_p^*, for any chart x.

A C^k *Riemannian manifold* is a C^k manifold V together with a function $p \to B_p$ on V such that for each $p \in S$, B_p is a positive definite bilinear form (inner product) on $V_p \times V_p$, and for any chart (U, x), $p \to B_p(\partial/\partial x_i|_p, \partial/\partial x_j|_p)$ is C^k on U.

For a Riemannian manifold (V, B), there is a natural isomorphism i_p of V_p onto V_p^* for each p, where $i_p(v)(w) = B_p(v, w)$ for all $v, w \in V_p$. (This is the usual isomorphism of a Hilbert space with its dual space.) We have the dual inner product B_p^* on $V_p^* \times V_p^*$ such that

$$B_p^*(u, u)^{1/2} = \sup\{|u(v)|\colon v \in V_p, B_p(v, v) = 1\}.$$

Let M be a C^k manifold and V a subset. Then V is called a C^k *submanifold* of M iff V has a C^k manifold structure of dimension s such that the natural injection i of V into M is a homeomorphism and is C^k with di everywhere of full rank s. Then, a Riemannian structure on M induces one on V.

If V is a C^k submanifold of some $\mathbf{R}^m$, $k \geqslant 1$, then at each $v \in V$, V has a *tangent flat* $F_v \subset \mathbf{R}^m$ defined as follows. Let (U, x) be a chart with $v \in U$ and $y = x^{-1}$. Then if $\dim V = s$, y maps an open set in $\mathbf{R}^s$ into $\mathbf{R}^m$, and

$$F_v := \{v + y'(x(v))(u)\colon u \in \mathbf{R}^s\}.$$

Also, $dy|_{x(v)}$ maps $\mathbf{R}^s$ linearly onto V_v, and

$$u \to dy|_{x(v)}\big(y'(x(v))^{-1}(u-v)\big)$$

is a 1-1 affine map of F_v onto V_v taking v into 0.

For further information on differential geometry see e.g. Auslander and Mackenzie [2], Bishop and Crittenden [5], Bourbaki [6], or Dieudonné [12].

4. Chi-squared tests of composite hypotheses

Again let $S = \{1, \ldots, m\}$, and let

$$P_m = \Big\{\{p_j\}_{j=1}^m\colon p_j \geqslant 0\ \forall j \text{ and } \sum_{j=1}^m p_j = 1\Big\}.$$

Then P_m is an $(m-1)$-dimensional simplex in $\mathbf{R}^m$ and represents the set of laws on S. Given observed i.i.d. $Y_1, \ldots, Y_n \in S$ with unknown law $p \in P_m$, where $Y_i = j$ for n_j values of $i \leqslant n$, let $r_j := n_j/n$. Then $r := \{r_j\}_{j=1}^m \in P_m$.

To test the *composite hypothesis* that p belongs to some subset V of P_m, we first *estimate* the unknown $p = v \in V$ by a function $v(r)$. One method of estimation is to maximize, insofar as possible, the multinomial probability

$$n!\,v_1^{n_1} \ldots v_m^{n_m}/n_1! \ldots n_m!.$$

For given r, noting that some factors are constants, it is equivalent to maximize the "log likelihood function" defined by

$$L(r, v) := \sum_{j=1}^m r_j \ln v_j, \qquad v \in V.$$

Then, we find the X^2 statistic:

$$X_e^2 := n\sum_{j=1}^m \big(r_j - v_j(r)\big)^2/v_j(r) = \sum_{j=1}^m \big(n_j - nv_j(r)\big)^2/nv_j(r). \tag{4.1}$$

Here the subscript e on X^2 indicates that we used an *estimated* $v(r)$ rather than a fixed p.

DEFINITION. A function $r \to v(r)$ from P_m into V will be called a *maximum likelihood estimator* (MLE) iff

$$L\big(r, v(r)\big) = \sup_{v \in V} L(r, v)$$

whenever the sup is attained on V, and then we say an MLE exists.

Note. If the sup is not attained on V, $v(r)$ may be an arbitrary element of V. We will use an MLE only on the countable set of $r \in P_m$ with r_j rational $(= n_j/n)$. All subsets of this countable set, hence all functions on it, are measurable.

DEFINITION. A set $V \subset P_m$ will be called a *Birch s-submanifold* iff (a) $v_j > 0$ for all j and all $v \in V$, and (b) for each $p \in V$, V has a tangent flat of dimension s at p, i.e. there is neighborhood U of 0 in $\mathbf{R}^s$ and a homeomorphism w of U onto a neighborhood W of p in V with $w(0) = p$ such that $w'(0)$ exists and has full rank s. (Since $V \subset \mathbf{R}^m$, $w'(0)$ can be defined, as well as $dw(0)$.)

Here w^{-1} can be considered as a chart. If w' can be taken continuous, then V is a C^1 manifold.

In previous literature, it was assumed that $W = V$ (e.g. Birch, [4]), so that one chart covered V; U was generally called Θ. But, for V a circle, sphere etc. (e.g. Mardia, [20]) we need more than one chart. Also, the choice of a chart is often somewhat arbitrary.

(4.2) THEOREM. *Let V be a Birch s-submanifold of P_m. If Y_j are* i.i.d. (p) *with* $p \in V$, *then*

$$\Pr\{\omega\colon \text{an MLE } v(r(\omega)) \text{ exists}\} \to 1 \quad \textit{as} \quad n \to \infty,$$

and

$$\mathcal{L}(X_e^2) \to \mathcal{L}(\chi^2_{m-s-1}).$$

Note. The theorem is applied in practice for $s \leqslant m-2$ in order to have a non-trivial limit law.

Historical notes. R. A. Fisher (1924) first stated a theorem like (4.2), and gave a non-rigorous proof. A proof by H. Cramér [11], assuming that w is C^2, has been criticized for some unclarity about the choice of estimates. C. R. Rao [25] gave a proof where w is C^1. M. Birch [4] reduced the hypothesis to simple differentiability, which seems to be the weakest possible assumption in this direction. The proof below is a simplified version of Birch's proof.

5. Proof of Birch's Theorem

We set $0\cdot\ln(x/0) = 0$ for all x and $x\ln(x/0) = +\infty$ for $x > 0$.

(5.1) LEMMA. *For any x and $y \in [0, 1]$,*

$$x\ln(x/y) \geqslant x-y+\tfrac{1}{2}(x-y)^2.$$

Proof. If x or y is 0, the result holds by our conventions. If $x > 0 < y$, then Taylor's Theorem with remainder gives

$$x\ln x = y\ln y+(1+\ln y)(x-y)+(x-y)^2/2w$$

for some w between x and y. Thus $1/w \geqslant 1$ and the result follows. ■

(5.2) LEMMA. *For any $r, v \in P_m$,*

$$\sum_{j=1}^{m} r_j\ln(r_j/v_j) \geqslant \tfrac{1}{2}|r-v|^2 := \tfrac{1}{2}\sum_{j=1}^{m}(r_j-v_j)^2.$$

Proof. By (4.1), for each j, $r_j\ln(r_j/v_j) \geqslant r_j-v_j+\frac{1}{2}(r_j-v_j)^2$. Then summing over j gives the Lemma since $\sum r_j-v_j = 1-1 = 0$. ■

(5.3) LEMMA. *For $p \in V$ and empirical $r(\omega) = r = \{r_j\}_{j=1}^{m}$ for p,*

$$\Pr(\text{an MLE } v(r) \in V \text{ exists}) \to 1 \quad \textit{as} \quad n \to \infty.$$

As $r \to p$, $v(r) \to p$.

Proof. We have $r \to p$ by the law of large numbers. For r close enough to p,

$$\sup_{v\in V} L(r, v) = \sup_{v\in W} L(r, v) > \sup_{v\notin W} L(r, v),$$

where W is a neighborhood of p in V, by (5.2). Such a W can be taken to be compact, so the sup is attained. As $r \to p$, $W \to p$ so $v(r) \to p$. ■

For each $p \in P_m$ with $p_j > 0$ for all j, we define an inner product

$$(x, y)_p := \sum_{j=1}^{m} x_j y_j/p_j \quad \text{and} \quad |x|_p := (x, x)_p^{1/2}.$$

For a fixed p, or for p with all p_j bounded away from 0, there is a constant $M < \infty$ such that

$$|x|/M \leqslant |x|_p \leqslant M|x|.$$

Thus in a statement like $|x_n|_p = o|y_n|_p$ as $n \to \infty$, the p subscript makes no difference.

(5.4) LEMMA. *As $r \to p$ and $v \to p$, $v \in V$,*

$$-2\sum_{i=1}^{m} r_i\ln(v_i/r_i) = |r-v|_p^2+o(|r-p|^2+|v-p|^2).$$

Proof. Since by assumption $p_j > 0$ for all j, we can assume $v_j > 0$ and $r_j > 0$ for all j. Then by the proofs of (5.1) and (5.2),

$$-2\sum_{i=1}^{m} r_i\ln(v_i/r_i) = \sum_{i=1}^{m}(v_i-r_i)^2/w_i,$$

where w_i is between v_i and r_i. Then $1/w_i = 1/p_i+o(1)$. Now $(r_i-v_i)^2 \leqslant 2(r_i-p_i)^2+ +2(p_i-v_i)^2$ since for all x and y, $(x+y)^2 \leqslant 2x^2+2y^2$. Thus $|r-v|_p^2 = o(|r-p|^2+ +|p-v|^2)$ and (5.4) follows. ■

Now let F be the tangent flat to V at p. Then $F = \{p+w'(0)(u)\colon u \in \mathbf{R}^s\}$ where w is as in the definition of Birch submanifold. For $x \in \mathbf{R}^m$ let $f(x) \in F$ be such that

$$|x-f(x)|_p = \min\{|x-y|_p\colon y \in F\}.$$

In other words, f is the "orthogonal projection into F" for the $(\cdot, \cdot)_p$ inner product.

(5.5) LEMMA. *As $v \to p$, $v \in V$, $|v-f(v)| = o(|v-p|)$.*

Proof. This follows directly from the definitions. ■

(5.6) LEMMA. *As $r \to p$ and $v \to p$, $v \in V$,*

$$-2\sum_{i=1}^{m} r_i\ln(v_i/r_i) = |r-f(r)|_p^2+|f(r)-f(v)|_p^2+o(|r-p|^2+|v-p|^2).$$

Proof. We apply (5.5) to obtain in (5.4)

$$|r-f(v)|_p^2+o(|r-p|^2+|v-p|^2).$$

Then since $r-f(r)$ is perpendicular to $F-F$ for $(\cdot, \cdot)_p$,

$$|r-f(v)|_p^2 = |r-f(r)|_p^2+|f(r)-f(v)|_p^2. \ ■$$

(5.7) LEMMA. (a) *For r close enough to p,*

$$|v(r)-p|_p \leqslant 2|r-p|_p,$$

(b) *As $r \to p$, $|f(r)-f(v(r))| = o(|p-r|)$, and*

(c) $|v(r)-f(r)| = o(|p-r|)$.

Proof. By (5.3), $v(r)$ exists and converges to p as $r \to p$. Then $f(r) = p+w'(0)(u)$ for some $u = u(r) \to 0$ in $\mathbf{R}^s$ as in the definition of Birch submanifold, and $|w(u)- -f(r)| = o(|u|)$. Then $o(|u|) = o(|f(r)-p|) = o(|r-p|)$. Thus $|f(w(u))-f(r)| = o(|r-p|)$.

In (5.6) the left side is minimized at $v = v(r)$ by definition, so it must be smaller there than at $v = w(u(r))$, where $|f(r)-f(v)|^2$ can be included in the $o(\cdot)$ error term. Hence

$$(*)\qquad |f(r)-f(v(r))|^2 = o(|r-p|^2+|v(r)-p|^2).$$

If (a) fails, take a sequence $r_n = r \to p$ with $|p-v(r)|_p > 2|p-r|_p$. Then $|f(r)-$ $-f(v(r))| = o(|p-v(r)|)$ by $(*)$, and $|f(v(r))-v(r)| = o(|v(r)-p|)$ by (5.5). Then $|f(r)-v(r)| = o(|p-v(r)|)$, and $|v(r)-p|_p$ is asymptotic to $|f(r)-p|_p \leqslant |r-p|_p$ as $r \to p$, a contradiction. Thus (a) is proved. By $(*)$, (b) follows.

Next, $v|(r)-f(v(r))| = o(|v(r)-p|) = o(|r-p|)$ by (5.3) and (5.5). Combining gives (c). ■

(5.8) LEMMA. *As $r \to p$, $w \to p$, and $v \to p$, $v \in V$,*

$$\sum_{i=1}^{m} (r_i-v_i)^2/w_i = |r-f(v)|_p^2+o(|r-p|^2+|v-p|^2).$$

Proof. Since $v(r) \to p$ by (5.3), (5.8) gives $Y^2 = |r-f(v(r))|_p^2+o(|r-p|^2)$ in view of (5.7) (a). By (5.7) (b), then, we are done.

Proof of Birch's Theorem (4.2). By the central limit theorem in $\boldsymbol{R}^m$,

$$\mathscr{L}(\{n(r_j-p_j)/(np_j)^{1/2}\}_{j=1}^m) \to N(0,\, I-\{(p_ip_j)^{1/2}\}_{i,j=1}^m)$$

as $n \to \infty$. Thus $no(|p-r|^2) \to 0$ in probability as $n \to \infty$. Then by (5.9), $X_e^2 = nY^2$ has the same limit law as $n|r-f(r)|_p^2$.

Now for any $r \in P_m$,

$$r = (r-f(r))+(f(r)-p)+p,$$

where the three summands are all orthogonal for $(\cdot\,,\,\cdot)_p$. In fact, for any $x, y \in P_m$, $(x-y, p)_p = 0$, and for any $a, b \in F$, $(r-f(r), a-b)_p = 0$. Thus,

$$|r-p|_p^2 = |r-f(r)|_p^2+|f(r)-p|_p^2.$$

Hence $\mathscr{L}(n|r-p|_p^2) \to \mathscr{L}(\chi_{m-1}^2)$ as $n \to \infty$.

Let $Z := \{z \in \boldsymbol{R}^m\colon \sum_{j=1}^{m} z_j = 0\}$, a linear subspace. For any $z \in Z$,

$$|z|_p^2 = |p+z-f(p+z)|_p^2+|f(p+z)-p|_p^2.$$

Let $g(z) := f(p+z)-p$. Then g is linear on Z, with range $F-p$ of dimension s. The map $z \to z-g(z)$ is also linear, and its range is orthogonal to $F-p$ and to p for $(\cdot\,,\,\cdot)_p$. Since $F \subset P_m$, F spans a linear subspace of dimension $s+1$. Thus $I-g$ on Z has rank at most $m-s-1$. By the Fisher–Cochran theorem (Scheffé, [28], Appendix VI) with $z = r-p$, we see that

$$\mathscr{L}(|r-f(r)|_p^2) \to \mathscr{L}(\chi_{m-1-s}^2) \quad \text{as} \quad n \to \infty. \ ■$$

The main difficulty in applying the Fisher–Cramér–Rao–Birch Theorem (4.2) in practice is the computation of the estimate $v(r)$. To evaluate the MLE one would first solve the "ML equations" $dL(r, v) = 0$, which in coordinates gives a system of non-linear equations. Methods of solution include "Newton's method" in several variables and Cauchy's "method of steepest descent"; cf. Ortega and Rheinboldt, [23], pp. 179–187, 240–247. Furthermore, the ML equations may have multiple solutions, even infinitely many which must be compared to find the actual MLE, and in general, one has no straightforward method of determining all solutions.

Neyman [21] proposed "best asymptotically normal" estimates, some of which are easier to compute, since the ML equations are replaced by linear ones (cf. also LeCam, [19], and Bickel, [3]). The linearization, however, may depend on the choice of chart. The method of Dzhaparidze and Nikulin [14] treated below is chart-free.

The function f as in (5.5) maps V onto a neighborhood of p in the tangent flat F by a theorem of Kronecker (Alexandroff and Hopf, [1], p. 468). Thus in the proof of (5.7) we could replace $w(u)$ by a $w \in V$ with $f(w) = f(r)$. In the C^1 case once could use the Implicit Function Theorem. But a direct proof from minimal assumptions seems preferable, although manifolds encountered in practice are usually C^∞.

6. A modified X^2 statistic for use with convenient estimates

As mentioned in Section 5, the multinomial MLE $v(r)$ may be hard to compute. In testing whether a distribution on $\boldsymbol{R}$ is normal, we may prefer to use the more convenient "ungrouped MLE" estimators $\bar{X} := (X_1+\ldots+X_n)/n$ for the mean and

$$n^{-1}\sum_{j=1}^{n}(X_j-\bar{X})^2 = n^{-1}\Big(\sum_{j=1}^{n}X_j^2\Big)-\bar{X}^2$$

for the variance.

In other cases as well, when a space X has been decomposed into cells for a χ^2 test, one can estimate the unknown law more accurately and easily by using the full original data rather than only the cell occupation numbers n_j. The main problem then is that X_e^2 no longer has a limiting χ^2 distribution (Chernoff and Lehmann, [8]). Following Dzhaparidze and Nikulin [14], and extending their result to more general manifolds (C^1 rather than C^2, or requiring more than one chart), we will modify the X^2 statistic to solve the problem. Recalling the log likelihood function

$$L(r, v) := \sum r_i \ln v_i,$$

for fixed r we take the differential at each $v \in V$, $dL(r, v) \in V_v^*$, where

$$(6.1)\qquad \begin{aligned} dL(r, v) &= \sum_{1\leqslant i\leqslant m} (r_i/v_i)\,dv_i \\ &= \sum_{1\leqslant i\leqslant m} \big(1+(r_i-v_i)/v_i\big)\,dv_i \\ &= \sum_{1\leqslant i\leqslant m} (r_i-v_i)\,dv_i/v_i \end{aligned}$$

since $\sum_{1\leqslant i\leqslant m} v_i = 1$ on P_m, so $\sum_{1\leqslant i\leqslant m} dv_i = 0$.

Now we will assume that V is a C^1 submanifold of $U_m := \{p \in P_m\colon p_j > 0$ for all $j\}$.

Here U_m is an open subset of the $(m-1)$-dimensional linear variety $\{x\colon \sum x_j = 1\}$. All tangent spaces $(U_m)_p$ of U_m, as a submanifold of $\boldsymbol{R}^m$, can be identified as vector spaces with the linear subspace Z of $\boldsymbol{R}^m$ defined by

$$Z := \Big\{x \in \boldsymbol{R}^m\colon \sum_{1 \leqslant j \leqslant m} x_j = 0\Big\}.$$

The inner product

$$B_p(x, y) := (x, y)_p = \sum_{1 \leqslant j \leqslant m} x_j y_j / p_j$$

restricted to Z thus defines a C^∞ Riemannian structure on U_m, and a C^1 Riemannian structure on V.

We also then have the dual inner product B_p^* on the cotangent spaces V_p^*. Let $\|u\|_{p*} := B_p^*(u, u)^{1/2}$.

(6.2). Lemma. *For any C^1 submanifold $V \subset U_m$, as $r \to p$ and $v \to p$,*

$$\|dL(r, v)\|_{v*}^2 = |v - v(r)|_v^2 + o(|r-p|^2 + |v-p|^2).$$

Proof. Note that here $|\cdot|_v^2$ is defined on $\boldsymbol{R}^m$, but $\|\cdot\|_{v*}^2$ on cotangent spaces V_v^*. We take a C^1 chart x defined on a neighborhood U of p in V. Then $dv_i = \sum_{j=1}^{s} (\partial v_i/\partial x_j) dx_j$ on U, where by the C^1 assumption, $\partial v_i/\partial x_j$ are continuous and hence locally bounded.

In (6.1) we write $r - v = (r - v(r)) + (v(r) - v)$.

Claim. $\sum_{1 \leqslant i \leqslant m} (r_i - v(r)_i) dv_i/v_i = o(|r-p|)$.

Proof of Claim. In terms of the chart x we have

$$\sum_{1 \leqslant i \leqslant m} (r_i - v(r)_i) dv_i/v_i = \sum_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant s} (r_i - v(r)_i)(\partial v_i/\partial x_j) dx_j/v_i$$
$$= \sum_{1 \leqslant j \leqslant s} (r - v(r), \partial v/\partial x_j)_v dx_j,$$

where $(\cdot, \cdot)_v$ is defined on $\boldsymbol{R}^m \times \boldsymbol{R}^m$, and $\partial v/\partial x_j := \{\partial v_i/\partial x_j\}_{i=1}^m$ is a vector in $\boldsymbol{R}^m$ parallel to the tangent flat F_v to V at v, i.e. $\partial v/\partial x_j \in F_v - F_v$. By the C^1 assumption, F_v converges to F_p as $v \to p$. Thus, the angle between $\partial v/\partial x_j$ and F_p approaches 0 as $v \to p$.

From Lemma (5.7) (c), $v(r) - f(r) = o(|p-r|)$, where $r - f(r)$ is orthogonal to $F_p - F_p$. Then as $v \to p$ and $r \to p$,

$$(r - v(r), \partial v/\partial x_j)_v = (r - v(r), \partial v/\partial x_j)_p + o(|p-r|)$$
$$= (r - f(r) + o(|p-r|), f_j + o(1))_p + o(|p-r|)$$
$$= o(|p-r|) + o(|r - f(r)|) = o(|p-r|),$$

where $f_j \in F_p - F_p$. The Claim is proved.

Thus $dL(r, v) = \sum_{i=1}^{m} (v(r)_i - v_i) dv_i/v_i + o(|r-p|)$. We have

$$\|dL(r, v)\|_{v*} = \sup\{|dL(r, v)(w)|\colon w \in V_v, |w|_v = 1\}.$$

Letting $w = \sum_{1 \leqslant i \leqslant s} w_i \partial/\partial x_i|_v$, we have $|w|_v^2 = \sum_{1 \leqslant i, j \leqslant s} w_i w_j C_{ij}(v)$, where

$$C_{ij}(v) = (\partial/\partial x_i, \partial/\partial x_j)_v = \sum_{k=1}^{m} \frac{1}{v_k} \frac{\partial v_k}{\partial x_i} \frac{\partial v_k}{\partial x_j}.$$

Thus our Riemannian metric is represented in a coordinate system by what statisticians call the "Fisher information" matrix. Now

$$dL(w) = \sum_{1 \leqslant i \leqslant m} (v(r)_i - v_i) w(v_i)/v_i + o(|r-p|)$$

as $r \to p$, $v \to p$, $w \in V_v$, and $|w|_v$ remains bounded.

For the natural mapping of V_v into $Z \subset \boldsymbol{R}^m$, w has components $w(v_i) = w_i$, $i = 1, \ldots, m$. Then

$$dL(r, v)(w) = (v(r) - v, w)_v + o(|r-p|)$$

as $r \to p$, $v \to p$, and $|w|_v$ stays bounded. Now since $v(r) \to p$, $v \to p$, and V is C^1,

$$v(r) - v = y(r, v) + o(|v-p| + |r-p|),$$

where $y(r, v) \in F_v - F_v \subset Z$. Let $W = y(r, v)/|y(r, v)|_v$. Then

$$|dL|_{v*} = |dL(w)| + o(|r-p| + |v-p|),$$

and

$$|dL(w)| = |v(r) - v|_v + o(|v-p| + |r-p|).$$

This gives Lemma (6.2). ∎

(6.3) Lemma. *As $r \to p$ and $v \to p$,*

$$|r - v|_v^2 = |r - v(r)|_{v(r)}^2 + |v(r) - v|_v^2 + o(|r-p|^2 + |v-p|^2).$$

Proof. As $r \to p$, $|p - v(r)|_p \leqslant 2|r-p|_p$ by Lemma (5.7) (a). Then all norms in (6.3) can be replaced by $|\cdot|_p$ norms as in Lemmas (5.4) and (5.8). Then $v(r)$ can be replaced by $f(r)$ using (5.7) (c), and v by $f(v)$ using (5.5). Then, the result follows by combining (5.4) and (5.6). ∎

Definition. An estimate $\hat{p} = \hat{p}(n, \omega) \in V$ of p, not necessarily a function of the r_j, will be called *$n^{1/2}$-consistent* iff $n^{1/2}|\hat{p} - p|$ is bounded in probability, i.e. for any $\varepsilon > 0$ there is an $M < \infty$ such that

$$\sup_n \Pr\{n^{1/2}|\hat{p} - p| > M\} < \varepsilon.$$

Let $Z^2 := Z_{\hat{p}}^2 := n[|r - \hat{p}|_{\hat{p}}^2 - \|dL(r, \hat{p})\|_{\hat{p}*}^2]$. Dzhaparidze and Nikulin [14] introduced Z^2 and proved the following result in case V is a C^2 image of an open set in $\boldsymbol{R}^s$.

(6.4) Theorem. *For any C^1 submanifold V of U_m with $\dim V = s$ and any $n^{1/2}$-consistent estimator $\hat{p} \in V$ of p, $\mathscr{L}(Z_{\hat{p}}^2) \to \mathscr{L}(\chi_{m-s-1}^2)$ as $n \to \infty$.*

Proof. By the $n^{1/2}$-consistency, $n\big(o(|\hat{p}-p|^2)\big) \to 0$ in probability as $n \to \infty$. We saw in Section 5 that $n\big(o(|p-r|^2)\big) \to 0$ in probability for the empirical $r = r(\omega, n)$. By Lemmas (6.2) and (6.3), $Z_{\hat{p}}^2$ has the same limit law as $n|r-v(r)|^2_{v(r)}$, namely $\mathcal{L}(\chi^2_{m-s-1})$ by Birch's Theorem. ■

To apply the theorem one will need to compute $\|dL(r,\hat{p})\|^2_{\hat{p}*}$ in terms of coordinates. For C a positive self-adjoint operator (matrix),

$$\begin{aligned}\sup\{|(x,y)|:\ (Cy,y)=1\} &= \sup\{|(x,y)|:\ \|C^{1/2}y\|=1\}\\ &= \sup\{|(x, C^{-1/2}z)|:\ \|z\|=1\}\\ &= \sup\{|(C^{-1/2}x, z)|:\ \|z\|=1\}\\ &= \|C^{-1/2}x\| = (C^{-1/2}x, C^{-1/2}x)^{1/2} = (C^{-1}x, x)^{1/2}.\end{aligned}$$

In our case, C is the Fisher information matrix, and

$$\|dL(r,v)\|^2_{v*} = \sum_{i,j=1}^{s} (C^{-1})_{ij}\big(\partial L(v)/\partial x_i\big)\big(\partial L(v)/\partial x_j\big)$$

which is evaluated at $v = \hat{p}$.

Note that e.g. for $s = 2$, inversion of C is not difficult.

Notes. Chentsov [7], pp. 173–182 treats the Riemannian metric $\sum_i (dp_i)^2/p_i$ we used above. He shows that it is the unique metric with a certain natural "equivariance" property, and notes its representation by the Fisher information matrix. Rao and Robson [26] also consider modified X^2 statistics for composite hypotheses.

Acknowledgement

I am grateful to O. Barndorff-Nielsen for telling me about the papers by Dzhaparidze and Nikulin and by Rao and Robson; to A. Holst Andersen for the opportunity to read some of his unpublished work; and to J. Kiefer for information about the cited works of Scheffé, Neyman, LeCam and Bickel.

References

[1] P. Alexandroff and H. Hopf, *Topologie*, I Band, Springer, Berlin 1935.
[2] L. Auslander and R. E. Mackenzie, *Introduction to differentiable manifolds*, McGraw-Hill, New York 1963.
[3] P. J. Bickel, *One-step Huber estimates in the linear model*, J. Amer. Statist. Assoc. 70 (1975), pp. 428–434.
[4] M. W. Birch, *A new proof of the Pearson–Fisher theorem*, Ann. Math. Statist. 35 (1964), pp. 817–824.
[5] R. L. Bishop and R. J. Crittenden, *Geometry of manifolds*, Academic Press, New York 1964.
[6] N. Bourbaki, *Variétés différentielles et analytiques, Fascicules de resultats*, par. 1–7 (1967), 8–15 (1971), Hermann, Paris.
[7] [N. N. Chentsov] Н. Н. Ченцов, *Статистические решающие правила и оптимальные выводы*, Наука, Москва 1972.

[8] H. Chernoff and E. L. Lehmann, *The use of maximum likelihood estimates in χ^2 tests for goodness of fit*, Ann. Math. Statist. 25 (1954), pp. 579–586.
[9] W. G. Cochran, *The distribution of quadratic forms in a normal system*, Proc. Cambridge Philos. Soc. 30 (1934), pp. 178–191.
[10] —, *The χ^2 test of goodness of fit*, Ann. Math. Statist. 23 (1952), pp. 315–345.
[11] H. Cramér, *Mathematical methods of statistics*, Almqvist and Wiksells, Uppsala 1945, also Princeton University Press.
[12] J. Dieudonné, *Treatise on analysis*, vol. 3 (1972), transl. by I. G. Macdonald, Academic Press, New York.
[13] R. Doornbos and H. J. Prins, *On slippage tests*, Proc. Koninkl. Akad. Wetensch. Ser. A (Math.) 61; Indag. Math. 20 (1958), pp. 438–455.
[14] K. O. Dzhaparidze and M. S. Nikulin, *On a modification of the standard statistics of Pearson*, Theor. Probability Appls. 19 (1974), pp. 851–863. [К. О. Джапаридзе, М. С. Никулин, *Об одном видоизменении стандартной статистики Пирсона*, Теория вероятностей и ее применения 19 (4) (1974), pp. 886–888.]
[15] R. A. Fisher, *The conditions under which χ^2 measures the discrepancy between observation and hypothesis*, J. Roy. Statist. Soc. 87 (1924), pp. 442.
[16] —, *Applications of Student's distribution*, Metron 5 (3) (1925), pp. 90–104.
[17] A. Hald, *Statistical theory with engineering applications*, Wiley, New York 1952, pp. 262–275.
[18] H. O. Lancaster, *The chi-squared distribution*, Wiley, New York 1969.
[19] L. LeCam, *On the asymptotic theory of estimation and testing hypotheses*, Proc. Third Berkeley Symp. Math. Statist. Prob. 1, Univ. of Calif. Press, Berkeley and Los Angeles 1956, pp. 129–156.
[20] K. W. Mardia, *Statistics of directional data*, Academic Press, London and New York 1972
[21] J. Neyman, *Contribution to the theory of the χ^2 test*, Proc. Berkeley Symp. Math. Statist. Prob. 1, Univ. of Calif. Press, Berkeley and Los Angeles 1949, pp. 239–273.
[22] W. F. Newns and A. G. Walker, *Tangent planes to a differentiable manifold*, J. London Math. Soc. 31 (1956), pp. 400–407.
[23] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York 1970.
[24] K. Pearson, *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*, Philosophical Magazine, ser. V 50 (1900), pp. 157–172.
[25] C. R. Rao, *Linear statistical inference and its applications*, Wiley, New York 1965, 1972.
[26] K. C. Rao and D. S. Robson, *A chi-square statistic for godness-of-fit tests within the exponential family*, Comm. Statist. 3 (1974), pp. 1139–1154.
[27] J. T. Roscoe and J. A. Byars, *An investigation of the restraints commonly imposed on the use of the chi-squared statistics*, J. Amer. Statist. Assoc. 66 (1971), pp. 775–759.
[28] H. Scheffé, *The analysis of variance*, Wiley, New York 1954.
[29] J. K. Yarnold, *The minimum expectation in X^2 goodness of fit tests and the accuracy of approximations for the null distribution*, J. Amer. Statist. Assoc. 65 (1970), pp. 864–886.

Presented to the Semester
Probability Theory
February 11–June 11, 1976