18.650, Oct. 19, 2015

χ^2 TESTS FOR COMPOSITE HYPOTHESES – ASYMPTOTIC DISTRIBUTIONS

Recall that we have a multinomial $(n, \pi_1, ..., \pi_k)$ distribution, where π_j is the probability of the jth of k possible outcomes on each of n independent trials. Thus $\pi_j \geq 0$ and $\sum_{j=1}^k \pi_j = 1$. Let X_j be the number of times that the jth outcome occurs in the n trials. In testing a composite hypothesis $\pi_j = \pi_j(\theta)$ indexed by θ in an m-dimensional parameter space Θ, where $\pi_j(\theta) > 0$ for all $j = 1, ..., k$ and all θ in Θ, and $m < k - 1$, we estimate θ by some $\widehat{\theta}$ and compute the chi-squared statistic

$$\hat{X}^2 = \sum_{j=1}^k \frac{(X_j - n\pi_j(\widehat{\theta}))^2}{n\pi_j(\widehat{\theta})}.$$

In this handout, we'll show that for two ways of estimating θ, maximum likelihood (based on the X_j) and minimizing $\widehat{X}^2$ ("minimum χ^2 estimation"), the distribution of $\hat{X}^2$ will converge as $n \to \infty$ to that of $\chi^2(k - m - 1)$. For other estimation methods, such as those based on continuous or otherwise ungrouped data, the limit distribution of $\widehat{X}^2$ can be different.

1. Preliminaries

1.1. **Notations $O(\cdot)$, O_p, $o(\cdot)$, o_p, $\sim$, and $\asymp$.** We'll be using notations defined as follows. For two functions f and g with $g > 0$ one says that $f = o(g)$ if $f(x)/g(x) \to 0$ or $f = O(g)$ if f/g remains bounded under some limiting condition, here where $g(x) \to 0$ (but it would be used similarly if $g(x) \to +\infty$). The same notations are also used for sequences, so that $x_n = O(y_n)$ means $y_n > 0$ and $x_n = o(y_n)$ means $x_n/y_n \to 0$ as $n \to \infty$, while $x_n = O(y_n)$ means x_n/y_n remains bounded. The notation $f = o(1)$ means $f(x) \to 0$ under some limiting condition, likewise $x_n = o(1)$ means $x_n \to 0$ as $n \to \infty$, and $x_n = O(1)$ means x_n is a bounded sequence.

For a sequence X_n of random variables, $X_n = O_p(1)$ will mean that X_n are *bounded in probability*, meaning that for any $\varepsilon > 0$ there is an $M < +\infty$ such that $P(|X_n| > M) < \varepsilon$ for all n. If $Y_n > 0$ are random variables then $X_n = O_p(Y_n)$ will mean that $X_n/Y_n = O_p(1)$. $X_n = o_p(1)$ will mean that $X_n \to 0$ in probability, i.e. for every $\varepsilon > 0$, $P(|X_n| > \varepsilon) \to 0$ as $n \to \infty$. If $Y_n > 0$ are random variables then $X_n = o_p(Y_n)$ will mean $X_n/Y_n = o_p(1)$.

Recall that $f \sim g$ means $f/g \to 1$, and likewise for sequences. Lastly $f \asymp g$ will mean that $f = O(g)$ and $g = O(f)$, in other words f and g

are of the same order of magnitude, under whatever limiting condition is given.

1.2. **Some linear algebra.** . For any $x = \{x_i\}_{i=1}^j \in \mathbb{R}^j$ and $j = 1, 2, 3, ...$, we take the usual Euclidean norm $|x| = (x_1^2 + \cdots + x_j^2)^{1/2}$. Let A be a $k \times m$ real matrix $A = \{A_{ij}\}_{1 \le i \le k, 1 \le j \le m}$. Taking elements $x = \{x_j\}_{j=1}^m$ of $\mathbb{R}^m$ as column vectors, A defines a linear transformation, also to be called A, of $\mathbb{R}^m$ into $\mathbb{R}^k$, by $y = Ax$ where $y_i = \sum_{j=1}^m A_{ij} x_j$. If $m \le k$, the *rank* of A is defined as the maximum number of linearly independent columns of A, in other words, the dimension of the linear subspace L of $\mathbb{R}^k$ onto which A transforms $\mathbb{R}^m$. A is said to have *full rank* if this rank is m, in other words, A defines a one-to-one linear transformation of $\mathbb{R}^m$ onto L. This implies that for some M with $1 \le M < +\infty$, for all $x \ne 0$ in $\mathbb{R}^m$,

$$1/M \le |Ax|/|x| \le M, \tag{1}$$

noting that $|Ax|$ is taken in $\mathbb{R}^k$ and $|x|$ in $\mathbb{R}^m$.

2. Chi-squared statistics using maximum likelihood estimates

A *maximum likelihood estimate* (MLE) of θ will be a $\widehat{\theta}$, if it exists, which maximizes $\prod_{j=1}^k \pi_j(\theta)^{X_j}$ with respect to $\theta \in \Theta$. If it is unique it is called *the* MLE. Let P_k be the set of all k-tuples $\pi = \{\pi_j\}_{j=1}^k$ such that $\pi_j > 0$ for each j and $\sum_{j=1}^k \pi_j = 1$ (mathematicians might describe this as the interior of a $(k-1)$-dimensional simplex). Let V be the set of all $v(\theta) = \{\pi_j(\theta)\}_{j=1}^k$ for θ in Θ. We will assume that v is a one-to-one function from Θ onto V, continuous with a continuous inverse (a homeomorphism). Moreover, we will want V to be a sufficiently smooth subset (surface or manifold) of P_k. It does not have to be very smooth. For those who know differential geometry, it would suffice for it to be a C^1 submanifold. In fact somewhat less suffices, as a statistician M. W. Birch proved in 1964. I had given expositions of Birch's theorem and proof in some 1976 lecture notes and a 1979 paper but seemingly not in this course until 2011. The set $V \subset P_k$ will be called a *Birch m-submanifold* of P_k if for all $v \in V$, V has a tangent hyperplane F at v. Specifically, there exist some $\gamma > 0$ and $\zeta > 0$ and a function w defined on the neighborhood $U = \{x : \ |x| < \zeta\}$ of 0 in $\mathbb{R}^m$ such that $w(0) = v$, and $w(\cdot)$ is a homeomorphism of U onto a subset of V including $\{w \in V : \ |w - v| < \gamma\}$. Assume also that the vector-valued function w has a first Fréchet derivative at 0, namely, the first partial

derivatives $A_{ij} := \partial w_i/\partial x_j|_{x=0}$ exist for all $i = 1, ..., k$ and $j = 1, ..., m$, and

$$|w(x) - v - Ax|/|x| \to 0 \tag{2}$$

as $|x| \to 0$. Further, assume that the matrix A_{ij} has full rank m.

Theorem 1 (Birch's theorem)**.** *Let V be any Birch m-submanifold of P_k where $0 < m < k - 1$, then for any $p \in V$, if $(X_1, ..., X_k)$ have a multinomial $(n, \{p_j\}_{j=1}^k)$ distribution, as $n \to \infty$, the probability that at least one MLE $\widehat{\theta}$ exists converges to 1, and for any choices $\widehat{\theta} = \widehat{\theta}_n$ of MLE, the distribution of $\hat{X}^2$ converges to that of $\chi^2(k - m - 1)$.*

Proof. Maximizing the likelihood is, as usual, equivalent to maximizing the log likelihood, which in this case is $\sum_{j=1}^k X_j \log(\pi_j(\theta))$. Letting $r_j := X_j/n$ for $j = 1, ..., k$, we have $r = \{r_j\}_{j=1}^k \in P_k$ provided that $X_j \geq 1$ for all j, which will occur with probability $\to 1$ as $n \to \infty$ since all $p_j > 0$.

Equivalently, we want to maximize $\sum_{j=1}^k r_j \log(v_j/r_j)$ with respect to $v = \{v_j\}_{j=1}^k \in V$ or to minimize $\sum_{j=1}^k r_j \log(r_j/v_j)$. The proof will be based on a sequence of lemmas. Let $0 \cdot \log(x/0) = 0$ for all $y \geq 0$ and $x \cdot \log(x/0) = +\infty$ for all $x > 0$.

Lemma 2. *For any x and y in $[0, 1]$, $x \cdot \log(x/y) \geq x - y + \frac{1}{2}(x - y)^2$.*

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

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

for some w between x and y. Thus $1/w \geq 1$ and the Lemma follows. Q.E.D.

Lemma 3. *For any r and v in P_k,*

$$L(r, v) := \sum_{j=1}^k r_j \log(r_j/v_j) \geq \frac{1}{2}|r - v|^2.$$

Proof. By Lemma 2, for each j,

$$r_j \log(r_j/v_j) \geq r_j - v_j + \frac{1}{2}(r_j - v_j)^2.$$

Then summing over j gives the conclusion since $\sum_{j=1}^k r_j - v_j = 1 - 1 = 0$, Q.E.D.

Lemma 4. *For any $p \in V$ and $(X_1, ..., X_k)$ multinomial $(n, p_1, ..., p_k)$, let $r = \{r_j\}_{j=1}^k$ as defined above. Then $r = r_n \to p$ with probability 1*

as $n \to \infty$, and the probability that an MLE $\widehat{\theta}$ (not necessarily unique) exists converges to 1 as $n \to \infty$. Letting $v = v(r) = v^{(n)} = \pi(\widehat{\theta})$, any such $v^{(n)}$ converge to p as $n \to \infty$.

Proof. Each X_j has a binomial (n, p_j) distribution, so $r_j = X_j/n$ converges to p_j with probability 1 as $n \to \infty$ by the law of large numbers, and $r \to p$. Given $\varepsilon > 0$, we will have $L(r, p) < \varepsilon^2/2$ and $|r - p| < \varepsilon$ for n large enough. Let $W := \{v \in V : \ |v - p| \le 2\varepsilon\}$. We have

$$\inf_{v \in V} L(r, v) \le \inf_{v \in W} L(r, v) \le L(r, p) < \varepsilon^2/2.$$

On the other hand, for $v \in V$ not in W, we have $|v - r| > \varepsilon$ and so $L(r, v) > \varepsilon^2/2$ by Lemma 3. Thus

$$\inf_{v \in V} L(r, v) \le \inf_{v \in W} L(r, v) < \inf_{v \in V, v \notin W} L(r, v).$$

Since W is a closed and bounded (compact) set, there exists a $v \in W$ at which the infimum is attained, giving an MLE. As n becomes large, and $\varepsilon \downarrow 0$, W shrinks down to p, so the MLEs $v^{(n)}$ converge to p, Q.E.D.

For each $p \in P_k$, an inner product (dot product) on $\mathbb{R}^k$ is defined by

$$(x, y)_p := \sum_{j=1}^{k} x_j y_j / p_j.$$

Let $|x|_p := (x, x)_p^{1/2}$. For a fixed p, or for p with all p_j bounded away from 0, there is a constant $M < \infty$ such that

$$|x|/M \le |x|_p \le M|x|$$

for all $x \in \mathbb{R}^k$. Thus in a statement such as $|x_n|_p = o(|y_n|_p)$, the p subscript makes no difference.

Lemma 5. *As $r \to p$ and $v \to p$ with $v \in V$,*

$$-2 \sum_{j=1}^{k} r_j \log(v_j/r_j) = |r - v|_p^2 + o(|r - p|^2 + |v - p|^2). \tag{3}$$

Proof. By assumption $p_j > 0$ and $v_j > 0$ for all j, and we can assume $r_j > 0$ for all j. Then by the proofs of Lemmas 2 and 3,

$$-2 \sum_{j=1}^{k} r_j \log(v_j/r_j) = \sum_{j=1}^{k} (v_j - r_j)^2 / w_j$$

where w_j is between r_j and v_j for each j. Then $1/w_j - 1/p_j \to 0$. Now $(r_j - v_j)^2 \le 2(r_j - p_j)^2 + 2(v_j - p_j)^2$ since for all real x and y,

$(x+y)^2 \le 2x^2+2y^2$. Thus $|r-v|_p^2 \le 2M^2(|r-p|^2+|p-v|^2)$ and Lemma 5 follows, Q.E.D.

Now let F be the tangent hyperplane to V at p. Then F is the set of all $p+w'(0)u$ for $u \in \mathbb{R}^s$ where $w'(0)$ is given by the $k \times m$ matrix A as in the definition (2) of Birch m-submanifold. For $x \in \mathbb{R}^k$ let $f(x) \in F$ be such that $|x-f(x)|_p = \min\{|x-y|_p : \ y \in F\}$. In other words, f is the orthogonal projection into F for the $(\cdot,\cdot)_p$ inner product.

Lemma 6. *As $v \to p$ with $v \in V$, $|v-f(v)| = o(|v-p|)$.*

Proof. By the definition of Birch m-submanifold and definitions in it applied to $v=p$, the function $w(\cdot)$ on U is a homeomorphism onto a subset of V including $\{v \in V : \ |v-p| < \gamma\}$. Letting $x(\cdot)$ be the inverse function of w, we have $x(v) \to 0$ if and only if $v \to p$, and by (2), then $|v-p-Ax(v)| = o(|x(v)|)$. Since A is of full rank m we have by (1) $|Ax(v)| \asymp |x(v)|$. It follows then that $|v-p| \sim |Ax(v)|$. By definition of f, we have $|v-f(v)| \le |v-p-Ax(v)| = o(|x(v)|)$ which we now see is $o(|v-p|)$, Q.E.D.

Lemma 7. *As $r \to p$ and $v \to p$ with $v \in V$,*
(4)

$$-2\sum_{j=1}^{k} r_j \log(v_j/r_j) = |r-f(r)|_p^2 + |f(r)-f(v)|_p^2 + o(|r-p|^2+|v-p|^2).$$

Proof. Apply Lemma 6 to get on the right in (3)

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

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

$$|r-f(v)|_p^2 = |r-f(r)|_p^2 + |f(r)-f(v)|_p^2,$$

and the conclusion follows, Q.E.D.

Lemma 8. *(a) For r close enough to p, $|v(r)-p|_p \le 2|r-p|_p$;*
(b) As $r \to p$, $|f(r)-f(v(r))| = o(|p-r|)$;
(c) $|v(r)-f(r)| = o(|p-r|)$.

Proof. By Lemma 4, $v(r)$ exists (not necessarily unique) and converges to p as $r \to p$. Then $f(r) = p+Au$ for some $u = u(r) \to 0$ as in the definition of Birch submanifold, and $|w(u)-f(r)| = o(|u|)$. By (1), $o(|u|) = o(|Au|) = o(|f(r)-p|) = o(|r-p|)$. Thus $|f(w(u))-f(r)| = o(|r-p|)$.

In (4), the left side is minimized at $v = v(r)$ by definition, so it must be smaller there than at $v = w = w(u(r))$, where as just shown,

$|f(r)-f(w)| = o(|r-p|)$, so on the right in (4), the term $|f(r)-f(w)|_p^2$ can be included in the $o(\cdot)$ error term. Thus

$$|f(r) - f(v(r))|^2 = o(|r-p|^2 + |v(r)-p|^2 + |w(u(r))-p|^2).$$

We have $|w(u(r))-p|^2 = O(|f(r)-p|^2) = O(|r-p|^2)$ and so

$$|f(r) - f(v(r))|^2 = o(|r-p|^2 + |v(r)-p|^2). \tag{5}$$

If (a) fails, take a sequence $r_n = r \to p$ with $|p-v(r)|_p > 2|p-r|_p$. Then $|f(r)-f(v(r))| = o(|v(r)-p|)$ by (5), and $|f(v(r))-v(r)| = o(|v(r)-p|)$ by Lemma 6. Then $|f(r)-v(r)| = o(|p-v(r)|)$, and $|v(r)-p|_p$ is asymptotic to $|f(r)-p|_p$ which is $\le |r-p|_p$ as $r \to p$, a contradiction. Thus (a) is proved. By (5), (b) follows.

Next, $|v(r)-f(v(r))| = o(|v(r)-p|) = o(|r-p|)$ by Lemma 6 and part (a). This gives part (c), proving Lemma 8, Q.E.D.

Lemma 9. *As $r \to p$, $w \to p$, and $v \to p$ with $v \in V$,*

$$|r-v|_w^2 = \sum_{j=1}^{k}(r_j - v_j)^2/w_j = |r-f(v)|_p^2 + o\left(|r-p|^2 + |v-p|^2\right).$$

Proof. $1/w_i = 1/p_i + o(1)$, so the proof of Lemma 5 applies. Also use Lemma 6 to replace v by $f(v)$. Q.E.D.

Lemma 10. *As $r \to p$,*

$$Y^2 := \sum_{j=1}^{k}(r_j - v_j(r))^2/v_j(r) = |r-f(r)|_p^2 + o(|r-p|^2).$$

Proof. By Lemma 4, $v(r) \to p$. Then by Lemma 9 and Lemma 8(a),

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

By Lemma 8(b) and since $(r-f(r), f(r)-f(v(r)))_p = 0$, the conclusion follows, Q.E.D.

Proof of Theorem 1. By the multidimensional central limit theorem, as we saw in the proof of the asymptotic $\chi^2(k-1)$ distribution of the X^2 statistic for a simple hypothesis, the k-variate distribution of $\{n(r_j-p_j)/\sqrt{np_j}\}_{j=1}^k$ converges as $n \to \infty$ to a normal distribution with mean vector 0 and covariance matrix $C_{ij} = \delta_{ij} - \sqrt{p_i p_j}$ for $i, j = 1, ..., k$. It follows that $no(|r-p|_p^2) \to 0$ in probability as $n \to \infty$. By Lemma 10, nY^2 has the same limit distribution as $n|r-f(r)|_p^2$. Now for any $r \in P_k$,

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

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

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

By the case of a simple hypothesis, we know that the distribution of $n|r-p|_p^2$ converges to that of $\chi^2(k-1)$ as $n\to\infty$. Let Z be the $(k-1)$-dimensional hyperplane of all $z = \{z_j\}_{j=1}^k$ with $\sum_{j=1}^k z_j = 0$. For any $z\in Z$,

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

Also, for any $y\in\mathbb{R}^k$, $y\in Z$ if and only if $(y,p)_p = 0$. Let $g(z) := f(p+z)-p$. Then g is linear on Z, with range $F-p := \{x-p : x\in F\}$ a linear subspace of dimension m. In the definition of Birch m-submanifold, we have $\sum_{i=1}^k w_i(x) = 1$ for all $x\in U$ and thus for each $j = 1,\ldots,m$,

$$\sum_{i=1}^k \partial w_i(x)/\partial x_j|_{x=0} = 0.$$

So the columns of A are in Z, the range of A is included in Z, and $F-p\subset Z$.

The map from z to $z-g(z)$ (the identity minus g) is also linear, and its range is orthogonal to $F-p$ and to p for $(\cdot,\cdot)_p$. Since $F\subset p+Z$ it does not contain 0, so it spans a linear subspace of dimension $m+1$. Thus $I-g$ has rank at most $k-m-1$. By the Fisher–Cochran Theorem (Corollary 15 in the Appendix below), we see that the distribution of $n|r-f(r)|_p^2$ converges to that of $\chi^2(k-1-m)$, proving Theorem 1 (Birch's theorem), Q.E.D.

3. Minimum χ^2 estimates

Given a composite hypothesis $V\subset P_k$, a *minimum χ^2 estimate* is a $\xi\in V$, if it exists, that minimizes $|r-\xi|_\xi^2$, or equivalently minimizes $n|r-\xi|_\xi^2$. Let $\widehat{X}^2_{\min} := \inf_{\xi\in V} n|r-\xi|_\xi^2$, which is the value of the statistic X^2 at a minimum χ^2 estimate if one exists. The value of X^2 at an MLE $v(r)$ is $n|r-v(r)|_{v(r)}^2$. It turns out that the two values of X^2 are asymptotically the same:

Theorem 11. *If $m<k-1$, V is a Birch m-submanifold of P_k, and if H_0: $p=(p_1,\ldots,p_k)\in V$ is true, then as $n\to\infty$,*

$$\inf_{\xi\in V} |r-\xi|_\xi^2 = |r-v(r)|_{v(r)}^2 + o_p(1/n), \tag{6}$$

or equivalently

$$\widehat{X}^2_{\min} = n|r - v(r)|^2_{v(r)} + o_p(1),$$

so that the distribution of $\widehat{X}^2_{\min}$ *also converges as* $n \to \infty$ *to that of* $\chi^2(k-m-1)$.

Proof. Under H_0 we have $Er_j = p_j$ for each j and $E((r_j - p_j)^2) = p_j(1-p_j)/n$. It follows that $E(|r-p|^2) < 1/n$. Since $p_j > 0$ for all k, there is an $M < \infty$ such that $1/p_j < M$ for all j. Thus $E(|r-p|^2_p) < M/n$ and $|r-p|^2 = O_p(1/n)$. By Lemma 10,

$$|r - v(r)|^2_{v(r)} = |r - f(r)|^2_p + o(|r-p|^2) \le |r-p|^2_p + o_p(1/n). \tag{7}$$

It follows that for any minimum χ^2 estimate ξ, or any $\xi \in V$ such that $|r-\xi|^2_\xi \le |r - v(r)|^2_{v(r)}$, we have $|r-\xi|^2 = O_p(1/n)$ so ξ is close to r with high probability as n becomes large, and both are close to p. Then for $w = v = \xi$ in Lemma 9,

$$\begin{aligned} |r-\xi|^2_\xi &= |r - f(\xi)|^2_p + o(|r-p|^2) \\ &= |r - f(r)|^2_p + |f(r) - f(\xi)|^2_p + o(|r-p|^2). \end{aligned} \tag{8}$$

Since $o(|r-p|^2) = o_p(1/n)$, in $\widehat{X}^2_{\min}$, to minimize the left side of (8), we need to choose $\xi \in V$ so that $|f(r) - f(\xi)|^2_p$ is also small, specifically, also $o_p(1/n)$. Then by (7)

$$\widehat{X}^2_{\min} = n|r - f(r)|^2_p + o_p(1) = n|r - v(r)|^2_{v(r)} + o_p(1),$$

proving the theorem, Q.E.D.

4. Appendix: Partition theorems for quadratic forms and χ^2 variables

On the finite-dimensional real vector space $\mathbb{R}^d$, consisting of points $x = \{x_j\}_{j=1}^d$, let $(x,y) := \sum_{j=1}^d x_j y_j$ be the usual inner product. Taking vectors $x \in \mathbb{R}^d$ as column vectors, for a $d \times d$ matrix A and vector x, the matrix product $y = Ax$ with entries $y_i = (Ax)_i = \sum_{j=1}^d A_{ij} x_j$ gives another column vector y. A *quadratic form* is a real-valued function Q which can be written as $Q(x) = (Ax, x) = \sum_{i=1}^d \sum_{j=1}^d A_{ij} x_i x_j$ for some $d \times d$ matrix $A = \{A_{ij}\}_{i,j=1}^d$. Clearly the notion of quadratic form doesn't depend on the choice of coordinates, although of course the coefficients A_{ij} do. We can always take A to be symmetric, namely with $A_{ij} = A_{ji}$ for all i and j, since we can replace both A_{ij} and A_{ji} by $(A_{ij} + A_{ji})/2$ without changing Q. A coordinate free definition of

symmetric (also called self-adjoint) is that $(Ax, y) = (x, Ay)$ for all x and y.

Proposition 12. *If Q is any quadratic form with $Q(x) = (Ax, x)$ for all $x \in \mathbb{R}^d$ and A is symmetric, then A is uniquely determined.*

Proof. For any x, y and symmetric B we have the polarization identity

$$4(Bx, y) = (B(x + y), x + y) - (B(x - y), x - y).)$$

Thus if $(Az, z) = (Cz, z)$ for all z where A and C are symmetric, let $B = A - C$. Then $(Bx, y) = 0$ for all x and y, which implies $Bx = 0$ (as a vector) for all x, which implies $B = 0$, so $A = C$, Q.E.D.

For any linear subspace J of $\mathbb{R}^d$ let $J^{(p)} = \{y : \ (x, y) = 0 \text{ for all } x \in J\}$. $J^{(p)}$ is sometimes called the *orthogonal complement* of J. Two linear subspaces J_1 and J_2 are called *orthogonal*, or $J_1 \perp J_2$, if and only if $(x, y) = 0$ for all $x \in J_1$ and $y \in J_2$. Also let $\ker(A)$ be defined as $\{x : \ A(x) = 0\}$. The *range* of A or $\operatorname{ran} A$ is defined as $\{Ax : \ x \in \mathbb{R}^d\}$. A symmetric matrix is called an *orthogonal projection* if $A(Ax) = Ax$ for all x (in other words $A^2 = A$ for matrix multiplication). Then $Ax = x$ for all $x \in \operatorname{ran}(A)$.

Proposition 13. *If A is symmetric then $\ker(A) = \operatorname{ran}(A)^{(p)}$.*

Proof. $Ax = 0$ if and only if $(Ax, y) = 0$ for all y, if and only if $(x, Ay) = 0$ for all y, i.e. $x \in \operatorname{ran}(A)^{(p)}$, Q.E.D.

Given a quadratic form $Q(x) \equiv (Ax, x)$ with A symmetric, the *rank* of Q or A is defined as the dimension of the range of A, which is some integer $r(A)$ with $0 \le r(A) \le d$.

Theorem 14. *Let $Q_1, ..., Q_K$ be quadratic forms on $\mathbb{R}^d$ with for each $j = 1, ..., K$, $Q_j(x) \equiv (A_j x, x)$ for A_j symmetric. Let r_j be the rank of Q_j and A_j for each j. Assume that for every $x \in \mathbb{R}^d$, $(x, x) = \sum_{j=1}^K Q_j(x)$. Then $\sum_{j=1}^K r_j = d$ if and only if the A_j are orthogonal projections onto orthogonal subspaces.*

Proof. By Proposition 12, $\sum_{j=1}^K A_j = I$, the identity $d \times d$ matrix. The "if" part is clear. To prove "only if," take $x \in H_1 := \cap_{j=2}^K \ker(A_j) = \big(\sum_{j=2}^K \operatorname{ran}(A_j)\big)^{(p)}$ by Proposition 13. Then $A_1 x = x$. Thus $\operatorname{ran}(A_1) \supset H_1$. Since the dimension $\dim\big(\sum_{j=2}^K \operatorname{ran}(A_j)\big) \le \sum_{j=2}^K r_j$ and $r_1 = d - \sum_{j=2}^K r_j$ it follows that $r_1 = \dim(H_1)$ and $\operatorname{ran}(A_1) = H_1$. Inductively, for each j, likewise, the subspaces $\operatorname{ran}(A_j)$ are all orthogonal, with sum $\mathbb{R}^d$. By Proposition 13, each A_j is the orthogonal projection onto its range, Q.E.D.

Corollary 15 (The Fisher–Cochran partition theorem for χ^2)**.** *Let P be a normal distribution with mean 0 on $\mathbb{R}^d$. Suppose Q_i for $i = 1, ..., K$ are quadratic forms on $\mathbb{R}^d$ of respective ranks d_i. Let $Q(x) := \sum_{i=1}^K Q_i$. Suppose Q has a $\chi^2(d)$ distribution. Then $\sum_{i=1}^K d_i = d$ if and only if the Q_i are jointly independent and have distributions $\chi^2(d_i)$.*

Proof. Again the "if" direction is clear. To prove "only if," we can choose coordinates $(x_1, \ldots, x_d)$ in which the covariance matrix C is diagonal, with diagonal entries $c_1 \geq c_2 \geq \cdots \geq c_d \geq 0$. Let J be the largest j with $c_j > 0$. Change coordinates again to $y_j = \sqrt{c_j} x_j$ if $c_j > 0$ ($j \leq J$) and $y_j = x_j$ otherwise. In the y_j coordinates C becomes an orthogonal projection, diagonal with first J diagonal entries equal to 1 and other entries 0. Thus $P(y_i = 0) = 1$ for all $i > J$. Let $Q(x) = (Ay, y)$ in the y coordinates. Without changing the distribution of Q we can assume that $A_{ij} = 0$ if $i > J$ or $j > J$. In the coordinates $y_1, \ldots, j_J$, C is the identity matrix, so by a rotation (not changing C) we can make A diagonal with diagonal entries $a_1, \ldots, a_J$. Then $Q = \sum_{j=1}^J a_j y_j^2$ where y_j are i.i.d. $N(0, 1)$. If any $a_j < 0$ then with positive probability, Q would have negative values, contradicting its $\chi^2(d)$ distribution. Also, any $a_j = 0$ can be omitted from the sum, so we can assume that $a_j > 0$ for all j.

The moment generating function of a $\chi^2(1)$ random variable is

$$E \exp(u\chi^2(1)) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(ux^2 - \frac{x^2}{2}\right) dx,$$

which is finite if and only if $u < 1/2$. By the substitution $y = \sqrt{1 - 2u}x$ we find that the moment generating function equals $(1 - 2u)^{-1/2}$ as is well known (a special case of the Γ generating function given by Rice, Appendix A, p. A2). Thus for $u < (1/2) \min(1, \min_j 1/a_j)$,

$$\prod_{j=1}^{J} (1 - 2a_j u)^{-1/2} = (1 - 2u)^{-d/2}$$

since Q has a $\chi^2(d)$ distribution. Taking the -2 power of both sides gives

$$\prod_{j=1}^{J} (1 - 2a_j u) = (1 - 2u)^d.$$

Two polynomials equal for all u in a half-line $(-\infty, c)$ for some $c > 0$ must be identically equal. The right side has a root only at $u = 1/2$. Thus $a_j = 1$ for all j and $J = d$. So in the y coordinates C is the identity $C = I$ and $Q(y) \equiv (y, y)$. Then application of Theorem 14 gives the corollary. Q.E.D.

NOTES

Although Theorem 1 allows non-unique maximum likelihood estimates, such cases would involve harder computations. One would find points where the gradient of the (log) likelihood is zero (the "likelihood equations" hold), then one to decide which of these are local maxima or minima or saddle points, and among the maxima, pick the global one(s). Often, the MLE is unique and fairly easily computed. Distributions for minimum χ^2 estimates can provide useful bounds even though such estimates are generally hard to compute.

Fisher (1924) first gave a statement of the conclusion of Theorem 1 with non-rigorous hypotheses and proof. H. Cramér (1945), in apparently the first mathematically rigorous statistics textbook, assumed that the function $w(\cdot)$ is twice continuously differentiable (C^2), i.e. all the second partial derivatives $\partial^2 w_i(x)/\partial x_j \partial x_s$ for $i = 1, ..., k$ and $j, s = 1, ..., m$ exist and are continuous. Cramér showed that there exists at least one sequence of solutions of the likelihood equations converging to the true value, but a criticism is, how does one recognize such solutions? C. R. Rao (1965, 1972) gave a proof assuming that w is just C^1 (once continuously differentiable). Birch's (1964) proof assumed only existence of a Fréchet derivative everywhere, not necessarily continuous. In practice, the manifolds considered are very smooth (C^∞). As assuming more smoothness seems to make the proof no easier, one was given here only under Birch's assumption.

Birch (1964) took Θ to be an open subset of $\mathbb{R}^m$. The definition is changed here because, first, one might want to consider other cases such as that Θ is a circle, sphere, or cylinder, e.g. Mardia (1972), Mardia and Jupp (2000). (Mardia's 1972 book has over 2,000 citations in published articles and books according to Google Scholar.) Second, in terms of differential geometry, the inverse of w is just a local coordinate system (chart), which is rather arbitrary.

In one case, an asymptotic distribution different from a χ^2 distribution was found. Let real (continuous) $X_1, \ldots, X_n$ be observed and let the hypothesis H_0 be that they are i.i.d. $N(\mu, \sigma^2)$ for some unknown μ and σ^2. Suppose μ and σ^2 are estimated by maximum likelihood from the given (ungrouped) data, namely by $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2$. Then decompose the line into some k groups (intervals, or half-lines) each of which has at least 5 expected observations according to $N(\hat{\mu}, \hat{\sigma}^2)$. If one does a χ^2 test, Chernoff and Lehmann (1954) showed that the approximate distribution of the X^2 statistic for large n is that of $\sum_{j=1}^{k-1} a_j G_j^2$ where G_j are i.i.d. $N(0, 1)$ and $a_j = 1$ for $j = 1, \ldots, k-3$ but $0 < a_j < 1$ for $j = k-2,\ k-1$. Here

$m = 2$, and the distribution is between that of $\chi^2(k-1-m)$ as for MLE based on the grouped data, and $\chi^2(k-1)$ as for a simple hypothesis. This way of testing normality is now outdated. The Shapiro–Wilk test for normality was published in 1965 and in recent years has become very accessible through R.

I first wrote a proof of Birch's theorem in lecture notes for a course in Aarhus, Denmark, in the spring of 1976. Later that same spring I visited the Banach Center in Warsaw and lectured on Birch's theorem. I had (re)discovered the "Fisher–Cochran" theorem, Corollary 15, and assumed it must be known but didn't know a reference. The late Jack Kiefer (1924–1981), a leading statistician to whom I'd sent a copy, informed me that the fact was indeed well known and was in the book of Scheffé (1954). The name given to this fact seems to be based on papers of Fisher (1925) and Cochran (1934). I was able to update references as the Banach Center lectures were not published until 1979.

The material in this handout on minimum χ^2 estimation was not in the 1976 (or 1979) notes. It was added in March 2011.

William G. Cochran (1909-1980), born in Scotland and educated in Britain, taught statistics in the United States at 5 different universities (Iowa State, Princeton, U. North Carolina, Johns Hopkins, Harvard) from 1939 until he retired from Harvard in 1976. The book *Statistical Methods* by George W. Snedecor, joined in editions after the first by Cochran, has had at least 8 editions including a posthumous one. During the 1970's it was the most cited item in all mathematical literature, although that is no longer the case. Snedecor (1881-1974) founded the first academic statistics department and statistical laboratory in the United States, both at Iowa State University.

REFERENCES

Birch, M. W. (1964). A new proof of the Pearson–Fisher theorem. *Annals of Mathematical Statistics* **35**, 817–824.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system. *Math. Proc. Cambridge Philos. Soc.* **30**, pp. 178-191.

Cramér, H. (1945). *Mathematical Methods of Statistics*. Uppsala, Almqvist and Wiksells; also Princeton University Press.

Dudley, R. M. (1976). *Probabilities and Metrics: Convergence of laws on metric spaces, with a view to statistical testing*. Aarhus Universitet, Lecture Notes Series, No. 45.

Dudley, R. M. (1979). On χ^2 tests of composite hypotheses. *Probability Theory, Banach Center Publications*, **5**, PWN—Polish Scientific Publishers, Warsaw, 1979, pp. 75–87.

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *J. Roy. Statist. Soc.* **87**, pp. 442–450. [Included in Fisher's works as published, and posted online, by Adelaide University.]

Fisher, R. A. (1925). Applications of "Student's" distribution. *Metron* **5**, 90–104. [Also included in Fisher's works, pub. by Adelaide University.]

Mardia, K. V. (1972). *Statistics of Directional Data*. London and New York, Academic Press.

Mardia, K. V., and Jupp, P. E. (2000). *Directional Statistics*. Wiley, New York [Second edition of Mardia, 1972].

Rao, C. R. (1965, 1972). *Linear Statistical Inference and its Applications*. New York, Wiley.

Scheffé, H., *The Analysis of Variance* (1954). Wiley, New York.

Snedecor, G. W., and Cochran, W. G. (1989). *Statistical Methods*, 8th ed., Iowa State University Press.