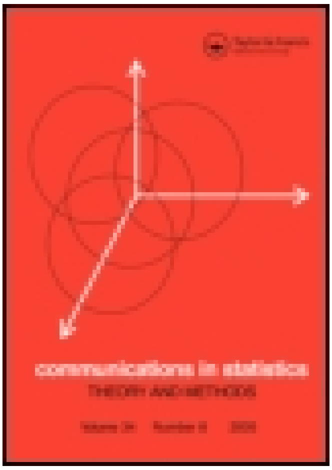

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

On the Asymptotic Distribution of the Likelihood Ratio Statistic

W. F. Scott [a]

[a] Department of Mathematical Sciences, University of Aberdeen, Aberdeen, UK

Published online: 28 Feb 2007.

Communications in Statistics—Theory and Methods, 36: 273–281, 2007

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920600974229

Inference

On the Asymptotic Distribution of the Likelihood Ratio Statistic

W. F. SCOTT

Department of Mathematical Sciences, University of Aberdeen, Aberdeen, UK

Likelihood ratio tests are among the most frequently used in practical statistics. Important references on this subject include Wilks (1962), Rao (1965, pp. 347–352), and Agresti (1990, pp. 434–435). We give proofs of some well-known results on the limiting distribution of $D = -2\log(\lambda)$, where λ is the likelihood ratio, under certain clearly-stated assumptions. We illustrate the theory by means of an example on contingency tables. Although results of this type are frequently used, the underlying theory is not always fully explained.

Keywords Contingency tables; Fisher's information; Maximum likelihood estimation.

Mathematics Subject Classification Primary 62F09; Secondary 62P25.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a variable X with density $f(x \mid \mathring{\boldsymbol{\theta}})$, where $\mathring{\boldsymbol{\theta}}$ is the true (but generally unknown) value of a parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)^\top$ ($\boldsymbol{\theta} \in H$, where H is an open set). We consider the asymptotic distribution (as $n \to \infty$) of the statistic

$$D = -2\log(\lambda) \tag{1.1}$$

where λ is the likelihood ratio

$$\lambda = \frac{\sup[L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in H_0]}{\sup[L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in H]} = \frac{\sup_{\boldsymbol{\theta}\in H_0}[\prod_{i=1}^n f(X_i \mid \boldsymbol{\theta})]}{\sup_{\boldsymbol{\theta}\in H}[\prod_{i=1}^n f(X_i \mid \boldsymbol{\theta})]}. \tag{1.2}$$

Here, $L(\boldsymbol{\theta})$ denotes the likelihood function and H_0 denotes the subset of H such that $\theta_{r-s+1} = \mathring{\theta}_{r-s+1}, \ldots, \theta_r = \mathring{\theta}_r$: that is, the final s components of $\boldsymbol{\theta}$ take their true

Received May 28, 2005; Accepted May 10, 2006
Address correspondence to W. F. Scott, School of Mathematics, University of Edinburgh, EH9 3J2, UK; E-mail: w.f.scott@ed.ac.uk

values. More generally, H_0 may be considered to be a subset of H subject to s independent constraints on $\theta_1, \ldots, \theta_r$, as one may change to a new set of parameters which satisfy the above restrictions. If $s = r$ we have $H_0 = \{\mathring{\boldsymbol{\theta}}\}$, this being referred to as a simple hypothesis.

It is well known that, under certain conditions, the asymptotic distribution of D is χ_s^2, but what are these conditions? Some of the proofs of this limiting result are sketchy and rather unclear: our main objective is to present rigorous proofs under clear conditions (similar to those given by Walker, 1969 and Scott, 2001). In addition, brief proofs of somewhat more general results will be given.

2. Notation and Assumptions

The notation of Scott (2001, 2003) will in general be used, but we shall write $\mathring{\boldsymbol{\theta}}$ rather than $\boldsymbol{\theta}_0$. Let

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \text{ the log-likelihood function} \\ &= \sum_{i=1}^{n} \log f(X_i \mid \boldsymbol{\theta}) \quad (\boldsymbol{\theta} \in H). \end{aligned} \tag{2.1}$$

We shall suppose that $f(x \mid \boldsymbol{\theta}) > 0$ for all $x \in B$ and $\boldsymbol{\theta} \in H$, where B is a Borel set containing all the possible values of X (this set being the same for all $\boldsymbol{\theta} \in H$). We now make the following assumptions (cf. Scott, 2001, 2003).

Assumption A. For each $\delta > 0$ such that $\{\boldsymbol{\theta} : |\boldsymbol{\theta} - \mathring{\boldsymbol{\theta}}| < \delta\}$ is a subset of H, there is $k(\delta) > 0$ such that, with probability tending to 1 as $n \to \infty$,

$$n^{-1}[L_n(\boldsymbol{\theta}) - L_n(\mathring{\boldsymbol{\theta}})] < -k(\delta).$$

Assumption B. The first and second partial derivatives of $\log f(x \mid \boldsymbol{\theta})$ with respect to $\theta_1, \theta_2, \ldots, \theta_r$ exist, and, uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\mathring{\boldsymbol{\theta}}$, are bounded in absolute value by integrable functions of x.

Assumption C. The third partial derivatives of $\log f(x \mid \boldsymbol{\theta})$ with respect to $\theta_1, \theta_2, \ldots, \theta_r$ exist, and there is a function $H(x)$ such that

$$\int_B H(x) f(x \mid \mathring{\boldsymbol{\theta}}) dx < \infty$$

and

$$\left| \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} \log f(x \mid \boldsymbol{\theta}) \right| \leq H(x)$$

for $i, j, k = 1, 2, \ldots, r$, for $\boldsymbol{\theta}$ in some neighborhood of $\mathring{\boldsymbol{\theta}}$ and for all x in B.

Assumption D. The matrix

$$\mathbf{J}(\mathring{\boldsymbol{\theta}}) = \left[E\left(\frac{\partial}{\partial\theta_i} \log f\big(X \mid \mathring{\boldsymbol{\theta}}\big) \frac{\partial}{\partial\theta_j} \log f\big(X \mid \mathring{\boldsymbol{\theta}}\big) \right) \right] \tag{2.2}$$

where $E(\cdot)$ denotes the expectation when $\boldsymbol{\theta} = \mathring{\boldsymbol{\theta}}$, is positive-definite.

Remark 2.1. Fisher's information for the sample is

$$I(\overset{\circ}{\boldsymbol{\theta}}) = n\mathbf{J}(\overset{\circ}{\boldsymbol{\theta}}).$$

Remark 2.2. Throughout this article, $\overset{\circ}{\boldsymbol{\theta}}$ is a fixed vector in H, and we may abbreviate the notation by writing

$$\mathbf{J} = \mathbf{J}(\overset{\circ}{\boldsymbol{\theta}})$$

and

$$J_{ij} = J_{ij}(\overset{\circ}{\boldsymbol{\theta}}) \quad (1 \leq i, j \leq r).$$

3. Some Lemmas

Lemma 3.1. *Let $\{X_n\}$ converge to X in distribution, and let $\{Y_n\}$ converge in probability to zero. Then $\{X_n Y_n\}$ converges in probability to zero.*

Proof. See Billingsley (1968, p. 28). □

Lemma 3.2. *Let $\hat{\boldsymbol{\theta}}^{(n)}$ (or $\hat{\boldsymbol{\theta}}$ when n is understood) denote the maximum likelihood estimator of $\overset{\circ}{\boldsymbol{\theta}}$, if this exists and is unique (if not, $\hat{\boldsymbol{\theta}}$ may be defined arbitrarily). Under the conditions given in Sec. 2, the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \overset{\circ}{\boldsymbol{\theta}})$ is $N(\mathbf{0}, \mathbf{J}^{-1})$, and $\hat{\boldsymbol{\theta}}$ converges in probability to $\overset{\circ}{\boldsymbol{\theta}}$.*

Proof. See (for example) Scott (2001). □

Lemma 3.3. *Let $\{\boldsymbol{\theta}^{(n)}\}$ converge in probability to $\overset{\circ}{\boldsymbol{\theta}}$. Under the conditions of Sec. 2, for each $i, j = 1, 2, \ldots, r$, the sequence $\left\{-n^{-1}\frac{\partial^2 L_n(\boldsymbol{\theta}^{(n)})}{\partial\theta_i\partial\theta_j}\right\}$ converges in probability to J_{ij}.*

Proof. This follows from Scott (2001, formula (3.7)). □

4. The Likelihood Ratio Test for a Simple Hypothesis

When H_0 is the single point $\{\overset{\circ}{\boldsymbol{\theta}}\}$, we have

$$D = 2\left[L_n(\hat{\boldsymbol{\theta}}) - L_n(\overset{\circ}{\boldsymbol{\theta}})\right]. \tag{4.1}$$

Theorem 4.1. *The limiting distribution of D is χ^2_r.*

Proof. It follows from Scott (2001, p. 50) that

$$D = \sqrt{n}(\hat{\boldsymbol{\theta}} - \overset{\circ}{\boldsymbol{\theta}})^\top \left[-n^{-1}\frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial\theta_i\partial\theta_j}\right]\sqrt{n}(\hat{\boldsymbol{\theta}} - \overset{\circ}{\boldsymbol{\theta}}) \tag{4.2}$$

where $\boldsymbol{\theta}^*$ (which depends on n) lies between $\hat{\boldsymbol{\theta}}$ and $\overset{\circ}{\boldsymbol{\theta}}$. It follows from Lemma 3.2 that $\{\boldsymbol{\theta}^*\}$ converges in probability to $\overset{\circ}{\boldsymbol{\theta}}$, and Lemma 3.3 shows that, for each

$i, j = 1, 2, \ldots, r$, the (i, j)th element of the matrix $\left[-n^{-1}\frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial\theta_i\partial\theta_j}\right]$ converges to J_{ij}. Using Lemma 3.1, we find that

$$D = \sqrt{n}(\hat{\boldsymbol{\theta}} - \mathring{\boldsymbol{\theta}})^\top \mathbf{J}\sqrt{n}(\hat{\boldsymbol{\theta}} - \mathring{\boldsymbol{\theta}}) + R_1 \tag{4.3}$$

where R_1 converges in probability to zero. It now follows from Lemma 3.2 that the limiting distribution of D is χ_r^2. □

5. The Likelihood Ratio Test for a Composite Hypothesis

We now consider the distribution of D when $H_0 = \{\boldsymbol{\theta} : \theta_{r-s+1} = \mathring{\theta}_{r-s+1}, \ldots, \theta_r = \mathring{\theta}_r\}$, where $1 \leq s \leq r - 1$. (The case $s = r$ is covered in the preceding section.)

Theorem 5.1. *The limiting distribution of D is* χ_s^2.

Proof. The argument involves the division of $\boldsymbol{\theta}$ into two "sections", i.e., we write

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \tag{5.1}$$

where $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ have dimensions $r - s$ and s, respectively. A similar notation is used for $\mathring{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$. We also introduce the block notation

$$\mathbf{J} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \tag{5.2}$$

where $\mathbf{B}_{11}$ is $(r - s) \times (r - s)$ and $\mathbf{B}_{22}$ is $s \times s$, and we note that $\mathbf{B}_{12}^\top = \mathbf{B}_{21}$ and $\mathbf{B}_{21}^\top = \mathbf{B}_{12}$.

Let us now introduce the vector $\tilde{\boldsymbol{\phi}}_1^{(n)}$, or $\tilde{\boldsymbol{\phi}}_1$ if n is understood, which is a maximum likelihood estimator of $\mathring{\boldsymbol{\phi}}_1$ when $\mathring{\boldsymbol{\phi}}_2$ is known. $\tilde{\boldsymbol{\phi}}_1$ refers to a "reduced" estimation problem, and it follows from Lemma 3.2, and the assumptions of Sec. 2, that $\{\tilde{\boldsymbol{\phi}}_1\}$ converges in probability to $\mathring{\boldsymbol{\phi}}_1$ and that $\sqrt{n}(\tilde{\boldsymbol{\phi}}_1 - \mathring{\boldsymbol{\phi}}_1)$ converges in distribution to $N(\mathbf{0}, \mathbf{B}_{11}^{-1})$.

We now define $\tilde{\boldsymbol{\phi}}_2 = \mathring{\boldsymbol{\phi}}_2$ and let $\tilde{\boldsymbol{\theta}} = \begin{pmatrix} \tilde{\boldsymbol{\phi}}_1 \\ \tilde{\boldsymbol{\phi}}_2 \end{pmatrix}$. Taylor's theorem shows that there is $\boldsymbol{\theta}^*$ between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ such that

$$D = 2\left[L_n(\hat{\boldsymbol{\theta}}) - L_n(\tilde{\boldsymbol{\theta}})\right] = \sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top\left[-n^{-1}\frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial\theta_i\partial\theta_j}\right]\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}).$$

It now follows by Lemmas 3.1–3.3 that

$$D = \sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top \mathbf{J}\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + R_1 \tag{5.3}$$

where R_1 converges in probability to zero. By Cramér (1946, Sec. 20.6), the limiting distribution of D is the same as that of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top \mathbf{J}\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = Q$, say.

We now make use of the block notation introduced earlier. It is clear that

$$Q = \sqrt{n}\big((\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)^\top, (\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2)^\top\big)\begin{bmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}\begin{pmatrix}\sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)\\ \sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2)\end{pmatrix}.$$

That is,

$$\begin{aligned}Q &= \sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)^\top\mathbf{B}_{11}\sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1) + \sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2)^\top\mathbf{B}_{21}\sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)\\ &\quad + \sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)^\top\mathbf{B}_{12}\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2) + \sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2)^\top\mathbf{B}_{22}\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2).\end{aligned} \tag{5.4}$$

The first mean value theorem shows that there are vectors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(r-s)}$, lying between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, such that

$$\begin{aligned}&(\hat{\theta}_1 - \tilde{\theta}_1)\frac{\partial^2 L_n(\boldsymbol{\theta}^{(1)})}{\partial\theta_1^2} + \cdots + (\hat{\theta}_r - \tilde{\theta}_r)\frac{\partial^2 L_n(\boldsymbol{\theta}^{(1)})}{\partial\theta_r\partial\theta_1} = \frac{\partial L_n(\hat{\boldsymbol{\theta}})}{\partial\theta_1} - \frac{\partial L_n(\tilde{\boldsymbol{\theta}})}{\partial\theta_1} = 0\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ &(\hat{\theta}_1 - \tilde{\theta}_1)\frac{\partial^2 L_n(\boldsymbol{\theta}^{(r-s)})}{\partial\theta_1\partial\theta_{r-s}} + \cdots + (\hat{\theta}_r - \tilde{\theta}_r)\frac{\partial^2 L_n(\boldsymbol{\theta}^{(r-s)})}{\partial\theta_r\partial\theta_{r-s}} = \frac{\partial L_n(\hat{\boldsymbol{\theta}})}{\partial\theta_{r-s}} - \frac{\partial L_n(\tilde{\boldsymbol{\theta}})}{\partial\theta_{r-s}} = 0.\end{aligned}$$

It follows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^\top\begin{bmatrix}\mathbf{B}_{11}\\ \mathbf{B}_{21}\end{bmatrix} + \mathbf{R}_2 = 0$$

where $\mathbf{R}_2$ converges in probability to zero (using the lemmas of Sec. 3). We thus have

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)^\top\mathbf{B}_{11} = -\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2)^\top\mathbf{B}_{21} - \mathbf{R}_2$$

and

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)^\top = -\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \tilde{\boldsymbol{\phi}}_2)\mathbf{B}_{21}\mathbf{B}_{11}^{-1} + \mathbf{R}_3 \tag{5.5}$$

where $\mathbf{R}_3$ converges in probability to zero. On substituting this expression for $\sqrt{n}(\hat{\boldsymbol{\phi}}_1 - \tilde{\boldsymbol{\phi}}_1)^\top$ into formula (5.4) we find that

$$Q = \sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \mathring{\boldsymbol{\phi}}_2)^\top[\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}]\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \mathring{\boldsymbol{\phi}}_2) + R_4 \tag{5.6}$$

where R_4 converges in probability to zero.

Let $\mathbf{A} = [\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}]^{-1}$. It follows by Graybill (1983, p. 184), that $\mathbf{A}$ equals $\mathbf{J}^{-1}$ with the first $r - s$ rows and columns eliminated, and it follows from Lemma 3.2 that the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \mathring{\boldsymbol{\phi}}_2)$ is $N(\mathbf{0}, \mathbf{A})$. We conclude that the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \mathring{\boldsymbol{\phi}}_2)^\top\mathbf{A}^{-1}\sqrt{n}(\hat{\boldsymbol{\phi}}_2 - \mathring{\boldsymbol{\phi}}_2)$ is χ^2_s. It now follows from formula (5.6) that the limiting distribution of Q is χ^2_s, and it has already been shown that the limiting distributions of Q and D are the same. This completes the proof. □

6. Discrete and Mixed Variables

The results given above may be proved for these cases with suitable modifications of the assumptions.

Suppose that X is discrete. We assume that the possible values of X lie on the same set $B = \{b_1, b_2, \ldots\}$ for all $\boldsymbol{\theta} \in H$, and that the probability mass function $f(x \mid \boldsymbol{\theta})$ is such that $f(x \mid \boldsymbol{\theta}) > 0$ for all x in B and $\boldsymbol{\theta}$ in H.

Theorem 6.1. *Let Assumptions A and D of Sec.* 2 *hold, and let Assumptions B and C be modified by the replacement of integrability with respect to x by summability. The limiting distribution of D is as given in Theorems* 4.1 *and* 5.1.

Proof. This is almost identical to the proofs of Theorems 4.1 and 5.1 (cf. Scott, 2003). □

Let us now consider "mixed" random variables X. Suppose that X has density $f(x \mid \boldsymbol{\theta})$ on a Borel set B_1 and probability mass function $f(x \mid \boldsymbol{\theta})$ on $B_2 = \{b_1, b_2, \ldots\}$, B_1 and B_2 being the same for all $\boldsymbol{\theta} \in H$. We also assume that $f(x \mid \boldsymbol{\theta}) > 0$ for all $x \in B_1 \cup B_2$ and $\boldsymbol{\theta} \in H$.

Theorem 6.2. *Let Assumptions A and D of Sec.* 2 *hold, and let Assumptions B and C be modified by the replacement of integrability with respect to x by integrability over B_1 and summability over B_2. The limiting distribution of D is as given in Theorems* 4.1 *and* 5.1.

Proof. This is almost identical to the proofs of Theorems 4.1 and 5.1 (cf. Scott, 2003). □

7. Extensions to the Theory

So far we have considered the n observations to be identically distributed. Let us now suppose that

$$n = n_1 + n_2 + \cdots + n_k$$

and consider random samples of $n_1, n_2, \ldots, n_k$ observations respectively from k random variables, each of which depends on the same vector of parameters, $\mathring{\boldsymbol{\theta}}$, where $\mathring{\boldsymbol{\theta}} \in H$. The log-likelihood function is

$$L_n(\boldsymbol{\theta}) = \sum_{m=1}^{n_1} \log f_1(x_m^{(1)} \mid \boldsymbol{\theta}) + \cdots + \sum_{m=1}^{n_k} \log f_k(x_m^{(k)} \mid \boldsymbol{\theta}) \tag{7.1}$$

where $f_t(x \mid \boldsymbol{\theta})$ refers to the tth random variable $(1 \le t \le k)$ and $\{x_m^{(t)} : m = 1, 2, \ldots, n_t\}$ denotes the observations from the tth variable.

Let $\lambda_t = \frac{n_t}{n}$ $(1 \le t \le k)$. Limiting theorems may now be proved under conditions similar to those considered above, under the further assumption that each λ_t remains fixed as $n \to \infty$. More specifically, $\mathbf{J}(\mathring{\boldsymbol{\theta}})$ is now defined as

$$\mathbf{J}(\mathring{\boldsymbol{\theta}}) = \lambda_1 \mathbf{J}_1(\mathring{\boldsymbol{\theta}}) + \cdots + \lambda_k \mathbf{J}_k(\mathring{\boldsymbol{\theta}}),$$

where (for $1 \leq t \leq k$) $\mathbf{J}_t(\overset{\circ}{\boldsymbol{\theta}})$ is defined as in formula (2.2) with $f(X \mid \overset{\circ}{\boldsymbol{\theta}}) = f_t(X \mid \overset{\circ}{\boldsymbol{\theta}})$. Fisher's information for the entire set of data is thus

$$\begin{aligned}\mathbf{I}(\overset{\circ}{\boldsymbol{\theta}}) &= n\mathbf{J}(\overset{\circ}{\boldsymbol{\theta}}) \\ &= \mathbf{I}_1(\overset{\circ}{\boldsymbol{\theta}}) + \cdots + \mathbf{I}_k(\overset{\circ}{\boldsymbol{\theta}}),\end{aligned}$$

where $\mathbf{I}_t(\overset{\circ}{\boldsymbol{\theta}})$ is Fisher's information in respect of the tth sample. The maximum likelihood estimator of $\overset{\circ}{\boldsymbol{\theta}}$ is again denoted by $\hat{\boldsymbol{\theta}}^{(n)}$, or $\hat{\boldsymbol{\theta}}$ if n is understood. We again find that the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \overset{\circ}{\boldsymbol{\theta}})$ is $N(\mathbf{0}, [\mathbf{J}(\overset{\circ}{\boldsymbol{\theta}})]^{-1})$, and analogs of Theorems 4.1 and 5.1 hold in the more general circumstances of this section.

8. Contingency Tables

The analysis of $I \times J$ contingency tables is a familiar topic in statistics. We regard the entries $\{n_{ij}\}$ as arising from n independent observations from a multinomial distribution with probabilities $\{p_{ij} : 1 \leq i \leq I, 1 \leq j \leq J\}$. The "full" model is such that

$$\log p_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{8.1}$$

where $\lambda_1^X = \lambda_1^Y = 0$ and $\lambda_{ij}^{XY} = 0$ if $i = 1$ or $j = 1$. This model has $r = IJ - 1$ parameters, i.e. $\lambda_i^X (2 \leq i \leq I)$, $\lambda_j^Y (2 \leq j \leq J)$ and $\lambda_{ij}^{XY} (2 \leq i \leq I, 2 \leq j \leq J)$. We do not regard λ as a parameter as it may be expressed in terms of the above parameters (though λ is included in the "saturated" model). A rather similar notation is used in the theory of loglinear models: see, for example, Agresti (1990, Ch. 5).

To test for independence of the row and column attributes, one may use the likelihood ratio test of the hypothesis

$$H_0 : \lambda_{ij}^{XY} = 0 \quad (2 \leq i \leq I, 2 \leq j \leq J) \tag{8.2}$$

The maximum likelihood estimator of the "reduced" vector of parameters is easily obtainable, and we find that

$$D = 2\left[\sum n_{ij} \log \hat{p}_{ij} - \sum n_{ij} \log(\hat{p}_i \cdot \hat{p}_j)\right] \tag{8.3}$$

where $\hat{p}_{ij} = n_{ij}/n$, $\hat{p}_i = (n_{i1} + \cdots + n_{iJ})/n$ and $\hat{p}_j = (n_{1j} + \cdots + n_{Ij})/n$. D is called G^2 in some books (see, for example, Agresti, 1990, p. 434; Powers and Xie, 2000, p. 105). Under H_0, and for large n, the distribution of D is approximately χ_s^2, where $s = (I - 1)(J - 1)$.

Example 8.1. The data (Table 1) are quoted by Agresti (1990, p. 273). It is found that D is 105.66, and the p-value is negligible (using the limiting χ^2-distribution with $(I - 1)(J - 1) = 4$ degrees of freedom).

Table 1
Political views

	Party affiliation		
Political ideology	Democrat	Independent	Republican
Liberal	143	119	15
Moderate	156	210	72
Conservative	100	141	127

We may also consider weaker hypotheses than independence. In particular, one may consider a "column effects" model:

$$\log p_{ij} = \lambda + \lambda_i + \lambda_j + \rho_j(u_i - u_1) \tag{8.4}$$

where $\{\rho_j\}$ are column effect terms (with $\rho_1 = 0$) and $\{u_i\}$ are fixed "scores" relating to the rows ($u_i = i$ in this example.) A test of the hypothesis that this model holds is provided by the likelihood ratio test of the hypothesis

$$H_0 : \lambda_{ij}^{XY} = \rho_j(u_i - u_1) \quad (2 \leq i \leq I, 2 \leq j \leq J). \tag{8.5}$$

This imposes $(I-1)(J-1) - (J-1) = (I-1)(J-2)$ constraints on the parameters, so the large-sample distribution of D under H_0 is approximately χ_s^2, where $s = (I-1)(J-2)$.

For the data of Example 8.1, we find that $D = 2.81$, with a p-value of 0.25. It thus appears that the column effects model fits better than the independence model (if one uses the p-value as the criterion of goodness of fit).

We may also test the independence model using the column effects model as the "full" model; that is, we may test $H_0 : \rho_j = 0$ $(2 \leq j \leq J)$. In this example,

$$D = 105.66 - 2.81 = 102.85$$

and the p-value is negligibly small (using the limiting χ^2-distribution with $J - 1 = 2$ degrees of freedom). This again indicates that the column effects model is preferable to the independence model.

Another example of this type may be found in Simonoff (2003, p. 55) (although the contingency table is sparse and the use of limiting distributions may require further justification).

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: John Wiley.

Cramér, H. C. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Graybill, F. A. (1983). *Matrices with Statistical Applications*. 2nd ed. Belmont, California: Wadsworth.

Powers, D. A., Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*. New York: John Wiley.
Scott, W. F. (2001). On the asymptotic distribution of the posterior distribution. *Math. Scientist* 26:46–55.
Scott, W. F. (2003). On the asymptotic distribution of the posterior distribution II. *Math. Scientist* 28:61–66.
Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York: Springer.
Walker, A. M. (1969). On the asymptotic behaviour of the posterior distribution. *J. Roy. Statist. Soc. B* 31:80–88.
Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons.