The Annals of Statistics
1984, Vol. 12, No. 3, 971–988

ADJUSTMENT BY MINIMUM DISCRIMINANT INFORMATION[1]

BY SHELBY J. HABERMAN

Hebrew University

Minimum discriminant information adjustment has primarily been used in the analysis of multinomial data; however, no such restriction is necessary. Let P be a distribution on R^a, and let $\mathscr{L}$ be a convex set of distributions on R^a. Let $\mathbf{X}_i$, $1 \le i \le n$, be independent and identically distributed observations with common distribution P. The minimum discriminant information adjustment (MDIA) of P relative to $\mathscr{L}$ is the element Q of $\mathscr{L}$ that is closest to P in the sense of Kullback–Leibler discriminant information. If $\bar{P}_n$ is the empirical distribution of the X_i, $1 \le i \le n$, and $\bar{Q}_n$ is the MDIA of $\bar{P}_n$ relative to $\mathscr{L}$, then $\bar{Q}_n$ is the maximum likelihood estimate in $\mathscr{L}$. Let $\mathscr{L}$ consist of distributions A on R^a such that $\int T\, dA = t$, where T is a measurable transformation from R^a to R^b and $t \in R^b$. It is shown that under mild regularity conditions $\bar{Q}_n$ converges weakly to Q, the MDIA of the true P, with probability 1 and that $\bar{E}_n(D) = \int D d\bar{Q}_n$ is an asymptotically normal and asymptotically unbiased estimate of $E(D) = \int D\, dQ$.

1. Introduction. Adjustment of distributions by the minimum discriminant information criterion has received attention from a variety of probabilists and statisticians, especially in the case of multinomial data. Among many relevant works are Good (1963), Ireland and Kullback (1968a, 1968b), Ireland, Ku, and Kullback (1969), Kullback (1959, 1971), and Csiszár (1975). Nonetheless, statistical properties of this adjustment technique are relatively unexplored except when observations are multinomial. In this paper, it is contended that adjustment by minimum discriminant information is also important when data are continuous, and statistical properties of this technique are explored.

The basic notion of minimum discriminant information adjustment is readily described. Let P be a probability distribution on R^a, $1 \le a \le \infty$, and let $\mathscr{L}$ be a convex set of probability distributions on R^a. For distributions A, B on R^a, let

$$(1.1)\qquad I(A,B) = \int dA \log(dA/dB), \quad A \ll B, \qquad = \infty \quad \text{otherwise},$$

denote the Kullback-Leibler discriminant information measure for A against B. The minimum discriminant information adjustment (MDIA) Q of P (relative to

Received September 1980; revised August 1983.
[1] Support for this research was provided in part by National Science Foundation Grant No. SOC76-80389 and by the John Simon Guggenheim Memorial Foundation. Part of this research took place while the author was a Lady Davis Visiting Professor at the Department of Statistics, Hebrew University, Jerusalem, Israel.
AMS 1980 *subject classifications.* Primary 62E99; secondary 62D05.
Key words and phrases. Consistency, asymptotic normality, empirical distribution, weighting.

$\mathscr{S}$) is the unique element of $\mathscr{S}$ such that

$$I(Q, P) = \min_{A \in \mathscr{S}} I(A, P). \tag{1.2}$$

That is, Q is the closest element of $\mathscr{S}$ to P, where closeness is defined in terms of discriminant information.

In this paper, estimation of Q is examined for P unknown and $\mathscr{S}$ known. Given independent observations $\mathbf{X}_i = \langle X_{hi}: 1 \le h \le a \rangle$ with common distribution P, Q is estimated by $\bar{Q}_n$, the MDIA relative to $\mathscr{S}$ of the empirical distribution $\bar{P}_n$ of the $\mathbf{X}_i$, $1 \le i \le n$. Since $\bar{P}_n$ is the maximum likelihood estimate of P in the sense of Kiefer and Wolfowitz (1956), $\bar{Q}_n$ may be regarded as the maximum likelihood estimate of Q. This paper considers computation and large-sample properties of $\bar{Q}_n$. To simplify results, $\mathscr{S}$ is always assumed to consist of the distributions A on R^a such that

$$\int \mathbf{T}\, dA = \mathbf{t}, \tag{1.3}$$

where $\mathbf{t}$ is a known element of R^b and $\mathbf{T}$ is a known measurable transformation from R^a to R^b.

To relate results in this paper to previous work, it is helpful to consider a classical example of Deming and Stephan (1940) which has received considerable attention in the literature. Let $p_{jk} > 0$, $1 \le j \le r$, $1 \le k \le s$, be cell probabilities in an r by s contingency table. Adjust the p_{jk} to form new cell probabilities q_{jk}, satisfying the constraints that the row and column marginal totals are specified positive values $q_{j\cdot}$ and $q_{\cdot k}$, respectively. In effect, Deming and Stephan (1940) find constants $a_j > 0$ and $b_k > 0$ such that

$$q_{jk} = a_j b_k p_{jk}, \quad 1 \le j \le r, \quad 1 \le k \le s. \tag{1.4}$$

As noted in Ireland and Kullback (1968a) and Csiszár (1975), their q_{jk} correspond to minimum discriminant information adjustment where P is specified by the $\langle p_{jk} \rangle$ and $\mathscr{S}$ is determined by the constraints on the marginal probabilities.

In practice, problems arise in which the cell probabilities p_{jk} are unknown, but the adjusted marginal totals $q_{j\cdot}$ and $q_{\cdot k}$ are given. Typically, independent observations $\mathbf{X}_i = \langle X_{1i}, X_{2i} \rangle$, $1 \le i \le n$, are available such that $\Pr(X_{1i} = j, X_{2i} = k) = p_{jk}$, $1 \le j \le r$, $1 \le k \le s$. Let the empirical distribution $\bar{P}_n$ of the $\mathbf{X}_i$, $1 \le i \le n$, assign probability $\bar{p}_{jkn}$ to $\langle j, k \rangle$. At least when all $\bar{p}_{jkn}$ are positive, the MDIA $\bar{Q}_n$ of $\bar{P}_n$ takes the form (1.4), i.e.,

$$\bar{q}_{jkn} = \bar{a}_{jn} \bar{b}_{kn} \bar{p}_{jkn}, \quad 1 \le j \le r, \quad 1 \le k \le s, \tag{1.5}$$

where $\bar{a}_{jn} > 0$ and $b_{kn} > 0$. Algorithms for computation of $\bar{q}_{jkn}$ are well known. For example, see Deming and Stephan (1940), Mosteller (1968), Ireland and Kullback (1968a), and Haberman (1974, Chapter 9; 1979, Chapter 9).

In practice, two rather distinct applications are encountered. In the first case, considered by Deming and Stephan (1940) and Ireland and Kullback (1968a), among others, the unknown true P is to be estimated, where marginal totals are known in advance. Thus $P \in \mathscr{S}$ and $Q = P$, so $\bar{q}_{jkn}$ is an estimate of the original probability p_{jk}. This estimate has the advantage over $\bar{p}_{jkn}$ that the supplementary information has been exploited. Large-sample properties are explored by Ireland and Kullback (1968a), who sketch a proof of asymptotic efficiency. More detailed proofs of large-sample properties appear in Haberman (1974, Chapter 9).

In the second case the true marginal totals are not known, but the relationships between the row and column variables of the table are to be examined in such a fashion that marginal distributions of row and column variables can be ignored. For example, Mosteller (1968) seeks to compare social mobility tables without concern for variations in the relative sizes of social classes. He suggests standardization to uniform marginal distributions, so that $q_{j\cdot} = 1/r$ and $q_{\cdot k} = 1/r$. (Since a mobility table is used, $r = s$). In this case, $p_{j\cdot}$ is not generally equal to $q_{j\cdot}$ and $p_{\cdot k}$ is not generally equal to $q_{\cdot k}$. Thus the cell probabilities q_{jk} are distinct from the p_{jk}, and the adjusted probabilities $\bar{q}_{jkn}$ estimate q_{jk} but not p_{jk}. In this case, asymptotic properties are also well known. See Haberman (1974, Chapter 9; 1979, Chapter 9) and Causey (1972).

A simple change in the problem of Deming and Stephan (1940) results in a less familiar problem. Suppose that instead of a contingency table, a continuous bivariate distribution P is to be studied. Let P be adjusted so that the new distribution Q has marginal means

$$\int\int x_1 \, dQ(x_1, x_2) = t_1 \tag{1.6}$$

and

$$\int\int x_2 \, dQ(x_1, x_2) = t_2. \tag{1.7}$$

Let $\mathbf{X}_i = \langle X_{1i}, X_{2i} \rangle$, $1 \leq i \leq n$, be independent observations with common distribution P. Then two questions arise: how should Q be defined, and how should Q be estimated? In this paper, the suggestion is made that it is often appropriate that Q be the MDIA of P relative to $\mathscr{L}$ and that Q be estimated by $\bar{Q}_n$. In contrast to the preceding example of Deming and Stephan (1940), this case does not appear to be considered in the literature. Thus the proposed approach to computation of $\bar{Q}_n$ and the discussion of large-sample properties of $\bar{Q}_n$ are both new for this case.

Once again, there exist two rather distinct versions of the adjustment problem. In one case, P is known to have marginal means $\int\int x_1 \, dP(x_1, x_2) = t_1$ and $\int\int x_2 \, dP(x_1, x_2) = t_2$, so that $P = Q$ and $\bar{Q}_n$ and $\bar{P}_n$ are both estimates of P. Here $\bar{Q}_n$ has the advantage over $\bar{P}_n$ that the supplemental information concerning P has been exploited. In the second case, the relationship between the two variables under study is to be examined without regard to marginal distributions. For instance, the relationship of height and weight might be examined for several groups of humans after an adjustment for variations between group marginal means of height and group marginal means of weight.

More generally, results in this paper will be previously known whenever P has all mass on a finite set. In the case $P \in \mathscr{L}$, work by Kullback and associates can be consulted. For the general case $P \notin \mathscr{S}$, Haberman (1974, Chapter 9; 1979, Chapter 9) provides general results. When P does not have support on a finite set, then results presented here apparently are new.

1.1. *Computation of adjustments.* As shown by Csiszár (1975), the essential relationship between an MDIA Q and an original distribution P involves the Radon-Nikodym derivative dQ/dP. One has a linear dependence of $\log[(dQ/dP)(x)]$ on $\mathbf{T}(\mathbf{x})$ for $\mathbf{x} \in M$, where M is a Borel set such that $Q(M) = 1$. More generally, M is assigned probability 1 by every distribution $A \in \mathscr{S}$ such that the discriminant information $I(A, P)$ is finite. Let $(\cdot, \cdot)$ denote the Euclidean inner product on R^b. Let I_M be the indicator function of M. Then for some $c > 0$ and $\boldsymbol{\theta} \in R^b$,

$$\frac{dQ}{dP} = cI_M \exp(\boldsymbol{\theta}, \mathbf{T}). \tag{1.8}$$

Since $Q \in \mathscr{S}$, one also has

$$c \int I_M \exp(\boldsymbol{\theta}, \mathbf{T})\, dP = 1 \tag{1.9}$$

and

$$c \int \mathbf{T} I_M \exp(\boldsymbol{\theta}, \mathbf{T})\, dP = \mathbf{t}. \tag{1.10}$$

If any $c > 0$, $\boldsymbol{\theta} \in R^b$, and M satisfy (1.9) and (1.10) and also the condition that $A(M) = 1$ for all $A \in \mathscr{S}$ such that $I(A, P) < \infty$, then Q exists and a version of dQ/dP satisfies (1.8).

Csiszár's (1975) results simplify dramatically when P is replaced by the empirical distribution $\bar{P}_n$. Let $\mathscr{B}$ be the class of Borel sets of R^a. Let I_B be the indicator function of $B \in \mathscr{B}$. Then

$$\bar{P}_n(B) = \sum_{i=1}^{n} \left(\frac{1}{n}\right) I_B(\mathbf{X}_i), \quad B \in \mathscr{B}, \tag{1.11}$$

only has mass at $\mathbf{X}_i$, $1 \le i \le n$. Thus $\bar{Q}_n$ only has mass at the $\mathbf{X}_i$, $1 \le i \le n$, and one may write $\bar{Q}_n$ as a weighted average

$$\bar{Q}_n(B) = \sum_{i=1}^{n} w_{in} \mathbf{I}_B(\mathbf{X}_i). \tag{1.12}$$

Since $\bar{Q}_n$ is a probability distribution and $I_B(\mathbf{X}_i) = 1$, $1 \le i \le n$, for $B = R^a$, it is obvious that

$$\sum_{i=1}^{n} w_{in} = 1. \tag{1.13}$$

Let $\mathbf{T}_i = \mathbf{T}(\mathbf{X}_i)$, $1 \le i \le n$. Since $\bar{Q}_n \in \mathscr{S}$, $\int \mathbf{T}\, d\bar{Q}_n$ satisfies

$$\sum_{i=1}^{n} w_{in} \mathbf{T}_i = \mathbf{t}. \tag{1.14}$$

From (1.8), (1.11), and (1.12), it is also clear that one may write

$$w_{in} = c_n I_{M(n)}(\mathbf{X}_i) \exp(\boldsymbol{\theta}_n, \mathbf{T}_i) \tag{1.15}$$

for some $c_n > 0$, $\boldsymbol{\theta}_n \in R^a$, and $M(n) \in \mathscr{B}$. Clearly $A(M(n)) = 1$ whenever $A \in \mathscr{S}$

and $I(A, P) < \infty$. Given (1.11), $A \in \mathscr{L}$ and $I(A, P) < \infty$ if and only if for some $z_i \geq 0$, $1 \leq i \leq n$,

$$\sum_{i=1}^{n} z_i = 1, \tag{1.16}$$

$$A(B) = \sum_{i=1}^{n} z_i I_B(\mathbf{X}_i), \quad B \in \mathscr{B}, \tag{1.17}$$

and

$$\sum_{i=1}^{n} z_i \mathbf{T}_i = \mathbf{t}. \tag{1.18}$$

If $A(M(n)) = 1$, then $\mathbf{X}_i \in M(n)$ whenever $z_i > 0$. Given (1.16) and (1.18), it follows that $\bar{Q}_n$ only exists if $\mathbf{t} \in K(n)$, the convex hull of the $\mathbf{T}_i$, $1 \leq i \leq n$. If $\mathbf{t} \in K(n)$ and $L(n)$ is the unique face of $K(n)$ that contains $\mathbf{t}$ in its relative interior (Rockafellar, 1970, page 164), then one may let $M(n) = \mathbf{T}^{-1}(L(n))$, so that (1.15) may be replaced by

$$w_{in} = c_n I_{L(n)}(\mathbf{T}_i)\exp(\boldsymbol{\theta}_n, \mathbf{T}_i), \quad 1 \leq i \leq n. \tag{1.19}$$

In typical examples, $\mathbf{t}$ is in the relative interior of $K(n)$, so that $I_{L(n)}(\mathbf{T}_i) = 1$, $1 \leq i \leq n$.

Equations (1.13)–(1.15) appear in the literature on log-linear models and exponential families. For example, see Dempster (1971). Solution by the Newton-Raphson algorithm is straightforward. At the beginning of iteration $\nu \geq 0$, let $\boldsymbol{\theta}_{n\nu}$ be the current approximation to $\boldsymbol{\theta}_n$. Then the current approximation to w_{in} is

$$w_{in\nu} = c_{n\nu} I_{L(n)}(\mathbf{T}_i)\exp(\boldsymbol{\theta}_{n\nu}, \mathbf{T}_i), \quad 1 \leq i \leq n, \tag{1.20}$$

where

$$c_{n\nu} = 1/\sum_{i=1}^{n} I_{L(n)}(\mathbf{T}_i)\exp(\boldsymbol{\theta}_{n\nu}, \mathbf{T}_i). \tag{1.21}$$

Given the weights, the $\mathbf{T}_i$, $1 \leq i \leq n$, have weighted mean

$$\mathbf{m}_{n\nu} = \sum_{i=1}^{n} w_{in\nu} \mathbf{T}_i \tag{1.22}$$

and weighted covariance matrix

$$\boldsymbol{\Sigma}_{n\nu} = \sum_{i=1}^{n} \sum_{i=1}^{n} w_{in\nu}(\mathbf{T}_i - \mathbf{m}_{n\nu})(\mathbf{T}_i - \mathbf{m}_{n\nu})' \tag{1.23}$$

($'$ denotes a transpose). A new parameter approximation $\boldsymbol{\theta}_{n(\nu+1)}$ is found by solving the equation

$$\boldsymbol{\Sigma}_{n\nu}(\boldsymbol{\theta}_{n(\nu+1)} - \boldsymbol{\theta}_{n\nu}) = \mathbf{t} - \mathbf{m}_{n\nu}. \tag{1.24}$$

It is not necessary that $\boldsymbol{\theta}_{n(\nu+1)}$ be uniquely determined by (1.24). Typically, $w_{in\nu}$ converges rapidly to w_{in}. If $L(n)$ has nonempty interior, then $\boldsymbol{\theta}_{n(\nu+1)}$ is uniquely determined and $\boldsymbol{\theta}_{n\nu} \to \boldsymbol{\theta}_n$ in typical cases. Use of the algorithm is described in Gokhale and Kullback (1980, Chapter 5) and Haberman (1979, Chapter 9). If $P \in \mathscr{L}$, then $\boldsymbol{\theta}_{n0}$ is usually chosen to be $\mathbf{0}$. This choice is often, but not always, helpful in the general case.

Much of the discussion in this paper concerns estimation of linear functionals of Q. Let D be a Borel-measurable real function of R^a, and let

$$E(D) = \int D\, dQ \tag{1.25}$$

be finite. Let $D_i = D(\mathbf{X}_i)$, $1 \le i \le n$. Then $E(D)$ has estimate

$$\bar{E}_n(D) = \int D\, d\bar{Q}_n = \sum_{i=1}^{n} w_{in} D_i. \tag{1.26}$$

Given the w_{in}, computation of $\bar{E}_n(D)$ is trivial.

1.2 *Strong consistency.* As shown in Section 3, the estimate $\bar{Q}_n$ is strongly consistent under general conditions. Let Q exist and satisfy $\int \exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$ for all $\mathbf{w}$ in a neighborhood N of $\mathbf{0}$. Then $\Pr(\bar{Q}_n \to_w Q) = 1$, where $\to_w$ denotes weak convergence. More generally, $\Pr(\bar{E}_n(D) \to E(D)) = 1$ if $\int (1 + |D|)\exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$ for some open neighborhood N of $\mathbf{0}$.

1.3 *Asymptotic normality.* As shown in Section 4, asymptotic normality results require slightly stronger conditions than those for strong consistency. Assume Q exists and assume $\int \exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$, $\mathbf{w} \in N$, for some open neighborhood N of $\mathbf{0}$. Let $\|\mathbf{x}\|^2 = (\mathbf{x}, \mathbf{x})$, $\mathbf{x} \in R^b$, be the squared Euclidean norm on R^b. Assume $\int \|\mathbf{T}\|^2 (dQ/dP)\, dQ < \infty$. Let $\int |D| \exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$ and $\int D^2 (dQ/dP)\, dQ < \infty$. Then

$$n^{1/2}[\bar{E}_n(D) - E(D)] \to_{\mathscr{D}} N(0, \tau^2(D)),$$

where $\to_{\mathscr{D}}$ denotes convergence in distribution. The asymptotic variance

$$\tau^2(D) = \int [D - c_D(\mathbf{T})]^2 \left(\frac{dQ}{dP}\right) dQ, \tag{1.27}$$

where $c_D(\mathbf{T})$ is the linear predictor

$$c_D(\mathbf{T}) = E(D) + (\boldsymbol{\delta}_D, \mathbf{T} - \mathbf{t}), \tag{1.28}$$

$\boldsymbol{\delta}_D$ is any solution of

$$\boldsymbol{\Sigma}\, \boldsymbol{\delta}_D = \int [D - E(D)](\mathbf{T} - \mathbf{t})\, dQ, \tag{1.29}$$

and

$$\boldsymbol{\Sigma} = \int (\mathbf{T} - \mathbf{t})(\mathbf{T} - \mathbf{t})'\, dQ \tag{1.30}$$

is the covariance matrix of $\mathbf{T}(\mathbf{X}^*)$, where $\mathbf{X}^*$ has distribution Q. A slight simplification occurs if $P \in \mathscr{S}$, so that $Q = P$. Then the condition

$$\int \|\mathbf{T}\|^2 \left(\frac{dQ}{dP}\right) dQ < \infty$$

is redundant and

$$\tau^2(D) = \tau_0^2(D) = \int [D - c_D(\mathbf{T})]^2 \, dP \tag{1.31}$$

is the residual variance from linear regression of D_1 on $\mathbf{T}_1$.

1.4 *Confidence intervals.* Using the above results, confidence intervals for an expected value $E(D)$ are readily obtained. If P is assumed in $\mathscr{L}$, then some simplifications occur. Assume $\int (D^2 + 1)\exp(\mathbf{w}, \mathbf{T}) \, dP < \infty$, $\mathbf{w} \in N$, for an open neighborhood N of $\mathbf{0}$. Assume $\tau_0^2(D) > 0$, so that D_1 is not equal with probability 1 to a linear function of $\mathbf{T}_1$. Let

$$\bar{\tau}_{0n}^2(D) = \sum_{i=1}^n w_{in}[D_i - \bar{E}_n(D) - (\bar{\delta}_{Dn}, \mathbf{T}_i - \mathbf{t})]^2, \tag{1.32}$$

where

$$\bar{\Sigma} \, \bar{\delta}_{Dn} = \sum_{i=1}^n w_{in} D_i(\mathbf{T}_i - \mathbf{t}) \tag{1.33}$$

and $\bar{\Sigma}_n$ is the estimated covariance matrix

$$\bar{\Sigma}_n = \sum_{i=1}^n w_{in}(\mathbf{T}_i - \mathbf{t})(\mathbf{T}_i - t)'. \tag{1.34}$$

Let $z_{\alpha/2}$ be the upper $-(\alpha/2)$ point of the standard normal distribution. For $0 < \alpha < 1$, $\Pr(\bar{E}_n(D) - z_{\alpha/2}\bar{\tau}_{0n}(D)/n^{1/2} \le E(D) \le \bar{E}_n(D) + z_{\alpha/2}\bar{\tau}_{0n}(D)/n^{1/2})$ approaches $1 - \alpha$ as $n \to \infty$. Thus

$$[\bar{E}_n(D) - z_{\alpha/2}\bar{\tau}_{0n}(D)/n^{1/2}, \bar{E}_n(D) + z_{\alpha/2}\bar{\tau}_{0n}(D)/n^{1/2}]$$

is an approximate level $-(1 - \alpha)$ confidence interval for $E(D)$.

If P need not be in $\mathscr{L}$, then

$$[\bar{E}_n(D) - z_{\alpha/2}\bar{\tau}_n(D)/n^{1/2}, \bar{E}_n(D) + z_{\alpha/2}\bar{\tau}_n(D)/n^{1/2}]$$

is an approximate level $-(1 - \alpha)$ confidence interval for $E(D)$ when

$$\int (1 + D^2)\exp(\mathbf{w})\mathbf{T})\left(\frac{dQ}{dP}\right) dQ < \infty, \quad \mathbf{w} \in N,$$

for an open neighborhood N of $\mathbf{0}$ and $D(\mathbf{X}^*)$ is not equal with probability 1 to a linear function of $\mathbf{T}(\mathbf{X}^*)$. Here

$$\bar{\tau}_n^2(D) = n \sum_{i=1}^n w_{in}^2[D_i -' \bar{E}_n(D) - (\bar{\delta}_n, \mathbf{T}_i - \mathbf{t})]^2. \tag{1.35}$$

1.5 *The form of Q.* A refinement of Csiszár (1975) is helpful in subsequent sections. The set M may be related to the convex hull K of the support of $P\mathbf{T}^{-1}$.

Obviously, (1.9) and (1.10) can only hold if $\mathbf{t} \in K$. If $\mathbf{t}$ is in the relative interior of the face L of K, then $\Pr(\mathbf{T}(\mathbf{X}_1) \in M) > 0$ and $\Pr(\mathbf{T}(\mathbf{X}_1) \in L \mid \mathbf{T}(\mathbf{X}_1) \in M) = 1$, as is evident from Rockafellar (1970, pages 162–164). Given the restraints on M, one may assume without loss of generality that $M \subset \mathbf{T}^{-1}(L)$.

A sharper result is available under the condition $\int \exp(\mathbf{w}, \mathbf{T}) \, dQ < \infty$, $\mathbf{w} \in N$, for an open neighborhood N of $\mathbf{0}$. One may then assume without loss of generality that $M = \mathbf{T}^{-1}(L)$. By Berk (1972), a set L', open relative to L, exists such that any $\mathbf{z} \in L'$ equals an expected value $\int \mathbf{T}\, dA$, where $dA/dP = c' \exp(\boldsymbol{\theta}', \mathbf{T})I_M$ for some $c' > 0$ and $\boldsymbol{\theta}' \in R^a$. Furthermore, $I(A, P) < \infty$. If $P(\mathbf{T}^{-1}(L) - M) > 0$, then a Borel set $U \subset \mathbf{T}^{-1}(L) - M$ exists with $P(U) > 0$ and $\mathbf{T}(U)$ bounded. Clearly, $I(A', P) < \infty$, where $dA'/dP = I_U/P(U)$. Thus for some α, $0 < \alpha < 1$, some $c' > 0$, some $\boldsymbol{\theta}' \in R^a$, and some A such that $dA/dP = c' \exp(\boldsymbol{\theta}', \mathbf{T})I_M$, $\int \mathbf{T}(\alpha A + (1 - \alpha)A') = \mathbf{t}$ and $I(\alpha A + (1 - \alpha)A', P) \le \alpha I(A, P) + (1 - \alpha)I(A', P) < \infty$. Since $A'(M) = 0$, $\alpha A(M) + (1 - \alpha)A'(M) = \alpha < 1$, a contradiction since $A + (1 - \alpha)A' \in \mathscr{L}$. Therefore, $P(\mathbf{T}^{-1}(L) - M) = 0$; i.e., one may assume without loss of generality that $M = \mathbf{T}^{-1}(L)$. This practice is followed throughout this paper.

2. Special cases. Several special cases help show the relationship between the general results of Section 1 and familiar results from the literature. Case 1, the unrestricted case, illustrates reduction of results to those for the empirical distribution. Case 2, the multinomial case, is the case considered by Kullback and his associates and by Haberman (1974, Chapter 9; 1979, Chapter 9). Case 3 describes a general approach to adjustment by sizes of overlapping strata, and Case 4 describes adjustment of moments.

CASE 1. *$\mathscr{L}$ unrestricted.* This case illustrates the case of the unadjusted empirical distribution $\bar{P}_n$. It arises if $b = 1$, $\mathbf{T} \equiv \mathbf{0}$, and $\mathbf{t} = \mathbf{0}$. Thus $\mathscr{L}$ contains all probability distributions on R^a. Since $\bar{P}_n \in \mathscr{L}$ and $P \in \mathscr{L}$, $\bar{Q}_n = \bar{P}_n$ and $Q = P$. The Newton-Raphson algorithm is unnecessary, and regularity conditions are trivial. As in the Glivenko-Cantelli theorem, $\Pr(\bar{P}_n \to_w P) = 1$. As in the strong law of large numbers, $\bar{E}_n(D) = \bar{D}_n = (1/n) \sum_{i=1}^n D_i \to E(D) = \int D\, dP$ with probability 1 whenever $\int |D|\, dP < \infty$. As in the central limit theorem, $\int D^2\, dP < \infty$ implies $n^{1/2}(\bar{D}_n - E(D)) \to_{\mathscr{D}} N(0, \sigma^2(D))$, where $\sigma^2(D) = \int [D - E(D)]^2\, dP$. For an approximate level $(1 - \alpha)$ confidence interval for $E(D)$, one has $[\bar{D}_n - z_{\alpha/2}\bar{\sigma}_n(D)/n^{1/2}, \bar{D}_n + z_{\alpha/2}\bar{\sigma}_n(D)/n^{1/2}]$, where $\bar{\sigma}_n^2(D) = (1/n) \sum_{i=1}^n (D_i - \bar{D}_n)^2$. This confidence interval is appropriate if $E(D_1^2) = \int D^2\, dP < \infty$ and PD^{-1}, the distribution of D_1, is not supported on a single point.

CASE 2. *The multinomial case.* Let P have all mass on a finite set Y. Then one obtains standard results of Kullback and associates and of Haberman (1974, Chapter 9; 1979, Chapter 9). To summarize results, let $p(\mathbf{y}) = P(\{\mathbf{y}\})$, $q(\mathbf{y}) = Q(\{\mathbf{y}\})$, $\bar{p}_n(y) = \bar{P}_n(\{\mathbf{y}\})$, and $\bar{q}_n(\mathbf{y}) = \bar{Q}_n(\{\mathbf{y}\})$ for $\mathbf{y} \in Y$.

The condition $\int \mathbf{T}\, dQ = \mathbf{t}$ is now the condition

$$\sum_{\mathbf{y} \in Y} \mathbf{T}(\mathbf{y})q(\mathbf{y}) = \mathbf{t},$$

and Q exists if and only if $\mathbf{t} \in K$, the convex hull of all $T(\mathbf{y})$ such that $P(y) > 0$. If Q exists and $\mathbf{t}$ is in the relative interior of the face L of K, then

$$q(\mathbf{y}) = cI_L(\mathbf{T}(\mathbf{y}))p(\mathbf{y})\exp(\boldsymbol{\theta}, \mathbf{T}(\mathbf{y}))$$

for some $c > 0$ and $\boldsymbol{\theta} \in R^a$. The constant c and $\boldsymbol{\theta}$ are determined by the equations

$$\sum_{\mathbf{y} \in Y} cI_L(\mathbf{T}(\mathbf{y}))\mathbf{T}(\mathbf{y})p(\mathbf{y})\exp(\boldsymbol{\theta}, \mathbf{T}(\mathbf{y})) = \mathbf{t}$$

and

$$\sum_{y \in Y} c I_L(\mathbf{T}(\mathbf{y})) p(\mathbf{y}) \exp(\theta, \mathbf{T}(\mathbf{y})) = 1.$$

Similarly, let $f_n(\mathbf{y}) = n\bar{p}_n(\mathbf{y})$ be the number of observations $\mathbf{X}_i$ equal to $\mathbf{y} \in Y$. Then $\bar{Q}_n$ exists whenever $\mathbf{t}$ is in the convex hull $K(n)$ of those $\mathbf{T}(\mathbf{y})$ for which $f_n(\mathbf{y}) > 0$. If $\mathbf{t} \in K(n)$, then

$$\bar{q}_n(\mathbf{y}) = c_n I_{L(n)}(\mathbf{T}(\mathbf{y})) f_n(\mathbf{y}) \exp(\theta_n, \mathbf{T}(\mathbf{y})),$$

where $L(n)$ is the face of $K(n)$ that contains $\mathbf{t}$ in its relative interior. One has

$$\sum_{\mathbf{y} \in Y} c_n I_{L(n)}(\mathbf{T}(\mathbf{y})) f_n(\mathbf{y}) \exp(\theta_n, \mathbf{T}(\mathbf{y})) = 1$$

and

$$\sum_{\mathbf{y} \in Y} c_n I_{L(n)}(\mathbf{T}(\mathbf{y})) \mathbf{T}(\mathbf{y}) f_n(\mathbf{y}) \exp(\theta_n, \mathbf{T}(\mathbf{y})) = t.$$

In the Newton-Raphson algorithm, (1.20)–(1.23) may be replaced by

$$q_{n\nu}(\mathbf{y}) = c_{n\nu} I_{L(n)}(\mathbf{T}(\mathbf{y})) f_n(\mathbf{y}) \exp(\theta_{n\nu}, \mathbf{T}(\mathbf{y})), \quad \mathbf{y} \in Y,$$

$$c_{n\nu} = 1/\sum_{\mathbf{y} \in Y} I_{L(n)}(\mathbf{T}(\mathbf{y})) f_n(\mathbf{y}) \exp(\theta_{n\nu}, \mathbf{T}(\mathbf{y})),$$

$$\mathbf{m}_{n\nu} = \sum_{\mathbf{y} \in Y} q_{n\nu}(\mathbf{y}) \mathbf{T}(\mathbf{y}),$$

$$\mathbf{\Sigma}_{n\nu} = \sum_{\mathbf{y} \in Y} q_{n\nu}(\mathbf{y}) [\mathbf{T}(\mathbf{y}) - \mathbf{m}_{n\nu}][\mathbf{T}(\mathbf{y}) - \mathbf{m}_{n\nu}]'.$$

Similarly,

$$E(D) = \sum_{\mathbf{y} \in Y} D(\mathbf{y}) q(\mathbf{y})$$

and

$$\bar{E}_n(D) = \sum_{\mathbf{y} \in Y} D(\mathbf{y}) \bar{q}_n(\mathbf{y}).$$

All regularity conditions are trivial. If Q exists,

$$\Pr(Q_n \to_w Q) = 1, \quad \Pr(q_n(y) \to q(\mathbf{y}), \mathbf{y} \in Y) = 1, \quad \Pr(\bar{E}_n(D) \to E(D)) = 1,$$

and

$$n^{1/2}[E_n(D) - E(D)] \to_{\mathscr{D}} N(0, \tau^2(D)).$$

One has

$$\tau^2(D) = \sum_{\mathbf{y} \in Y} [p(\mathbf{y})]^{-1} [q(\mathbf{y})]^2 [D(\mathbf{y}) - E(D) - (\boldsymbol{\delta}_D, \mathbf{T}(\mathbf{y}) - \mathbf{t})]^2$$

for

$$\mathbf{\Sigma} = \sum_{\mathbf{y} \in Y} q(\mathbf{y}) [\mathbf{T}(\mathbf{y}) - \mathbf{t}][\mathbf{T}(\mathbf{y}) - \mathbf{t}]'$$

and

$$\mathbf{\Sigma}\, \boldsymbol{\delta}_D = \sum_{\mathbf{y} \in Y} [D(\mathbf{y}) - E(D)][\mathbf{T}(\mathbf{y}) - \mathbf{t}] q(\mathbf{y}).$$

If $\int \mathbf{T}\, dP = \sum_{\mathbf{y} \in Y} \mathbf{T}(\mathbf{y}) p(\mathbf{y}) = \mathbf{t} (P \in \mathscr{L})$, then

$$\tau^2(D) = \tau_0^2(D) = \sum_{\mathbf{y} \in Y} p(\mathbf{y}) [D(\mathbf{y}) - E(D) - (\boldsymbol{\delta}_D, \mathbf{T}(\mathbf{y}) - \mathbf{t})]^2.$$

To find confidence intervals, note that

$$\bar{\tau}_n^2(D) = \sum_{\mathbf{y}\in Y} \left\{\frac{[\bar{q}_n(\mathbf{y})]^2}{\bar{\mathbf{p}}_n(\mathbf{y})}\right\}[D(\mathbf{y}) - \bar{E}_n(D) - (\bar{\boldsymbol{\delta}}_{Dn}, \mathbf{T}(\mathbf{y}) - \mathbf{t})]^2,$$

$$\bar{\tau}_{0n}^2(D) = \sum_{\mathbf{y}\in Y} \bar{q}_n(\mathbf{y})[D(\mathbf{y}) - \bar{E}_n(D) - (\boldsymbol{\delta}_{Dn}, \mathbf{T}(\mathbf{y}) - \mathbf{t})]^2,$$

$$\textstyle\sum_n = \sum_{\mathbf{y}\in Y} \bar{q}_n(\mathbf{y})[\mathbf{T}(\mathbf{y}) - \mathbf{t}][\mathbf{T}(\mathbf{y}) - \mathbf{t}]',$$

and

$$\textstyle\sum_n \bar{\boldsymbol{\delta}}_{Dn} = \sum_{\mathbf{y}\in Y}[D(\mathbf{y}) - E(D)][\mathbf{T}(\mathbf{y}) - \mathbf{t}]\bar{q}_n(\mathbf{y}).$$

All results presented here are found in Haberman (1974, Chapter 9; 1979, Chapter 9). Kullback and associates also provide some sketches of proofs, and Causey (1972) presents a special case. Kullback and associates note that MDIA leads to asymptotically efficient estimates in this example under the model $\int \mathbf{T}\, dP = \mathbf{t}$, even though MDIA and maximum likelihood estimation need not coincide in such a model. If $p(\mathbf{y}) > 0$, $\mathbf{y} \in Y$, this claim is supported by results of Haberman (1977).

The example of Deming and Stephan (1940) presented earlier is included in this case. Other examples of constraints on marginal totals of contingency tables are considered by Deming and Stephan (1940), Ireland and Kullback (1968a), Mosteller (1968), Causey (1972), and Haberman (1974, Chapter 9; 1979, Chapter 9), among others. Applications to symmetry assumptions are considered in Ireland and Kullback (1968b) and Kullback (1971), who also provide other applications. See also Gokhale and Kullback (1980).

CASE 3. *Overlapping strata.* Suppose that the distribution P is to be adjusted so that some selected probabilities have desired values. Thus for Borel sets $G(k)$, $1 \le k \le b$, P is to be adjusted so that the new distribution Q assigns probability t_k to $G(k)$, $1 \le k \le b$.

It is quite acceptable to have $G(k)$ that are not disjoint. The earlier example of Deming and Stephan (1940) illustrates such a problem. In that case, P had mass at $\langle j, k\rangle$, $1 \le j \le r$, $1 \le k \le s$, and the $G(k)$ were the sets $\{\langle j, k\rangle : 1 \le k \le s\}$, $1 \le j \le r$, and $\{\langle j, k\rangle : 1 \le j \le r\}$, $1 \le k \le s$. Nonetheless, this example is meaningful even if P is a continuous distribution. For example, a sample might be available in which income and age are accurately observed; however, in the reference population all that is available are relative sizes of age groups and income groups, with no reference cross-classification of age group by income group.

In this example, $\int \mathbf{T}\, dP = \mathbf{t}$ if $t = \langle t_k : 1 \le k \le b\rangle$ and $\mathbf{T} = \langle I_{G(k)} : 1 \le k \le b\rangle$. Since $T_k = I_{G(k)}$ is bounded, regularity conditions are little problem here. The adjustment Q exists if and only if $\mathbf{t}$ is in K, the convex support of $P\mathbf{T}^{-1}$. The adjustment $\bar{Q}_n$ exists if and only if for some $z_{in} \ge 0$, $1 \le i \le n$, $\sum_{i=1}^n z_{in} = 1$ and $\sum_{i=1}^n z_{in} I_{G(k)}(\mathbf{X}_i) = t_k$, $1 \le k \le b$.

Use of the Newton-Raphson algorithm is routine here; however, the iterative proportional fitting algorithm is also available. For descriptions, see Deming and Stephan (1940).

Given that Q exists, $\Pr(\bar{Q}_n \to_w Q) = 1$ and $\Pr(\bar{E}_n(D) \to E(D)) = 1$ whenever $\int D\,dP$ is finite. If $\int D^2\,dP < \infty$, then $n^{1/2}[\bar{E}_n(D) - E(D)] \to_{\mathscr{D}} N(0, \tau^2(D))$ and confidence intervals may be obtained in the usual fashion. No special simplifications are found in formulas.

CASE 4. *Moment adjustment.* The continuous analogous of Section 1 to the Deming-Stephan problem illustrates adjustment by moments. Let the distribution P be adjusted so that the mean μ_k of X_{k1} is transformed to t_k, $1 \le k \le a = b$. Then one requires that $\int \mathbf{X}\,dQ = \mathbf{t}$; i.e., $\mathbf{T}$ is the identity. No special simplifications occur here. One simply replaces $\mathbf{T}$ by $\mathbf{X}$ and $\mathbf{T}_i$ by X_i in all formulas.

3. Strong consistency.

To prove strong consistency, it is important to reduce the estimation problem to one of maximization of strictly concave functions. To begin, the behavior of the convex hull $K(n)$ of $\mathbf{T}_i$, $1 \le i \le n$, will be considered.

THEOREM 1. *Assume* $\mathbf{t} \in K$. *Then with probability* 1, $\bar{Q}_n$ *exists for all but at most a finite number of* n.

PROOF. By Section 1.1, it suffices to show that with probability 1, $\mathbf{t} \in K(n)$ for n sufficiently large. Since $K(n) \subset K(n+1)$, one may consider $K(\infty) = \bigcup_{n=1}^{\infty} K(n)$. One must show that $\Pr(\mathbf{t} \in K(\infty)) = 1$. The vector $\mathbf{t}$ is in K if and only if for some points $\mathbf{x}_j$, $1 \le j \le c$, $\mathbf{t} = \sum_{j=1}^{c} \alpha_j \mathbf{x}_j$, where $\sum_{j=1}^{c} \alpha_j = 1$, $\alpha_j > 0$, $1 \le j \le c$, and for every open neighborhood U of an $\mathbf{x}_j$, $1 \le j \le c$, $P(U) > 0$. For some open U_j, $1 \le j \le c$, one thus has $P(U_j) > 0$, $1 \le j \le c$, with $\mathbf{t}$ in the relative interior of the convex hull of $\mathbf{y}_j$, $1 \le j \le c$, for any $\mathbf{y}_j \in U_j$, $1 \le j \le c$. The probability that no $\mathbf{T}_i \in U_j$, $1 \le i < \infty$, is 0, so with probability 1, some $\mathbf{T}_i$ is in each U_j, $1 \le j \le c$, and $\mathbf{t} \in K(\infty)$. □

The same method of proof may be used in the following theorem.

THEOREM 2. *Assume* $\mathbf{t} \in K$. *Then the probability is* 1 *that* $L(n)$ *is a relatively open subset of* L *for all sufficiently large* n.

At this point, a related maximization problem needs to be considered. Assume Q exists. Let $H = \operatorname{span}\{\mathbf{u} - \mathbf{t} : \mathbf{u} \in L\}$. Assume $\int \exp(\mathbf{w}, \mathbf{T})\,dQ < \infty$ for $\|\mathbf{w}\| < \varepsilon$, for some $\varepsilon > 0$. Let

$$\begin{aligned} \ell(\mathbf{w}) &= -\log \textstyle\int_M \exp(\mathbf{w}, \mathbf{T})\,dP + (\mathbf{w}, \mathbf{t}) \\ &= -\log \textstyle\int_L \exp(\mathbf{w}, \mathbf{u})\,dP\mathbf{T}^{-1}(\mathbf{u}) + (\mathbf{w}, \mathbf{t}), \quad \mathbf{w} \in H \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} \ell_n(\mathbf{w}) &= -\log[(1/n) \textstyle\sum_{i=1}^{n} \exp(\mathbf{w}, \mathbf{T}_i) I_M(\mathbf{X}_i)] + (\mathbf{w}, \mathbf{t}) \\ &= -\log[(1/n) \textstyle\sum_{i=1}^{n} I_L(\mathbf{T}_i) \exp(\mathbf{w}, \mathbf{T}_i)] + (\mathbf{w}, \mathbf{t}), \quad \mathbf{w} \in H. \end{aligned} \tag{3.2}$$

The function ℓ is a strictly concave function (Berk, 1972). In (1.9) and (1.10), addition to or subtraction from θ of any vector δ in the orthogonal complement 7 of H has no effect on the equations, so one may assume without loss of generality that $\theta \in H$. As in Berk (1972), ℓ has its unique maximum at θ. Similarly, if $L(n)$ is open relative to L, θ_n may be assumed to be in H and ℓ_n is a strictly concave function with unique maximum at θ_n. Given this relationship of ℓ to θ and ℓ_n to θ_n, the following basic consistency result can be proven.

THEOREM 3. *Assume Q exists and* $\int \exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$, $\mathbf{w} \in N$, *where N is an open neighborhood of θ. Let $\theta \in H$ and $\theta_n \in H$ for* ri $L(n)$ *open relative to L. Then* $\Pr(\theta_n \to \theta) = 1$.

PROOF. By the strong law of large numbers,

$$n^{-1} \textstyle\sum_{i=1}^{n} I_M(\mathbf{X}_i)\exp(\mathbf{w}, \mathbf{T}_i) \to \displaystyle\int_M \exp(\mathbf{w}, \mathbf{T})\, dP \tag{3.3}$$

with probability 1 for each $\mathbf{w} \in H$. For any countable dense subset Z of H, (3.3) implies $\Pr\{\ell_n(\mathbf{w}) \to \ell(\mathbf{w}),\ \mathbf{w} \in Z\} = 1$. Let $J \subset H$ be closed and convex and let θ be in the relative interior of J. Let ri J be open relative to H, and let $J - \theta = \{w - \theta : \mathbf{w} \in J\} \subset N$. Since ℓ_n is concave, Rockafellar (1970, page 90, Theorem 10.8) implies that $\max_{\mathbf{w}\in J} | \ell_n(\mathbf{w}) - \ell(\mathbf{w}) | \to 0$ with probability 1. Let ∂J be the relative boundary of J. Then $\max_{\mathbf{w}\in\partial J} \ell(\mathbf{w}) < \ell(\theta)$ and with probability 1, $\max_{\mathbf{w}\in\partial J} \ell_n(\mathbf{w}) < \ell n(\theta)$ for all but a finite number of n. Thus the probability is 1 that $\theta_n \in$ ri J for n sufficiently large. Since J is arbitrary, $\Pr(\theta_n \to \theta) = 1$. □

The remaining consistency results follow from Theorem 3. An important preliminary result is Lemma 1.

LEMMA 1. *Let D be a real measurable function on R^a such that* $\int (1 + |D|)\exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$, *for all $\mathbf{w}$ in an open neighborhood N of θ. Assume Q exists. Then*

$$F_n(D) = \frac{1}{n} \textstyle\sum_{i=1}^{n} I_M(\mathbf{X}_i) D_i \exp(\theta_n, \mathbf{T}_i) \to F(D) = \displaystyle\int_M D \exp(\theta, \mathbf{T})\, dQ \tag{3.4}$$

with probability 1.

PROOF. Consider

$$F_{0n}(D) = (1/n) \textstyle\sum_{i=1}^{n} I_M(\mathbf{X}_i) D_i \exp(\theta, \mathbf{T}_i). \tag{3.5}$$

By the strong law of large numbers,

$$F_{0n}(D) \to F(D) \tag{3.6}$$

with probability 1. By Taylor's theorem,

$$F_n(D) - F_{0n}(D) = (\theta_n - \theta, (1/n) \textstyle\sum_{i=1}^{n} I_M(\mathbf{X}_i) D_i \exp(\theta_n^*, \mathbf{T}_i)\mathbf{T}_i) \tag{3.7}$$

for some θ_n^* on a line segment between θ and θ_n. Let θ be in the interior of the convex hull of $\mathbf{w}_k$, $1 \le k \le b+1$, where $\int |D| \exp(\mathbf{w}_k, \mathbf{T})\, dP < \infty$. Then the probability is 1 that for all sufficiently large n,

$$\theta_n^* = \textstyle\sum_{k=1}^{b+1} \alpha_{kn}\mathbf{w}_k \tag{3.8}$$

for some $\alpha_{kn} \ge 0$, $1 \le k \le b+1$, such that $\sum_{k=1}^{b+1} \alpha_{kn} = 1$. For such n,

$$\exp(\theta_n^*, \mathbf{T}_i) \le \textstyle\sum_{k=1}^{b+1} \alpha_{kn}\exp(\mathbf{w}, \mathbf{T}_i) < \sum_{k=1}^{b+1} \exp(\mathbf{w}, \mathbf{T}_i) \tag{3.9}$$

and

$$|(1/n) \textstyle\sum_{k=1}^{n} I_M(\mathbf{X}_i)D_i\exp(\theta_n^*, \mathbf{T}_i)| < \sum_{k=1}^{b+1} (1/n) \sum_{i=1}^{n} |D_i| \exp(\mathbf{w}_k, \mathbf{T}_i). \tag{3.10}$$

With probability 1, the right-hand side of (3.10) has a finite limit as $n \to \infty$. Since $\theta_n - \theta \to 0$ with probability 1, (3.7)–(3.10) imply that $F_n(D) - F_{0n}(D) \to 0$ with probability 1. The conclusion of the lemma is immediate. □

The first consequence of Lemma 1 and Theorem 1 is the following result concerning strong consistency of $\bar{E}_n(D)$.

THEOREM 4. *Assume Q exists and D satisfies the conditions of Lemma* 2. *Then*

$$\bar{E}_n(D) \to E(D) \tag{3.11}$$

with probability 1.

PROOF. Clearly

$$\bar{E}_n(D) = F_n(D)/F_n(1) \tag{3.12}$$

and

$$E(D) = F(D)/F(1), \tag{3.13}$$

where $F(1) > 0$. The conditions of Lemma 1 apply to D and to the constant function 1. Thus $F_n(D) \to F(D)$ and $F_n(1) \to F(1)$ with probability 1. Hence (3.11) holds. □

One special case, the indicator function $D = I_B$, leads to the desired result concerning weak convergence. One has

COROLLARY 1. *Let B be a Borel set in R^a. Let the conditions of Theorem* 1 *hold. Then $\bar{Q}_n(B) \to Q(B)$ with probability* 1.

PROOF. Just note that $\bar{Q}_n(B) = \bar{E}_n(I_B)$ and $E(I_B) = Q(B)$. Since

$$\int |I_B| \exp(\mathbf{w}, \mathbf{T})\, dP \le \int exp(\mathbf{w}, \mathbf{T})\, dP$$

is assumed finite in a neighborhood of θ, Theorem 2 applies. □

The final result involves weak convergence.

THEOREM 5. *Assume Q exists and $\int \exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$ for $\mathbf{w}$ in an open neighborhood of* $\mathbf{0}$. *Then* $\Pr(\bar{Q}_n \to_w Q) = 1$.

PROOF. Note that R^a is a separable metric space. By Corollary 1 and Parthasarthy (1967, pages 47, 53), the result follows. □

Results of this section are easily derived from Haberman (1974, Chapter 9) if P has support on a finite number of points. They may be generalized somewhat in two ways. First, use of R^a matters little. Results apply to any probability measure P on a separable metric space. Second, use of independent and identically distributed observations is of only limited importance. The key requirement is that $(1/n) \sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}, \mathbf{T}_i) \to \int_M \exp(\mathbf{w}, \mathbf{T})\, dP$ with probability 1 for $\mathbf{w}$ in a neighborhood of $\boldsymbol{\theta}$.

4. Asymptotic normality. Asymptotic normality results of Section 1.3 may be proven fairly easily given the basic structure established in Section 3. To begin, consider the asymptotic behavior of $\boldsymbol{\theta}_n$.

THEOREM 6. *Assume Q exists and $\int \exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$ for $\mathbf{w}$ in an open neighborhood N of* $\mathbf{0}$. *Assume $\int \|\mathbf{T}\|^2 (dQ/dP)\, dQ < \infty$. Let*

$$\Sigma^* = \int (\mathbf{T} - \mathbf{t})(\mathbf{T} - \mathbf{t})' \left(\frac{dQ}{dP}\right) dQ. \tag{4.1}$$

Let $\boldsymbol{\theta}_n \in H$ whenever ri $L(n)$ *is open relative to L. Let Σ^- be the Moore-Penrose generalized inverse of Σ. Then*

$$n^{1/2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \to_{\mathscr{D}} N(\boldsymbol{\theta}, \Sigma^- \Sigma^* \Sigma^-). \tag{4.2}$$

REMARK. In the special case of $\boldsymbol{\theta} = \mathbf{0} (P = Q)$, $\Sigma^- \Sigma^* \Sigma^- = \Sigma^-$.

PROOF. For $\mathbf{w} \in R^b$, let

$$\mathbf{m}_n(\mathbf{w}) = \left[\frac{1}{n} \sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}, \mathbf{T}_i)\right]^{-1} \frac{1}{n} \sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}, \mathbf{T}_i)\mathbf{T}_i \tag{4.3}$$

and

$$\Sigma_n(\mathbf{w}) = \left[\frac{1}{n} \sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}, \mathbf{T}_i)\right]^{-1} \cdot \frac{1}{n} \sum_{i=1}^n [\mathbf{T}_i - \mathbf{m}_n(\mathbf{w})][T_i - \mathbf{m}_n(\mathbf{w})]' I_M(\mathbf{X}_i)\exp(\mathbf{w}, \mathbf{T}_i). \tag{4.4}$$

By Taylor's theorem, the probability is 1 that for n sufficiently large and for some functions $\mathbf{w}_n$ from R^b to R^b and α_n from R^b to $(0, 1)$,

$$\mathbf{w}_n(\mathbf{x}) = \alpha_n(\mathbf{x})\boldsymbol{\theta} + [1 - \alpha_n(\mathbf{x})]\boldsymbol{\theta}_n, \quad \mathbf{x} \in R^b, \tag{4.5}$$

and

$$(\mathbf{x}, \mathbf{m}_n(\theta)) + (\mathbf{x}, \textstyle\sum_n(\mathbf{w}_n(\mathbf{x}))(\theta_n - \theta)) = (\mathbf{x}, \mathbf{m}_n(\theta_n)) = (\mathbf{x}, \mathbf{t}), \quad \mathbf{x} \in R^b. \tag{4.6}$$

As in the proof of Lemma 1,

$$\frac{1}{n} \textstyle\sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}_n(\mathbf{x}), \mathbf{T}_i) \to \displaystyle\int_M \exp(\theta, \mathbf{T})\, dP \tag{4.7}$$

$$\frac{1}{n} \textstyle\sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}_n(\mathbf{x}), \mathbf{T}_i)\mathbf{T}_i \to \displaystyle\int_M \exp(\theta, \mathbf{T})\mathbf{T}\, dP, \tag{4.8}$$

and

$$\frac{1}{n} \textstyle\sum_{i=1}^n I_M(\mathbf{X}_i)\exp(\mathbf{w}_n(\mathbf{x}), \mathbf{T}_i)\mathbf{T}_i\mathbf{T}_i' \to \displaystyle\int_M \exp(\theta, \mathbf{T})\mathbf{T}\mathbf{T}'\, dP. \tag{4.9}$$

Thus

$$\textstyle\sum_n(\mathbf{w}_n(\mathbf{x})) \to \sum \tag{4.10}$$

with probability 1. Since

$$\mathbf{t} = \frac{\int_M \exp(\theta, \mathbf{T})\mathbf{T}\, dP}{\int_M \exp(\theta, \mathbf{T})\, dP}, \tag{4.11}$$

elementary large sample theory implies that

$$n^{1/2}[\mathbf{t} - \mathbf{m}_n(\theta)] \to_{\mathscr{D}} N(0, \textstyle\sum^*). \tag{4.12}$$

Let $\mathbf{x}_k \in R^b$, $1 \le k \le b$, satisfy

$$\begin{aligned}(\mathbf{x}_k, \textstyle\sum \mathbf{x}_{k'}) &= 1, \quad k = k', \\ &= 0, \quad k \neq k'.\end{aligned} \tag{4.13}$$

Then

$$\begin{aligned}\textstyle\sum_{k=1}^b (\mathbf{x}_k, \mathbf{t} - \mathbf{m}_n(\theta)) &= \textstyle\sum_{k=1}^b (\mathbf{x}_k, \sum_n(\mathbf{w}_n(\mathbf{x}_k))(\theta_n - \theta)) \\ &= (\theta_n - \theta, \textstyle\sum (\theta_n - \theta))^2 + \varepsilon_n,\end{aligned} \tag{4.14}$$

where $\varepsilon_n/(\theta_n - \theta, \sum (\theta_n - \theta))^2 \to 0$ with probability 1. Thus $\theta_n - \theta$ is of order $n^{-1/2}$. By (4.6), (4.10), and (4.12),

$$(\mathbf{x}, \textstyle\sum n^{1/2}(\theta_n - \theta)) \to_{\mathscr{D}} N(0, (\mathbf{x}, \sum \mathbf{x})). \tag{4.15}$$

The conclusion of the theorem follows since $\theta_n - \theta \in H$ for n sufficiently large with probability 1 and since $\mathbf{x}$ is arbitrary (Rao, 1973, page 128). □

LEMMA 2. *Let D be a real measurable function on R^a such that $\int D^2(dQ/dP)\, dQ < \infty$, $\int \| T \|^2(dQ/dP)dQ < \infty$, and $\int (1 + D^2)\exp(\mathbf{w}, \mathbf{T})\, dQ < \infty$ for $\mathbf{w}$ in an* *open neighborhood of* **0**. *Let*

$$\Sigma\, \gamma_D = \int_M D \exp(\boldsymbol{\theta}, \mathbf{T})\mathbf{T}\, dP \tag{4.16}$$

and

$$\sigma^2(D) = \int [DI_M \exp(\boldsymbol{\theta}, \mathbf{T}) - F(D) - cI_M \exp(\boldsymbol{\theta}, \mathbf{T})(\mathbf{T} - \mathbf{t}, \boldsymbol{\theta}_D)]^2\, dP. \tag{4.17}$$

Then

$$n^{1/2}[F_n(D) - F(D)] \to_{\mathscr{D}} N(0, \sigma^2(D)). \tag{4.18}$$

PROOF. The proof of Theorem 6 also implies that

$$n^{1/2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) - n^{1/2}\, \Sigma^{-} \frac{1}{n} \sum_{i=1}^n cI_M(\mathbf{X})\exp(\boldsymbol{\theta}, \mathbf{T}_i)(\mathbf{T}_i - \mathbf{t}) \to_P 0, \tag{4.19}$$

where $\to_p$ denotes convergence in probability. By (3.5) and (3.7)

$$\begin{aligned} n^{1/2}[F_n(D) - F(D)] &= n^{1/2}\left[\frac{1}{n} \sum_{i=1}^n D_i I_M(\mathbf{X}_i)\exp(\boldsymbol{\theta}, \mathbf{T}_i) - F(D)\right] \\ &= \left(n^{1/2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}), \frac{1}{n} \sum_{i=1}^n D_i I_M(\mathbf{X}_i)\exp(\boldsymbol{\theta}_n^*, \mathbf{T}_i)\mathbf{T}_i\right). \end{aligned} \tag{4.20}$$

By the same argument used in Lemma 1, one has

$$\frac{1}{n} \sum_{i=1}^n D_i I_M(\mathbf{X}_i)\exp(\boldsymbol{\theta}_n^*, \mathbf{T}_i)\mathbf{T}_i \to \int_M D \exp(\boldsymbol{\theta}, \mathbf{T})\mathbf{T}\, dP \tag{4.21}$$

with probability 1. By (4.19) – (4.21),

$$n^{1/2}[F_n(D) - F(D)] - n^{-1/2} \sum_{i=1}^n C_i(D) \to_p 0, \tag{4.22}$$

where

$$C_i(D) = I_M(\mathbf{X}_i)D_i \exp(\boldsymbol{\theta}, T_i) - F(D) - I_M(\mathbf{X}_i)\exp(\boldsymbol{\theta}, \mathbf{T}_i)(\mathbf{T}_i - \mathbf{t}, \gamma_D) \tag{4.23}$$

has mean 0 and variance $\sigma^2(D)$. By the central limit theorem

$$n^{-1/2} \sum_{i=1}^n C_i(D) \to_{\mathscr{D}} N(0, \sigma^2(D)). \tag{4.24}$$

By (4.22) and (4.24), (4.18) must hold. □

Given Lemma 2, Theorem 7, the principal result of this section, readily follows.

THEOREM 7. *Let the conditions of Lemma* 2 *hold. Then*

$$n^{1/2}[\bar{E}_n(D) - E(D)] \to_{\mathscr{D}} N(0, \tau^2(D)). \tag{4.25}$$

PROOF. Given Lemma 2, (3.12), and (3.13), it follows from standard large-sample theory that

$$n^{1/2}[\bar{E}_n(D) - E(D)] - n^{1/2}\left\{\frac{F_n(D)}{F(1)} - \frac{F(D)F_n(1)}{[F(1)]^2}\right\} \to_p 0. \tag{4.26}$$

Since

$$\frac{F_n(D)}{F(1)} - \frac{F(D)F_n(1)}{[F(1)]^2} = F_n(D') - F(D') \tag{4.27}$$

for

$$D' = \frac{[D - E(D)]}{F(1)} = c[D - E(D)] \tag{4.28}$$

and since

$$F(D') = 0, \tag{4.29}$$

Lemma 2 and (4.26) imply that (4.25) holds. □

4.1. *Confidence intervals.* Given Theorem 7, asymptotic confidence intervals for the expected value $E(D)$ are readily constructed, as noted in Section 1.4. By use of Lemma 1, one readily finds that $\bar{\tau}_n^2(D) \to \tau^2(D)$ with probability 1 whenever D satisfies the conditions of Lemma 1, the conditions of Theorem 3 hold, and, in addition,

$$\int (1 + D^2)\exp(\mathbf{w}, \mathbf{T})\left(\frac{dQ}{dP}\right) dQ < \infty \tag{4.30}$$

for $\mathbf{w}$ in an open neighborhood N of $\mathbf{0}$. Provided that $\tau^2(D) > 0$, one then has

$$\Pr\left\{\bar{E}_n(D) - \frac{z_{\alpha/2}\bar{\tau}_n(D)}{n^{1/2}} \le E(D) \le \bar{E}_n(D) + \frac{z_{\alpha/2}\bar{\tau}_n(D)}{n^{1/2}}\right\} \to 1 - \alpha \tag{4.31}$$

for any α, $0 < \alpha < 1$. The condition $\tau^2(D) > 0$ is satisfied if for no $\alpha \in R$ and $\beta \in R^b$ is $\Pr\{D(\mathbf{X}^*) = \alpha + (\beta, \mathbf{T}(\mathbf{X}^*))\} = 1$ for $\mathbf{X}^*$ with distribution Q. If one assumes in advance that $P = Q$, then $\bar{\tau}_n^2(D)$ may be replaced by $\bar{\tau}_{0n}^2(D)$ without changing results.

4.2. *Remarks.* As in Section 3, the results of this section can be further generalized, although generalizations are not trivial. The main case of interest involves complex sampling procedures in which the $\mathbf{X}_i$ are not independent and identically distributed. Typically, results are still available, but asymptotic variances are changed.

REFERENCES

BERK, R. H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43** 193–204.

CAUSEY, B. (1972). Sensitivity of raked contingency table totals to changes in problem conditions. *Ann. Math. Statist.* **43** 656–658.

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

DEMING, W. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11** 423–444.

DEMPSTER, A. P. (1971). An overview of multivariate data analysis. *J. Multivariate Anal.* **1** 316–346.

GOKHALE, D. V. and KULLBACK, S. (1978). *The Information in Contingency Tables.* Dekker, New York.

GOOD, I. J. (1963). Maximum entropy for hypothesis formation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data.* University of Chicago Press.

HABERMAN, S. J. (1977). Product models for frequency tables involving indirect observation. *Ann. Statist.* **6** 1124–1147.

HABERMAN, S. J. (1979). *Analysis of Qualitative Data*, Vol. 2. Academic, New York.

IRELAND, C. T., KU, H. N. and KULLBACK, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Amer. Statist. Assoc.* **64** 1123–1341.

IRELAND, C. T. and KULLBACK, S. (1968a). Contingency tables with given marginals. *Biometrika* **55** 179–188.

IRELAND, C. T. and KULLBACK, S. (1968b). Minimum discriminant information estimation. *Biometrics* **24** 707–713.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

KULLBACK, S. (1959). *Information Theory and Statistics.* Wiley, New York.

KULLBACK, S. (1971). Marginal homogeneity of multidimensional contingency tables. *Ann. Math. Statist.* **42** 594–606.

MOSTELLER, F. (1968). Association and estimation in contingency tables. *J. Amer. Statist. Assoc.* **63** 1–28.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces.* Academic, New York.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton University Press.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
91905 JERUSALEM
ISRAEL