COMMUN. STATIST.-THEORY METH., 18(5), 1843-1861 (1989)

BAYESIAN ESTIMATION OF PROPORTIONS WITH A CROSS-ENTROPY PRIOR

Arthur T. Denzau
Economics Dept.
Washington Univ.
St. Louis, MO 63130

Patrick C. Gibbons
Physics Dept.
Washington Univ.
St. Louis, MO 63130

Edward Greenberg
Economics Dept.
Washington Univ.
St. Louis, MO 63130

Key words and Phrases: Bayesian methods, contingency tables, IPF, log-linear model, logit model, maximum entropy.

ABSTRACT

This paper suggests estimators of the frequencies (N_s) or proportions (N_s/N) of N distinguishable objects contained in S categories, given various types of information. We consider information in the form of exact constraints on the N_s, sample frequencies, and frequencies of related data. The analysis uses Bayesian methods, where the prior distribution is assumed to be a function of the cross-entropy between the N_s and a reference distribution. We show the relationship between our estimator and the log-linear and logit models and also present a sampling experiment to compare our proposed estimator with the iterated proportional fitting estimator.

1. INTRODUCTION

We are concerned with the problem of estimating finite population parameters, specifically the vector

$$\vec{N} = (N_1, \ldots, N_S),$$

where N_s is the number of objects in category s $(s = 1, \ldots, S)$ and $\sum N_s = N$. The Bayesian method described in the paper is applicable to a situation in which the statistician possesses prior estimates that are considered to be quite accurate. It is shown that this assumption leads to a prior distribution for $\vec{N}$ that is a function of the cross-entropy between $\vec{N}$ and $\vec{K} = (K_1, \ldots, K_S)$,

$$H(\vec{N}|\vec{K}) = -\sum (N_s/N) \log[(N_s/N)/(K_s/K)], \tag{1}$$

where $K = \sum K_s$.

The problem is examined in two settings. In the first, it is assumed that a set of linear constraints involving the N_s are known exactly; for example, some or all of the row and column totals may be known in census data. In the second, along with a set of linear constraints, sample data are available for individual cells. U. S. Census data provide an example: a complete enumeration is taken for such variables as race and age, while income information is collected only for a sample.

In Section 2 we show how (1) can be derived from a hypothetical hyper-population with parameter $\vec{K}$. The derivation requires the assumption that $K_s \gg N_s$, which corresponds to considerable confidence on the part of the statistician that the assumed values of K_s are accurate.

The approach is applied in Section 3 to the situation in which a set of linear constraints involving the N_s is known exactly, but no sample information is available. The mode of the prior distribution for the set of N_s that satisfy the constraints is suggested as a point estimator, and the relationship between this estimator and the log-linear and logit approaches to the problem is explained. We also consider an example in which the elements of a $2 \times 3 \times 4$ table are estimated with several different priors to explore sensitivity of results to the particular prior utilized.

In Section 4 the information base is assumed to include sample proportions as well as exact linear constraints. The posterior distribution for the N_s is obtained by multiplying the prior by the likelihood, and its mode is suggested as a point

estimator. Results of a sampling experiment are reported that compares the proposed estimator with the iterated proportional fitting technique. Section 5 contains conclusions.

2. THE CROSS-ENTROPY PRIOR

Following Jaynes [1986], we begin by sketching an argument that implies a prior density for $\vec{N}$ that is proportional to $\exp\{NH(\vec{N}|\vec{K})\}$. Assume an exchangeable, infinite sequence $\{X_i\}$ of which the first N terms are $\{X_1, \ldots, X_N\}$, let $y_i = f(X_i)$ have range $(1, \ldots, S)$, and $N_s = \#(y_i = s)$. Then by the de Finetti theorem on exchangeable sequences there exists a $g(\vec{\rho})$ such that

$$P(\vec{N}) = \frac{N!}{\prod N_s!} \int \cdots \int \prod (\rho_s^{N_s}) g(\vec{\rho})\, d\vec{\rho}, \tag{2}$$

where $\sum \rho_s = 1$ and $\vec{\rho} = (\rho_1, \ldots, \rho_S)$. Equation (2) may be interpreted to mean that the N_s are determined by multinomial sampling from a distribution with parameters ρ_s, where $g(\vec{\rho})$ is a prior distribution for the ρ_s. We assume the Dirichlet form for $g(\vec{\rho})$,

$$g(\vec{\rho}) = \frac{\Gamma(K)}{\prod \Gamma(K_s)} \prod \rho_s^{K_s - 1}, \tag{3}$$

because of its familiarity and convenience. See Zabell [1982] for further discussion of this distribution.

From (2) and (3), we have

$$P(\vec{N}|\vec{K}) = \frac{N!\Gamma(K)}{\Gamma(N+K)} \prod \frac{\Gamma(N_s + K_s)}{N_s!\Gamma(K_s)}. \tag{4}$$

To interpret $\vec{K}$ we note that (4) implies

$$P(y_{N+1} = s|\vec{N}, \vec{K}) = (N_s + K_s)/(N + K), \tag{5}$$

which is a weighted average of N_s/N and K_s/K, with weights $N/(N+K)$ and $K/(N+K)$, respectively. This result is equivalent to a situation in which we start with prior information that the proportion of objects in cell s is K_s/K and then a sample of size N yields the proportion N_s/N. Thus, K may be regarded as the weight placed on the prior information, which is contained in $\vec{K}$, measured in units of number of observations.

In some situations it may be valid to assume that $K \gg N$, i. e. the statistician has great confidence in the values of the K_s. This might happen in working with slowly changing demographic data for a time period soon after a complete enumeration or with medical data where a large information base has been previously compiled. When $K \gg N$, (4) can be approximated by

$$P(\vec{N}|\vec{K}) = \frac{N!}{\prod_s N_s!} \prod_s \left(\frac{K_s}{K}\right)^{N_s - 1},$$

i. e. the N_s have a multinomial distribution with parameters K_s/K. Another useful way of writing (4) for $K \gg N$ is shown by Jaynes [1986]; it utilizes the Stirling approximation for $m!$, permitting us to treat the N_s as continuous variables rather than integers. The approximation is given by

$$\log P(\vec{N}|\vec{K}) \sim NH(\vec{N}|\vec{K}) + \text{constant}, \tag{6}$$

We call this the cross-entropy prior because of its relation to the Shannon [1948] expression for entropy.

Although we are primarily concerned with $K \gg N$ in this paper, Jaynes [1986] also considers $N \gg K$. In that case, (4) goes into

$$\log P(\vec{N}|\vec{K}) \sim \sum K_s \log N_s + \text{constant}, \tag{7}$$

which is a form of entropy used by Burg [1967].

3. ESTIMATION WITH EXACT INFORMATION

3.1 The Maximum Entropy Estimate of $\vec{N}$

For our first application, we consider the problem of estimating the cell values of a multidimensional contingency table, given exact information about some or all of the lower-order contingency tables derivable from the complete table. We adopt the prior density $P(\vec{N}|\vec{K})$ that is proportional to $\exp\{NH(\vec{N}|\vec{K})\}$, and then estimate $\vec{N}$ by the value that maximizes that prior density, subject to any exact constraints involving the $\vec{N}$; this estimator is called the maxent estimator. For large N this prior will tend to be sharply peaked; hence the mode of the prior should be close to its mean. We assume linear constraints of the form

$$\sum a_{rs} N_s = a_r, \; r = 1, \ldots, R. \tag{8}$$

To obtain the estimates it is convenient to define

$$P_s = N_s/N, \; \vec{P} = (P_1, \ldots, P_s), \; Q_s = K_s/K, \text{and } \vec{Q} = (Q_1, \ldots, Q_S).$$

Then, after rewriting the constraints, we maximize the Lagrangian expression,

$$\begin{aligned} L \;=\; & -\sum_s P_s \log(P_s/Q_s) + \sum_r \lambda_r \sum_s (a_{rs} P_s - a_r/N) + \\ & (1+\lambda_0)(\sum_s P_s - 1) + \text{constant}, \end{aligned} \tag{9}$$

and obtain

$$\hat{P}_s = \frac{Q_s \exp\{\sum_r \lambda_r a_{rs}\}}{\sum_t Q_t \exp\{\sum_r \lambda_r a_{rt}\}}, \tag{10}$$

in which the λ_r are such that the P_s satisfy (8). The λ_r can be found by methods described by Gokhale [1973].

The estimator in the case of equal $\{Q_s\}$ is shown in Figure 1 for $S = 3$. The triangle represents the set of P_1, P_2, and P_3 that are nonnegative and sum to one. The line represents a constraint of the form

$$a_{11}P_1 + a_{12}P_2 + a_{13}P_3 = a_1/N.$$

Constant prior probability contours are approximately concentric circles centered at (1/3,1/3,1/3), and the estimate of $\vec{P}$ is denoted by $\hat{\vec{P}}$, where a probability contour is tangent to the constraint. The general case of unequal $\{Q_s\}$ is depicted in Figure 2. Constant probability contours are approximately ellipses centered at $\vec{Q}$, and the estimate is found where a contour is tangent to the constraint.

Although only row and column sums have been emphasized in the above discussion, it should be clear that other linear or nonlinear constraints could be employed. For example, known values of means, variances, or covariances may be written as linear functions of the N_s. Knowledge about medians or other fractiles imply linear inequality constraints that can be included in the maxent estimation procedure. For example, knowledge of the cell in which the median appears can be written as the inequality constraints,

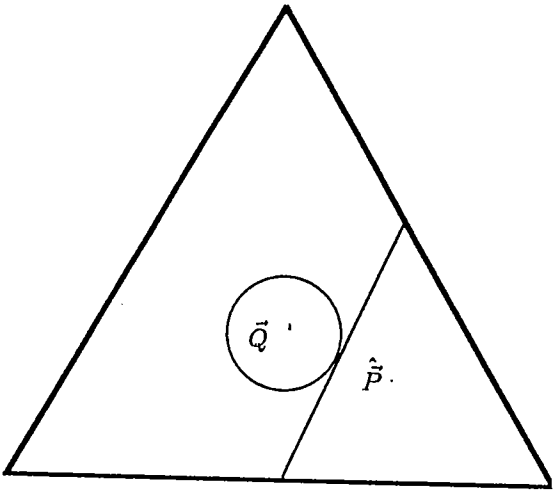

FIG 1 Maxent Estimator $\hat{\vec{P}}$ with Uniform $\vec{Q}$

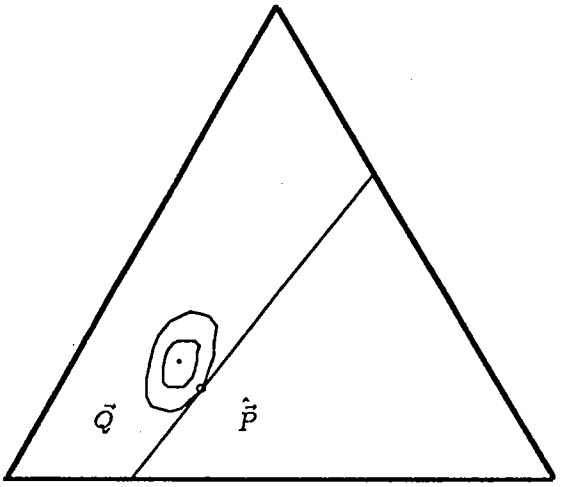

FIG 2 Maxent Estimator $\hat{\vec{P}}$ with Nonuniform $\vec{Q}$

$$\sum_{1}^{m-1} N_s < N/2 \text{ and } \sum_{1}^{m} N_s \geq N/2,$$

where the m^{th} cell is known to contain the median.

The maxent estimator has been advocated as the best way to assign the probability of choosing a member of the s^{th} cell, given knowledge of the exact constraints and no other information (other than that there are S cells); see for example Rosenkrantz [1977, Chapter 3]. In such cases, it is desirable to assign probabilities that introduce as little additional information as possible to the known constraints. Shannon [1948] shows that the negative of entropy is the unique, well behaved measure of information; hence, choosing the maxent distribution realizes this goal. As an example of the way in which little information is introduced, note that no N_s will be estimated as zero unless this is implied by the constraints.

3.2 Relation Between Maxent, Log-linear and Logit Models

The relation between the above procedure and the log-linear model in the problem of table reconstruction with partial information is next illustrated with an example contained in Bishop, Fienberg, and Holland [1975, pp. 107-111]. They estimate a 5-way table, where one variable is membership in the Communist Party of the U.S.S.R. and the other four are demographic variables. The 4-way table among the demographic variables is known, as are three of the four 2-way tables between membership and the demographic variables. In accordance with the log-linear model, they postulate that $\log M_{ijklm}$, where M_{ijklm} is the expected value of the distribution that is assumed to have generated the number of observations in cell $ijklm$, may be expressed as the sum of an overall mean, individual classification effects, and the effects of interactions between classifications. Although most applications of the log-linear model are concerned with testing for the presence of various interactions when complete information is available, in the present example the interactions specified are limited to those for which data are available. While this is a reasonable way to use the data, it should be noted that the omission of interactions for which no data are available does not follow from any theoretical considerations.

TABLE I

Actual, Qs, and Estimated Proportions

Category*	Kentucky Actual	Census Qs	South Qs	All one-way constraints: Uniform Qs	All one-way constraints: Census Qs	All one-way constraints: South Qs	Race-residence constraints: Uniform Qs	Race-residence constraints: Census Qs	Race-residence constraints: South Qs
1-U-W	0.0918	0.0807	0.1700	0.1566	0.1114	0.1479	0.1159	0.0648	0.1482
1-U-B	0.0228	0.0155	0.0650	0.0101	0.0128	0.0179	0.0123	0.0125	0.0293
2-U-W	0.1639	0.1798	0.1780	0.1687	0.1634	0.1699	0.1159	0.1444	0.1552
2-U-B	0.0173	0.0273	0.0320	0.0109	0.0148	0.0096	0.0123	0.0220	0.0144
3-U-W	0.1272	0.1735	0.1090	0.1007	0.1232	0.1053	0.1159	0.1393	0.0950
3-U-B	0.0068	0.0120	0.0080	0.0065	0.0051	0.0024	0.0123	0.0097	0.0036
4-U-W	0.0808	0.1435	0.0750	0.0558	0.0802	0.0588	0.1159	0.1153	0.0654
4-U-B	0.0021	0.0060	0.0040	0.0036	0.0020	0.0010	0.0123	0.0048	0.0018
1-RNF-W	0.1373	0.0650	0.0830	0.1029	0.1080	0.0974	0.0821	0.0742	0.1052
1-RNF-B	0.0052	0.0137	0.0320	0.0066	0.0135	0.0119	0.0022	0.0052	0.0052
2-RNF-W	0.1159	0.1116	0.0870	0.1109	0.1220	0.1119	0.0821	0.1274	0.1103
2-RNF-B	0.0027	0.0070	0.0160	0.0072	0.0045	0.0065	0.0022	0.0027	0.0026
3-RNF-W	0.0531	0.0719	0.0530	0.0662	0.0614	0.0690	0.0821	0.0821	0.0672
3-RNF-B	0.0007	0.0018	0.0040	0.0043	0.0009	0.0016	0.0022	0.0007	0.0007
4-RNF-W	0.0220	0.0391	0.0360	0.0367	0.0263	0.0380	0.0821	0.0446	0.0456
4-RNF-B	0.0003	0.0007	0.0020	0.0024	0.0003	0.0001	0.0022	0.0003	0.0003
1-RF-W	0.0662	0.0155	0.0120	0.0458	0.0737	0.0452	0.0368	0.0560	0.0491
1-RF-B	0.0017	0.0020	0.0040	0.0030	0.0056	0.0048	0.0007	0.0021	0.0015
2-RF-W	0.0496	0.0142	0.0120	0.0494	0.0444	0.0496	0.0368	0.0513	0.0491
2-RF-B	0.0007	0.0005	0.0020	0.0032	0.0010	0.0026	0.0007	0.0005	0.0008
3-RF-W	0.0210	0.0075	0.0070	0.0295	0.0183	0.0293	0.0368	0.0271	0.0286
3-RF-B	0.0002	a	0.0010	0.0019	0.0001	0.0013	0.0007	0.0001	0.0004
4-RF-W	0.0105	0.0036	0.0050	0.0163	0.0069	0.0170	0.0368	0.0130	0.0205
4-RF-B	0.0001	a	b	0.0011	0.0001	0.0003	0.0007	a	0.0001
Cross-Entropy	---	-0.1240	-0.1975	-0.1764	-0.0192	-0.0549	-0.1489	-0.0539	-0.0462
Squared Deviations	0.0000	0.0167	0.0182	0.0079	0.0017	0.0070	0.0151	0.0082	0.0069

a. Less than .00005 b. Less than .005

*1: Income<$5000; 2: $5000<Income<$10000; 3: $10000<Income<$15000; 4: $15000<Income. U: Urban; RNF: Rural nonfarm; RF: Rural farm. W: White; B: Black.

set greatly improves the reconstruction. Overall, the best reconstruction utilizes the U.S. Q_s and the 1-way constraints. Unfortunately, the latter are not usually available because the Census Bureau does not collect complete information on income. Performance with the race-residence constraints, which are known for the decennial census, does not seem to depend greatly on the set of Q_s that is used. With the exception of perhaps three cells in each case, the reconstructions are good.

As a check on the sensitivity of results to the assumption that $K \gg N$, we computed estimates with the Burg prior (7). For this example, it turned out that results were virtually identical to those presented in Table I. Since estimates depend in a complex way on the prior and the constraints, further research is necessary to understand fully the contribution of each.

4. ESTIMATION WITH EXACT AND SAMPLE INFORMATION

4.1 The Posterior Mode Estimator

We return to the contingency table problem, but now assume that data are available for a sample of size n, with n_s observations in the s^{th} cell, and $\sum n_s = n$. Let

$$\vec{n} = (n_1, \ldots, n_S),$$

$$p_s = n_s/n,$$

$$\vec{p} = (p_1, \ldots, p_S),$$

and

$$m = n/N.$$

As above, the prior distribution for $\vec{N}$ is based on (6) for $\vec{N} \in \Omega$, where Ω is the set of $\vec{N}$ that satisfy any exact constraints that may be known. Good [1963, p. 931] mentions this prior as a possible approach to the problem discussed in this section. Two other prior distributions have been used in this type of problem. Sarndal [1965] employs the Bose-Einstein uniform distribution, and Janardan

[1976] uses the unified multivariate hypergeometric. The cross-entropy prior is given by

$$P(\vec{N}|\Omega,\vec{Q}) = \begin{cases} C_1 \exp\{-\sum N_s \log(N_s/Q_s)\}, & \vec{N} \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The likelihood function, assuming sampling without replacement, is proportional to the multivariate hypergeometric form for the admissible $\vec{N}$,

$$L(\vec{N}|\vec{n}) = \begin{cases} C_2 \prod \binom{N_s}{n_s}, & \vec{N} \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The posterior distribution is therefore given by

$$P(\vec{N}|\Omega,\vec{Q},\vec{n}) = \begin{cases} C_3 \exp\{-\sum N_s \log(N_s/Q_s)\} \prod \binom{N_s}{n_s}, & \vec{N} \in \Omega \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

We take as a point estimator the $\vec{N} \in \Omega$ that maximizes (15), the posterior mode (PM) estimator. Since the posterior distribution is often sharply peaked, this estimator is likely to be near the posterior mean, which is the Bayes estimator relative to a quadratic loss function.

We next examine the PM estimator in the case of linear constraints:

$$\Omega = \{\vec{N} : \sum_s a_{rs} N_s = a_r;\ r = 1, \ldots, R\}.$$

It is convenient to work with the logarithm of (15) and with proportions. Accordingly, for $\vec{N} \in \Omega$, we have

$$\log P(\vec{P}|\Omega,\vec{Q},\vec{n}) = C_4 + N \sum P_s \log Q_s - N \sum (P_s - m p_s) \log(P_s - m p_s), \tag{16}$$

where we have used Stirling's approximation for the factorials and C_4 is independent of P_s.

As an aid in interpreting results, it is convenient to make a further transformation. Define

$$\alpha_s = \frac{N_s - n_s}{N - n} = \frac{P_s - m p_s}{1 - m},$$

which is the proportion of the unsampled population in category s; now rewrite the constraints as

$$\begin{aligned} a_r/N &= \sum a_{rs}P_s = \sum a_{rs}[(1-m)\alpha_s + mp_s] \\ &= (1-m)\sum a_{rs}\alpha_s + m\sum a_{rs}p_s, \end{aligned}$$

which implies that

$$\sum a_{rs}\alpha_s = \frac{a_r/N - m\sum a_{rs}p_s}{1-m} \equiv b_r,\ r = 1,\ldots,R. \tag{17}$$

With these definitions, (16) becomes

$$\log P(\vec{\alpha}|\Omega, \vec{K}, \vec{n}) = C_5 - (1-m)N\sum \alpha_s \log(\alpha_s/Q_s),$$

which is maximized subject to (17). The constraints on the unsampled population are derived by excluding the sample observations from the original population constraints. The resulting $\hat{\alpha}_s$ may therefore be interpreted as the maxent estimate of α_s (the proportion of the unsampled population in category s), and we find $\hat{P}_s$ from

$$\hat{P}_s = mp_s + (1-m)\hat{\alpha}_s. \tag{18}$$

Thus, the PM estimator based on the cross-entropy prior is a weighted average of the sample proportion in category s (weighted by the proportion of the population sampled) and the maxent estimate of the number in category s contained in the unsampled population. As m grows, the influence of the sample proportions increases, and as $m \to 1$, $\hat{P}_s \to P_s$.

Figure 3 illustrates the approach geometrically for $S = 3$. The triangle represents P_s satisfying $\sum P_s = 1$, the solid line represents a constraint satisfied by the true $\vec{P}$, the dashed line represents the constraints on the unsampled proportions (which is parallel to the original constraint because ratios of the a_{rs} are unchanged after sampling), and point $\vec{p}$ is the sample proportions. If $\hat{\vec{\alpha}}$ is the maxent point on the dashed line, the maximum of the posterior distribution is at $\hat{\vec{P}}$, which is on the straight line between $\vec{p}$ and $\hat{\vec{\alpha}}$ because

$$\hat{\vec{P}} = m\vec{p} + (1-m)\hat{\vec{\alpha}},$$

and $\hat{\vec{P}}$ satisfies the original constraints. (The diagram assumes a uniform $\vec{Q}$.)

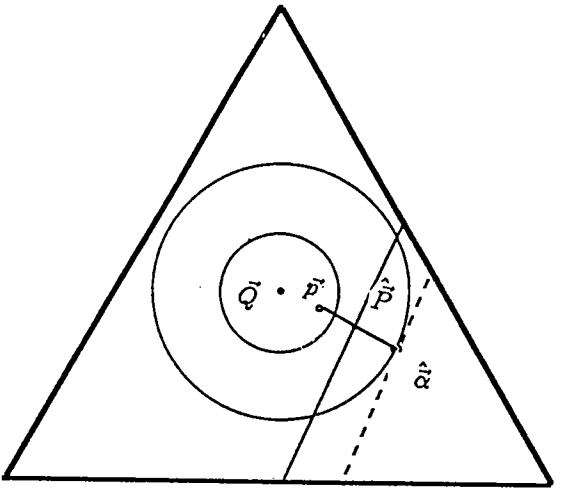

FIG 3 PM Estimator $\hat{\vec{P}}$ with Uniform $\vec{Q}$ and Sample Proportions $\vec{p}$: $\hat{\vec{\alpha}}$ is Maxent Estimator of Unsampled Proportions

Reconciliation of sample and census data would be an important application for the PM; the U.S. Census Bureau now uses a minimum chi-squared method for this purpose. This and other methods of estimating the P_s when marginal totals are known are reviewed by Ireland and Kullback [1968]. Causey [1983] compares the small sample performance of four non-Bayesian estimators in the 2×2 case under simple random sampling, and Causey [1984] compares several non-Bayesian estimators under more complicated sampling schemes.

A method commonly used for estimating cell values when marginal totals are known is "iterated proportional fitting" (IPF), which is one of the estimators studied by Causey [1983, 1984] and is also discussed by Bishop, Fienberg, and Holland [1975]. The latter note that IPF preserves in the final estimates any interactions present in the values used to start the iterations, whether or not they are present in the population. In contrast, PM utilizes both sample information, which may contain such interactions, and maxent estimates of proportions in the unsampled population, which contain only those interactions imposed by the exact constraints and $\vec{Q}$.

TABLE II
Cell Proportions used in Sampling Experiment

	β				
Cell	0	.005	.01	.03	.07
1	.1115	.1137	.1158	.1245	.1418
2	.1552	.1565	.1578	.1629	.1730
3	.0449	.0469	.0490	.0572	.0738
4	.1568	.1566	.1564	.1558	.1546
5	.0317	.0337	.0357	.0435	.0591
6	.0119	.0142	.0166	.0260	.0447
7	.0100	.0123	.0147	.0241	.0429
8	.4500	.4365	.4229	.3688	.2606
9	.0280	.0296	.0311	.0372	.0495
Distance from $\beta = 0$ Distribution					
MSE ($\times 10^4$)		2.3240	9.3681	84.1088	457.563
-Cross-entropy ($\times 10^4$)		8.3771	33.1557	269.3503	1316.542

4.2 Example

In order to explore the behavior of the IPF and PM estimators we conducted a sampling experiment in which we assumed a population of 10,000 arranged in a 3×3 table and sampled without replacement. The experiment is concerned with examining the effects of 1) changing the sampling proportion m and 2) using an inaccurate set of Q_s. For the former, the sampling proportions were set, respectively, at 1%, 2%, 5%, 10%, and 20%. For the latter, we generated a set of P_s indexed by a parameter, β. Table II contains the values used to generate the samples. In all cases we used the $\beta = 0$ values for the Q_s in the PM estimation. Thus, as β increases, the prior distribution is concentrated at an increasingly inaccurate set of values. The distances from the $\beta = 0$ distribution are noted at the end of Table II in terms of cross-entropy and mean squared errors. It should be noted that the assumed population proportions grow more uniform as β increases. For both IPF and PM estimates, the constraints employed are those for the parameters that generated the sample.

In addition to comparing the performance of the PM and IPF estimators, the sampling experiment permits us to investigate the consequences of wrongly

assuming that a prior distribution is accurate. As we discussed above, the cross-entropy prior rests on the assumption that the statistician has considerable confidence in the chosen prior. It is therefore important to determine whether the method allows the data to offset an incorrect prior; i. e. we are concerned with the robustness of the PM estimator based on a cross-entropy prior to inaccurate prior information.

From (18) we expect that for small sampling proportions and β close to 0, the PM estimator will have both small variance and small squared bias, hence small mean squared error. As the sampling proportion increases, for a constant β we expect that the variance will first increase and then decrease because of the increase in weight given to the sample proportions; the squared bias should decrease for the same reason. The mean squared error may increase or decrease, depending on the changes in variances and biases. For IPF, we expect the mean squared error to decrease with the sampling proportion because both variance and squared bias should fall. Finally, since the distributions become more uniform as β increases, we expect the mean squared error for the IPF estimator to increase with β.

Table III summarizes the results under a mean squared error criterion: the PM estimator outperforms the IPF estimator when the assumed value of $\vec{Q}$ is close to $\vec{P}$ (see the $\beta = 0$ and $\beta = .005$ columns) and for small samples up to $\beta = .03$ (see the $m = .01$ and .02 rows). These results are consistent with our expectations. It is interesting to consider the source of improvement, and we take the $\beta = .01$ case as an example. An investigator who has no exact constraints and no sample proportions, and who takes the $\beta = 0$ distribution as an estimator, would obtain a MSE$\times 10^4$ of 9.3681. With constraints and no sample information, the MSE$\times 10^4$ of the PM estimator falls to 1.258. As the sample proportion is increased from .01 to .20, the MSE$\times 10^4$ drops from 1.241 to .839. Thus, although the constraints are responsible for a considerable amount of the explanatory power, the PM estimator improves as the sampling proportion increases. Similar patterns may be found for the other values of β.

TABLE III

IPF and PM Mean Squared Errors ($\times 10^4$)[a]

	β				
m	0.000	0.005	0.010	0.030	0.070
0.00	n.a	n.a	n.a	n.a	n.a
	0.000	0.309	1.258	11.667	69.025
0.01	18.471	19.074	22.943	31.338	35.691
	0.000	0.305	1.241	11.428	67.665
0.02	9.278	9.489	11.621	14.221	16.665
	0.001	0.301	1.215	11.211	66.399
0.05	2.928	3.387	3.895	5.278	6.663
	0.007	0.282	1.141	10.529	62.430
0.10	1.612	1.657	1.872	2.277	3.518
	0.017	0.265	1.044	9.502	56.101
0.20	0.730	0.756	0.882	1.059	1.647
	0.028	0.223	0.839	7.490	44.362

[a]First line is IPF and second is PM.

This sampling experiment is evidence that PM performs better than IPF in situations of small sampling proportions and accurate, but by no means exact, beliefs about the values of the parameters. This is evidence that the PM estimator is somewhat robust to poor prior information. Moreover, when the hyperparameters are close to the true values, PM dominates IPF for sampling proportions as large as 20%. It should be noted that properties other than the mode of the posterior distribution, such as the covariance matrix, may be more sensitive to the quality and the assumed strength of the prior information.

5. CONCLUSIONS

Both on logical grounds and on the basis of the sampling experiment reported above, use of a cross-entropy prior in estimating proportions is warranted when the investigator has strong beliefs about the hyperparameters of the prior. In future work, we shall investigate the use of other forms of the entropy expression

for situations when beliefs are not very strong and examine the performance of PM with general sample information, i.e. sample information that is not in the form of estimates of cell proportions. We shall also examine properties of the posterior distribution other than the mode.

ACKNOWLEDGMENTS

We are pleased to acknowledge comments we received from seminar participants at the University of Michigan, the Technion, Bar-Ilan University, Tel Aviv University, the University of Bergamo, and IIASA. Technion and Bergamo provided the third author with helpful support during a sabbatical. We are also grateful for discussions with E. T. Jaynes and for written comments from I. J. Good, R. B. Frieden, and several referees, none of whom are responsible for our errors and omissions. Many thanks to Karen Rensing for her skillful typing and monumental patience.

BIBLIOGRAPHY

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis,* Cambridge, Massachusetts: The MIT Press.

Burg, J. P. (1967), "Maximum Entropy Spectral Analysis," presented at the 37th Annual Meeting of the Society of Exploration Geophysicists, Oklahoma City, Oklahoma. Reprinted in D. G. Childers (ed.), *Modern Spectrum Analysis,* New York: IEEE Press, John Wiley and Sons, 1978, pp. 34–41.

Causey, B. D. (1983), "Estimation of Proportions for Multinomial Contingency Tables Subject to Marginal Constraints," *Communications in Statistics (A), 12,* 22, pp. 2581–2587.

Causey, B. D. (1984), "Estimation under Generalized Sampling of Cell Proportions for Contingency Tables Subject to Marginal Constraints," *Communications in Statistics (A), 13,* 20, pp. 2487–2494.

Gokhale, D. V. (1973), "Approximating Discrete Distributions, with Applications," *Journal of the American Statistical Association, 68,* 344, pp. 1009–1012.

Good, I. J. (1963), "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *Annals of Mathematical Statistics, 34,* pp. 911–934.

Ireland, C. T. and S. Kullback (1968), "Contingency Tables with Given Marginals," *Biometrika, 55,* pp. 179–188.

Janardan, K. G. (1976), "Certain Estimation Problems for Multivariate Hypergeometric Models," *Annals of the Institute of Statistical Mathematics, 28,* Part A, pp. 429–444.

Jaynes, E. T. (1986), "Monkeys, Kangaroos, and N," in J. H. Justice (ed.), *Maximum Entropy and Bayesian Methods in Applied Statistics,* Cambridge: Cambridge University Press, pp. 26–58.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl and T. C. Lee (1985), *The Theory and Practice of Econometrics,* 2nd Edition, New York: John Wiley and Sons.

Rosenkrantz, R. D. (1977), *Inference, Method and Decision,* Dordrecht, Holland: D. Reidel Pub. Co.

Sarndal, C.–E. (1965), "Derivation of a Class of Frequency Distributions via Bayes' Theorem," *Journal of the Royal Statistical Society, B, 27,* pp. 290–300.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal, 27,* pp. 379–423, 623–656.

Zabell, S. L. (1982), "W. E. Johnson's 'Sufficientness' Postulate," *The Annals of Statistics, 10,* pp. 1091–1099.

Zellner, A. and P. E. Rossi (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics, 25,* pp. 365–393.

Received May 1988; Revised January 1989.

Recommended by K. V. Mardia, University of Leeds, United Kingdom.

Refereed by A. O'Hagan, University of Warwick, Coventry, United Kingdom.