The Annals of Statistics
1985, Vol. 13, No. 1, 246–261

ADAPTIVE ESTIMATORS FOR SIMULTANEOUS ESTIMATION OF POISSON MEANS[1]

BY H. M. HUDSON

Macquarie University

A new estimator for smoothing towards log-linear models for Poisson means is introduced. The estimator is a generalization of one of Peng (1975). The theoretical basis for the choice of estimator is developed from an approximation to the mean square error of any estimator of Poisson means. The new estimator and its competitors are evaluated in simulations. The method has widespread application in contingency table analysis.

1. Introduction. In contingency table analysis we may often be able to improve estimation of cell means by "borrowing strength", accepting some bias in return for reduced variability of the estimates. Use of a parsimonious model for cell means is one method of increasing the precision of estimation. But model selection introduces bias.

Biased estimation has gained wide practical acceptance. A common technique is to eliminate as many interaction terms and main effects as possible, based on the results of hypothesis tests, in a complex model for cell means. The resulting reduced model is then used to determine "smoothed" cell estimates. See Bishop, Fienberg and Holland (1975), for illustrations. Bias is then a consequence of the preliminary hypothesis testing, and the possibility that the reduced model is incorrect.

This "preliminary test" approach to estimation is unsatisfactory from several points of view. The real presence of interactions, when relatively small in nature, may not sufficiently bias a simpler main effects model to compensate for the loss of precision in trying to estimate the many extra interaction parameters. From a decision theoretic viewpoint the discontinuity in estimation brought about by the hypothesis testing dichotomy means this procedure cannot be admissible. Sclove, Morris and Radhakrishnan (1972) investigated the performance of preliminary test estimators in linear models. The preliminary test procedure often produced poor estimates. The risk function of the preliminary test estimator attested to this poor performance. The risk exceeded the minimax bound (the risk from use of the unadjusted data) over a substantial region of the parameter space. By contrast, procedures amongst a class of minimax estimators introduced by James and Stein (1961) achieved low risk when the reduced model was correct without sacrificing precision when the adequacy of the model was uncertain.

Received June 1980; revised September 1984.

[1] This work was supported in part by a Macquarie University research grant.

AMS 1980 *subject classification.* Primary 62C15.

Key words and phrases. Empirical Bayes, empty cells, log-linear model, minimax, multinomial, multiparameter estimation, Poisson, random effects estimation, Stein estimator.

In a contingency table context Fienberg and Holland (1970) and Sutherland, Fienberg and Holland (1974) have proposed empirical Bayes procedures somewhat analogous to the Stein estimator, though in a product multinomial setting. In "large sparse" tables in which the multinomial distributions approach independent Poisson laws, an asymptotic minimax property of these empirical Bayes estimates was established. The cell estimates obtained by this method are a shrinkage, to an extent determined by the data, of the raw cell counts towards a set of expected cell counts, these derived from a reduced model.

Peng (1975), and Clevenson and Zidek (1975), have obtained direct generalizations of the Stein estimator when cell counts are independent Poisson variables and shrinkage is towards 0. Hudson and Tsui (1981) consider arbitrary nonzero a priori values. The latter result permits immediate generalization to the case of shrinkage towards data determined expected cell counts.

This paper introduces this new class of estimators and illustrates its use in contingency tables. Theory is developed indicating the basis for use. When cell counts are independent Poisson variables, the estimators are (almost) minimax. Estimates are shrunk to conform with models linear in the logarithm of the cell count. The new class is flexible and easy to use. Estimates may be calculated simply.

The procedure is defined in Section 2. In Section 3 an approximation to the risk function is developed which demonstrates the near minimax property the new class possesses. In Section 4 the results above are shown to apply also to multinomial data. Two examples of applications are contained in Section 5.

Section 6 contrasts the precision and robustness of the several approaches discussed above in a simulation study.

In Section 7 it is shown that the new procedures may be derived from stochastic models for the cell means. Thus the new procedures may be placed in an empirical-Bayes context (Maritz, 1969). Leonard (1972) presented related Bayesian theory for binomial data, and extended this analysis to other exponential families in unpublished notes.

2. Estimators of cell means in contingency tables. Tests of hypothesis for contingency table analysis are prevalent in the literature. Suppose $X_1, \cdots, X_p$ are independent Poisson variates with means $\mu_1, \cdots, \mu_p$. The preliminary test method suggestion of Section 1 may be based on log-linear models for the data. Then for $\{\hat{\mu}_{0i}\}$ computed from a reduced log-linear model, the preliminary test estimator is

$$\hat{\mu}_i = \begin{cases} \hat{\mu}_{0i} & \text{if the reduced model is acceptable at level } \alpha \\ X_i & \text{if the reduced model is rejected at level } \alpha. \end{cases}$$

In simulations we used a significance level $\alpha = .15$. We comment on more stringent choices in Section 7. Sclove, Morris and Radhakrishnan (1972) have proved that the Stein estimator dominates the preliminary test estimator based on linear models in multiparameter Gaussian estimation, for significance levels above $\alpha = .50$.

The Fienberg–Holland rule is defined as

$$\hat{\mu}_i = \hat{\mu}_{0i} + (1 - 1/(K + 1))(X_i - \hat{\mu}_{0i})$$

where $K = Q/[p\bar{X} - \sum X_i^2/p\bar{X}]$, where $Q = \sum (X_i - \hat{\mu}_{0i})^2$.

A detailed description of this procedure and its properties is available in Bishop, Fienberg and Holland (1975), Chapter 12. While the presentation is appropriate for multinomial data, the nondegenerate asymptotic results provided there for the Fienberg–Holland rule are based on special asymptotics for sparse multinomials in which the counts X within the table follow independent Poisson distributions. Thus these results apply to Poisson counts also.

If $Y_1, \cdots, Y_p$ is a sample of independent normal random variables with mean $\theta_1, \cdots, \theta_p$ and common variance σ^2, the Stein estimator appropriate for shrinkage toward any reduced linear model $\theta = A\beta$ dominates the coordinate estimator $(Y_1, \cdots, Y_p)$. Here A, $p \times q$, is prespecified and β, $q \times 1$, $q < p - 2$, will be estimated by least squares if a mean square error risk criterion is used. Suppose the vector of predicted means in the reduced model is written $\hat{\theta}_0$; so $\hat{\theta}_0 = A(A'A)^{-1}A'Y$. Then the appropriate Stein estimator is

$$\hat{\theta}_0 + (1 - (p - q - 2)/S_A)(Y - \hat{\theta}_0) \tag{2.1}$$

where $S_A = \sum_{i=1}^p (Y_i - \hat{\theta}_{0i})^2/\sigma^2$ is proportional to the residual sum of squares for the reduced model. The gain in MSE at θ is given by $p^{-1}E_\theta\{(p - q - 2)^2/S_A\} \times 100\%$, and will be large if the reduced model is substantially correct. See James and Stein (1961), Hudson (1974), and Efron and Morris (1972).

Thus one expects that after a variance stabilizing transformation, the Stein estimator could be applied with useful gains to Poisson counts. We shall use the standard positive part adaption of (2.1) in simulations which follow.

The last estimator we consider is a generalization of one introduced in Hudson and Tsui (1981), there shown to dominate X. We again shrink towards a reduced model of the log-linear form. Specifically the procedure is:

1. Transform the cell entries according to

$$H_i = h(X_i) = \sum_{j=1}^{X_i} 1/j, \quad i = 1, \cdots, p. \tag{2.2}$$

It is helpful to note that $\log((x + 0.56)/0.56)$ is a very satisfactory approximation to $h(x)$.

2. (2.3) Find the least squares fitted values $\hat{H}_1, \cdots, \hat{H}_p$ appropriate to the log-linear model being considered, i.e., define $\hat{H} = A(A'A)^{-1}A'H$, for some design matrix A, $p \times q$.

3. (2.4) Define $R = (p - N_0 - q - 2)_+$ where $a_+ = \max(a, 0)$, N_0 denotes the number of observed zeros, and q is the rank of A. Let $S = \sum_{i=1}^p (H_i - \hat{H}_i)^2$ and $\hat{X}_i = 0.56(\exp(\hat{H}_i) - 1)$, if $\hat{H}_i \geq 0$, or 0 otherwise, for $i = 1, \cdots, p$.

4. Then set

$$\begin{aligned} \hat{\mu}_i &= X_i - (R/S)(H_i - \hat{H}_i) \quad \text{if} \quad (X_i + .56) > (R/S), \\ &= \hat{X}_i \qquad\qquad\qquad\qquad\quad \text{otherwise,} \end{aligned} \tag{2.5}$$

for $i = 1, \cdots, p$.

3. Approximations to the risk function. The application of Stein's integration by parts method (Stein, 1981) to Poisson estimation (Hudson, 1978) leads to an expression for the mean square error of any estimator of the mean vector μ as an improvement $E_\mu\{\psi_0(X)\}$ of the mean square error of the unsmoothed estimator X. For any functions $g_1, \cdots, g_p$ defined on p-dimensional tables X of nonnegative counts, and such that $E \mid g_i(X) \mid < \infty$, for $i = 1, \cdots, p$, extend definition of $g_1, \cdots, g_p$ to tables including negative integer counts by $g_i(x) = 0$, say, if any coordinate $x_j < 0$. Then

$$\begin{aligned} &E_\mu\{\textstyle\sum_{i=1}^p (X_i - \mu_i)^2\} - E_\mu\{\sum_{i=1}^p (X_i + g_i(X) - \mu_i)^2\} \\ &\qquad = E_\mu\{-2 \textstyle\sum_{i=1}^p X_i[g_i(X) - g_i(X - e_i)] - \sum g_i^2(X)\} \\ &\qquad = E_\mu\{\psi_0(X)\}. \end{aligned} \tag{3.1}$$

Here the vectors $e_1, \cdots, e_p$ are unit vectors of the form $(0 \cdots 010 \cdots 0)$, so that $X - e_i$ denotes data with X_i reduced by one. The function ψ_0, dependent on the alternative estimator $X + g(X)$ but not on μ, is defined as the interior of the next to last expectation above. We shall usually write $\psi_0(X)$ instead of $\psi_0(X, g)$ except when the context requires the full notation. The equality (3.1) expresses the advantage in MSE to be obtained by the alternative estimator as an expectation of a known function. The estimators considered hereafter all satisfy the necessary conditions stated above for (3.1) to apply.

Mean square error is used to evaluate precision of estimation here. Other weightings of component errors might be considered, in particular the alternative $\sum (\hat{\mu}_i - \mu_i)^2/\mu_i$. Unfortunately this weighting implies great deference must be paid to zero counts, limiting the possibility of smoothing, since the possibility of very small μ_i cannot be ignored. It is also difficult to derive an unbiased estimator of risk, as in (3.1), for other loss functions. However, recent results of Hwang (1982) may make a similar approach to that outlined here feasible with other weightings.

The identity (3.1) allows the risk reduction of an alternative estimator to be assessed. The mean square error of any estimator can thus be determined. This approach is used in simulations in Section 6. An estimator with the property $\psi_0 \geq 0$ would be guaranteed by (3.1) to have a smaller MSE than X. For the estimator to be of practical value ψ_0 would need to be large for many data sets.

Such an estimator has been proposed by Peng (1975). His procedure yields useful gains only when all the means $\mu_1, \cdots, \mu_p$ are near 0. Because of the complexity of (3.1) it is difficult to obtain improved estimators for larger means

by this approach. We therefore consider the approximation

$$\psi_1 = -2 \sum_{i=1}^{p} x_i(\partial g_i/\partial x_i) - \sum_{i=1}^{p} g_i^2 \tag{3.2}$$

for ψ_0, which replaces a first order difference by a derivative, for suitable functions $g_1, \cdots, g_p$ which agree with the estimation rule on integers. It is sufficient for left-sided derivatives to exist in (3.2) for these to replace terms of the form $[g_i(x) - g_i(x - e_i)]$.

A heuristic solution, g, to $\psi_1(X, g) \geq 0$, based on similar inequalities for continuous exponential families in Hudson (1978), suggests the choice $g_i(X) = -(R/S)(H_i - \hat{H}_i)$, for $i = 1, \cdots, p$. Here R, S, H_i and $\hat{H}_i$ are defined in (2.2)–(2.4).

The modification shown in (2.5) is then suggested by noting the approximate convex form of the estimator on a log-like scale. For

$$\begin{aligned} h(X_i + g_i(X)) &\doteq \log(X_i + 0.56 + g_i(X)) - \log(0.56) \\ &= \log\left(\frac{X_i + 0.56}{0.56}\right) + \log\left(1 + \frac{g_i(X)}{X_i + 0.56}\right) \\ &\doteq h(X_i) + \frac{g_i(X)}{X_i + 0.56} \qquad (3.3) \\ &= H_i - \frac{R}{(X_i + 0.56)S}(H_i - \hat{H}_i) \\ &= z_i\hat{H}_i + (1 - z_i)H_i \end{aligned}$$

where the "credibility" weight $z_i = R/(X_i + 0.56)S$. The third equality depends on the condition that $|g_i(X)|/(X_i + 0.56)$ be small. Thus, when this condition is met, the estimation procedure involves shrinking the transformed cell counts towards a linear model.

It is reasonable to insist that (3.3) represents a convex combination of the raw data H_i and the smoothed estimate $\hat{H}_i$ by requiring that $z_i \leq 1$, $i = 1, \cdots, p$. Otherwise the estimator "overshoots" the fitted value. The procedure adopted in (2.5) replaces any z_i exceeding 1 by 1.

Another heuristic solution g to $\psi_1(X, g) \geq 0$ is based on Theorem 2 of Hudson (1978), which generalizes Stein's result (Stein, 1981, Section 6) for the normal distribution. The shrinkage of extreme counts can be truncated. Then one or more extreme observations will not affect the estimator unduly. Risk properties of this estimator have not been pursued here however.

We now consider the approximate risk properties of the log-linear estimator. We use (3.1), and assume that the table is such that only small relative changes in S are possible when any count is diminished by 1.

Assume then that $1 - (S_i/S)$ is small in absolute magnitude, for each i, where $S_i = S(X - e_i)$. Let $Q = A(A'A)^{-1}A'$, so that Q is a projection matrix with diagonal elements $\{q_{ii}\}$ whose sum is q, the rank of A. Note that $\hat{H} = QH$, from

(2.3). Then

$$S_i = S - (H_i - \hat{H}_i)^2 + \left(H_i - \hat{H}_i - \frac{1}{X_i}(1 - q_{ii})\right)^2$$

$$= S\left\{1 - \frac{1 - q_{ii}}{X_i S}\left[2(H_i - \hat{H}_i) - \frac{1}{X_i}(1 - q_{ii})\right]\right\}.$$

The assumption made therefore requires the second term within the parentheses to be small, in which case

$$\frac{1}{S_i} \doteq \frac{1}{S}\left\{1 + \frac{1 - q_{ii}}{X_i S}\left[2(H_i - \hat{H}_i) - \frac{1}{X_i}(1 - q_{ii})\right]\right\}.$$

Consider now the term $X_i[g_i(X) - g_i(X - e_i)]$ in the expression for the risk benefit, for $g_i(X) = -(R/S)(H_i - \hat{H}_i)$. Here $R = (p - N_0 - q - 2)_+$. Then it is straightforward to demonstrate that, for $X_i \geq 1$,

$$X_i[g_i(X) - g_i(X - e_i)]$$

$$= \frac{R(1 - q_{ii})}{S} + \frac{R}{X_i}\left(\frac{1}{S} - \frac{1}{S_i}\right)\left(H_i - \hat{H}_i - \frac{1}{X_i}(1 - q_{ii})\right),$$

from which, using the expansion of S_i^{-1} above, one may obtain

$$\frac{X_i[g_i(X) - g_i(X - e_i)]}{R/S}$$

$$= (1 - q_{ii}) - \frac{2(1 - q_{ii})}{S}(H_i - \hat{H}_i)^2 + \frac{3(1 - q_{ii})^2}{X_i S}(H_i - \hat{H}_i) - \frac{(1 - q_{ii})^3}{X_i^2 S}$$

$$\geq (1 - q_{ii}) - \frac{2(H_i - \hat{H}_i)^2}{S} + \frac{3}{2}\frac{1 - q_{ii}}{X_i S}\left[2(H_i - \hat{H}_i) - \frac{1}{X_i}(1 - q_{ii})\right],$$

since $0 \leq q_{ii} \leq 1$. The last term above is negligible, by assumption, and summation of the remaining terms yields R^2/S as the risk reduction.

These results require only that any change which reduces by 1 the count in one cell of a contingency table, engenders a small relative change in the lack of fit, S. It is clear that as the effect of any one observation on S diminishes, as occurs as p increases, the approximation improves. The introduction of a lower bound for S, in (2.5), further aids the adequacy of the approximation in circumstances in which the data and model agree. On the other hand, the approximation may not be very satisfactory for small p, if counts X_i are small. Nevertheless it appears that, for many models, $\psi_0(X)$ exceeds R^2/S over a large region of the data space, implying near minimaxity of the log-linear estimator and significant risk reduction.

The general form of this estimator, and heuristic support for it, was first given in Brown (1979), through considerations of asymptotic expansions of the risk function, as cell means increased. Our development includes shrinking to

data dependent values, and suggests that minimaxity is achieved as $p \to \infty$ even with small cell means.

The estimate R^2/S, of the reduction in MSE points to substantial benefits when the fitted log-linear model is simple (so R is large) and provides reasonably accurate estimates of cell means. Hence the choice of design matrix A (or equivalently, the model) is of great importance.

Two special cases considered previously are: Peng's estimator, in which $A = 0$; and Hudson and Tsui's estimator for the model $\mu_1 = \cdots = \mu_p$, for which $A = (1, 1, \cdots, 1)'$.

4. Application to multinomial data. When a multinomial model is appropriate for contingency table data there will be situations for which the Poisson theory developed in Section 3 remains valid.

Multinomial observations are often distributed approximately as independent Poisson variates. For one such situation, see Feller (1950) Exercise VI.10.38. The approximation also appears appropriate with "large-sparse" contingency tables.

Let $(X_1, \cdots, X_p)$ have the multinomial distribution with parameters $(n, \Pi_1, \cdots, \Pi_p)$, where $\sum \Pi_j = 1$. Then the term "large-sparse" tables would refer to tables where the following asymptotics are relevant:

(a) $n \to \infty$
(b) $n\Pi_i \to \mu_i$ as $n \to \infty$, (hence $p \to \infty$ also).

It is immediate, from Feller's result, that the joint distribution of counts within given cells will, under these asymptotics, be the joint distribution of independent Poisson variables. In particular, the marginal distribution of X_i is Poisson with mean μ_i, and the counts in different cells are pairwise independent, in asymptotics.

In these cases, it is therefore natural to use Poisson estimation theory, which ignores the multinomial constraint on a row total. However, the argument above is not wholly compelling as convergence in distribution does not necessarily imply the convergence of risk functions that is at issue. A direct examination of this convergence is possible.

This is because an identity for multinomial observations is equivalent for large n, to the Poisson identity leading to (3.1), the property on which the new estimator's near minimaxity depends. For fixed $(\Pi_1, \cdots, \Pi_p)$ denote expectation with respect to the multinomial distribution above by E^n. Then it is easy to show that, with $\mu_j = n\Pi_j$,

$$\mu_j E^n g_j(X) = (n/(n+1))E^{n+1}X_j g_j(X - e_j), \quad j = 1, \cdots, p, \tag{4.1}$$

corresponding to a Poisson identity

$$\mu_j E g_j(X) = EX_j g_j(X - e_j), \quad j = 1, \cdots, p, \tag{4.2}$$

used to derive (3.1).

In (4.1) suppose $(n/(n+1))E^{n+1}X_j g_j(X - e_j)$ may be replaced by $E^n X_j g_j(X - e_j)$, for $j = 1, 2, \cdots, p$, to a suitable degree of accuracy. From

(4.1) this is equivalent to replacing $E^n g_j(X)$ by $(n/(n-1))E^{n-1}g_j(X)$, for $j = 1, \cdots, p$. Then (4.1) may be written

$$\mu_j E^n g_j(X) \doteq E^n X_j g_j(X - e_j), \quad j = 1, \cdots, p,$$

and hence all results of Section 3 based on (3.1) continue to apply to the difference in risk

$$E^n \sum (X_i - \mu_i)^2 - E^n \sum (X_i + g_i - \mu_i)^2.$$

In the asymptotics above—which include "large-sparse" tables—many estimators g will have the required property, namely that, for any j,

$$|(n/(n-1))E^{n-1}g_j(X) - E^n g_j(X)| \to 0, \quad \text{as} \quad n \to \infty.$$

This is because the addition of one observation to the table at random will, for many estimation procedures, have a negligible effect on any one cell estimate. Subject to this check, the Poisson theory will also apply to multinomials with large marginal totals.

5. Examples of the use of the "log-linear" estimator. In this section we give two examples of the use of log-linear estimators. The first illustrates the use of the estimator with Poisson counts in a two way cross-classification in which the reduced model specifies a simple relationship between cell means. The second example treats a pure multinomial sampling scheme in which the total count in the table is fixed. The cells are arranged in a 3×2^3 factorial arrangement and the reduced model corresponds to the absence of some effects.

The first set of data examined records the frequencies of entry of stroke patients with given severity of condition initially, and on discharge. The data is tabulated in Plackett (1981) and is shown below.

		Final rating				
		A	**B**	**C**	**D**	**E**
Initial	**E**	11	23	12	5	8
rating	**D**	9	10	4	1	
	C	6	4	4		
	B	4	5			
	A	5				

In the absence of specific prior information on distribution of stroke patients' initial and final ratings, a beta form of distribution for each suggested the simple model

$$\log \mu_{ij} = \mu + \alpha_1 \log i + \alpha_2 \log(6 - i) + \beta_1 \log j + \beta_2 \log(6 - j)$$

where i and j denote row and column numbers of the cell, for $i \leq j = 1, 2, \cdots, 5$. The intrinsic ordering of the ratings A–E is thereby recognized in both initial and final ratings, and with the assumption of pseudo-independence (conditional independence of initial and final ratings, given that the final rating is not lower than the initial rating), leads to a flexible model with relatively few parameters.

We obtain the least squares fit to the counts transformed to

$$H = \log[(X + 0.56)/0.56]$$

as:

$$\hat{H}_{ij} = 3.4 - 1.4 \log i - 0.7 \log(6 - i) - 0.1 \log j + 0.7 \log(6 - j)$$

with $S = 2.12$, by multiple regression analysis. Hence $R/S = (15 - 5 - 2)/2.12 = 3.77$. From the fitted values, $\hat{H}$, predictions $\hat{X} = 0.56(\exp(\hat{H}) - 1)$ are obtained as:

	Final rating				
Initial	18.2	14.4	11.1	7.9	4.4
rating	7.9	6.2	4.7	3.3	
	5.4	4.2	3.1		
	4.7	3.6			
	5.6				

The log-linear estimates of cell means $\hat{\mu}_{ij} = X_{ij} - (R/S)(H_{ij} - \hat{H}_{ij})$ are:

	Final rating				
Initial	12.8	21.3	11.7	6.6	5.9
rating	8.6	8.3	4.6	4.4*	
	5.6	4.1	3.2		
	4.5	3.9			
	5.4				

The estimate 4.4 would be replaced by 3.3 in accordance with (2.5). The calculations are very simple using many statistical computing systems.

The estimator's effect is to smooth the cell counts, where small, towards the fitted values; the relative change is less for larger counts.

The estimated lower bound on the improvement in MSE, $R^2/S = 30.2$, is only one-fourth of the estimated MSE of unmodified counts, $\sum X_i = 121$, an apparent risk reduction of 25%. The unreliable small counts may be observed to be shrunk by a considerable factor. This seems a very desirable behavior for an estimator, and suggests a useful rule for the log-linear procedure when zero counts present a problem.

Our second example concerns multinomial data. Bishop, Fienberg and Holland (1975) examined data, collected by Ries and Smith (1963) on detergent preferences, in their Table 4.4-6. This data is reproduced in Table 1. The design is a 3×2^3 factorial in which the four variables considered are water softness (at 3 levels), previous use of the detergent of interest, wash temperature, and brand preference. Their analysis of cell counts indicated a strong main effect on temperature and a strong interaction between previous use and brand preference; smaller interactions between temperature and water softness, and temperature and brand preference, were considered to be present also. In order to estimate the importance of these effects we shall obtain the log-linear estimates of cell means using a reduced model in which only previous use and brand preference

variables and their interaction are fitted in addition to the temperature main effect.

This example is chosen because it highlights the precision to be gained from estimates based on models involving relatively few parameters, when extra terms in the model add significant, but limited, explanation. Several alternative procedures could also be applied: the Stein estimator would be suitable after a variance stabilizing transformation. Other estimators with Bayesian justifications have been proposed for multinomial data (see particularly, Leonard (1977), and references to work of Good contained therein), but their risk properties are unknown. The reader is invited to compare his favoured method with our approach. After transformation of the counts by the natural logarithm (equivalent to transformation by h), observations may now be pooled over the water softness categories, since the reduced model does not include variation attributable to this factor. The fitted values in Table 2 are then obtained by standard analysis of variance techiques for the model with no temperature interactions. The residual sum of squares of this model is $S = .574$ on 19 d.f. Hence the shrinkage factor R/S in (2.4) is 29.6. We can thus calculate the revised cell estimates as $X - 29.6 \log(X/\hat{X})$; these are as given in Table 3. The risk benefit is estimated to be $R^2/S = 17^2/.574 = 504$—compare with 1008, the estimated risk of X.

The requirement for Poisson theory to apply in this case was discussed in Section 4. It is that the expectation of the change in the log-linear estimator obtained when a single random count is added to the table is negligible, in every cell. This condition is clearly met here, as fitted values (such as in Table 2) would be virtually unaffected by the additional count, and with high probability the count will not affect the observations in the cell being considered. Thus a substantial MSE reduction is expected to apply to the estimates of Table 3.

TABLE 1
Observed data of Ries and Smith

brand preference	water softness	previous use		no previous use	
		temperature			
		high	low	high	low
X	Soft	19	57	29	63
	Medium	23	47	33	66
	Hard	24	37	42	68
M	Soft	29	49	33	66
	Medium	47	55	23	50
	Hard	43	52	30	42

TABLE 2
Fitted values for the reduced model (log scale)

	previous use		no previous use	
	high	low	high	low
X	3.18	3.74	3.58	4.14
M	3.52	4.09	3.29	3.86

TABLE 3
*Log-linear estimates***

brand preference	water softness	previous use		no previous use	
		temperature			
		high	low	high	low
X	Soft	23.9	48.0	35.7	62.9
	Medium	23.9 (23.9*)	43.7 (42.3)	35.5 (35.7)	74.5 (63.1)
	Hard	23.9	40.8	37.3	75.6
M	Soft	33.8	54.9	26.9	49.7
	Medium	37.2 (33.8)	57.4 (59.7)	26.9 (26.9)	48.5 (47.5)
	Hard	35.9	56.1	26.7	45.6

* Fitted values under the reduced model are shown in parentheses.
** Obtained by use of (2.2)–(2.5).

Table 3 indicates that the estimates obtained from the model may be regarded as generally accurate—with the possible exception of the soft water, low temperature, brand X preference of previous users of M. The brand M preference of previous users in low temperature conditions may be somewhat underestimated, and in hot temperature conditions somewhat overestimated, but the size of this effect is minimal.

The estimated preference for brand X in each cell, shown below, clearly depends on previous use of brand M.

	previous use		no previous use	
	high	low	high	low
Soft	41%	47%	57%	56%
Medium	39%	43%	57%	57%
Hard	40%	42%	58%	59%

Those with no previous use of brand M show very little difference in preference whatever conditions apply. Previous users' estimated preference for brand X appear similar except in low temperature, soft water conditions. These conclusions are not unlike those of Cox and Lauh (1967).

6. Simulation results. For tables containing $p = 5$ or $p = 20$ cells, 864 Poisson data sets were generated in a number of steps.

First, 16 replicate samples of size p were selected with distribution

$$\mu = \begin{cases} \mu_0 & \text{with probability } 1 - 2\nu \\ \mu_0(1+\alpha)^2 & \text{with probability } \nu \\ \mu_0/(1+\alpha)^2 & \text{with probability } \nu. \end{cases}$$

Cases studied were $\mu_0 = 4$, $\nu = 0.10$ and $\mu_0 = 2$, $\nu = 0.05$—for $\alpha = 0, 0.5, 1.96$ sets of p means (not all distinct) were thus generated, 16 each from six prior distributions.

Then, for each set μ, a sample X of p independent Poisson counts with these means were generated. Nine replicates of X for each configuration μ were obtained.

Taking the sum of squared errors of the unbiased estimator X as a base, the reduction in squared error achieved by each of the multiparameter estimators shown in Table 4 was calculated in two forms:

$$L_1 = \psi_0(X, g)/\sum \mu_i$$

and

$$L_2 = \frac{\{\sum_{i=1}^{p} (X_i - \mu_i)^2 - \sum_{i=1}^{p} (X_i + g_i(X) - \mu_i)^2\}}{\sum \mu_i}.$$

L_1 and L_2 provide unbiased antithetic estimates of $1 - \text{MSE}(X + g, \mu)/\sum \mu_i$, the reduction in risk achieved by the estimator for mean vector μ. In Table 5 summary values of the 144 pairs (L_1, L_2) are shown for each prior, together with the maximum attainable reduction—achieved by use of the Bayes estimator for that distribution.

Table 5a shows the attainable MSE reduction, expressed as a proportion of $\sum \mu_i$. For three distributions, use of the Bayes estimator (which assumed knowledge of the prior) resulted in nearly 100% reduction in MSE; in the other three cases it was possible to reduce MSE to about one-third of the risk of X. Savings in MSE of these magnitudes are substantial: a 95% reduction is equivalent to increasing sample size by a factor of 20, and a 67% reduction is equivalent to increasing sample size by a factor of 3. Table 5b gives the risk reduction of each multiparameter estimator.

In small tables ($p = 5$ cells) with little data (priors 1–3, for which $\mu_0 = 2$) the preliminary test and Fienberg–Holland rule were comparable in achieving risk reductions of around 50% of that attainable. These estimators were superior in this regard to the Stein and log-linear estimator. These had very similar performance in the case of tables with small means. This is because, for $0 \leq x \leq 10$ the relationship between $h(x)$ and the variance stabilizing transformation $\sqrt{x + 3/8}$ is near linear, the only minor deviation from linearity occurring when $x = 0$. Thus fitted values and shrinkage factors are virtually identical for data in this range. The equivalence does not extend to data sets in which some larger counts are present.

With $p = 5$, $\mu_0 = 4$ (priors 4–6) the Fienberg–Holland estimator recorded the most consistent risk reductions, slightly higher than Stein and log-linear results.

In larger tables ($p = 20$ cells) the risk reductions attained by the Stein and

TABLE 4
Estimators considered in the risk computations

no.	description	comments
1.	Preliminary test	Hypothesis tested was equality of means. Significance level $\alpha = .15$ (see Section 2).
2.	Fienberg–Holland rule	With $\hat{\mu}_{0i} = \bar{X} = \sum_{i=1}^{20} X_i/20$.
3.	Positive part Stein estimator, square root transformed data	Estimates of $\sqrt{\mu_i + 3/8}$, $i = 1, \cdots, 20$, obtained from (2.1) with $\sigma^2 = 1/4$, and A, 20×1, having elements $a_i \equiv 1$, for $i = 1, \cdots, 20$. Invert transformation to obtain estimates of μ.
4.	Log-linear estimator	Estimator as given in (2.2)–(2.5), with A chosen as for estimator 3.

TABLE 5
Simulation results

a. *Risk reductions attainable**

p	prior 1	2	3	4	5	6
5	1.00**	.88	.78	1.00	.56	.64
20	.98	.92	.67	.99	.67	.65

* Tabled is the Bayes estimator's average risk reduction in simulations.
** MSE scaled by division by $\sum \mu_i$.

b. *Percentage of attainable risk reduction achieved*

prior	mean	estimator P. test	F. H.	Stein	L. L.
$p = 5$					
1	79*, 79**	53, 54	52, 51	39, 37	37, 36
2	77, 78	45, 54	50, 54	35, 39	34, 38
3	6, 19	32, 41	45, 46	33, 35	35, 35
4	76, 83	46, 54	52, 53	50, 48	47, 47
5	24, 40	23, 25	59, 66	49, 51	46, 48
6	−172, −208	44, 30	39, 31	35, 29	41, 34
$p = 20$					
1	92, 98	68, 75	67, 70	82, 85	80, 82
2	94, 91	71, 63	72, 71	84, 85	82, 77
3	48, 32	61, 60	70, 66	70, 61	99, 91
4	91, 98	61, 72	66, 70	83, 88	77, 81
5	57, 57	18, 25	76, 76	81, 84	84, 83
6	−235, −258	0, 0	34, 31	17, 12	44, 40

* Average of $\psi_0(x)/\sum \mu_i$ as a percentage of attainable risk reduction.
** Average of $[\sum(x_i - \mu_i)^2 - (\hat{\mu}_i - \mu_i)^2]/\sum \mu_i$ as a percentage of attainable risk reduction.

c. *Minimum risk gain estimate**

prior	mean	estimator P. test	F. H.	Stein	L. L.
$p = 5$					
1	−1.32	−4.20	−.10	−.49	−.24
2	−1.44	−5.76	−.08	−.31	−.21
3	−8.02	−4.50	−.12	−.43	−.36
4	−1.20	−4.84	−.05	−.29	−.02
5	−5.15	−3.68	−.10	−.13	−.04
6	−11.39	−3.78	−.09	−.21	−.04
$p = 20$					
1	−0.42	−2.31	.29	.29	.17
2	−0.41	−2.87	.28	.08	.17
3	−2.47	−2.27	.08	−.49	.16
4	−0.56	−2.58	.26	.24	.21
5	−2.35	−2.27	.20	.00	.16
6	−4.93	−2.49	.06	−.45	.08

* Tabled is minimum value of $\psi_0(X)/\sum \mu_i$ obtained in 144 simulations from the prior.

log-linear estimators were consistently over 80% of that achievable. The Stein estimator's performance was slightly the better of the two when all means were identical, while the log-linear estimator was slightly more effective when the model was mis-specified. These estimators dominated the preliminary test and Fienberg–Holland rule for each prior.

Since the distributions considered put weight on only a limited region in the p-dimensional parameter space, we may have weighted down some poor cases for an estimator. To gain a wider perspective, we examined $\psi_0(X)$ in each of the 1728 data sets X generated. (The configurations of counts were almost entirely distinct and varied considerably even when the means were similar). When $\psi_0(X)$ is negative, it is probable that there is a corresponding configuration of means for which the risk exceeds the minimax bound.

Table 5c shows the worst excess risk estimate in any of the data sets X generated for each prior. The preliminary test procedure produces risk estimates as much as 5 times the minimax risk, and increased risk estimates occur frequently, particularly with data exhibiting borderline evidence against the reduced model. With the Stein estimator, the risk estimate indicated a possible 50% increase in risk for some data configurations, generally those including a single extreme count. The Fienberg–Holland estimator exhibited a near minimax performance, with at worst a 10% increase in risk estimated, and the log-linear estimator confirmed its near minimaxity property of Section 3 in large tables, or in small tables with expected counts of 4 per cell. With 10 counts in the table, the excess risk was estimated to be at worst one-third of the minimax risk (again this occurred when a single extreme outlier was present).

In both Bayes gain and minimax risk, the overall performance of the preliminary test approach is disappointing, and no fine tuning would appear to improve it. Worst case (minimax) risk would increase if more stringent significance levels were used, while average risk will degrade if a less stringent level were chosen. A strong case has been made for an alternative procedure.

When both risk criteria are considered, it appears that the Fienberg–Holland estimator is the procedure of choice in small tables, while the log-linear estimator is preferable for larger tables, when equality of cell means is a plausible hypothesis.

7. Discussion. The simulations of the previous section were designed with the intent of examining the average risk of each estimator for various departures from the hypothesized model. Additionally, data configurations for which excess risk was apparent were to be examined, and an estimate of the minimax risk obtained. Average risk and minimax risk criteria for evaluating the robustness of an estimator to mis-specification of prior information are discussed in Berger (1982).

As demonstrated elsewhere for normally distributed data, our results show preliminary test estimators to be deficient with respect to both criteria above. By contrast, the simulations confirm that with Poisson data too, Stein type procedures with near minimax risk can make substantial use of prior information,

even when imperfect. This is particularly true with tables large enough for the adequacy of the model to be evaluated reliably. The robustness of an estimator to an imperfect model is important. It means simple models may be used without attendant chance of substantial error in estimation.

Through such simple models, expected counts are obtained that are very stable. Even oversimplification can be justified with robust estimators, as illustrated for the Ries–Smith data (refer to Section 5). Robustness is the guarantee that the estimates will not suffer from a poorly chosen model.

A key element in the assessment of Bayesian robustness of an estimator is the decision theoretic examination of its risk function (Berger, 1980). The near minimaxity of log-linear estimators, provided under appropriate conditions by the approximations of Section 3, permits confidence in expecting such robustness in these estimators.

Log-linear estimators are Stein-type estimators. They involve shrinkage toward a model determined value, after appropriate transformation of cell counts. Indeed, with very small counts only, the transformation is variance stabilizing. Log-linear and Stein estimates are then very similar. Attention has been drawn to the need to slightly modify Stein and log-linear estimators to avoid problems caused by extreme outliers in small tables, but the exact methodology requires further development.

Another derivation of the log-linear form of estimator may add insight and suggest an empirical Bayes interpretation of this class of estimators. We formulate a stochastic model in which the means μ_i in the table are themselves random. Let

$$\alpha_i = \log \mu_i = a_i'\beta + \varepsilon_i, \quad i = 1, \cdots, p \tag{7.1}$$

where the ε_i are assumed to be independent normal errors with mean 0 and common variance σ^2, and a_i' denotes the ith row of the design matrix A. Models similar to this are considered by Leonard in unpublished notes. Given β and σ, the maximum likelihood estimates of $\alpha_1, \cdots, \alpha_p$ may be obtained as the solutions of the equations

$$\exp(\alpha_i) = X_i - \sigma^{-2}(\alpha_i - a_i'\beta), \quad i = 1, \cdots, p. \tag{7.2}$$

If $\mu = e^{\alpha}$, and $f(\alpha) = \exp(\alpha) - x + \sigma^{-2}(\alpha - \alpha_0)$, then an approximate solution to $f(\alpha) = 0$ is $\alpha = z\alpha_0 + (1 - z)\log x$, with $z = 1/(1 + \sigma^2\mu)$. Note the relationship between this solution and the log-linear form (3.3), which differs only in the choice of shrinkage factor, z.

A similarity is thus observed between log-linear estimators and Leonard's Bayesian class of estimators. We may infer from this that when cell means accord with the model (7.1), the log-linear estimator will achieve a considerable risk reduction.

The use of estimators smoothed towards log-linear models provides a very flexible technique for estimation in contingency tables.

Acknowledgements. Computational results of Section 6 were conducted with the assistance of Mr. E. Ranson. The author wishes to thank the Editor, Associate Editor and referees for suggestions that improved this paper.

REFERENCES

BERGER, J. O. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods.* Springer-Verlag, New York.

BERGER, J. O. (1982). Bayesian robustness and the Stein effect. *J. Amer. Statist. Assoc.* **77** 358–368.

BISHOP, Y. M. M., FIENBERG, S. E., and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, Mass.

BROWN, L. D. (1979). A heuristic method of determining admissibility of estimators—with applications. *Ann. Statist.* **7** 960–994.

CLEVENSON, M. L. and ZIDEK, J. V. (1975). Simultaneous estimation of the mean of independent Poisson laws. *J. Amer. Statist. Assoc.* **70** 698–705.

COX, D. R. and LAUH, E. (1967). A note on the graphical analysis of multidimensional contingency tables. *Technometrics* **9** 481–488.

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators, part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.

FELLER, W. (1950). *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd edition. Wiley, New York.

FIENBERG, S. E. and HOLLAND, P. W. (1970). Methods for eliminating zero counts in contingency tables. In *Random Counts on Models and Structures.* G. P. Patil, ed. Pennsylvania State Univ. Press.

HUDSON, H. M. (1974). *Empirical Bayes Estimation.* Tech. Report No. 58, Dept. Statist., Stanford University.

HUDSON, H. M. (1978). A natural identity for exponential families, with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484.

HUDSON, H. M. and TSUI, K. W. (1981). Simultaneous Poisson estimators for a priori hypotheses about means. *J. Amer. Statist. Assoc.* **76** 182–187.

HWANG, J. T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *Ann. Statist.* **10** 857–867.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, vol. 1. J. Neyman, ed. Univ. of California Press.

LEONARD, T. (1972). Bayesian methods for binomial data. *Biometrika* **59** 581–589.

LEONARD, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Assoc.* **72** 869–874.

LEONARD, T. The Bayesian analysis of categorical data. Unpublished notes.

MARITZ, J. S. (1969). Empirical Bayes estimation for the Poisson distribution. *Biometrika* **56** 349–359.

PENG, J. (1975). *Simultaneous estimation of Poisson means.* Tech. Report No. 58, Dept. Statist., Stanford Univ.

PLACKETT, R. L. (1981). *The Analysis of Categorical Data.* Griffin's Statistical Monograph No. 35, A. Stuart ed., 2nd edition. London.

RIES, P. N. and SMITH, H. (1963). The use of chi-square for preference testing in multidimensional problems. *Chem. Eng. Progress* **59** 39–43.

SCLOVE, S., MORRIS, C. and RADHAKRISHNAN, R. (1972). Non-optimality of the preliminary test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.* **43** 1481–1490.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

SUTHERLAND, M., HOLLAND, P. and FIENBERG, S. E. (1974). Combining Bayes and frequency approaches to estimate a multinomial parameter. In *Studies in Bayesian Econometrics and Statistics*, S. Fienberg and A. Zellner, ed. North Holland, Amsterdam.

SCHOOL OF ECONOMIC AND FINANCIAL STUDIES
MACQUARIE UNIVERSITY
NORTH RYDE, NEW SOUTH WALES 2113
AUSTRALIA