Ariel Transmission Coversheet

www.infotrieve.com/ariel

ILL Request:19161754 **FREE** **Borrower:**ORE

MATH LIB.
QA279.4 .A87 1994
AVAILABLE

Ariel Address: OSU-ILL.library.orst.edu

CAN YOU SUPPLY? YES NO CONDITIONAL FUTURE DATE

Affiliation: OCLC Western, GWLA (BTP), ORBIS Cascade Alliance

APR 13

Status: PENDING 20060410 Request Date: 20060410 Need Before: 20060510
OCLC: 29026658 Source: ILLiad Due Date:
Lender: *ORU, UBY, AZU, AZS, IQU
CALLNO:

Title: Aspects of uncertainty : a tribute to D.V. Lindley /
Article: Leonard, T. and Hsu, J. S. J.: The Bayesian analysis of categorical data - A selective review
Date: 1994
Pages: 283-310
Imprint: Chichester ; New York : Wiley, c1994.Series:Wiley series in probability and mathematical statistics
ISBN: 0471943479 :

Verified: <TN:188993> OCLC
Patron: Bulatov, Yaroslav
Bill To: same.O005911 ** GWLA MEMBER **
Ship Via: ARIEL; OSU-ILL.library.orst.edu when possible
Maximum Cost: IFM - $21.25
Copyright Compliance: CCL
Fax: 541-737-1328
Email: valley.ill@oregonstate.edu
Borrowing Notes:
Lending Charges:
Shipped:
Lending Notes:
Lending Restrictions:
Return To:
Return Via:

Library-ILL
Oregon State University
121 The Valley Library
Cross Streets Jefferson Way & Waldo Pl
Corvallis, OR 97331-4501

de Finetti, B. (1962) Does it make sense to speak of good probability appraisers? *The Scientist Speculates: An Anthology of Partially Baked Ideas*, I. J. Good, ed., London, Methuen.
de Finetti B (1974) *Theory of Probability*, vol. 1, London, John Wiley & Sons, pp. 143–6.
Fisher, R. (1925) *Statistical Methods for Research Workers*, Edinburgh, Oliver and Boyd.
Fisher, R. (1935) *The Design of Experiments*, Edinburgh, Oliver and Boyd.
Good, I. J. (1957) The appropriate mathematical tools for describing and measuring uncertainty, reprinted in I. J. Good, *Good Thinking: The Foundations of Probability and its Applications*, 1983, Minneapolis, University of Minnesota Press.
Lindley, D. (1953) Statistical Inference, *Journal of the Royal Statistical Society*, **B15**, 30–76.
Lindley, D. (1956) On a measure of the information provided by an experiment *Annals of Mathematical Statistics*, **27**, 986–1005.
Lindley, D. (1971) *Making Decisions*, London, John Wiley & Sons.
Lindley, D. (1982) Scoring rules and the inevitability of probability, *International Statistical Review*, **50**, 1–26.
Pilz, J. (1991) *Bayesian Estimation and Experimental Design in Linear Regression Models*, New York, John Wiley & Sons.
Verdinelli, I. (1992) Advances in Bayesian experimental design, *Bayesian Statistics 4*, J. O. Berger, J. Bernardo, A. P. Dawid, and A. F. M. Smith, eds, Oxford, Oxford University Press.

Department of Mathematics
University of Canterbury
Christchurch
NEW ZEALAND

CHAPTER 18

The Bayesian Analysis of Categorical Data—a Selective Review

Tom Leonard[†] and John S. J. Hsu[‡]
[†]*University of Wisconsin-Madison,*
[‡]*University of California at Santa Barbara*

18.1 THE FOUNDATIONS OF THE 1960s

Lindley (1964) proposed a Bayesian analysis for $r \times s$ contingency tables. Suppose that cell frequencies $\{y_{ij};\ i = 1, \ldots, r; j = 1, \ldots, s\}$ are taken to possess a multinomial sampling distribution, with respective unconditional cell probabilities $\{\theta_{ij};\ i = 1, \ldots, r; j = 1, \ldots, s\}$, satisfying

$$\sum_{rs} \theta_{kg} = 1, \quad \text{and samples size} \quad n = \sum_{kg} y_{kg}.$$

Consider multivariate logits $\{\gamma_{ij};\ i = 1, \ldots, r; j = 1, \ldots, s\}$, satisfying

$$\theta_{ij} = \exp(\gamma_{ij}) / \sum_{kg} \exp(\gamma_{kg}), \tag{1.1}$$

and any log contrast, taking the form

$$\lambda = \sum_{ij} a_{ij} \log \theta_{ij} = \sum_{ij} a_{ij} \gamma_{ij}, \tag{1.2}$$

where $\sum_{ij} a_{ij} = 0$.

Suppose that the prior distribution of the θ_{ij} is Dirichlet, with parameters $\{\alpha_{ij};\ i = 1, \ldots, r; j = 1, \ldots, s\}$, so that the posterior distribution is also Dirichlet, but with updated parameters

Aspects of Uncertainty edited by P. R. Freeman and A. F. M. Smith.

$$\{\alpha_{ij}^* = \alpha_{ij} + y_{ij};\ i = 1, \ldots, r;\ j = 1, \ldots, s\}.$$

Then Lindley proved that λ is distributed, in the posterior assessment, as a linear combination of independent log-Gamma variates. Equivalently

$$\lambda = \sum_{ij} a_{ij} \log u_{ij}, \tag{1.3}$$

where the u_{ij} are independent chi-squared variates, with respective degrees of freedom $2\alpha_{ij}^* = 2(\alpha_{ij} + y_{ij})$. If $\alpha_{ij}^* \geqslant 5$, $\log u_{ij}$ may be taken to be approximately normally distributed with mean $\log 2 + \log(\alpha_{ij} + y_{ij})$, and variance $(\alpha_{ij} + y_{ij})^{-1}$. Hence, in many practical situations, the posterior distribution of λ will be approximately normal with mean

$$\lambda^* = \sum_{ij} a_{ij} \log(\alpha_{ij} + y_{ij}), \tag{1.4}$$

and variance

$$v^* = \sum_{ij} a_{ij} (\alpha_{ij} + y_{ij})^{-1}. \tag{1.5}$$

Bloch and Watson (1967) proposed some refinements to this approximation. However, Leonard *et al.* (1989) and Hsu (1990) show that Lindley's original approximation is remarkably accurate, when compared with the exact result. Next consider a 2×2 table, set $r = s = 2$, and let

$$\lambda = \log\theta_{11} + \log\theta_{22} - \log\theta_{12} - \log\theta_{21} = \gamma_{11} + \gamma_{22} - \gamma_{12} - \gamma_{21}, \tag{1.6}$$

denote the log measure of association. Then (e.g. Lindley 1965, p. 150) λ is exactly distributed in the posterior assessment as

$$\lambda = \log\alpha_{11}^* + \log\alpha_{22}^* - \log\alpha_{12}^* - \log\alpha_{21}^* + \log F_{2\alpha_{12}^*}^{2\alpha_{11}^*} + \log F_{2\alpha_{21}^*}^{2\alpha_{22}^*}, \tag{1.7}$$

where $F_{2\alpha_{12}^*}^{2\alpha_{11}^*}$ and $F_{2\alpha_{21}^*}^{2\alpha_{22}^*}$ are independent F-variates, with the notationally obvious degrees of freedom. Hence the posterior distribution of λ is approximately normal, with mean

$$\lambda^* = \log\alpha_{11}^* + \log\alpha_{22}^* - \log\alpha_{12}^* - \log\alpha_{21}^*,$$

and variance

$$v^* = (\alpha_{11}^*)^{-1} + (\alpha_{22}^*)^{-1} + (\alpha_{12}^*)^{-1} + (\alpha_{211}^*)^{-1}. \tag{1.8}$$

In situations involving vague prior information, Lindley (1964) suggests setting $\alpha_{ij} = 0$ for $i = 1, 2$ and $j = 1, 2$, and proposes

$$B^2 = (\lambda^*)^2 / v^*, \tag{1.9}$$

as an excellent alternative to the usual chi-squared statistic, for testing $\lambda = 0$, on one degree of freedom. The choices $\alpha_{ij} = 0.5$, for $i = 1, 2$ and $j = 1, 2$, however, provide the 'Jeffreys prior' and may be preferable.

Let $\xi_1 = \theta_{11}/(\theta_{11} + \theta_{12})$ and $\xi_2 = \theta_{21}/(\theta_{21} + \theta_{22})$, denote the conditional probabilities given the rows. Then the log-measure of association λ in (1.6) satisfies

$$\lambda = \text{logit}(\xi_1) - \text{logit}(\xi_2), \tag{1.10}$$

where $\text{logit}(\xi) = \log(\xi) - \log(1 - \xi)$. Note that, in the posterior assessment, ξ_1 and ξ_2 possess independent beta distributions with

$$\xi_j \sim \text{Beta}(\alpha_{1j}^*, \alpha_{2j}^*) \qquad (j = 1, 2). \tag{1.11}$$

Altham (1969) exploits this representation to develop an explicit expression, involving a hypergeometric series, for the posterior probability, that $\lambda > 0$. For fixed p, let

$$I_p(\alpha_1, \alpha_2) = \text{prob}(\xi \leqslant p \mid \xi \sim \text{Beta}(\alpha_1, \alpha_2)). \tag{1.12}$$

Then, if a is an integer, the cumulative distribution function (1.12) satisfies

$$I_p(a, n - a + 1) = \sum_{j=a}^{n} {}^nC_j p^j (1 - p)^{n-j} \tag{1.13}$$

$$= p(Y \geqslant a \mid Y \text{ possesses a binomial distribution with probability } p \text{ and sample size } n). \tag{1.14}$$

Altham uses (1.14) to obtain a representation for the posterior distribution of ξ_2 in (1.10), and then develops an expression for the posterior probability that $\xi_1 > \xi_1$, via an integration with respect to ξ_1. In the special case, when $\alpha_{11} = 0$, $\alpha_{12} = 1$, $\alpha_{21} = 1$, and $\alpha_{22} = 0$, her posterior probability reduces to a significance probability for Fisher's exact test. However, Jeffreys' choices $\alpha_{ij} = 0.5$ for and $i = 1, 2$ and $j = 1, 2$ also possess some frequency justification, since, whenever a is an integer

$$\sum_{j=a+1}^{n} {}^nC_j p^j (1 - p)^{n-j} \leqslant I_p(a + 1/2, n - a + 1/2) \leqslant \sum_{j=a}^{n} {}^nC_j p^j (1 - p)^{n-j}. \tag{1.15}$$

For example, the posterior probability that ξ_j is less than some hypothesized value $\xi_j^{(0)}$ (i.e. a 'Bayesian significance probability') will, under the Jeffreys prior, always lie between two frequency-based significance probabilities i.e., (1) the probability that y_{1j} is strictly greater than the value of y_{1j} actually observed, if indeed $\xi_j = \xi_j^{(0)}$, (2) a similar probability, but with weak inequality. It is possible to develop extensions of this argument, as a frequency-based justification for the Jeffreys prior ($\alpha_{ij} \equiv 0.5$), for a general $r \times s$ contingency table. For a 2×2 table, the choices $\alpha_{11} = 1$, $\alpha_{12} = 0$, $\alpha_{21} = 0$ and $\alpha_{22} = 1$, would also lead to a significance probability for Fisher's exact test. The Jeffreys prior provides a compromise between the two frequency-based significance probabilities, involving weak and strong inequalities.

For an $r \times s$ table, Lindley's (1964) Bayesian approach can be regarded as formalizing a non-Bayesian approach developed by Goodman (1964). Goodman's full rank interaction model, assumes that the multivariate logits γ_{ij} satisfy

$$\gamma_{ij} = \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad (i = 1, \ldots, r; j = 1, \ldots, s), \tag{1.16}$$

where λ_i^A denotes the ith row effect, λ_j^B denotes the jth column effect, and λ_{ij}^{AB} denotes the (i, j)th interaction effect. Subject to the constraints $\lambda_\cdot^A = \lambda_\cdot^B = \lambda_{i\cdot}^{AB} = \lambda_{\cdot j}^{AB} = 0$, where the dot notation denotes average with respect to that subscript, we have

$$\lambda_i^A = \gamma_{i\cdot} - \gamma_{\cdot\cdot}, \tag{1.17}$$

$$\lambda_j^B = \gamma_{\cdot j} - \gamma_{\cdot\cdot}, \tag{1.18}$$

and

$$\lambda_{ij}^{AB} = \gamma_{ij} - \gamma_{i\cdot} - \gamma_{\cdot j} + \gamma_{\cdot\cdot}. \tag{1.19}$$

Since each expression in (1.17)–(1.19) is a linear contrast, the preceding developments (e.g. (1.4) and (1.5)) may be used to provide a normal approximation to the posterior distribution of any parameter of interest. It is, for example, straightforward to calculate an approximate Bayesian significance probability to investigate whether any particular interaction effect can reasonably be set equal to zero. Non-zero interactions can be used to indicate cells of possible interest in the contingency table. One possible choice of an overall test statistic, for independence of rows and columns, is

$$B^2 = \frac{(r-1)(s-1)}{rs} \sum_{ij} (\tilde{\lambda}_{ij}^{AB})^2 / v_{ij}^{AB}, \tag{1.20}$$

where $\tilde{\lambda}_{ij}^{AB}$ and v_{ij}^{AB} are respectively the approximate posterior mean and variance of λ_{ij}, under a Jeffreys prior. The observed value of B^2 may be compared, as a first-order asymptotic approximation, with appropriate upper percentage points of the chi-squared distribution with $(r-1)(s-1)$ degrees of freedom.

Irving Jack Good (1965, 1967) also developed the analysis of contingency tables, using Dirichlet priors. For example, under our above specification the posterior mean of θ_{ij} may be represented in the form

$$\theta_{ij}^* = \frac{np_{ij} + \alpha\mu_{ij}}{n + \alpha} \tag{1.21}$$

where $p_{ij} = y_{ij}/n$ denotes the corresponding cell proportion, $\mu_{ij} = \alpha_{ij}/\alpha$ is the prior mean of θ_{ij}, and $\alpha = \sum_{kg} \alpha_{kg}$ denotes the 'prior sample size'. Good argued that the choice of α should be based upon the data, either via a hierarchical Bayes, or bootstrap Bayes approach. This is with the intention

of representing possible smoothness of the contingency table, and purposefully violates Johnson's sufficientness postulate (Johnson and Braithwaite 1932). The genesis of Good's ideas can be found in his historical collaboration with Alan Turing, the father of machine intelligence, during the Second World War, when the problem of breaking the Nazi codes was solved via a cryptanalysis based on the estimation of letter frequencies in the German language.

18.2 NUMERICAL EXAMPLES

Consider, firstly the gene frequency data, reported by Lindley (1965, p. 180). Out of $n = 106$ individuals, $y_{11} = 6$ possess both gene A and gene B, $y_{22} = 61$ possess neither gene, $y_{12} = 13$ possess gene A, but not gene B, and $y_{21} = 16$ possess gene B, but not gene A.

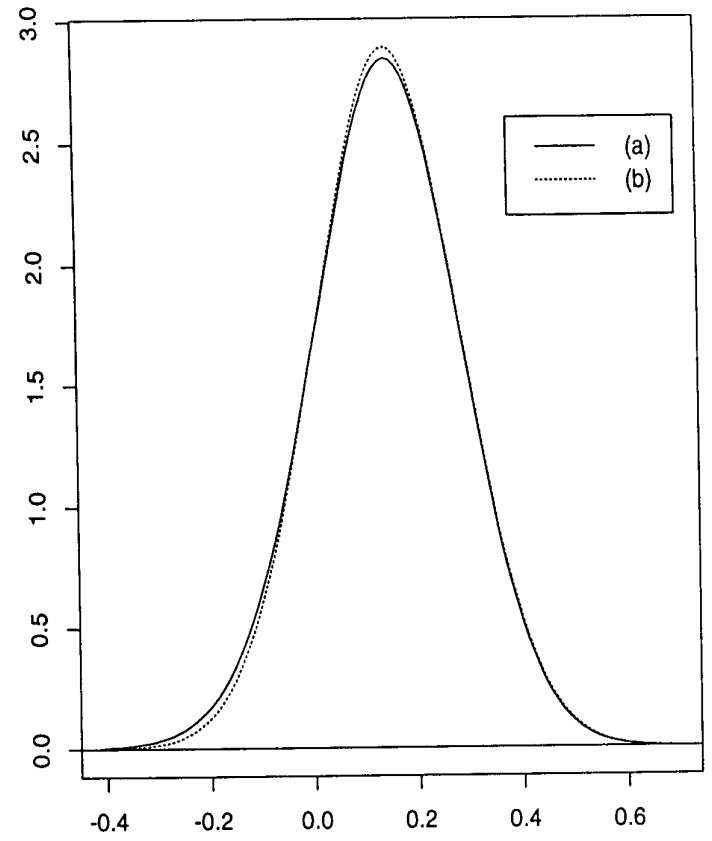

Figure 18.1 Posterior density of log measure of association gene data: (a) exact and BCM; (b) normal approximation.

Curve (a) of Figure 18.1 discribes the exact posterior density of the log measure of association (1.6), using a Jeffreys prior for the unconditional cell frequencies, and curve (b) describes Lindley's normal approximation, which is remarkably accurate. The exact Bayesian significance probability i.e. posterior probability that $\lambda < 0$, is 0.153, while the normal approximation gives 0.140. The test statistic in (1.10) is evaluated as $B^2 = 1.12$, on one degree of freedom, while the usual chi-squared statistic for testing $\lambda = 0$, gives $X^2 = 0.87$.

Table 18.1. The engineering apprentice data.

Section head's assessment	Written test result			
	A	B	C	D
Excellent	26	29	21	11
	(0.034)	(0.311)	(0.721)	(0.916)
	[0.036]	[0.301]	[0.726]	[0.907]
Very good	33	43	35	20
	(0.251)	(0.402)	(0.560)	(0.743)
	[0.248]	[0.390]	[0.548]	[0.752]
Average	47	71	72	45
	(0.860)	(0.703)	(0.197)	(0.231)
	[0.860]	[0.699]	[0.204]	[0.248]
Needs to improve	7	12	11	9
	(0.851)	(0.586)	(0.460)	(0.109)
	[0.842]	[0.612]	[0.484]	[0.108]

The 4×4 contingency table in Table 18.1 was reported by Lindley (1965, p. 180), and cross-classifies $n = 492$ engineering apprentices, according to their section head's assessment, and their grade on a written test. Under Goodman's full rank interaction model (1.17), histogram (a), of Figure 2, represents the exact posterior density of the interaction effect λ_{11}^{AB} for the (1, 1)th cell, obtained by Monte Carlo simulation, and curve (c) gives the normal approximation. The bracketed entries of Table 18.1 describe the exact Bayesian significance probabilities (i.e. posterior probabilities that the interaction effect is less than zero under a Jeffreys prior), for each of the 16 cells; the figures in square brackets were instead obtained using the normal approximation, and they are very close to the exact values. The B^2 statistic (1.21) becomes $B^2 = 6.76$, while $X^2 = 7.51$, on nine degrees of freedom.

When analysing any two-way contingency table, it is always important to consider a summary of the cells with significant interaction effects, since this is comparable with the analysis of residuals, in regression analysis. By highlighting the important cells in the table, and considering patterns of important cells, the statistician is able to infer the main conclusions from the data. In this case, cells (1, 1) and (4.4) give low significance probabilities, while cells (1, 4) and (4.4) give high significance probabilities. This enables us to infer the (in this case obvious) conclusion that there is some, but not overwhelming, positive association between the section head's assessment and the written test result. While it would be possible to perfectly fit the significant cells using Goodman's (1968) quasi-independence procedures, this can frequently give a model which overfits, but with less practical interpretability. A full rank interaction analysis of a 12×8 table (the Marine Corps data) is proposed by Leonard (1985) and Leonard and Novick (1986), using a more informative prior.

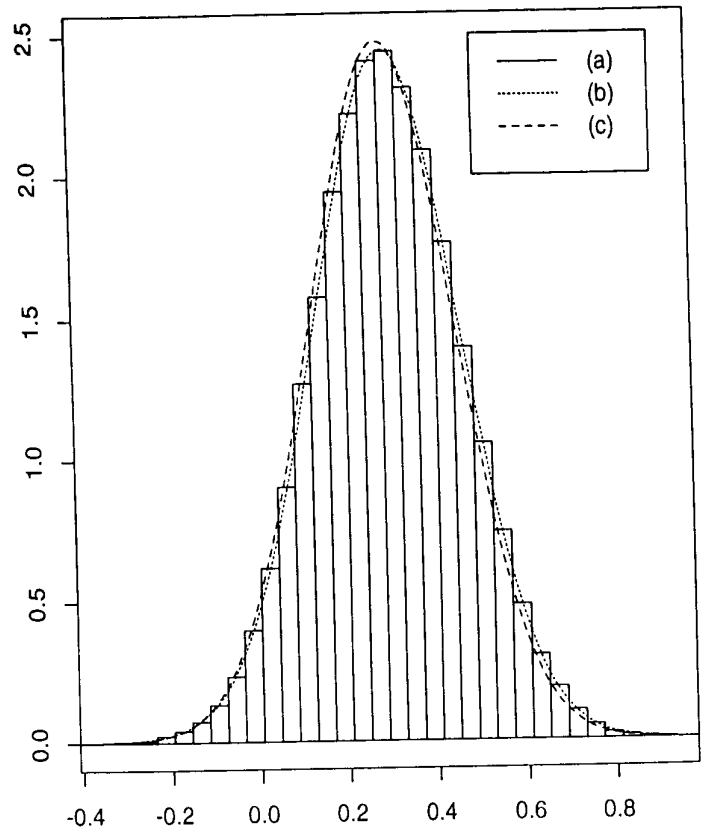

Figure 18.2 Posterior density of (1, 1)th interaction effect (engineering apprentice data): (a) exactly simulated histogram; (b) BCM; (c) normal approximation.

18.3 BAYES–STEIN METHODS OF THE 1970s

When the first co-author commenced graduate studies at University College London, in September 1970, as a 22-year-old Masters student, Dennis

Curve (a) of Figure 18.1 discribes the exact posterior density of the log measure of association (1.6), using a Jeffreys prior for the unconditional cell frequencies, and curve (b) describes Lindley's normal approximation, which is remarkably accurate. The exact Bayesian significance probability i.e. posterior probability that $\lambda < 0$, is 0.153, while the normal approximation gives 0.140. The test statistic in (1.10) is evaluated as $B^2 = 1.12$, on one degree of freedom, while the usual chi-squared statistic for testing $\lambda = 0$, gives $X^2 = 0.87$.

Table 18.1. The engineering apprentice data.

Section head's assessment	Written test result			
	A	B	C	D
Excellent	26	29	21	11
	(0.034)	(0.311)	(0.721)	(0.916)
	[0.036]	[0.301]	[0.726]	[0.907]
Very good	33	43	35	20
	(0.251)	(0.402)	(0.560)	(0.743)
	[0.248]	[0.390]	[0.548]	[0.752]
Average	47	71	72	45
	(0.860)	(0.703)	(0.197)	(0.231)
	[0.860]	[0.699]	[0.204]	[0.248]
Needs to improve	7	12	11	9
	(0.851)	(0.586)	(0.460)	(0.109)
	[0.842]	[0.612]	[0.484]	[0.108]

The 4×4 contingency table in Table 18.1 was reported by Lindley (1965, p. 180), and cross-classifies $n = 492$ engineering apprentices, according to their section head's assessment, and their grade on a written test. Under Goodman's full rank interaction model (1.17), histogram (a), of Figure 2, represents the exact posterior density of the interaction effect λ_{11}^{AB} for the (1, 1)th cell, obtained by Monte Carlo simulation, and curve (c) gives the normal approximation. The bracketed entries of Table 18.1 describe the exact Bayesian significance probabilities (i.e. posterior probabilities that the interaction effect is less than zero under a Jeffreys prior), for each of the 16 cells; the figures in square brackets were instead obtained using the normal approximation, and they are very close to the exact values. The B^2 statistic (1.21) becomes $B^2 = 6.76$, while $X^2 = 7.51$, on nine degrees of freedom.

When analysing any two-way contingency table, it is always important to consider a summary of the cells with significant interaction effects, since this is comparable with the analysis of residuals, in regression analysis. By highlighting the important cells in the table, and considering patterns of important cells, the statistician is able to infer the main conclusions from the data. In this case, cells (1, 1) and (4.4) give low significance probabilities, while cells (1, 4) and (4.4) give high significance probabilities. This enables us to infer the (in this case obvious) conclusion that there is some, but not overwhelming, positive association between the section head's assessment and the written test result. While it would be possible to perfectly fit the significant cells using Goodman's (1968) quasi-independence procedures, this can frequently give a model which overfits, but with less practical interpretability. A full rank interaction analysis of a 12×8 table (the Marine Corps data) is proposed by Leonard (1985) and Leonard and Novick (1986), using a more informative prior.

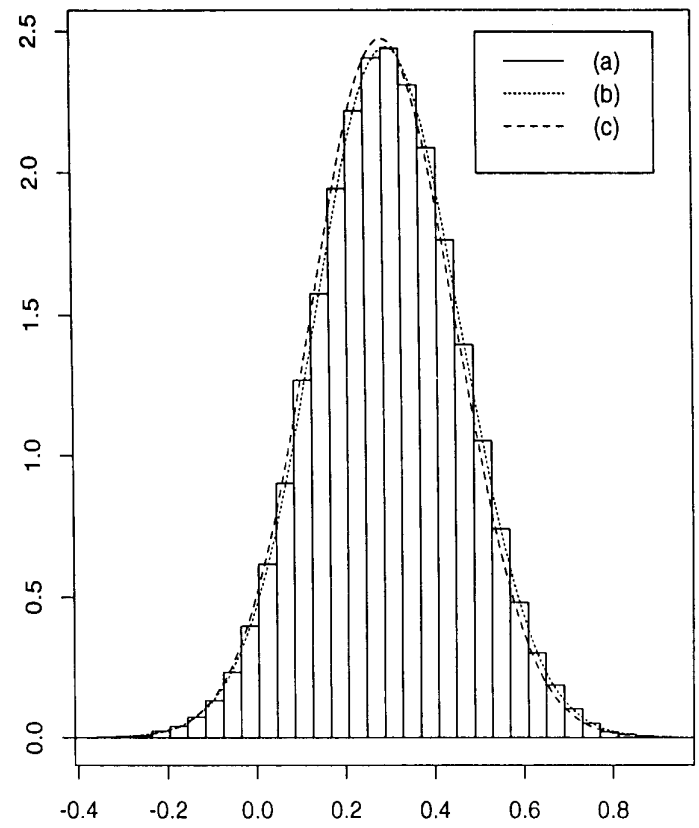

Figure 18.2 Posterior density of (1, 1)th interaction effect (engineering apprentice data): (a) exactly simulated histogram; (b) BCM; (c) normal approximation.

18.3 BAYES–STEIN METHODS OF THE 1970s

When the first co-author commenced graduate studies at University College London, in September 1970, as a 22-year-old Masters student, Dennis

Lindley gave him a preprint of Lindley (1971) and asked him to 'do the same thing for the binomial distribution'. Lindley and Smith were at the time developing their research on Bayes–Stein estimators for several normal means, and the linear model, (e.g. Lindley and Smith 1972). The corresponding analysis of categorical data turned out to be a question of finding a convenient transformation, and then handling the non-normality in the posterior distribution. This led to choices of non-conjugate prior distributions, other than the Dirichlet, and a general approach to the Bayesian analysis of categorical data (Leonard 1972, 1973a, b, 1975, 1976, 1977a; Laird 1978) using either multivariate normal prior distributions, or mixtures thereof, for logistic transformations of the parameters. The basic idea was simple though very unconventional at the time, as convention required conjugacy: if everybody is interested in normal approximations to the posterior distribution of the logits, then why not assume normality in the prior?

To illustrate this approach, we describe some hitherto unpublished research, for a generalization of Leonard's original thesis problem, i.e. the simultaneous estimation of the parameters of the product multinomial model (e.g. simultaneously smoothing several histograms). For $i = 1, \ldots, m$, consider cell frequencies $y_{i1}, \ldots, y_{is}$, which are taken to possess a multinomial distribution, with respective cell probabilities, $\theta_{i1}, \ldots, \theta_{is}$ satisfying $\sum_g \theta_{ig} = 1$, and sample size $n_i = \sum_g y_{ig}$. Given the θ_{ij}, the vectors $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{is})^T$ are taken to be independent. For $i = 1, \ldots, m$, consider m sets of multivariate logits $\{\gamma_{i1}, \ldots, \gamma_{is}\}$, satisfying

$$\theta_{ij} = \exp(\gamma_{ij}) / \sum_{g=1}^{s} \exp(\gamma_{ig}) \qquad (j = 1, \ldots, s). \tag{3.1}$$

Suppose that the parameters of different multinomial distributions are a priori *exchangeable*, and that given $\boldsymbol{\mu}$ and $\boldsymbol{C}$, the $\boldsymbol{\gamma}_i = (\gamma_{ij}, \ldots, \gamma_{is})^T$ are independent and possess multivariate normal distributions with common mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$. This is a much more general specification than available (e.g. Leonard 1977b, c) via Dirichlet distributions, for the cell probabilities, since very general prior dependencies are permitted between the parameters. The $\boldsymbol{\gamma}_i$ are a posteriori independent, give $\boldsymbol{\mu}$ and $\boldsymbol{C}$, and the posterior density of $\boldsymbol{\gamma}_i$ is

$$\pi(\boldsymbol{\gamma}_i \mid \boldsymbol{y}_i) \propto \exp\{\boldsymbol{y}_i^T \boldsymbol{\gamma}_i - n_i D(\boldsymbol{\gamma}_i) - (\boldsymbol{\gamma}_i - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{\gamma}_i - \boldsymbol{\mu})/2\} \tag{3.2}$$

where

$$D(\boldsymbol{\gamma}_i) = \log \sum_{g=1}^{g} \exp(\gamma_{ig}) \tag{3.3}$$

The posterior density (3.2) is non-normal owing to the presence of the $D(\boldsymbol{\gamma}_i)$ term (3.3). Therefore, in the 1970s an exact posterior analysis was

not available. However, two approximations (A and B) are applicable. Method (A) is based upon the normal approximation (Leonard 1973b):

$$\mathscr{L}^*(\boldsymbol{\gamma}_i \mid \boldsymbol{y}_i) = B(\boldsymbol{y}_i) \exp\{-(\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i) \boldsymbol{R}_i (\boldsymbol{\gamma}_i - \hat{\boldsymbol{\gamma}}_i)/2\}, \tag{3.4}$$

$$= B(\boldsymbol{y}_i) \exp\{-\sum_j y_{ij} (\gamma_{ij} - \log y_{ij} - d_j)^2/2\}, \tag{3.5}$$

to the likelihood of $\boldsymbol{\gamma}_i$, given $\boldsymbol{y}_i$. Here,

$$\hat{\boldsymbol{\gamma}}_i = (\log y_{i1}, \ldots, \log y_{is})^T, \tag{3.6}$$

$$\boldsymbol{R}_i = \operatorname{diag}(y_{i1}, \ldots, y_{is}) - n_i^{-1} \boldsymbol{y}_i \boldsymbol{y}_i^T, \tag{3.7}$$

$$B(\boldsymbol{y}_i) = n_i! n_i^{-n_i} y_{ij}^{-y_{ij}} / y_{ij}!, \tag{3.8}$$

and

$$d_j = \sum_j y_{ij} (\gamma_{ij} - \log y_{ij}). \tag{3.9}$$

The likelihood approximation (3.4) is reasonable if $y_{ij} \geq 5$ for all i, j. Consequently, the posterior distribution of $\boldsymbol{\gamma}_i$, given $\boldsymbol{\mu}$ and $\boldsymbol{C}$, is approximately normal with mean vector

$$\boldsymbol{\gamma}_i^* = (\boldsymbol{R}_i + \boldsymbol{C}^{-1})^{-1} (\boldsymbol{R}_i \hat{\boldsymbol{\gamma}}_i + \boldsymbol{C}^{-1} \boldsymbol{\mu}), \tag{3.10}$$

and covariance matrix

$$\boldsymbol{D}_i = (\boldsymbol{R}_i + \boldsymbol{C}^{-1})^{-1}. \tag{3.11}$$

The expression in (3.10) expresses $\boldsymbol{\gamma}_i^*$ in the form of a matrix weighted average of $\boldsymbol{\gamma}_i$ and $\boldsymbol{\mu}$. A more precise procedure, (method B) which is suitable even if some of the y_{ij} are zero, involves taking the posterior distribution of $\boldsymbol{\gamma}_i$, given $\boldsymbol{\mu}$ and $\boldsymbol{C}$, to be approximately normal with mean vector $\tilde{\boldsymbol{\gamma}}_i$ and covariance matrix $\tilde{\boldsymbol{D}}_i$, where $\tilde{\boldsymbol{\gamma}}_i$ and $\tilde{\boldsymbol{D}}_i$ are respectively the exact posterior mode vector and dispersion matrix of $\boldsymbol{\gamma}_i$. These satisfy

$$n_i \tilde{\boldsymbol{\theta}}_i = \boldsymbol{y}_i - \boldsymbol{C}^{-1} (\tilde{\boldsymbol{\gamma}}_i - \boldsymbol{\mu}), \tag{3.12}$$

and

$$\boldsymbol{D}_i = n_i \{\operatorname{diag}(\tilde{\theta}_{i1}, \ldots, \tilde{\theta}_{is}) - \tilde{\boldsymbol{\theta}}_i \tilde{\boldsymbol{\theta}}_i^T\} + \boldsymbol{C}^{-1}, \tag{3.13}$$

where $\tilde{\boldsymbol{\theta}}_i = (\tilde{\theta}_{i1}, \ldots, \tilde{\theta}_{is})^T$, with

$$\tilde{\theta}_{ij} = \exp(\tilde{\gamma}_{ij}) / \sum_{g=1}^{g} \exp(\tilde{\gamma}_{ig}). \tag{3.14}$$

Note that (3.12) and (3.14) should be solved, using standard Newton–Raphson techniques for the $\tilde{\boldsymbol{\gamma}}_i$. Both $\boldsymbol{\gamma}_i^*$ and $\tilde{\boldsymbol{\gamma}}_i$ are, of course, highly

dependent upon $\boldsymbol{\mu}$ and $\boldsymbol{C}$. Following I. J. Good's philosophy Fienberg and Holland (1973) showed in a much simpler context that it is possible to obtain estimators for the θ_{ij} with excellent mean squared error properties, by estimating the prior parameters empirically, from the current data set. Under a hierarchical Bayes procedure (e.g. Lindley and Smith 1972) it would be possible to assign further distributions to $\boldsymbol{\mu}$ and $\boldsymbol{C}$, at the second stage of the prior model. Before 1977, it seemed impossible, in the categorical data context, to find an estimation procedure for $\boldsymbol{\mu}$ and $\boldsymbol{C}$ which worked in practical terms, if some of the cell frequencies were zero. For example, joint modal procedures (Leonard 1972, 1973a, b, 1975) tend to over-collapse the estimates of the variance components, as noted by Leonard (1976, 1977a).

However, the EM algorithm, summarized by Dempster, *et al.* (1977), and Laird (1978), tells us that we may sensibly estimate $\boldsymbol{\mu}$ and $\boldsymbol{C}$ by equating prior and posterior expectations of $\boldsymbol{\gamma}.\boldsymbol{\gamma}.^T$ and $\sum_i \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T$, where $\boldsymbol{\gamma}. = m^{-1}\sum_i \boldsymbol{\gamma}_i$. If performed exactly, this would lead to estimates $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{C}}$ for $\boldsymbol{\mu}$ and $\boldsymbol{C}$ maximizing their 'integrated likelihood', obtained from the joint distribution of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ conditional only upon $\boldsymbol{\mu}$ and $\boldsymbol{C}$. Under the approximations developed above, for our method *B*, the EM algorithm yields the succinct expressions

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\gamma}}., \tag{3.15}$$

and

$$\tilde{\boldsymbol{C}} = (m-1)^{-1}\sum_{i=1}^{m}(\tilde{\boldsymbol{\gamma}}_i - \tilde{\boldsymbol{\gamma}}.)(\tilde{\boldsymbol{\gamma}}_i - \tilde{\boldsymbol{\gamma}}.)^T + m^{-1}\sum_{i=1}^{m}\tilde{\boldsymbol{D}}_i, \tag{3.16}$$

which may be solved by cyclic substitution, combined with Newton–Raphson, for (3.12) and (3.14). Note the vital importance of the second term in (3.16) which avoids the difficult over-collapsing of the estimates of the variance components, mentioned earlier.

A numerical example of this procedure is described in section 18.4. Most of the first co-author's contributions of the 1970s follow naturally from Lindley (1964), together with Lindley and Smith (1972). For example, Lindley (1964), suggested smoothing the probabilities in a histogram by employing a multivariate normal prior distribution for a set of linearly independent log-contrasts, thus motivating Leonard (1973a, 1978). Both Leonard (1975) and Laird (1978) assign normal prior distributions to the parameters in Goodman's full rank interaction model (1.17). Nazaret (1987) later applied similar techniques to the analysis of three-way tables, but did not include Laird's important adjustments, paralleling (3.16) to the estimates of the variance components.

During the 1970s the Dirichlet prior approach was further developed (e.g. Good 1975, 1976; Crook and Good 1980; Lochner 1975). Gunel and Dickey (1974) combined Dirichlet priors with positive prior probabilities

on null hypotheses of independence. Hickman and Miller (1977) further developed histogram smoothing, on an actuarial procedure for graduating mortality tables, and this is perhaps one of the most sucessful real applications of Bayesian procedures.

18.4 SMOOTHING GRADE DISTRIBUTIONS FOR 40 LONDON HIGH SCHOOLS

The observed percentages in the first six columns of Table 18.2 describe the percentages of students obtaining grades 1–6 in a mathematics test at 40 London schools. The underlying data were regarded as numerical realizations of the frequencies of a product multinomial model, i.e. 40 multinomial distributions, each with six cells. The smoothed percentages in the last six columns of Table 18.2 were calculated from $100\tilde{\boldsymbol{\theta}}_{ij}$, where the $\tilde{\boldsymbol{\theta}}_{ij}$ are the smoothed cell probabilities discussed in section 18.3. The sample sizes n_i in the 13th column describe the numbers of students taking the test at the 40 schools.

Table 18.2. Observed and smoothed percentages

	Observed						Smoothed						
i \ j	1	2	3	4	5	6	1	2	3	4	5	6	n_i
1	6.7	17.8	24.4	28.9	6.7	15.6	6.6	18.8	24.2	28.2	7.2	15.0	45
2	0.0	21.6	24.3	18.9	13.5	21.6	3.8	16.9	22.2	26.2	9.9	20.9	37
3	5.3	15.8	42.1	26.3	5.3	5.3	8.3	21.4	29.1	26.5	5.8	8.9	19
4	22.2	25.9	29.6	11.1	7.4	3.7	19.5	26.9	26.5	17.9	4.2	4.9	27
5	16.7	33.3	11.1	16.7	5.6	16.7	12.7	25.4	22.8	22.3	5.7	11.1	18
6	5.9	7.4	26.5	33.8	8.8	17.6	4.8	12.9	23.8	31.5	8.9	18.0	68
7	31.7	17.1	14.6	19.5	9.8	7.3	24.1	22.1	21.7	19.1	5.7	7.4	41
8	4.5	4.5	22.7	31.8	9.1	27.3	3.6	12.4	20.2	29.7	9.6	24.5	22
9	16.1	45.2	19.4	9.7	3.2	6.4	17.5	33.6	23.8	16.8	3.4	5.0	31
10	13.0	31.5	31.5	24.1	0.0	0.0	14.9	30.3	29.1	20.5	2.4	2.9	54
11	23.5	35.3	20.6	20.6	0.0	0.0	22.7	31.7	24.3	16.5	2.2	2.6	34
12	22.8	26.3	21.1	15.8	3.5	10.5	19.8	26.2	23.0	18.7	4.4	7.8	57
13	7.1	14.2	14.2	32.1	10.7	21.4	5.3	15.7	20.5	29.0	9.0	20.4	28
14	13.9	33.3	25.0	19.4	0.0	8.3	14.2	29.2	25.6	20.9	3.7	6.5	38
15	0.0	18.2	13.6	54.5	9.1	4.5	4.2	17.2	23.2	34.7	7.4	13.4	22
16	12.5	31.3	18.8	18.8	6.3	12.5	11.3	24.8	24.6	23.5	5.7	10.1	16
17	0.0	9.4	25.0	50.0	9.4	6.3	3.3	14.3	24.4	36.3	8.0	13.7	32
18	0.0	5.3	15.8	21.1	26.3	31.6	2.0	9.1	16.1	26.6	13.9	32.3	18
19	0.0	3.7	7.4	14.8	3.7	70.4	0.7	4.6	8.4	20.0	9.4	57.1	27
20	2.4	7.3	19.5	36.6	4.8	29.3	2.7	11.3	18.8	31.9	8.4	27.0	42
21	29.2	16.7	8.3	37.5	4.2	4.2	19.5	23.5	22.8	23.3	4.4	6.5	24
22	14.2	21.4	42.9	21.4	0.0	0.0	14.8	26.4	29.0	21.2	3.7	4.8	14
23	0.0	17.4	34.8	17.4	17.4	13.0	4.8	17.2	25.5	26.8	9.7	15.9	23
24	5.6	19.4	30.1	22.2	11.1	11.1	7.1	19.9	26.5	26.1	7.9	12.5	36

25	31.3	18.8	28.1	15.6	3.1	3.1	25.6	25.3	25.5	16.7	3.2	3.7	32
26	9.5	19.0	42.9	14.3	0.0	14.3	10.1	22.9	28.4	23.5	5.2	9.8	21
27	5.6	37.0	22.2	25.9	3.7	5.6	8.8	30.0	25.4	24.5	4.3	7.1	54
28	0.0	30.8	7.7	30.8	0.0	30.8	4.4	17.7	20.2	28.7	7.4	21.6	13
29	4.0	12.0	16.0	32.0	8.0	28.0	3.7	13.6	19.3	29.5	9.1	24.8	25
30	0.0	16.7	27.8	33.3	5.6	16.7	4.5	17.1	23.9	30.1	7.6	16.8	18
31	3.7	11.1	37.0	37.0	0.0	11.1	5.8	18.0	27.6	30.9	5.8	11.9	27
32	0.0	42.9	28.6	21.4	7.1	0.0	9.6	27.6	27.4	23.6	4.8	6.9	14
33	0.0	14.3	28.6	21.4	21.4	14.3	4.5	15.8	23.2	28.1	10.1	18.2	14
34	19.5	39.0	17.1	22.0	0.0	2.4	19.4	33.0	23.4	18.1	2.5	3.5	41
35	37.9	13.8	37.9	6.9	3.4	0.0	31.8	24.6	26.2	12.8	2.4	2.1	28
36	0.0	18.8	6.2	25.0	6.2	43.8	2.5	11.9	15.9	27.2	9.4	33.1	16
37	13.3	20.0	20.0	26.7	6.7	13.3	9.8	21.6	24.5	25.7	6.4	11.8	15
38	16.7	37.5	41.7	4.2	0.0	0.0	20.3	32.4	28.3	14.3	2.2	2.3	24
39	18.3	31.7	21.7	15.0	6.7	6.7	17.5	28.7	24.0	18.4	4.8	6.5	67
40	0.0	11.1	11.1	33.3	22.2	22.2	3.4	13.3	19.7	29.4	10.5	23.7	9

The EM algorithm procedures of section 18.3 provided an empirical estimate

$$\tilde{\boldsymbol{\mu}} = (-0.51, 0.41, 0.56, 0.59, -0.82, -0.24)^T \tag{4.1}$$

for the prior mean vector $\boldsymbol{\mu}$. This corresponds to a common prior estimate

$$\tilde{\boldsymbol{\xi}} = (\exp[\tilde{\boldsymbol{\mu}}_1], \ldots, \exp[\tilde{\boldsymbol{\mu}}_6])^T / \sum_{j=1}^{6} \exp(\tilde{\boldsymbol{\mu}}_j)$$

$$= (0.087 \quad 0.217 \quad 0.255 \quad 0.262 \quad 0.064 \quad 0.115)^T, \tag{4.2}$$

for the $\boldsymbol{\theta}_i$. An empirical estimate $\tilde{\boldsymbol{C}}$ was also obtained for the prior covariance matrix $\boldsymbol{C}$. This possessed diagonal terms

$$\text{diag}(\tilde{\boldsymbol{C}}) = (1.08 \quad 0.39 \quad 0.25 \quad 0.25 \quad 0.26 \quad 0.46 \quad 0.92), \tag{4.3}$$

with corresponding correlation matrix.

$$\tilde{\boldsymbol{B}} = \begin{pmatrix} 1 & 0.79 & 0.57 & -0.07 & -0.38 & -0.57 \\ 0.79 & 1 & 0.79 & 0.29 & -0.14 & 0.31 \\ 0.57 & 0.79 & 1 & 0.61 & 0.21 & -0.01 \\ -0.07 & 0.29 & 0.61 & 1 & 0.68 & 0.62 \\ -0.38 & -0.14 & 0.21 & 0.68 & 1 & 0.83 \\ -0.57 & -0.31 & -0.01 & 0.62 & 0.83 & 1 \end{pmatrix}. \tag{4.4}$$

The correlation matrix $\tilde{\boldsymbol{B}}$ indicates moderately high correlations for adjacent cells and negative correlations for all cells, more than two cells apart. The consequent smoothed percentages in Table 18.2 are quite complex. They combine empirical Bayes–Stein shrinkages which take into account the common estimate (4.2), across all schools, with histogram-style smoothing of the percentages for each individual school, which refer to the correlation matrix (4.4).

18.5 COMPUTATIONAL TECHNIQUES OF THE 1980s

Zellner and Rossi (1984) made perhaps the key contribution of the decade by pioneering exact Bayesian inferences for the linear logistic model (under either a uniform or multivariate normal prior), using a generalization of Monte Carlo simulation, referred to as 'importance sampling'. Geweke (1988, 1989) describes some precise theorems regarding the convergence of importance sampling. Similar techniques may be used to compute exact posterior inferences for many of the models reviewed in the current chapter. Hsu *et al.* (1991) apply the methodology to general families of discrete distributions. Furthermore, the 'Gibbs sampler' is particularly useful for multinomial-Dirichlet models (see Gelfand and Smith 1990). Duffy and Santner (1988) also provide a Bayesian analysis for the linear logistic model.

Consider the multinomial model, analysed in section 18.3, and the posterior density (3.2), when $\boldsymbol{\mu}$ and $\boldsymbol{C}$ are specified. It is virtually impossible to simulate a $\boldsymbol{\gamma}_i$ vector with density equal to (3.2). However, paralleling Zellner and Rossi, we can simulate a sequence of $\boldsymbol{\gamma}_i$ vectors from a multivariate t-distribution with ν (say $\nu = 20$), degrees of freedom, mean vector $\tilde{\boldsymbol{\gamma}}_i$, and precision matrix $\nu \tilde{\boldsymbol{D}}_i^{-1}(\nu + s)$, where $\tilde{\boldsymbol{\gamma}}_i$ and $\tilde{\boldsymbol{D}}_i^{-1}$ are calculated by the techniques of section 18.3.

Then importance sampling permits the computation of the exact posterior expectation, or posterior probabilities, for any parameter of interest, by reference to a simple reweighting formula, involving the ratio of the exact posterior density of $\boldsymbol{\gamma}_i$ and the above multivariate t-approximation. The degrees of freedom ν may be chosen pragmatically, to speed the convergence. In particular, the posterior expectations of $\boldsymbol{\gamma}_\cdot \boldsymbol{\gamma}_\cdot^T$ and $(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_\cdot)(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_\cdot)^T$ can be exactly computed, permitting the exact calculation of the integrated likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{C}$.

Rather than simulate, it is also possible to approximate closely the marginal posterior density of any parameter of interest by a continuous curve, with saddlepoint accuracy, using Laplacian/conditional maximization techniques (e.g. Leonard 1982; Tierney and Kadane 1986; Leonard, *et al.*, 1989). Extensions to hierarchical Bayesian models are described by Kass and Steffey (1989), who in particular discuss the logistic/multivariate normal prior formulation of section 18.3. Leonard, *et al.* use Laplacian techniques to approximate the posterior density for a general measure of association, proposed by Altham (1970), for an $r \times s$ contingency table. We refer to the general Laplacian/conditional maximization approach as BCM (Bayesian conditional maximization).

Curve (a) of Figure 18.1 is also the BCM approximation to the posterior density of the log measure of association, and this is identical, to within three decimal point accuracy, of the exact curve. The BCM curve (a) for Figure 18.2 is even closer to the histogram (c), representing the exact curve, when compared to the normal approximation (b).

A variety of extensions of Good's Dirichlet prior approach are discussed by Albert (1983, 1985a, b, 1987a, b, 1988, 1990) and Albert and Gupta (1980, 1981, 1982, 1983a, b, c, 1985).

18.6 THREE-WAY TABLES AND SIMPSON'S PARADOX

Consider the data reported by Radelet (1981), and described in Table 18.3. Overall 11.88% of white defendants receive the death penalty, compared with 10.24% of black defendants. However, in cases with white victims these percentages should be replaced by 12.58% and 17.46%, and in the remaining cases, with black victims, they should be replaced by 0% and 5.83%. This is Simpson's paradox (as for example discussed by Lindley and Novick 1981). The conclusion of discrimination against white defendants suggested by the overall table is invalidated by a 'lurking variable' i.e. colour of victim; opposite conclusions are obtained in each of the two subtables.

Table 18.3 Racial characteristics, and imposition of death penalty

	Overall		White victim		Black victim	
	Death	Not death	Death	Not death	Death	Not death
White defendant	19	141	19	132	0	9
Black defendant	17	149	11	52	6	97

Simpson's paradox can be used to refute many analyses of contingency tables based upon non-randomized data. There may always exist a lurking variable which would lead to consistently opposite conclusions, if the contingency table was split according to that variable. However, if the data are appropriately randomized (e.g. if the white defendants and black defendants could have been chosen at random from subpopulations of white defendants and black defendants) then any lurking variable is much less likely to affect the analysis. Lindley and Novick alternatively argue that the problem can be minimized if the individuals can be subjectively regarded as 'exchangeable' members of a population. In any case, any conclusion obtained from a contingency table, based on non-randomized data, can at best be regarded as subjective.

It is difficult to think and interpret data in three dimensions, and in 1971, D. V. Lindley suggested a 'three-directional approach' to the first co-author. The co-authors have implemented this approach while teaching Statistics 421 at the University of Wisconsin, and it provides an alternative to log-linear models for three-way tables (e.g. Nazaret 1987) which we find a bit more difficult to interpret. We argue that the three-directional approach can be used to extract most of the important real-life conclusions from the data. Note that, for the data in Table 18.3, few of the normal

approximations of section 18.1 are valid, owing to the occurrence of a zero cell frequency.

Direction 1 (Split according to 'colour of victim' variable).

Step A1 Investigate inequality of the unconditional cell probabilities of the two subtables in Table 18.3. One way of doing this is to consider the 2×4 table, with first row 19 132 11 52, and second row 0 9 6 97. Then investigate whether the interaction effects are zero, under Goodman's full rank interaction model (1.16). Curves (a), (b), (c) and (d) of Figure 18.3, describe our BCM approximation to the posterior densities of the interaction effects for the (1, 1), (1, 2), (1, 3), and (1, 4) cells of this 2×4 table, under a Jeffreys prior. The BCM approximations are virtually exact. Histograms representing exact, simulated results, are available from the authors. As zero lies in the extreme tail of three of the densities, we

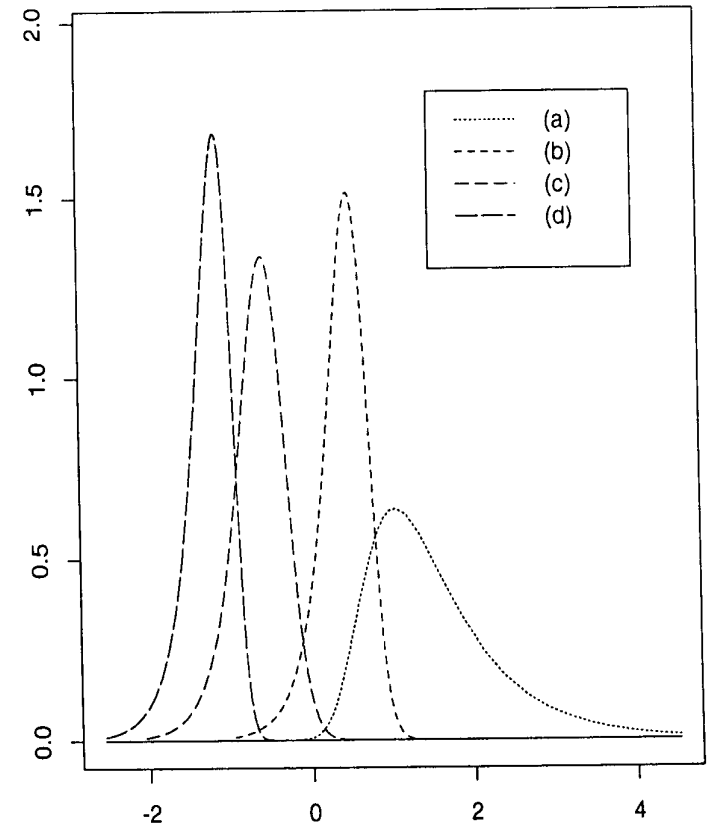

Figure 18.3 Posterior densities of interaction effects (split on colour of victim): (a) (1, 1) cell; (b) (1, 2) cell; (c) (1, 3) cell; (d) (1, 4) cell.

conclude that three of the interaction effects may be non-zero, and therefore that the unconditional cell probabilities in the two subtables are not identical.

Step B1 If step A1 concludes that each of the two subtables is unequal, analyse the two subtables, separately. Otherwise, just analyse the overall table in this direction. Curves (a) and (b) of Figure 18.4 describe the exact posterior densities of the log-measure of associations for our two subtables (white victim and black victim). We conclude that while there is some evidence of a negative association between colour of defendant, and death penalty, in each subtable, this evidence is not particularly significant, as zero does not lie in the tail of either density. Curve (c) describes the posterior density of the log-measure of association for the overall table, and it is not really relevant to consider this curve as the two subtables

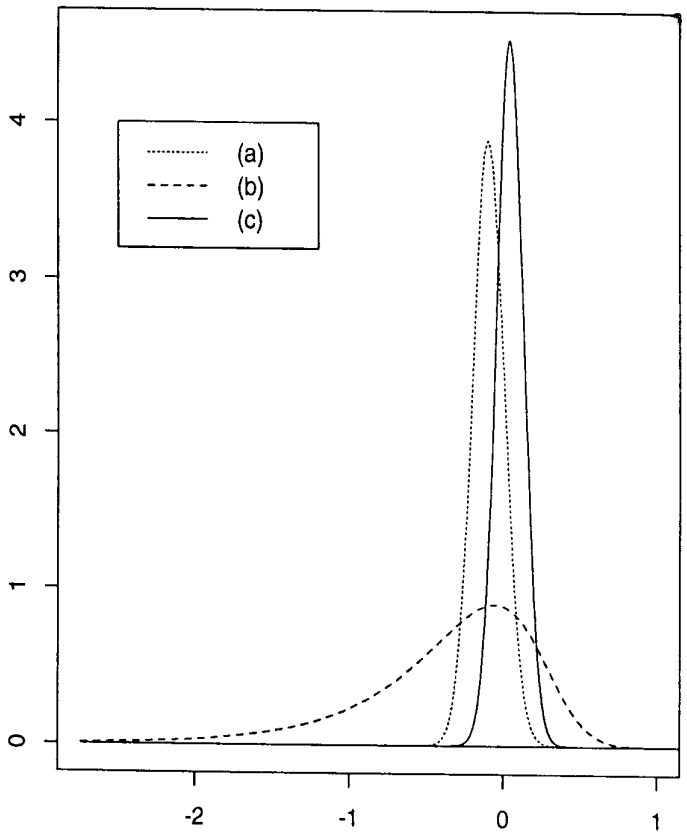

Figure 18.4 Posterior densities of log measures of association (split on colour of victim); (a) White victims; (b) Black victims; (c) overall.

have already been taken to be unequal. However, it does indicate a positive association, thus again highlighting Simpson's paradox.

Direction 2 (Repeat steps of direction 1, but with split according to death penalty variable. Each table or subtable cross-classifies colour of defendant against colour of victim).

Step A2 The locations of the posterior densities of the interaction effects in Figure 18.5 suggest than the data cannot strongly refute equality of the unconditional cell probabilities of the two subtables.

Step B2 Posterior density (c) of the log-measure of association for the overall table suggests that there is a noticeable negative association between colour of victim and colour of defendant. Curves (a) and (b) suggest that the death penalty variable does not affect this conclusion. Note that,

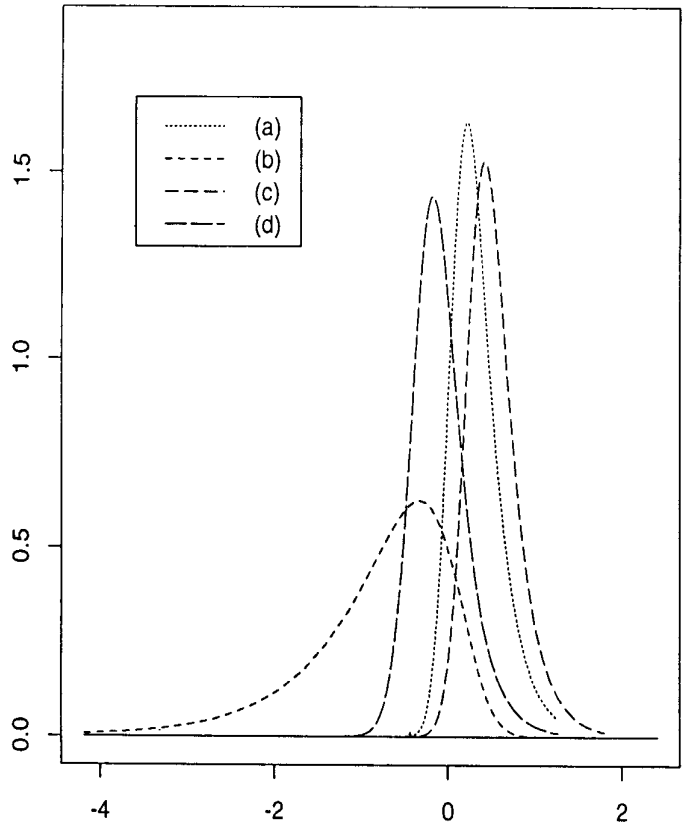

Figure 18.5 Posterior densities of interation effects (split on death penalty): (a) (1, 1) cell; (b) (1, 2) cell; (c) (1, 3) cell; (d) (1, 4) cell.

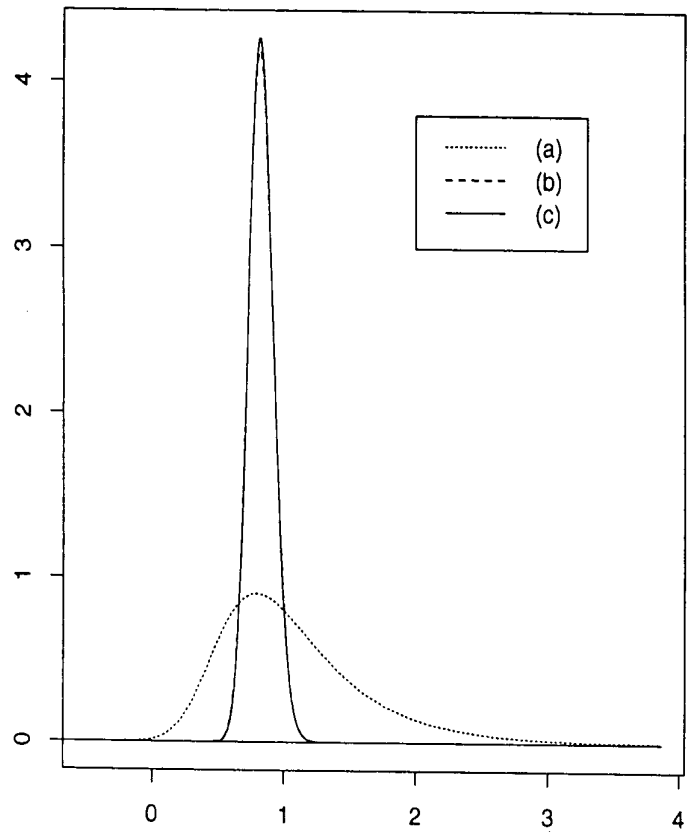

Figure 18.6 Posterior densities of log measures of association (split on death penalty): (a) death penalty; (b) no death penalty; (c) overall.

overall, 94.38% of white defendants have white victims, while 59.43% of black defendants have white victims i.e. black defendants have a stronger propensity to find victims of their own colour.

Direction 3 (Repeat steps of previous directions, but with split according to 'colour of defendant' variable. Each table as subtable cross-classifies colour of victim against death penalty variable).

Step A3 The posterior densities of the four indication effects, in Figure 18.7, indicates substantial differences between the two subtables.

Step B3 The posterior densities of the log-measures of association, in Figure 18.8, indicate that there is a strong positive association between colour of victim, and the death penalty, which is not refuted in either

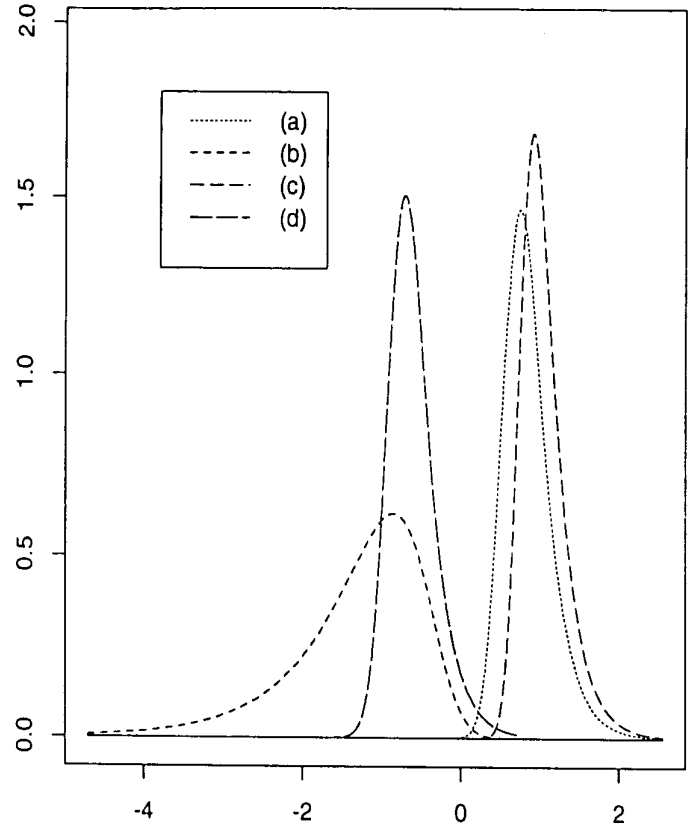

Figure 18.7 Posterior densities if interaction effects (split on colour of defendant): (a) (1, 1) cell; (b) (1, 2) cell; (c) (1, 3) cell; (d) (1, 4) cell.

subtable. Note that, overall, 94.38% of defendants with white victims, receive the death penalty, while only 61.17% of defendants with black victims receive the death penalty.

The key conclusions have been extracted from the data at steps B2 and B3 of our three-directional approach, with the help of virtually exact inferential techniques which are unavailable to non-Bayesians. Note that a fourth variable, e.g. socio-economic status of victim, might change the conclusions based upon the first three variables.

Next consider the Berkeley admissions data, reported by Freedman *et al.* (1978, p. 17). Out of 2691 men applying to the six largest graduate programme at Berkeley in 1973 44.5% were admitted, and out of 1835 women 30.4% were admitted. This apparent sex bias is spurious, as demonstrated by applying our three-directional approach to the $6 \times 2 \times 2$ table (graduate major versus gender of applicants versus acceptance/

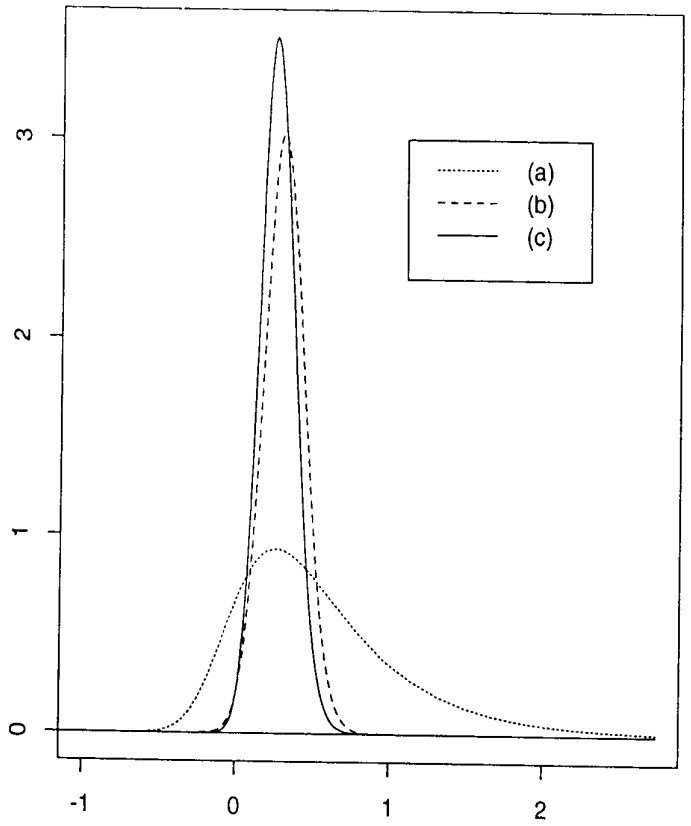

Figure 18.8 Posterior densities of log measures of association (split on colour of defendant); (a) Black defendant; (b) White defendant; (c) overall.

non-acceptance) approach. As all the cell frequencies are greater than 5, we do not report exact posterior densities, but instead refer to B^2, in (1.9) and (1.20), together with associated approximate normal statistics.

Direction 1 (Split according to major, each 2×2 subtable cross-classifies gender of applicant against acceptance/non-acceptance).

Step A1 Equality of the unconditional cell probabilities of the 2×2 tables for the six graduate majors A, B, C, D, E and F was investigated by calculating B^2, in (1.20) and under a Jeffreys prior for the 6×4 table, whose six rows each comprised the four entries for a 2×2 table for one of the graduate majors. This gave $B^2 = 1285.72$, with 15 degrees of freedom, clearly suggesting inequality of the unconditional cell probabilities of the six 2×2 tables, and hence indicating that any conclusion from the overall 2×2 table might be misleading.

Step A2 We analysed each of the 2×2 tables by calculating the approximate normal test statistic $B = \lambda^*/(v^*)^{1/2}$, for each table, where λ^* and v^* are defined in section 18.1. The B statistic for majors A, B, C, D, E and F, was evaluated, under a Jeffreys prior, as -3.97, -0.44, -1.77, -0.55, 1.01 and -0.62. Hence, there is only a significant association for major A, and this is negative (62% males admitted versus 82% females). All other majors give non-significant negative associations, apart from major E which is non-significant and positive. Hence, a version of Simpson's paradox occurs.

Direction 2 (Split according to gender, each 6×2 subtable cross-classifies graduate major against acceptance/non-acceptance).

Step A2 Equality of the unconditional cell probabilities of the 6×2 subtables for men and for women was investigated by calculating B^2 in (1.20) for an approximate 2×12 table, giving $B^2 = 861.35$ with 11 degrees of freedom, and hence refuting equality.

Step B2 The 6×2 table for men applicants, was analysed by calculating the approximate normal test statistics $b_{ij} = \tilde{\lambda}_{ij}/(v_{ij}^{AB})^{1/2}$, where $\tilde{\lambda}_{ij}$ and v_{ij}^{AB} are the approximate mean and variance of the interaction effects, introduced in section 18.1. For the first column of this 6×2 table, the b_{ij} were respectively 15.02, 13.84, -1.83, 0.11, -1.67 and -10.95. This suggests that, among the men, significantly more applicants were accepted for majors A and B, and significantly less were accepted for major F. For the first column of the corresponding 6×2 table for women the b_{ij} were respectively 8.88, 3.37, -1.62, -1.18, -5.56 and -10.77. This suggests that, among the women, significantly more applicants were accepted for majors A and B, and significantly less were accepted for majors E and F.

Direction 3 (Split according to acceptance non-acceptance each 6×2 subtable cross-classifies graduate major against gender).

Step A3 Equality of the unconditional cell probabilities, of the 6×2 subtables, for accepted applicants, and non-accepted applicants, was investigated by calculating B^2 in (1.20), for an appropriate 2×12 table, giving $B^2 = 581.33$, with 11 degrees of freedom, and hence refuting equality.

Step B3 The 6×2 table for accepted applicants, was analysed by calculating the b_{ij}. For the first column of this 6×2 table, the b_{ij} were respectively 9.39, 11.20, -9.23, -4.39, -7.35, -2.75, suggesting that significantly more men were accepted for majors A and B, and significantly more women were accepted for majors C, D, E and F. For the first column of the 6×2 table for non-accepted applicants, the b_{ij} were respectfully 9.35,

7.90, − 11.53, − 6.97, − 11.43 and − 7.68, suggesting that significantly more men were not accepted for majors A and B, and significantly more women were not accepted for majors C, D, E and F.

When performing the three-dimensional approach, it is important to also carefully investigate the raw data in each direction, to facilitate the extraction of all real-life conclusions from the data, and to compare conclusions from each direction. Anyway, it is clear that more men apply to the majors (A and B) with higher admission rates, and that the college perhaps tries to compensate for this by admitting a higher proportion of women to these majors. Majors C, D, E and F have much lower admission rates, but most of the women apply to these four majors.

18.7 FURTHER PROBLEMS WITH NON-RANDOMIZED DATA

Cell frequencies $y_1, \ldots, y_s$ may be taken to possess a multinomial distribution, with cell probabilities $\theta_1, \ldots, \theta_s$, and sample size n, if the n individuals were chosen at random, e.g. without replacement from a much larger population, and if $\theta_1, \ldots, \theta_s$ denote appropriate population proportions. Conversely, without randomization at the experimental design stage, a multinomial assumption may be both incorrect and misleading. However, Bayesian hierarchical models indicate possible ways of generalizing the multinomial sampling distribution which are given under the headings below.

18.7.1 The multinomial–Dirichlet distribution

Following Leonard (1977b, c) and Paul and Plackett (1978) suppose that, conditional on $\theta_1, \ldots, \theta_s$, the frequencies $y_1, \ldots, y_s$ possess a multinomial distribution, with cell probabilities $\theta_1, \ldots, \theta_s$, and sample size n, but where $\theta_1, \ldots, \theta_s$ possess a Dirichlet distribution, with parameters $\alpha\xi_1, \ldots, \alpha\xi_s$. Here $\xi_1 + \xi_2 + \ldots + \xi_s = 1$, and ξ_j is the expectation of the corresponding θ_j. Then the first and second moments of the y_j satisfy

$$E(y_j) = n\xi_j \qquad (j = 1, \ldots, s), \tag{7.1}$$

$$\operatorname{cov}(y_j, y_k) = n\tau^{-1}(\xi_j\delta_{jk} - \xi_j\xi_k) \qquad (j = 1, \ldots, s;\ k = 1, \ldots, s), \tag{7.2}$$

where δ_{jk} denotes the Kronecker delta function, and $\tau = (1 + \alpha)/(n + \alpha)$. The covariance structure (7.2) is similar to the covariance structure for a multinomial distribution, but an over-dispersion factor $\tau = (n + \alpha)/(1 + \alpha)$ is also included. Hence the multinomial–Dirichlet model may provide a better fit to non-randomized data, when this possesses larger dispersion. Leonard and Novick (1986) show that the over-dispersion factor can be combined with a log-linear model for the ξ_j.

Note that, as n gets large, with τ fixed, the sampling distribution of τX^2 approaches a chi-squared distribution, with $s - 1$ degrees of freedom, where

$$X^2 = \sum_j (y_j - n\xi_j)^2/(n\xi)$$

denotes the usual chi-squared statistic. Since $\tau < 1$, goodness-of-fit tests based upon the multinomial–Dirichlet distribution tend to be less likely to yield significant results, when compared with tests based upon the multinomial distribution.

18.7.2 The multinomial logit–multivariate normal distribution

Under the multinomial assumptions in (A), conditional on $\theta_1, \ldots, \theta_s$, consider instead multivariate logits $\gamma_1, \ldots, \gamma_s$ satisfying $\theta_j = \exp(\gamma_j)/\sum\exp(\gamma_s)$, for $j = 1, \ldots, s$; and let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_s)^T$ possess a multivariate normal distribution, with mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{C}$.

This sampling distribution also allows for over-dispersion, and possesses the advantage of permitting complex interdependencies between the frequencies, much more general than permitted by the restrictive covariance structure (7.2). This is in the spirit of other suggestions, permitting serial correlation, recommended by Tavaré and Altham (1983), which lead to alternative adjustments to the distribution of the chi-squared statistic. The methodology, in section 18.3, for the product–multinomial distribution, under a hierarchical prior, effectively provides a method for estimating $\boldsymbol{\mu}$ and $\boldsymbol{C}$, when there are several replications from the above distribution. For example, our analysis in section 18.4, of the London high school data could be regarded as a non-Bayesian analysis under a choice of sampling distribution which generalizes the multinomial, to compensate for the non-randomized nature of the data.

18.7.3 Lurking variables

In the State of Wisconsin the estimation of population gene frequencies, for purposes of HLA blood group testing (e.g. in parcentage and criminal cases) is based upon a non-random sample of 5500 white males attending blood-testing clinics in Milwaukee, plus a non-random sample of about 2000 patients at University of Wisconsin hospitals and clinics. This means that (a) binomial or multinomial assumptions are inappropriate, so that no valid statistical properties or standard errors are available for the estimation procedures, and (b) the conclusions are very much subject to the problem of lurking variables. Similar world-wide problems for the subject of genetics are summarized by Leonard (1991). For DNA testing, a non-random sample of about $n = 7500$ is used to estimate the USA population distribution of the allele lengths, for any particular DNA probe, with similar problems. For both HLA blood group testing, and DNA testing, a serious misapplication of Bayes's theorem is then used to calculate

an alleged probability of guilt. Genetic testing, however, provides excellent real-life examples to illustrate the general point that, when considering categorical data, aspects of experimental design and choice of sampling distribution are often much more important than the precise details of a Bayesian analysis. For the gene frequency example of section 18.2 it is difficult to find a valid Bayesian or non-Bayesian analysis, unless the $n = 106$ individuals in the sample are assumed to be drawn at random from a population. In 1985 the first co-author was headlined in the Madison press as 'Statistics Professor asks "What do we really know?" ' in connection with a similar analysis of the local AIDS–HIV data.

In 1992 the Department of Sociology at the University of Wisconsin-Madison used logistic regression to investigate the gender equity data for men and women at the university. Data for the entire population were analysed, but some important variables, e.g. measuring merit, were not included in the analysis. Nevertheless, the conclusion was reached that 'women faculty are underpaid by 2.8% when compared with men'. Using the above concepts, we advised the Senate that 'any conclusion drawn from these data is at best subjective'. Clearly, the problems of non-randomization (e.g. reducing the apparent significance of any conclusion drawn from classical significance tests) and lurking variables, are overwhelming. This type of refutation of non-randomized data is not new; see, for example Fisher's 1936 treatise regarding Mendel's pea-breeding data.

ACKNOWLEDGEMENTS

The first co-author's contributions to this area would not have been possible without generous help and advice from Patricia Altham, Jack Good, Jim Hickman, Adrian Smith and Dennis Lindley. Both authors are indebted to Kam-Wah Tsui for his advice during the 1980s, and to two John Woods, the first for providing the London High School data, and the second for his advice on contingency tables and genetic testing.

REFERENCES

Albert, J. H. (1983) Bayesian estimation methods for incomplete two-way contingency tables using prior beliefs of association, *ASA Proc. of Survey Rsch. Methods Sect.*, pp. 738–42.

Albert, J. H. (1985a) Bayesian estimation methods for incomplete two-way contingency tables using prior beliefs of association, *Bayes Stat.* **2**, 589–602.

Albert, J. H. (1985b) Simultaneous estimation of Poisson means under exchangeable and independence models, *J. Statist. Computation and Simulation*, **23**, 1–14.

Albert, J. H. (1987a) Empirical Bayes estimation in contingency tables, *Communications in Stat., Part A, Th. and Meth.*, **16**, 2459–85.

Albert, J. H. (1987b) Bayesian estimation of odds ratios under prior hypotheses of independence and exchangeability, *J. Statist. Computation and Simulation*, **27**, 251–68.

Albert, J. H. (1988) Bayesian estimation of Poisson means using a hierarchical log-linear model, *Bayes Stat.*, **3**, 519–31.

Albert, J. H. (1990) A Bayesian test for two-way contingency table using independence priors, *Canadian J. Stat.*, **18**, 347–63.

Albert, J. H. and Gupta, A. K. (1980) Bayesian estimation in 2×2 contingency tables, *ASA Proc. of Social Stat. Sect.*, pp. 461–6.

Albert, J. H. and Gupta, A. K. (1981) Mixtures of Dirichlet distributions and estimation in contingency tables, *ASA Proc. of Social Stat. Sect.*, pp. 189–93.

Albert, J. H. and Gupta, A. K. (1982) Mixtures of Dirichlet distributions and estimation in contingency tables, *Ann. of Statistics*, **10**, 1261–8.

Albert, J. H. and Gupta, A. K. (1983a) Models for reflecting prior beliefs of association in two-way contingency tables, *Communications in Stat., Part A, Th. and Meth.*, **12**, 1241–59.

Albert, J. H. and Gupta, A. K. (1983b) Bayesian estimation methods for 2×2 contingency tables using mixtures of Dirichlet distributions, *J. Amer. Statist. Assn.*, **78**, 708–17.

Albert, J. H. and Gupta, A. K. (1983c) Estimation in contingency tables using prior information, *JRSS-B*, **45**, 60–9.

Albert, J. H. and Gupta, A. K. (1985) Bayesian methods for binomial data with applications to a nonresponse problem, *J. Amer. Statist. Assn.*, **80**, 167–74.

Altham, P. M. E. (1969) Exact Bayesian analysis of a 2×2 contingency table, and Fisher's exact test, *J. Roy. Statist. Soc.*, **B31**, 261–9.

Altham, P. M. E. (1970) The measurement of association of rows and columns of an $r \times s$ contingency table, *J. Roy. Statist. Soc.*, **B32**, 63–73.

Bloch, D. A. and Watson, G. (1967) A Bayesian study of the multinomial distribution, *Ann. Math. Stat.*, **38**, 1423–35.

Crook, J. F. and Good, I. J. (1980) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables: Part II, *Annals of Statistics*, **8**, 1198–1218.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc.*, **B39**, 1–38.

Duffy, D. E. and Santner, T. J. (1988) Estimating logistic regression probabilities, *V*, **1**, 177–94.

Fienberg, S. E. and Holland, P. W. (1973) Simultaneous estimation of multinomial cell probabilites, *J. Amer. Statist. Assoc.*, **68**, 683–9.

Fisher, R. A. (1936) Has Mendel's work been rediscovered? *Ann. Sci.*, **1**, 115–37.

Freedman, D., Pisani, R., Purves, R. and Adhikari, A. (1978) *Statistics*, 2nd edn, Norton.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities, *J. Amer. Statist. Assn.*, **85**, 398–409.

Geweke, J. (1988) Antithetic acceleration of Monte Carlo integration in Bayesian inference. *J. Econometrics*, **38**, 73–90.

Geweke, J. (1989) Bayesian inference in econometric models using Monte Carlo integration, *Econometrica*, **57**, 1311–70.

Good, I. J. (1965) *The Estimation of Probabilities*, MIT Press.

Good, I. J. (1967) A Bayesian significance test for multinomial distributions (with discussion), *J. Roy. Statist. Soc.*, **B29**, 399–431.

Good, I. J. (1975) The Bayes factor against equiprobability of a multinomial distribution using a symmetric Dirichlet prior, *Annals of Statistics*, **3**, 246–50.

Good, I. J. (1976) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, *Annals of Statistics*, **4**, 1159–89.

Goodman, L. A. (1964) Interactions in multidimensional contingency tables, *Ann. of Math. Stat.*, **35**, 632–46.

Goodman, L. A. (1968) The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries, *J. Amer. Statist. Assn.*, **63**, 1091–1131.

Gunel, E. and Dickey, J. M. (1974) Bayes factors for independence in contingency tables, *Biometrika*, **61**, 545–57.

Hickman, J.C. and Miller, R. B. (1977) Notes on Bayesian graduation (with discussion), *Trans. Soc. Actuaries*, **29**, 7–49.

Hsu, J. S. J. (1990) Bayesian inference and marginalization, PhD thesis, University of Wisconsin-Madison.

Hsu, J. S. J., Leonard, T. and Tsui, K. W. (1991) Statistical inference for multiple choice tests, *Psychometrika*, **56**, 327–48.

Johnson, W. E. and Braithwaite, R. B. (1932) Appendix to 'Probability, deductive, and inductive Problems', *Mind*, **41**, 421–3.

Kass, R. E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models), *J. Amer. Statist. Assn.*, **84**, 717–26.

Laird, N. M. (1978) Empirical Bayes methods for two-way contingency tables, *Biometrika*, **65**, 581–90.

Leonard, T. (1972) Bayesian methods for binomial data, *Biometrika*, **59**, 581–9.

Leonard, T. (1973a) A Bayesian method for histograms, *Biometrika*, **60**, 197–308.

Leonard, T. (1973b) Bayesian methods for the simultaneous estimation of several parameters, PhD thesis, University of London.

Leonard, T. (1975) Bayesian estimation methods for two-way contingency tables, *J. Roy. Statist. Soc.*, **B37**, 23–37.

Leonard, T. (1976) Some alternative approaches to multi-parameter estimation, *Biometrika*, **63**, 69–75.

Leonard, T. (1977a) An alternative Bayesian approach to the Bradley–Terry model for paired comparisons, *Biometrics*, 121–30.

Leonard, T. (1977b) Bayesian simultaneous estimation for several multinomial distributions, *Comm. Statist.*, **A6**, 610–30.

Leonard, T. (1977c) A Bayesian approach to some multinomial estimation and pre-testing problems, *J. Amer. Statist. Assn.*, **72**, 865–8.

Leonard, T. (1978) Density estimation, stochastic processes, and prior information (with discussion), *J. Roy. Statist. Soc.*, **B40**, 113–46.

Leonard, T. (1982) Comment on the paper by Lejeune and Faulkenberry, *J. Amer. Statist. Assn.*, **77**, 657–8.

Leonard, T. (1985) Comment on the paper by Diaconis and Efron, *Annals of Statistics*, **13**, 893–8.

Leonard, T. (1991) Commentary on 'Paternity probability; an unnecessary artifact', *J. Undergraduate Mathematics and Its Application*, **12**, 69–72.

Leonard, T., Hsu, J. S. J. and Tsui, K. W. (1989) Bayesian marginal inference, *J. Amer. Statist. Assn.*, **84**, 1051–8.

Leonard, T. and Novick, M. R. (1986) Bayesian full rank marginalization for two-way contingency tables, *J. of Educ. Stat.*, **11**, 33–56.

Lindley, D. V. (1964) The Bayesian analysis of contingency tables, *Am. Math. Statist.*, **35**, 1622–43.

Lindley, D. V. (1965) *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part II: Inference*, CUP.

Lindley, D. V. (1971) The estimation of many parameters, *in Foundations of Statistical Inference*, V. Godamke and D. A. Sprott, eds, Holt, Rinehart and Winston, pp. 435–55.

Lindley, D. V. and Novick, M. R. (1981) The role of exchangeability in inference, *Ann. of Statistics*, **9**, 45–58.

Lindley, D. V., and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion), *J. Roy. Statist. Soc.*, **B34**, 1–41.

Lochner, R. H. (1975) A generalized Dirichlet distribution in life testing, *JRSSB*, **37**, 103–13.

Nazaret, W. A. (1987) Bayesian log linear estimates for three-way contingency tables, *Biometrika*, **74**, 401–10.

Paul, S. R. and Plackett, R. L. (1978) Inference sensitivity for Poisson mixtures, *Biometrika*, **65**, 591–602.

Radelet, M. (1981) Racial characteristics and imposition of the death penalty, *American Sociological Review*, **46**, 918–27.

Ritter, C. (1992). Modern inference in non-linear least squares regression, PhD thesis, University of Wisconsin-Madison.

Tavaŕe, S. and Altham, P. M. E. (1983) Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics, *Biometrika*, **70**, 139–44.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities, *J. Amer. Statist. Assn.*, **81**, 82–6.

Zellner, A. and Rossi, P. E. (1984) Bayesian analysis of dichotomous quantal response models, *J. of Econometrics*, **25**, 365–93.

Department of Statistics
University of Wisconsin-Madison
1210 West Dayton Street
Madison
WI 53706-1693
USA

Department of Statistics and Applied Probability
University of California at Santa Barbara
Santa Barbara
CA 93106-3110
USA

CHAPTER 19

Conflicting Information and a Class of Bivariate Heavy-tailed Distributions

Anthony O'Hagan and **Huiling Le**
University of Nottingham

19.1 HEAVY-TAILED BAYESIAN MODELLING

If x is normally distributed with mean μ and variance 1, and if the prior distribution of μ is $N(0, 1)$, then the posterior distribution is $N(x/2, 1/2)$. Elementary introductions to Bayesian statistics often present this example as typical of how the two sources of information, prior and data, are synthesized by Bayes's theorem. In particular, the posterior mean is a compromise between the prior mean and the data estimate. The compromise seems natural enough, but is questionable when the two sources conflict. If $x = 10$, for instance, the prior which asserts that μ is very unlikely to lie outside $[-3, 3]$ conflicts with the observation which is very unlikely for μ outside $[7, 13]$. The posterior distribution $N(5, 1/2)$ then claims that μ is almost certain to lie in $[3, 7]$, a region which is not supported by either prior or likelihood.

Normal-theory models, and all exponential family models with conjugate priors, have this property of compromising between the various sources of information, even when they conflict. Replacing normal distributions by distributions with heavier tails, such as t-distributions, will cause the posterior distribution to respond very differently to conflict. In the above example, if we replace both the prior distribution and the likelihood by

Aspects of Uncertainty edited by P. R. Freeman and A. F. M. Smith.