Bayesian Inference for Categorical Data Analysis: A Survey

Alan Agresti
Department of Statistics
University of Florida
Gainesville, Florida, USA 32611-8545

Phone USA (352) 392-1941, Fax (352) 392-5175

e-mail aa@stat.ufl.edu

David B. Hitchcock
Department of Statistics
University of South Carolina
Columbia, SC, USA 29208

e-mail hitchcock@stat.sc.edu

Bayesian Inference for Categorical Data Analysis: A Survey

Summary

This article surveys Bayesian methods for categorical data analysis, with primary emphasis on contingency table analysis. Early innovations were proposed by Good (1953, 1956, 1965) for smoothing proportions in contingency tables and by Lindley (1964) for inference about odds ratios. These approaches primarily used conjugate beta and Dirichlet priors. Altham (1969, 1971a) presented Bayesian analogs of small-sample frequentist tests for 2×2 tables using such priors. An alternative approach using normal priors for logits received considerable attention in the 1970s by Leonard and others (e.g., Leonard 1972). Adopted usually in a hierarchical form, the logit-normal approach allows greater flexibility and scope for generalization. The 1970s also saw considerable interest in Stein-influenced Bayesian shrinkage methods and in loglinear modeling. The advent of modern computational methods since the mid-1980s has led to a growing literature on fully Bayesian analyses with models for categorical data, with main emphasis on generalized linear models such as logistic regression for binary and multi-category response variables and hierarchical generalizations.

Key words: Beta distribution; Binomial distribution; Dirichlet distribution; Empirical Bayes; Graphical models; Hierarchical models; Logistic regression; Loglinear models; Markov chain Monte Carlo; Matched pairs; Multinomial distribution; Odds ratio; Smoothing.

1 Introduction

1.1 A brief history up to 1965

The purpose of this article is to survey Bayesian methods for analyzing categorical data. The starting place is the landmark work by Bayes (1763) and by Laplace (1774) on estimating a binomial parameter. They both used a uniform prior distribution for the binomial parameter. Stigler (1986, pp. 100-136) summarized this work, Stigler (1982) discussed what Bayes implied by his use of a uniform prior, and Hald (1998) discussed later developments.

For contingency tables, the sample proportions are ordinary maximum likelihood (ML) estimators of multinomial cell probabilities. When data are sparse, these can have undesirable features. For instance, for a cell with a sampling zero, 0.0 is usually an unappealing estimate. Early applications of Bayesian methods to contingency tables involved smoothing cell counts to improve estimation of cell probabilities with small samples.

Much of this appeared in various works by I. J. Good. Good (1953) used a uniform prior distribution over several categories in estimating the population proportions of animals of various species. Good (1956) used log-normal and gamma priors in estimating *association factors* in contingency tables. For a particular cell, the association factor is defined to be the probability of that cell divided by its probability assuming independence (i.e., the product of the marginal probabilities). This has similar intent as later work on modeling interaction parameters of loglinear models (Leonard 1975). Good's (1965) monograph summarized the use of Bayesian methods for estimating multinomial probabilities in contingency tables, using a Dirichlet prior distribution. Good also was innovative in his early use of hierarchical and empirical Bayesian approaches. His interest in this area apparently evolved out of his service as the main statistical assistant in 1941 to Alan Turing on intelligence issues during World War II (e.g., see Good 1980 and Hodges 1983, pp. 196-197.)

In an influential article, Lindley (1964) focused on estimating summary measures of association in contingency tables. For instance, using a Dirichlet prior (and a related improper prior) distribution for the multinomial probabilities, he found the posterior distribution of contrasts of log probabilities, such as the log odds ratio. Early critics of the Bayesian approach included R. A. Fisher (Interestingly, according to David (1998), Fisher was the first to use the term "Bayesian," starting in 1950.). For instance, in his book *Statistical Methods and Scientific Inference* in 1956, he challenged the use of a uniform prior for the binomial parameter, noting that uniform priors on other scales would lead to different results.

1.2 Outline of this article

Leonard and Hsu (1994) selectively reviewed the growth of Bayesian approaches to categorical data analysis since the groundbreaking work by Good and by Lindley. Much of this review focused on research in the 1970s by Leonard that evolved naturally out of Lindley (1964). They also discussed Stein-influenced Bayesian shrinkage methods of the 1970s, and the more computationally intensive methods that extended Bayesian methods since the mid-1980s. An encyclopedia article by Albert (2004) also provided a brief review of Bayesian methods for contingency tables. He focused on more recent developments, such as model selection issues. Of the many books published in recent years on the Bayesian approach, the most complete coverage of categorical data analysis is the chapter of O'Hagan and Forster (2004) on discrete data models.

The purpose of our article is to provide a somewhat broader overview. We organize our presentation according to the structure of the data. Section 2 begins with estimation of binomial and multinomial parameters, continuing into estimation of cell probabilities in contingency tables and related parameters based on loglinear models (Section 3). Section 4 discusses Bayesian analogs of some classical confidence intervals and significance tests. Section 5 deals with extensions to the regression modeling of categorical response variables. Computational aspects are covered in detail in other sources, and we touch on them only briefly, in Section 6.

2 Estimating Binomial and Multinomial Parameters

2.1 Prior distributions for a binomial parameter

Let y denote a binomial random variable for n trials and parameter π, and let $p = y/n$. The conjugate prior density for π is the beta density, which is proportional to $\pi^{\alpha-1}(1-\pi)^{\beta-1}$ for some choice of parameters $\alpha > 0$ and $\beta > 0$. It has $E(\pi) = \alpha/(\alpha+\beta)$. The posterior density

$h(\pi|y)$ of π is proportional to

$$h(\pi|y) \propto [\pi^y(1-\pi)^{n-y}][\pi^{\alpha-1}(1-\pi)^{\beta-1}] = \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1},$$

for $0 < \pi < 1$ and is also beta. Specifically,

- π has the beta distribution with parameters $\alpha^* = y+\alpha$ and $\beta^* = n-y+\beta$. Equivalently, this is the distribution of

  $$\frac{\left(\frac{y+\alpha}{n-y+\beta}\right)F}{1+\left(\frac{y+\alpha}{n-y+\beta}\right)F}$$

  where F is a F random variable with $df_1 = 2(y+\alpha)$ and $df_2 = 2(n-y+\beta)$.

- $\left(\frac{n-y+\beta}{y+\alpha}\right)\frac{\pi}{1-\pi}$ has the F distribution with $df_1 = 2(y+\alpha)$ and $df_2 = 2(n-y+\beta)$.

The mean of the beta posterior distribution for π is

$$\begin{aligned} E(\pi|y) &= \alpha^*/(\alpha^*+\beta^*) = (y+\alpha)/(n+\alpha+\beta) \\ &= w(y/n) + (1-w)[\alpha/(\alpha+\beta)], \end{aligned}$$

where $w = n/(n+\alpha+\beta)$. This is a weighted average of the sample proportion $p = y/n$ and the mean of the prior distribution. The weight given the sample increases as n increases. The variance of the posterior distribution equals

$$\text{Var}(\pi|y) = \alpha^*\beta^*/(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1),$$

which is approximately $\sqrt{p(1-p)/n}$ for large n.

The ML estimator $p = y/n$ results from the improper prior with $\alpha = \beta = 0$, which corresponds to a uniform prior over the real line for the log odds, $\text{logit}(\pi) = \log[\pi/(1-\pi)]$. Haldane (1948) proposed this, arguing it was reasonable for genetics applications in which one expects $\log(\pi)$ to be roughly uniform for parameter values close to 0 (e.g., according to Haldane, "If we are trying to estimate a mutation rate, ... we might perhaps guess that such a rate would be about as likely to lie between 10^{-5} and 10^{-6} as between 10^{-6} and 10^{-7}.") The posterior distribution in that case is improper if $y = 0$ or n. See Novick (1969) for related arguments supporting this prior, and see the discussion of that paper by W. Perks for a summary of criticisms that he, Jeffreys, and others had about that choice.

The uniform prior distribution is the beta distribution with $\alpha = \beta = 1$. For it, the posterior distribution has the same shape as the binomial likelihood function and has mean

$$E(\pi|y) = (y+1)/(n+2).$$

Suggested by Laplace (1774), this may be the first example of a shrinkage estimate, shrinking the sample proportion toward 1/2. Geisser (1984) advocated the uniform prior in the context of suitability for predictive inference, and discussants of his paper gave arguments for other priors. Other than the uniform, the most popular prior with Bayesians for binomial inference is the Jeffreys prior. This is proportional to the square root of the determinant of the Fisher information matrix for the parameters of interest, for a single observation. In the binomial case, this prior is the beta with $\alpha = \beta = 0.5$.

Bernardo and Ramón (1998) presented an informative survey article about Bernardo's *reference analysis* approach (Bernardo 1979), which optimizes a limiting entropy distance criterion. This attempts to derive non-subjective posterior distributions that satisfy certain natural criteria such as invariance, consistent frequentist performance (e.g., coverage probability of confidence intervals close to the nominal level for large sample sizes), and admissibility. With this approach, a primary purpose is to use a prior distribution such that, even for small sample sizes, the information provided by the data dominates the prior information. The specification of the reference prior is often computationally complex, but for the binomial parameter, it is the Jeffreys prior (Bernardo and Smith 1994, p. 315).

The beta distribution sometimes does not provide sufficient flexibility for a prior. Chen and Novick (1984) introduced a generalized three-parameter beta distribution. Among various properties, it can more flexibly account for heavy tails or skewness, and it reduces to the ordinary beta distribution for certain parameter choices. The resulting posterior distribution is a four-parameter type of beta. Chen and Novick (1984) provided tables as evidence for its usefulness.

An alternative two-parameter approach specifies a normal prior for logit(π), the natural parameter for the exponential family representation of the binomial distribution. Although used occasionally in the 1960s (e.g., Cornfield 1966), this was first strongly promoted by T. Leonard, in work instigated by D. Lindley (e.g., Leonard 1972; see the Leonard and Hsu 1994 discussion on p. 290 about the initiation of this approach). With a $N(0, \sigma^2)$ prior distribution for logit(π), the prior density function for π is the logistic-normal density,

$$f(\pi) = \frac{1}{\sqrt{2(3.14)\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}\Big(\log\frac{\pi}{1-\pi}\Big)^2\Big\} \frac{1}{\pi(1-\pi)}, \quad 0 < \pi < 1.$$

On the probability (π) scale the shape of this density is symmetric, being unimodal when $\sigma^2 \leq 2$ and bimodal when $\sigma^2 > 2$, but always tapering off toward 0 as π approaches 0 or 1. It is mound-shaped for $\sigma = 1$, roughly uniform except near the boundaries when $\sigma \approx 1.5$, and with more pronounced peaks for the modes when $\sigma = 2$. The peaks for the modes get closer to 0 and 1 as σ increases further, and the curve has essentially a U-shaped appearance when $\sigma = 3$ that is similar to that of a beta(.5,.5) prior. For $\sigma = 1$, 2, 3, the standard deviations of these priors are 0.21, 0.31, and 0.37, similar to the values of 0.22, 0.29, and 0.35 for beta priors with $\alpha = \beta = 2.0$, 1, 0.5 (Agresti and Min 2004). The logistic-normal prior with $\mu = 0$ and $\sigma = 2.67$ matches the Jeffreys prior in the first two moments, and the logistic-normal prior with $\mu = 0$ and $\sigma = 1.69$ matches the uniform prior in the first two moments. With the logistic-normal prior, the posterior density function for π is not tractable, as an integral needs to be numerically evaluated for the normalizing constant.

2.2 Bayesian inference about a binomial parameter

Walters (1985) used the uniform prior and its implied posterior distribution in constructing a confidence interval for a binomial parameter π (in Bayesian terminology, a "credible region"), essentially formulating in modern terms Bayes's (1763) approach. He noted how the bounds were contained in the Clopper and Pearson classical 'exact' confidence bounds based on inverting two frequentist one-sided binomial tests (e.g., the lower bound π_L of a 95% Clopper-Pearson interval satisfies $.025 = P_{\pi_L}(Y \geq y)$). Brown, Cai, and Das Gupta (2001, 2002) showed that the posterior distribution generated by the Jeffreys prior yields a confidence interval for π with better frequentist performance. It approximates the narrower adaptation of the Clopper and Pearson 'exact' confidence interval based on inverting two binomial frequentist one-sided tests, when one uses the mid P-value in place of the ordinary P-value.

(The mid P-value is the null probability of more extreme results plus *half* the null probability of the observed result.) See also Leonard and Hsu (1999, pp. 142-144).

For a test of H_0: $\pi \geq \pi_0$ against H_a: $\pi < \pi_0$, a Bayesian P-value is the posterior probability, $P(\pi \geq \pi_0|y)$. Routledge (1994) showed that with the Jeffreys prior and $\pi_0 = 1/2$, this approximately equals the one-sided mid P-value for the frequentist binomial test.

Much of the literature about Bayesian inference for a binomial parameter deals with decision-theoretic results. For estimating a parameter θ using estimator T with loss function $(T-\theta)^2$, the Bayesian estimator is the mean of the posterior distribution. With loss function $w(\theta)(T-\theta)^2$, the Bayes estimator is $E[\theta w(\theta)|y]/E[w(\theta)|y]$ (Ferguson 1967, p. 47). With loss function $(T - \pi)^2/[\pi(1 - \pi)]$ and uniform prior distribution, the Bayes estimator of π is the ML estimator $p = y/n$. For this loss function, the risk function is constant, so the Bayes estimator is minimax; that is, its maximum risk is no greater than the maximum risk for any other estimator. Johnson (1971) showed that the sample proportion is an admissible estimator, for standard loss functions. Brown, Chow and Fong (1992) showed that the ML estimator of the binomial variance is also admissible. Sharma (1975) formed a Bayesian confidence interval using a beta prior based on a decision rule having loss function proportional to the length of the interval when the interval contains the parameter and equal to that loss plus a constant when the interval does not contain the parameter. Jones (1974) considered sequential estimation of a binomial parameter, with a stopping boundary based on a quadratic loss function with constant cost for observations. Rukhin (1988) introduced a loss function that combines the estimation error of a statistical procedure with a measure of its accuracy, an approach that motivates a beta prior with parameter settings between those for the uniform and Jeffreys priors, converging to the uniform as n increases and to the Jeffreys as n decreases.

Diaconis and Freedman (1990) investigated the degree to which posterior distributions put relatively greater mass close to the sample proportion as the sample size n increases. They showed that the posterior odds for an interval of fixed length centered at the sample proportion is bounded below by a term of form ab^n with computable constants $a > 0$ and $b > 1$. They noted that Laplace considered this problem with a uniform prior in 1774. Related work deals with the consistency of Bayesian estimates. Freedman (1963) showed consistency under general conditions for sampling from discrete distributions such as the multinomial. He also showed asymptotic normality of the posterior assuming a local smoothness assumption about the prior. For early work about the asymptotic normality of the posterior distribution for a binomial parameter, see Bernstein (1934, p. 406) and von Mises (1964, Chapter VIII, Section C). (The asymptotic normality is sometimes referred to as the Bernstein – von Mises theorem.)

Draper and Guttman (1971) explored Bayesian estimation of the binomial sample size n based on r independent binomial observations, each with parameters n and π. They considered both π known and unknown. The π unknown case arises in capture-recapture experiments for estimating population size n. One difficulty there is that different models can fit the data well yet yield quite different projections. A later extensive Bayesian literature on the capture-recapture problem includes Smith (1991), George and Robert (1992), Madigan and York (1997), and King and Brooks (2001a, 2002). For a hierarchical model, George and Robert (1992) used Gibbs sampling to simulate the posterior distribution of n. Madigan and York (1997) explicitly accounted for model uncertainty by placing a prior distribution over a discrete set of models as well as over n and the cell probabilities for the table of the capture-recapture observations for the repeated sampling. Fienberg, Johnson and Junker (1999) surveyed other Bayesian and classical approaches to this problem, focusing on ways to permit heterogeneity in catchability among the subjects. Dobra and Fienberg (2001) used a fully Bayesian specification of the Rasch model (discussed in Section 5.1) to estimate the size of the World Wide Web.

Joseph, Wolfson, and Berger (1995) addressed sample size calculations for binomial experiments, using criteria such as attaining a certain expected width of a confidence interval. DasGupta and Zhang (2004) reviewed inference for binomial and multinomial parameters, with emphasis on decision-theoretic results.

2.3 Bayesian estimation of multinomial parameters

Results for the binomial with beta prior distribution generalize to the multinomial with a Dirichlet prior distribution (Lindley 1964, Good 1965). With c categories, suppose cell counts $(n_1, \ldots, n_c)$ have a multinomial distribution with $n = \sum n_i$ and parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)'$.

Let $\{p_i = n_i/n\}$ be the sample proportions. The likelihood is proportional to

$$\prod_{i=1}^{c} \pi_i^{n_i}.$$

The conjugate density is the Dirichlet, expressed in terms of gamma functions as

$$g(\boldsymbol{\pi}) = \frac{\Gamma\left(\sum \alpha_i\right)}{\left[\prod_i \Gamma(\alpha_i)\right]} \prod_{i=1}^{c} \pi_i^{\alpha_i - 1} \quad \text{for } 0 < \pi_i < 1 \text{ all } i, \quad \sum_i \pi_i = 1,$$

where $\{\alpha_i > 0\}$. Let $K = \sum \alpha_j$. The Dirichlet has $E(\pi_i) = \alpha_i/K$ and $\text{Var}(\pi_i) = \alpha_i(K - \alpha_i)/[K^2(K+1)]$. The posterior density is also Dirichlet, with parameters $\{n_i + \alpha_i\}$, so the posterior mean is

$$E(\pi_i | n_1, \ldots, n_c) = (n_i + \alpha_i)/(n + K).$$

Let $\gamma_i = E(\pi_i) = \alpha_i/K$. This Bayesian estimator equals the weighted average

$$[n/(n+K)]p_i + [K/(n+K)]\gamma_i,$$

which is the sample proportion when the prior information corresponds to K trials with α_i outcomes of type i, $i = 1, \ldots, c$.

Good (1965) referred to K as a *flattening constant*, since with identical $\{\alpha_i\}$ this estimate shrinks each sample proportion toward the equi-probability value $\gamma_i = 1/c$. Greater flattening occurs as K increases, for fixed n. Good (1980) attributed first use of $\{\alpha_i = 1\}$ to De Morgan, whose use in 1847 of the estimate $(n_i + 1)/(n + c)$ for π_i extended Laplace's estimate to the multinomial case. Perks (1947) suggested $\{\alpha_i = 1/c\}$, noting the coherence with the Jeffreys prior for the binomial (See also his discussion of Novick 1969). The Jeffreys prior sets all $\alpha_i = 0.5$. Lindley (1964) gave special attention to the improper limiting case with $\{\alpha_i = 0\}$, also considered by Novick (1969). The discussion of Novick (1969) shows the lack of consensus about what 'noninformative' means for a prior for this fundamental case.

Stein's results for estimating multivariate normal means suggest lower total mean squared error may be possible with Bayesian estimators that shrink the sample proportions (Efron and Morris, 1975). However, Bayesian estimators of multinomial parameters are not uniformly better than ML estimators for all possible parameter values. For instance, if a true cell probability equals 0, the sample proportion equals 0 with probability one, so the sample proportion is better than any other estimator. Meeden et al. (1998) showed admissibility for ML estimators of cell probabilities in "decomposable" loglinear models, for which cell probabilities factor into products and ratios of certain marginal probabilities and closed-form expressions exist for the ML estimates. Nonetheless, the shrinkage form of estimator combines good characteristics of sample proportions and model-based estimators. Like sample proportions and unlike model-based estimators, they are consistent even when a particular model (such as equi-probability) does not hold. The weight given the sample proportion increases to 1.0 as the sample size increases. Like model-based estimators and unlike sample proportions, the Bayes estimators smooth the data. The resulting estimators, although slightly biased, usually have smaller total mean squared error than the sample proportions.

Hoadley (1969) examined Bayesian estimation of multinomial probabilities when the population of interest is finite, of known size N. He argued that a finite-population analogue of the Dirichlet prior is a compound multinomial prior, which leads to a translated compound multinomial posterior. Let $\mathbf{N}$ denote a vector of nonnegative integers such that its i-th component N_i is the number of objects (out of N total) that are in category $i, i = 1, \ldots, c$. If conditional on the probabilities and N, the cell counts have a multinomial distribution, and if the multinomial probabilities themselves have a Dirichlet distribution indexed by parameter $\boldsymbol{\alpha}$ such that $\alpha_j > 0$ for all j with $K = \sum \alpha_j$, then unconditionally $\mathbf{N}$ has the compound multinomial mass function,

$$f(\mathbf{N}|N; \boldsymbol{\alpha}) = \frac{N!\,\Gamma(K)}{\Gamma(N+K)} \prod_{j=1}^{c} \frac{\Gamma(N_j + \alpha_j)}{N_j!\Gamma(\alpha_j)}.$$

This serves as a prior distribution for $\mathbf{N}$. Given cell count data $\{n_j\}$ in a sample of size n, the posterior distribution of $\mathbf{N}$ - $\mathbf{n}$ is compound multinomial with N replaced by $N - n$ and $\boldsymbol{\alpha}$ replaced by $\boldsymbol{\alpha}$ + $\mathbf{n}$. Hoadley discussed properties of the compound multinomial, pointing out that the Bayesian approach permits an interval estimator of a function of the cell probabilities (of interest in a telephone line performance study) when no classical confidence interval can be derived. Ericson (1969) gave a general Bayesian treatment of the finite-population problem, including theoretical investigation of the compound multinomial.

The Dirichlet distribution is restricted by having relatively few parameters. For instance, one can specify the means through the choice of $\{\gamma_i\}$ and the variances through the choice of K, but then there is no freedom to alter the correlations. As an alternative to the Dirichlet, Leonard (1973), Aitchison (1985), Goutis (1993), and Forster and Skene

(1994) proposed using a multivariate normal prior distribution for multinomial logits. This induces a multivariate logistic-normal distribution for the multinomial parameters. Specifically, if $\mathbf{X} = (X_1, \ldots, X_c)$ has a multivariate normal distribution, then $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$ with $\pi_i = \exp(X_i)/\sum_{j=1}^{c} \exp(X_j)$ has the logistic-normal distribution. This can provide extra flexibility. For instance, when the categories are ordered and one expects similarity of probabilities in adjacent categories, one might use an autoregressive form for the normal correlation matrix. Leonard (1973) suggested this approach in estimating a histogram.

Here is a summary of other Bayesian literature about the multinomial: Good and Crook (1974) discussed the idea of a Bayes / non-Bayes compromise by using Bayesian methods to generate criteria for frequentist significance testing, illustrating for the test of multinomial equiprobability. An example of such a criterion is the Bayes factor given by the prior odds of the null hypothesis divided by the posterior odds. See Good (1967) for related comments. Dickey (1983) discussed nested families of distributions that generalize the Dirichlet distribution, and argued that they were appropriate for contingency tables. Sedransk, Monahan, and Chiu (1985) considered estimation of multinomial probabilities under the constraint $\pi_1 \leq ... \leq \pi_k \geq \pi_{k+1} \geq ... \geq \pi_c$, using a truncated Dirichlet prior and possibly a prior on k if it is unknown. Under various quadratic loss functions, Nayak and Naik (1989) found the Bayes and empirical Bayesian estimators in sampling from a multinomial or from several multinomials. In their landmark paper on using Gibbs sampling to calculate marginal densities, Gelfand and Smith (1990) discussed a multinomial genetic linkage model that had previously been considered by many authors (e.g., Tanner and Wong 1987). Delampady and Berger (1990) derived lower bounds on Bayes factors in favor of the null hypothesis of a point multinomial probability, and related them to P-values in chi-squared tests. Bernardo and Ramón (1998) illustrated Bernardo's reference analysis approach by applying it to the problem of estimating the ratio π_i/π_j of two multinomial parameters. The posterior distribution of the ratio (which is a beta distribution of the second kind) depends on the counts in those two categories but not on the overall sample size or the counts in other categories, which need not be the case with conventional prior distributions. The posterior distribution of $\pi_i/(\pi_i + \pi_j)$ is the beta with parameters $n_i + 1/2$ and $n_j + 1/2$, the Jeffreys posterior for the binomial parameter.

2.4 Hierarchical Bayesian estimates of multinomial parameters

We mentioned in the previous subsection that Dirichlet priors do not always provide sufficient flexibility. Good (1965, 1967, 1976, 1980) noted this and adopted a hierarchical approach of specifying distributions for the Dirichlet parameters. This approach treats the $\{\alpha_i\}$ in the Dirichlet prior as unknown and specifies a second-stage prior for them. Good (1965) considered the symmetric case with $\{\alpha_i = k\}$. For instance, in one example (p. 38) he assumed that $\log k$ has a symmetrical distribution about 0 such that $P(\gamma < k < 1/\gamma) = 1-\gamma$, which provides the density for k of $1/2$ for $0 < k < 1$ and $1/2k^2$ for $k \geq 1$. He also suggested that one could obtain more flexibility with prior distributions by using a weighted average of Dirichlet distributions. See Albert and Gupta (1982) for later work on hierarchical Dirichlet priors. They let the prior parameters take certain structure natural for a contingency table, as discussed below in Section 3.2.

These approaches gain greater generality at the expense of giving up the simple conjugate Dirichlet form for the posterior. Once one departs from the conjugate case, there are advantages of computation and of ease of more general hierarchical structure by using a multivariate normal prior for logits, as in Leonard's work in the 1970s discussed in Section 3 in particular contexts.

2.5 Empirical Bayesian methods

In his discussion of Good (1980), Novick stated "Bayesian methods are very difficult to employ and sometimes very sensitive because they require the assessment of certain probabilities that are, indeed, very difficult to assess." Having to select a prior distribution is regarded as a stumbling block by many when they first consider the Bayesian approach. The use of an improper prior sometimes results in improper posteriors. Instead of choosing particular parameters for a prior distribution, the empirical Bayesian approach uses the data to determine parameter values for use in the prior distribution. This approach traditionally uses the prior density that maximizes the marginal probability of the observed data, integrating out with respect to the prior distribution of the parameters. Good (1965) referred to this as "Type II maximum likelihood."

Good (1956) may have been the first to use an empirical Bayesian approach with contingency tables, estimating parameters in gamma and log-normal priors for association factors. Good (1965) used it to estimate the parameter value for a symmetric Dirichlet prior for multinomial parameters, the problem for which he also considered the above-mentioned hierarchical approach. Later research on empirical Bayesian estimation of multinomial parameters includes Fienberg and Holland (1973), Leonard (1977b), and Nayak and Naik (1989). Most of the empirical Bayesian literature applies in a context of estimating multiple parameters (such as several binomial parameters), and we will discuss it in such contexts in Section 3.

A disadvantage of the empirical Bayesian approach is not accounting for the source of variability due to substituting estimates for prior parameters. It is increasingly preferred to use the hierarchical approach in which those parameters themselves have a second-stage prior distribution, as mentioned in the previous subsection.

3 Estimating Cell Probabilities in Contingency Tables

Bayesian methods for multinomial parameters apply to cell probabilities for a contingency table. With contingency tables, however, typically one attempts to model the cell probabilities in some way. It often does not make sense to regard the cell probabilities as exchangeable. In addition, in many applications it is more natural to assume independent binomial or multinomial samples rather than a single multinomial over the entire table.

3.1 Estimating several binomial parameters

For several (say r) independent binomial samples, the contingency table has size $r \times 2$. For simplicity, we denote the binomial parameters by $\{\pi_i\}$ (realizing that this is somewhat of an abuse of notation, as we've just used $\{\pi_i\}$ to denote multinomial probabilities).

Much of the early literature on estimating multiple binomial parameters used an empirical Bayesian approach. Griffin and Krutchkoff (1971) assumed an unknown prior on parameters for a sequence of binomial experiments. They expressed the Bayesian estimator in a form that does not explicitly involve the prior but is in terms of marginal probabilities of events involving binomial trials. They substituted ML estimates $\hat{\pi}_1, \ldots, \hat{\pi}_r$ of these marginal probabilities into the expression for the Bayesian estimator to obtain an empirical Bayesian estimator. They tabulated the estimate for a variety of different priors and numerically showed that its risk compares favorably to that of the ML estimate. See Copas (1972) for related work. Albert (1984) considered interval estimation as well as point estimation with the empirical Bayesian approach. Martz and Lian (1974) performed a simulation study that compared the risks of eight empirical Bayesian estimators and the ML estimator.

Much of the literature on multiple binomials has used a hierarchical approach, starting with Leonard (1972). At stage 1, given μ and σ, Leonard assumed that $\{\text{logit}(\pi_i)\}$ are independent from a $N(\mu, \sigma^2)$ distribution. At stage 2, he assumed an improper uniform prior for μ over the real line and assumed an inverse chi-squared prior distribution for σ^2. Specifically, he assumed that $\nu\lambda/\sigma^2$ is independent of μ and has a chi-squared distribution with df $= \nu$, where λ is a prior estimate of σ^2 and ν is a measure of the sureness of the prior conviction. For simplicity, he used a limiting improper uniform prior for $\log(\sigma^2)$. Integrating out μ and σ^2, his two-stage approach corresponds to a multivariate t prior for $\{\text{logit}(\pi_i)\}$. For sample proportions $\{p_j\}$, the posterior mean estimate of $\text{logit}(\pi_i)$ is approximately a weighted average of $\text{logit}(p_i)$ and a weighted average of $\{\text{logit}(p_j)\}$.

Leonard mentioned that his preference for logit-normal priors over beta priors for $\{\pi_i\}$ was partly for computational reasons. With this hierarchical approach, using the beta prior was intractable because of the factorials in the proportionality constant. Teather (1984) extended Leonard (1972) by considering a family of symmetric prior densities for $\{\text{logit}(\pi_i)\}$ that includes the normal and the double exponential. He also used a hierarchical approach with a vague distribution for the prior location parameter and an imprecise but proper conjugate prior for the scale parameter.

Berry and Christensen (1979) took the prior distribution of $\{\pi_i\}$ to be a Dirichlet process prior (Ferguson 1973). For instance, with $r = 2$, one form of this is a measure on the unit square that is a weighted average of a product of two beta densities and a beta density concentrated on the line where $\pi_1 = \pi_2$. The posterior is a mixture of Dirichlet processes. When $r > 2$ or 3, calculations were complex and numerical approximations were given and compared to empirical Bayesian estimators.

Albert and Gupta (1983a) used a hierarchical approach with independent beta$(\alpha, K-\alpha)$ priors on the binomial parameters $\{\pi_i\}$ for which the second-stage prior had discrete uniform form,

$$\pi(\alpha) = 1/(K-1),\ \ \alpha = 1, \ldots, K-1,$$

with K user-specified. In the resulting marginal prior for $\{\pi_i\}$, the size of K determines the extent of correlation among $\{\pi_i\}$. Albert and Gupta (1985) suggested a related hierarchical approach in which α has a noninformative second-stage prior.

Consonni and Veronese (1995) considered examples in which prior information exists about the way various binomial experiments cluster. They assumed exchangeability within certain subsets according to some partition, and allowed for uncertainty about the partition using a prior over several possible partitions. Conditionally on a given partition, beta priors were used for $\{\pi_i\}$, incorporating hyperparameters.

Crowder and Sweeting (1989) considered a sequential binomial experiment in which a trial is performed with success probability $\pi_{(1)}$ and then, if a success is observed, a second-stage trial is undertaken with success probability $\pi_{(2)}$. They showed the resulting likelihood can be factored into two binomial densities, and hence termed it a bivariate binomial. They derived a conjugate prior that has certain symmetry properties and reflects independence of $\pi_{(1)}$ and $\pi_{(2)}$. Polson and Wasserman (1990) employed the reference prior method of Bernardo (1979) for this model. The expression for the prior depends on the parameter of interest, and Polson and Wasserman listed priors for four possibilities: $\pi_{(1)}, \pi_{(2)}, \pi_{(1)}\pi_{(2)}$, and $\pi_{(1)}(1-\pi_{(2)})/(1-\pi_{(1)}\pi_{(2)})$.

Here is a brief summary of other work with multiple binomial parameters: Bratcher and Bland (1975) extended Bayesian decision rules for multiple comparisons of means of normal populations to the problem of ordering several binomial probabilities, using beta priors. Sobel (1993) presented Bayesian and empirical Bayesian methods for ranking binomial parameters, with hyperparameters estimated either to maximize the marginal likelihood or to minimize a posterior risk function. Alvo and Cabilio (1982) considered the allocation of sample size n between two samples in order to minimize expected squared error loss in estimating the difference between two binomial parameters, using beta priors. Springer and Thompson (1966) derived the posterior distribution of the product of several binomial parameters (which has relevance in reliability contexts) based on beta priors. Franck et al. (1988) considered estimating posterior probabilities about the ratio π_2/π_1 for an application in which it was appropriate to truncate beta priors to place support over $\pi_2 \leq \pi_1$. Sivaganesan and Berger (1993) used a nonparametric empirical Bayesian approach assuming that a set of binomial parameters come from a completely unknown prior distribution.

3.2 Estimating multinomial cell probabilities

Next, we consider arbitrary-size contingency tables, under a single multinomial sample. The notation will refer to two-way $r \times c$ tables with cell counts $\mathbf{n} = \{n_{ij}\}$ and probabilities $\boldsymbol{\pi} = \{\pi_{ij}\}$, but the ideas extend to any dimension. A Bayesian approach can compromise between sample proportions and model-based estimators. A Bayesian estimator can shrink the sample proportions $\mathbf{p} = \{p_{ij}\}$ toward a set of proportions satisfying a model.

Fienberg and Holland (1970, 1972, 1973) proposed estimates of $\{\pi_{ij}\}$ using data-dependent priors. For a particular choice of Dirichlet means $\{\gamma_{ij}\}$ for the Bayesian estimator

$$[n/(n+K)]p_{ij} + [K/(n+K)]\gamma_{ij},$$

they showed that the minimum total mean squared error occurs when

$$K = \left(1 - \sum \pi_{ij}^2\right) / \left[\sum (\gamma_{ij} - \pi_{ij})^2\right].$$

The optimal $K = K(\boldsymbol{\gamma}, \boldsymbol{\pi})$ depends on $\boldsymbol{\pi}$, and they used the estimate $K(\boldsymbol{\gamma}, \mathbf{p})$ plugging in the sample proportion $\mathbf{p}$. As $\mathbf{p}$ falls closer to the prior guess $\boldsymbol{\gamma}$, $K(\boldsymbol{\gamma}, \mathbf{p})$ increases and the prior guess receives more weight in the posterior estimate. They selected $\{\gamma_{ij}\}$ based on the fit of a simple model. For two-way tables, they used the independence fit $\{\gamma_{ij} = p_{i+}p_{+j}\}$ for the sample marginal proportions. For extensions and further elaboration, see Chapter 12 of Bishop, Fienberg, and Holland (1975). When the categories are ordered, improved performance usually results from using the fit of an ordinal model, such as one that adds a linear-by-linear association term to the independence model (Agresti and Chuang 1989).

Epstein and Fienberg (1992) suggested two-stage priors on the cell probabilities, first placing a Dirichlet$(K, \boldsymbol{\gamma})$ prior on $\boldsymbol{\pi}$ and using a loglinear parametrization of the prior means $\{\gamma_{ij}\}$. The second stage places a multivariate normal prior distribution on the terms in the loglinear model for $\{\gamma_{ij}\}$. Applying the loglinear parametrization to the prior means $\{\gamma_{ij}\}$ rather than directly to the cell probabilities $\{\pi_{ij}\}$ permits the analysis to reflect uncertainty about the loglinear structure for $\{\pi_{ij}\}$. This was one of the first uses of Gibbs sampling to calculate posterior densities for cell probabilities.

Albert and Gupta wrote several articles in the early 1980s exploring Bayesian estimation for contingency tables. Albert and Gupta (1982) used hierarchical Dirichlet$(K, \boldsymbol{\gamma})$ priors for $\boldsymbol{\pi}$ for which $\{\gamma_{ij}\}$ reflect a prior belief that the probabilities may be either symmetric or independent. The second stage places a noninformative uniform prior on $\boldsymbol{\gamma}$. The precision parameter K reflects the strength of prior belief, with large K indicating strong belief in symmetry or independence. Albert and Gupta (1983a) considered 2×2 tables in which the prior information was stated in terms of either the correlation coefficient ρ between the two variables or the odds ratio $(\pi_{11}\pi_{22}/\pi_{12}\pi_{21})$. They suggested specifying as prior information about the odds ratio a separate 2×2 table of counts having odds ratio equal to the prior guess for its value. The sizes of the counts in this separate table reflect the strength of prior belief. Albert and Gupta (1983b) used a Dirichlet prior on $\{\pi_{ij}\}$, but instead of a second-stage prior, they reparametrized so that the prior is determined entirely by the prior guesses for the odds ratio and K. They showed how to make a prior guess for K by specifying an interval covering the middle 90% of the prior distribution of the odds ratio.

Albert (1987b) discussed derivations of the estimator of form $(1-\lambda)p_{ij} + \lambda\tilde{\pi}_{ij}$, where $\tilde{\pi}_{ij} = p_{i+}p_{+j}$ is the independence estimate and λ is some function of the cell counts. The conjugate Bayesian multinomial estimator of Fienberg and Holland (1973) shown above has such a form, as do estimators of Leonard (1975) and Laird (1978). Albert (1987b) extended Albert and Gupta (1982, 1983b) by suggesting empirical Bayesian estimators that use mixture priors. For cell counts $\mathbf{n} = \{n_{ij}\}$, Albert derived approximate posterior moments

$$E(\pi_{ij}|\mathbf{n}, K) \approx (n_{ij} + Kp_{i+}p_{+j})/(n + K)$$

that have the form $(1-\lambda)p_{ij} + \lambda\tilde{\pi}_{ij}$. He suggested estimating K from the marginal density $m(\mathbf{n}|K)$ and plugging in the estimate to obtain an empirical Bayesian estimate. Alternatively, a hierarchical Bayesian approach places a noninformative prior on K and uses the resulting posterior estimate of K. Albert numerically compared the posterior behavior of his estimators with Laird's (1978) empirical Bayesian estimators, arguing that his approach had superior ease of computation for large tables.

3.3 Estimating loglinear model parameters in two-way tables

The Bayesian approaches presented so far focused directly on estimating probabilities, with prior distributions specified in terms of them. One could instead focus on association parameters. Lindley (1964) did this with $r \times c$ contingency tables, using a Dirichlet prior distribution (and its limiting improper prior) for the multinomial. He showed that contrasts of log cell probabilities, such as the log odds ratio, have an approximate (large-sample) joint normal posterior distribution. This gives Bayesian analogs of the standard frequentist results for two-way contingency tables. Using the same structure as Lindley (1964), Bloch and Watson (1967) provided improved approximations to the posterior distribution and also considered linear combinations of the cell probabilities.

As mentioned previously, a disadvantage of a one-stage Dirichlet prior is that it does not allow for placing structure on the probabilities, such as corresponding to a loglinear model. Leonard (1975), based on his thesis work, considered loglinear models, focusing on parameters of the saturated model

$$\log[E(n_{ij})] = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

using normal priors. Leonard argued that exchangeability within each set of loglinear parameters is more sensible than the exchangeability of multinomial probabilities that one gets with a Dirichlet prior. He assumed that the row effects $\{\lambda_i^X\}$, column effects $\{\lambda_j^Y\}$, and interaction effects $\{\lambda_{ij}^{XY}\}$ were a priori independent. For each of these three sets, given a mean μ and variance σ^2, the first-stage prior takes them to be independent and $N(\mu, \sigma^2)$. As in Leonard's 1972 work for several binomials, at the second stage each normal mean is assumed to have an improper uniform distribution over the real line, and σ^2 is assumed to have an inverse chi-squared distribution. For computational convenience, parameters were estimated by joint posterior modes rather than posterior means. The analysis shrinks the log counts toward the fit of the independence model.

Laird (1978), building on Good (1956) and Leonard (1975), estimated cell probabilities using an empirical Bayesian approach with the loglinear model. Her basic model differs somewhat from Leonard's (1975). She assumed improper uniform priors over the real line for the main effect parameters and independent N(0, σ^2) distributions for the interaction parameters. For computational convenience, as in Leonard (1975) the loglinear parameters were estimated by their posterior modes, and those posterior modes were plugged into the loglinear formula to get cell probability estimates. The empirical Bayesian aspect occurs from replacing σ^2 by the mode of the marginal likelihood, after integrating out the loglinear parameters. This provides yet another way of inducing shrinkage towards the independence model. As $\sigma \to \infty$, the estimates converge to the sample proportions; as $\sigma \to 0$, they converge to the independence estimates, $\{p_{i+}p_{+j}\}$. The fitted values have the same row and column marginal totals as the observed data. She noted that the use of a symmetric Dirichlet prior results in estimates that correspond to adding the same count to each cell, whereas her approach permits considerable variability in the amount added or subtracted from each cell to get the fitted value.

In related work, Jansen and Snijders (1991) considered the independence model and used lognormal or gamma priors for the parameters in the multiplicative form of the model, noting the better computational tractability of the gamma approach. This followed Leonard and Novick (1986), who had used independent gamma priors for Poisson means of cell counts in a two-way table. More generally, Albert (1988) used a hierarchical approach for estimating a loglinear Poisson regression model, assuming a gamma prior for the Poisson means and a noninformative prior on the gamma parameters.

Square contingency tables with the same categories for rows and columns have extra structure that can be recognized through models that are permutation invariant for certain groups of transformations of the cells. Forster (2004b) considered such models and discussed how to construct invariant prior distributions for the model parameters. As mentioned previously, Albert and Gupta (1982) had used a hierarchical Dirichlet approach to smoothing toward a prior belief of symmetry. Vounatsou and Smith (1996) analyzed certain structured contingency tables, including symmetry, quasi-symmetry and quasi-independence models for square tables and for triangular tables that result when the category corresponding to the (i, j) cell is indistinguishable from that of the (j, i) cell (a case also studied by Altham 1975). They assessed goodness of fit using distance measures and by comparing sample predictive distributions of counts to corresponding observed values. The Bradley-Terry model for preference outcomes with pairs of items has a square table representation. Davidson and Solomon (1973) considered a Bayesian approach with a conjugate prior for the preference parameters that can be expressed as an infinite mixture of Dirichlet distributions.

Here is a summary of some other Bayesian work on loglinear-related models for two-way tables. Leighty and Johnson (1990) used a two-stage procedure that first locates full and reduced loglinear models whose parameter vectors enclose the important parameters and then uses posterior regions to identify which ones are important. Evans, Gilula, and Guttman (1993) considered Goodman's generalization of the independence model that has multiplicative row and column effects, called the RC model. Based on independent normal priors (with large variances) for the loglinear parameters for the saturated model, they used a posterior distribution for loglinear parameters to induce a marginal posterior distribution on the RC sub-model. As a by-product, this yields a statistic based on the posterior expected distance between the RC association structure and the general loglinear association structure in order to check how well the RC model fits the data. Evans, Gilula, and Guttman (1989) noted that latent class analysis in two-way tables usually encounters identifiability conditions, which can be overcome with a Bayesian approach putting prior distributions on the latent parameters.

3.4 Extensions to multi-dimensional tables

Nazaret (1987) extended Leonard's approach with loglinear models to three-way contingency tables. He used exchangeable priors of the same form as Leonard's for each set of loglinear effect parameters. He argued, however, against Leonard's approach of taking the degrees of freedom term ν in the inverse chi-squared distribution for a variance arbitrarily close to 0. Marginally, integrating out the variance hyperparameter, he noted that each set of loglinear effect parameters has a multivariate t distribution with df $= \nu - 1$, and one needs $\nu > 1$ to get a positive definite covariance matrix for the effects.

Knuiman and Speed (1988) generalized Leonard's approach further by taking a multivariate normal prior for all parameters collectively rather than univariate normal priors on individual parameters. They noted that this permits separate specification of prior information for different interaction terms, and they applied this to unsaturated models. They computed the posterior mode and used the curvature of the log posterior at the mode to measure precision. King and Brooks (2001b) also specified a multivariate normal prior on the loglinear parameters, which induces a multivariate log-normal prior on the expected cell counts. They derived the parameters of this distribution in an explicit form and stated the corresponding mean and covariances of the cell counts.

For frequentist methods, it is well known that one can analyze a multinomial loglinear model using a corresponding Poisson loglinear model (before conditioning on the sample size), in order to avoid awkward constraints. Following Knuiman and Speed (1988), Forster (2004a) considered corresponding Bayesian results, also using a multivariate normal prior on the model parameters. He adopted prior specification having invariance under certain permutations of cells (e.g., not altering strata). Under such restrictions, he discussed conditions for prior distributions such that marginal inferences are equivalent for Poisson and multinomial models. These essentially allow the parameter governing the overall size of the cell means (which disappears after the conditioning that yields the multinomial model) to have an improper prior. Forster also derived necessary and sufficient conditions for the posterior to then be proper, and he related them to conditions for maximum likelihood estimates to be finite. An advantage of the Poisson parameterization is that MCMC methods are typically more straightforward to apply than with multinomial models.

Loglinear model selection, particularly using Bayes factors, now has a substantial literature. Spiegelhalter and Smith (1982) gave an approximate expression for the Bayes factor for a multinomial loglinear model with an improper prior (uniform for the log probabilities) and showed how it related to the standard chi-squared goodness-of-fit statistic. Raftery (1986) noted that this approximation is indeterminate if any cell is empty but is valid with a Jeffreys prior. He also noted that, with large samples, -2 times the log of this approximate Bayes factor is approximately equivalent to Schwarz's BIC model selection criterion. More generally, Raftery (1996) used the Laplace approximation to integration to obtain approximate Bayes factors for generalized linear models. Madigan and Raftery (1994) proposed a strategy for loglinear model selection with Bayes factors that takes account of the true model uncertainty by averaging over a relatively small set of models. See also Raftery (1996) and Dellaportas and Forster (1999) for related work. Adapting the George and McCulloch (1993) stochastic search variable selection method, Ntzoufras, Forster and Dellaportas (2000) developed a MCMC algorithm for loglinear model selection. Albert (1996) suggested partitioning the loglinear model parameters into subsets and testing whether specific subsets are nonzero. Using normal priors for the parameters, he examined the behavior of the Bayes factor under both normal and Cauchy priors, finding that the Cauchy was more robust to misspecified prior beliefs.

An interesting recent application of Bayesian loglinear modeling is to issues of confidentiality (Fienberg and Makov 1998). Agencies often release multidimensional contingency tables that are ostensibly confidential, but the confidentiality can be broken if an individual is uniquely identifiable from the data presentation. Fienberg and Makov considered loglinear modeling of such data, accounting for model uncertainty via Bayesian model averaging. They generated "population tables" from draws from the distribution of the sample data under the presumed model. They then estimated probabilities of having "population-unique" individuals, given various sample sizes, to determine the proper fraction of the population that the sample size should reflect in order not to compromise confidentiality.

Considerable literature has dealt with analyzing a set of 2×2 contingency tables, such as often occur in meta analyses or multi-center clinical trials comparing two treatments on a binary response. Maritz (1989) derived empirical Bayesian estimators for the log-odds ratios, based on a Dirichlet prior for the cell probabilities and estimating the hyperparameters using data from the other tables. See Albert (1987a) for related work. Wypij and Santner (1992) considered the model of a common odds ratio and used Bayesian and empirical Bayesian arguments to motivate an estimator that corresponds to a conditional ML estimator after adding a certain number of pseudotables that have a concordant or discordant pair of observations. Skene and Wakefield (1990) modeled multi-center studies using a model that allows the treatment–response log odds ratio to vary among centers. Meng and Dempster (1987) considered a similar model, using normal priors for main effect and interaction parameters in a logit model, in the context of dealing with the multiplicity problem in hypothesis testing with many 2×2 tables. Warn, Thompson, and Spiegelhalter (2002) considered meta analyses for the difference and the ratio of proportions. This relates essentially to identity and log link analogs of the logit model, in which case it is necessary to truncate normal prior distributions so the distributions apply to the appropriate set of values for these measures. Efron (1996) outlined empirical Bayesian methods for estimating parameters corresponding to many related populations, exemplified by odds ratios from 41 different trials of a surgical treatment for ulcers. His method permits selection from a wide class of priors in the exponential family. Casella (2001) analyzed data from Efron's meta-analysis, estimating the hyperparameters as in an empirical Bayes analysis but using Gibbs sampling to approximate the posterior of the hyperparameters, thereby gaining insight into the variability of the hyperparameter terms. Casella and Moreno (2003) gave another approach to the meta-analysis of contingency tables, employing intrinsic priors. Wakefield (2004) discussed the sensitivity of various hierarchical approaches for ecological inference, which involves making inferences about the associations in the separate 2×2 tables when one observes only the marginal distributions.

3.5 Graphical models

Much attention has been paid in recent years to graphical models. These have certain conditional independence structure that is easily summarized by a graph with vertices for the variables and edges between vertices to represent a conditional association. For contingency tables the graphical models correspond to decomposable loglinear models. The cell probabilities can be expressed in terms of marginal and conditional probabilities, and independent Dirichlet prior distributions for them induce independent Dirichlet posterior distributions. See O'Hagan and Forster (2004, Chap. 12) for discussion of the usefulness of graphical representations for a variety of Bayesian analyses.

Dawid and Lauritzen (1993) introduced the notion of a probability distribution defined over probability measures on a multivariate space that concentrate on a set of such graphs. A special case includes a *hyper Dirichlet* distribution that is conjugate for multinomial sampling and that implies that certain marginal probabilities have a Dirichlet distribution. Madigan and Raftery (1994) and Madigan and York (1995) used this family for graphical model comparison and for constructing posterior distributions for measures of interest by averaging over relevant models. Giudici (1998) used a prior distribution over a space of graphical models to smooth cell counts in sparse contingency tables, comparing his approach with the simple one based on a Dirichlet prior for multinomial probabilities.

3.6 Dealing with nonresponse

Several authors have considered Bayesian approaches in the presence of nonresponse. Using the terminology of Little and Rubin (2002), nonresponse is called *ignorable* if the indicator of being a nonrespondent is independent of the unobserved response or *nonignorable* if the nonresponse depends on the unobserved response. Modeling nonignorable nonresponse has mainly taken one of two approaches: Introducing parameters that control the extent of nonignorability into the model for the observed data and checking the sensitivity to these parameters, or modeling of the joint distribution of the data and the response indicator. Forster and Smith (1998) reviewed these approaches and cited relevant literature.

Basu and Pereira (1982) suggested a multinomial model, with nonresponse being another category in the contingency table, and used a conjugate Dirichlet prior on the multinomial parameters. They also considered a multivariate hypergeometric model, suitable for finite-population data. Dickey, Jiang, and Kadane (1987) used a generalization of the Dirichlet family proposed by Dickey (1983) to develop a conjugate Bayesian analysis that allows for nonresponse. Park (1998), extending Park and Brown (1994), smoothed the observed counts by placing a conjugate Dirichlet prior on the conditional probabilities of the possible outcomes, given the response indicator.

Forster and Smith (1998) considered models having categorical response and categorical covariate vector, when some response values are missing. They investigated a Bayesian method for selecting between nonignorable and ignorable nonresponse models, pointing out that the limited amount of information available makes standard model comparison methods inappropriate. Other works dealing with missing data for categorical responses include Albert and Gupta (1985), Kadane (1985), Paulino and Pereira (1995), Cowles, Carlin and Connett (1996), Bradlow and Zaslavsky (1999), Lee et al. (2001), and Soares and Paulino (2001). Viana (1994) and Prescott and Garthwaite (2002) studied misclassified multinomial and binary data, respectively, with applications to misclassified case-control data.

4 Tests and Confidence Intervals in Two-Way Tables

We next consider Bayesian analogs of frequentist significance tests and confidence intervals for contingency tables. For 2×2 tables, with multinomial Dirichlet priors or binomial beta priors there are connections between Bayesian and frequentist results.

4.1 Confidence intervals for association parameters

For 2×2 tables resulting from two independent binomial samples with parameters π_1 and π_2, the measures of usual interest are $\pi_1 - \pi_2$, the relative risk π_1/π_2, and the odds ratio $[\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$. It is most common to use a beta(α_i, β_i) prior for π_i, $i = 1, 2$, taking them to be independent. Alternatively, one could use a correlated prior. An obvious possibility is the bivariate normal for [logit(π_1), logit(π_2)]. Howard (1998) instead amended the independent beta priors and used prior density function proportional to

$$e^{-(1/2)u^2}\pi_1^{a-1}(1-\pi_1)^{b-1}\pi_2^{c-1}(1-\pi_2)^{d-1},$$

where

$$u = \frac{1}{\sigma}\log\left(\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}\right).$$

Howard suggested $\sigma = 1$ for a standard form. This has a relatively strong dependence. For the amended Jeffreys priors ($a = b = c = d = 0.5$) the correlation between the binomial parameters is 0.84, 0.59, and 0.41 when $\sigma = 1$, 2, and 3 (Agresti and Min 2004).

Priors for π_1 and π_2 induce corresponding priors for the measures themselves. For instance, with uniform priors, $\pi_1 - \pi_2$ has a symmetrical triangular density over (-1, +1), $r = \pi_1/\pi_2$ has density $g(r) = 1/2$ for $0 \leq r \leq 1$ and $g(r) = (1/2)r^2$ for $r > 1$, and the log odds ratio has the Laplace density (Nurminen and Mutanen 1987). The posterior distribution for (π_1, π_2) induces posterior distributions for the measures. For the independent beta priors, Hashemi, Nandram and Goldberg (1997) and Nurminen and Mutanen (1987) gave integral expressions for the posterior distributions for the difference, ratio, and odds ratio.

Hashemi et al. (1997) formed Bayesian highest posterior density (HPD) confidence intervals for these three measures. With the HPD approach, the posterior probability equals the desired confidence level and the posterior density is higher for every value inside the interval than for every value outside of it. The HPD interval lacks invariance under parameter transformation. This is a serious liability for the odds ratio and relative risk, unless the HPD interval is computed on the log scale. For instance, if (L, U) is a $100(1-\alpha)\%$ HPD interval using the posterior distribution of the odds ratio, then the $100(1-\alpha)\%$ HPD interval using the posterior distribution of the inverse of the odds ratio (which is relevant if we reverse the identification of the two groups being compared) is not (1/U, 1/L). The "tail method" $100(1-\alpha)\%$ interval consists of values between the $\alpha/2$ and $(1-\alpha/2)$ quantiles. Although longer than the HPD interval, it is invariant.

Agresti and Min (2004) discussed Bayesian confidence intervals for association parameters in 2×2 tables. They argued that if one desires good coverage performance (in the frequentist sense) over the entire parameter space, it is best to use quite diffuse priors. Even uniform priors are often too informative, and they recommended the Jeffreys prior.

4.2 Tests comparing two independent binomial samples

Using independent beta priors, Novick and Grizzle (1965) focused on finding the posterior probability that $\pi_1 > \pi_2$ and discussed application to sequential clinical trials. Using diffuse, conjugate "indifference" priors, they also considered multinomial and Poisson sampling and argued that the sequential design is more flexibly addressed by a Bayesian analysis than by a frequentist analysis.

Cornfield (1966) also examined sequential trials from a Bayesian viewpoint, focusing on stopping-rule theory. He used prior densities of a type suggested by Jeffreys (1948) for hypothesis testing, which concentrate some nonzero probability at the null hypothesis point. His test assumed normal priors for $\mu_i = \text{logit}(\pi_i), i = 1, 2$, putting a nonzero prior probability λ on the null $\mu_1 = \mu_2$. This gives the joint prior density for μ_1 and μ_2:

$$(1-\lambda)\frac{1}{\sigma^2}\phi(\mu_1/\sigma)\phi(\mu_2/\sigma),\ \mu_1 \neq \mu_2;$$

$$\frac{\lambda}{\sigma}\phi(\mu_1/\sigma),\ \mu_1 = \mu_2$$

where $\phi(\cdot)$ is the standard normal density. From this, Cornfield derived the posterior probability that $\mu_1 = \mu_2$ and showed connections with stopping-rule theory. For other work on sequential comparison of binomial proportions, see Jones and Madhi (1986).

Altham (1969) discussed Bayesian testing for 2×2 tables in a multinomial context. She treated the cell probabilities $\{\pi_{ij}\}$ as multinomial parameters having a Dirichlet prior with parameters $\{\alpha_{ij}\}$. For testing H$_0$: $\theta = \pi_{11}\pi_{22}/\pi_{12}\pi_{21} \leq 1$ against H$_a$: $\theta > 1$ with cell counts $\{n_{ij}\}$ and the posterior Dirichlet distribution with parameters $\{\alpha'_{ij} = \alpha_{ij} + n_{ij}\}$, she showed that

$$P(\theta \leq 1 | \{n_{ij}\}) = \sum_{s=\max(\alpha'_{21}-\alpha'_{12},0)}^{\alpha'_{21}-1} \binom{\alpha'_{+1}-1}{s} \binom{\alpha'_{+2}-1}{\alpha'_{2+}-1-s} \Big/ \binom{\alpha'_{++}-2}{\alpha'_{1+}-1}.$$

This posterior probability equals the one-sided P-value for Fisher's exact test, when one uses the improper prior hyperparameters $\alpha_{11} = \alpha_{22} = 0$ and $\alpha_{12} = \alpha_{21} = 1$, which correspond to a prior belief favoring the null hypothesis. That is, the ordinary P-value for Fisher's exact test corresponds to a Bayesian P-value with a conservative prior distribution, which some have taken to reflect the conservative nature of Fisher's exact test. If $\alpha_{ij} = \gamma, i, j = 1, 2$, with $0 \leq \gamma \leq 1$, Altham showed that the Bayesian P-value is smaller than the Fisher P-value. The difference between the two is no greater than the null probability of the observed data.

Altham's results extend to comparing independent binomials with corresponding beta priors. In that case, see Irony and Pereira (1986) for related work comparing Fisher's exact test with a Bayesian test. Howard (1998) showed that with Jeffreys priors the posterior probability that $\pi_1 \leq \pi_2$ approximates the one-sided P-value for the large-sample z test using pooled variance (i.e., the signed square root of the Pearson statistic) for testing $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 > \pi_2$. Altham (1971a) considered a Bayesian test of independence for the special case of a 2×2 table in which the two off-main-diagonal cells are indistinguishable.

Little (1989) argued that if one believes in conditioning on approximate ancillary statistics, then the conditional approach leads naturally to the likelihood principle and to a Bayesian analysis such as Altham's. Zelen and Parker (1986) considered Bayesian analyses for 2×2 tables that result from case-control studies. They argued that the Bayesian approach is well suited for this, since such studies do not represent randomized experiments or random samples from a real or hypothetical population of possible experiments. Later Bayesian work on case-control studies includes Ghosh and Chen (2002), Müller and Roeder (1997), Seaman and Richardson (2001, 2004), and Sinha, Mukherjee, and Ghosh (2004). For instance, Seaman and Richardson (2004) extend to Bayesian methods the equivalence between prospective and retrospective models in case-control studies. See Berry (2004) for a recent exposition of advantages of using a Bayesian approach in clinical trials.

Weisberg (1972) extended Novick and Grizzle (1965) and Altham (1969) to the comparison of two multinomial distributions with ordered categories. Assuming independent Dirichlet priors, he obtained an expression for the posterior probability that one distribution is stochastically larger than the other. In the binary case, he also gave the posterior distribution of the relative risk.

Kass and Vaidyanathan (1992) studied sensitivity of Bayes factors to small changes in prior distributions. Under a certain null orthogonality of the parameter of interest and the nuisance parameter, and with the two parameters being independent a priori, they showed that small alterations in the prior for the nuisance parameter have no effect on the Bayes factor up to order n^{-1}. They illustrated this for testing equality of binomial parameters. With equal sample sizes in a test that the log odds ratio equals 0, the null orthogonal nuisance parameter is the average of the logits of the two binomial parameters. Numerical evaluations were given for normal and Cauchy priors for the nuisance parameter.

Walley, Gurrin, and Burton (1996) suggested using a large class of prior distributions to generate upper and lower probabilities for testing a hypothesis. These are obtained by maximizing and minimizing the probability with respect to the density functions in that class. They applied their approach to clinical trials data for deciding which of two therapies is better. See also Walley (1996) for discussion of a related "imprecise Dirichlet model" for multinomial data.

Brooks (1987) used a Bayesian approach for the design problem of choosing the ratio of sample sizes for comparing two binomial proportions. Matthews (1999) also considered design issues in the context of two-sample comparisons. In that simple setting, he was able to present closed forms of the optimal Bayesian design for the estimation of the log odds ratio, and he also studied the effect of the specification of the prior distributions.

4.3 Testing independence in two-way tables

Gunel and Dickey (1974) considered independence in two-way contingency tables under the Poisson, multinomial, independent multinomial, and hypergeometric sampling models.

Conjugate gamma priors for the Poisson model induce priors in each further conditioned model. They showed that the Bayes factor for independence itself factorizes, highlighting the evidence residing in the marginal totals.

Good (1976) also examined tests of independence in two-way tables based on the Bayes factor, as did Jeffreys for 2×2 tables in later editions of his book. As in some of his earlier work, for a prior distribution Good used a mixture of symmetric Dirichlet distributions. Crook and Good (1980) developed a quantitative measure of the amount of evidence about independence provided by the marginal totals and discussed conditions under which this is small. See also Crook and Good (1982) and Good and Crook (1987).

Albert (1990) gave an approximation for a high-dimensional integral necessary in the calculation of the Bayes factor, improving on an approximation of Leonard (1977a). Albert compared numerically his method with previous Bayes factor approaches of Gunel and Dickey (1974) and Good and Crook (1987). Albert (1997) generalized Bayesian methods for testing independence and estimating odds ratios to other settings, extending Albert (1996). He used a prior distribution for the loglinear association parameters that reflects a belief that only part of the table reflects independence (a "quasi-independence" prior model) or that there are a few "deviant cells," without knowing where these outlying cells are in the table. Quintana (1998) proposed a nonparametric Bayesian analysis for developing a Bayes factor to assess homogeneity of several multinomial distributions, using Dirichlet process priors. The model has the flexibility of assuming no specific form for the distribution of the multinomial probabilities.

Intrinsic priors, introduced for model selection and hypothesis testing by Berger and Pericchi (1996), allow a conversion of an improper noninformative prior into a proper one. For testing independence in contingency tables, Casella and Moreno (2002), noting that many common noninformative priors cannot be centered at the null hypothesis, suggested the use of intrinsic priors. Typically in hypothesis tests concerning, say, π, the intrinsic prior for π is given conditional on the null value π_0:

$$g^I(\pi|\pi_0) = g(\pi)E_\pi[m_0(\mathbf{y})/m_1(\mathbf{y})]$$

where $g(\pi)$ is a default prior for π, and $m_i(\mathbf{y}), i = 0, 1$ represent the marginals under the null and the total parameter space, respectively. Casella and Moreno (2002) derived intrinsic priors for testing independence in two-way tables under several sampling models.

4.4 Comparing two matched binomial samples

There is a substantial literature on comparing binomial parameters with independent samples, but the dependent-samples case has attracted less attention. Altham (1971b) seems to have been the first to develop Bayesian analyses for matched-pairs data with a binary response. Consider the simple model in which the probability π_{ij} of response i for the first observation and j for the second observation is the same for each subject. Using the Dirichlet($\{\alpha_{ij}\}$) prior and letting $\{\alpha'_{ij} = \alpha_{ij} + n_{ij}\}$ denote the parameters of the Dirichlet posterior, she showed that the posterior probability of a higher probability of success for the first observation is

$$P[\pi_{12}/(\pi_{12} + \pi_{21}) > 1/2 | \{n_{ij}\}] = \sum_{s=0}^{\alpha'_{12}-1} \binom{\alpha'_{12} + \alpha'_{21} - 1}{s} (\frac{1}{2})^{\alpha'_{12}+\alpha'_{21}-1}.$$

This equals the frequentist one-sided P-value using the binomial distribution when the prior parameters are $\alpha_{12} = 1$ and $\alpha_{21} = 0$. As in the independent samples case studied by Altham (1969), this is a Bayesian P-value for a prior distribution favoring H_0. If $\alpha_{12} = \alpha_{21} = \gamma$, with $0 \le \gamma \le 1$, Altham showed that this is smaller than the frequentist P-value, and the difference between the two is no greater than the null probability of the observed data.

Altham (1971b) also considered the logit model in which the probability varies by subject but the within-pair effect is constant. She showed that the Bayesian evidence against the null is weaker as the number of pairs $(n_{11} + n_{22})$ giving the same response at both occasions increases, for fixed values of the numbers of pairs giving different responses at the two occasions. This differs from the analysis in the previous paragraph and the corresponding conditional likelihood result for this model, which do not depend on such "concordant" pairs. Ghosh et al. (2000a) showed related results.

Antelman (1972) considered applications in which binary matched pairs are supplemented by two sets of observations observed only marginally. Focusing on finding a posterior probability that one success probability exceeds the other, he argued that Dirichlet priors were insufficiently rich and did not recognize the structural aspect of the 2×2 table. He used a Dirichlet-beta prior that combined a Dirichlet form for the joint probabilities and a beta form for the marginal probabilities.

Altham (1971b) also considered logit models for cross-over designs with two treatments, adding two strata for the possible orders. She showed approximate correspondences with classical inferences in the case of great prior uncertainty. For cross-over designs, Forster (1994) used a multivariate normal prior for a loglinear model. One can incorporate prior beliefs about the existence of a carry-over effect, and check the posterior sensitivity to such assumptions. For obtaining the posterior, the non-conjugacy was handled by Gibbs sampling. This has the facility to deal easily with cases in which the data are incomplete, such as when subjects are observed only for the first period.

5 Regression Models for Categorical Responses

5.1 Binary regression

Bayesian approaches to estimating binary regression models took a sizable step forward with Zellner and Rossi (1984). They examined the binary generalized linear models (GLMs) $h[E(y_i)] = \mathbf{x}_i'\boldsymbol{\beta}$, where $\{y_i\}$ are independent binary random variables, $\mathbf{x}_i$ is a vector of covariates for y_i, and $h(\cdot)$ is a link function such as the probit or logit. They derived approximate posterior densities both for an improper uniform prior on $\boldsymbol{\beta}$ and for a general class of informative priors, giving particular attention to the multivariate normal. Their approach is discussed further in Section 6.

Ibrahim and Laud (1991) considered the Jeffreys prior for $\boldsymbol{\beta}$ in a GLM, giving special attention to its use with logistic regression. They showed that it is a proper prior, and all joint moments are finite, as is also true for the posterior distribution. See also Poirier (1994).

Wong and Mason (1985) extended logistic regression modeling to a multilevel form of model. They assumed that "micro observations" were embedded in various "macro observations". They modeled the micro observations with a logistic regression model whose coefficients could vary across macro observations, but with the same regressors in each. For micro observation i in macro observation j, the binary observation y_{ij} has mean π_{ij} satisfying

$$\log[\pi_{ij}/(1-\pi_{ij})] = \mathbf{x}_{ij}'\boldsymbol{\beta}_j.$$

The logit model parameters are assumed to satisfy

$$\boldsymbol{\beta}_j = \mathbf{G}_j\boldsymbol{\eta} + \boldsymbol{\alpha}_j, \text{ with } \boldsymbol{\alpha}_j \sim N(\mathbf{0}, \boldsymbol{\Gamma}),$$

where $\mathbf{G}_j$ has measurements on macro variable j, $\boldsymbol{\eta}$ are macro coefficients and $\boldsymbol{\Gamma}$ is the macro covariance matrix. Assuming a diffuse multivariate normal prior for $\boldsymbol{\eta}$, they used empirical Bayesian methodology to estimate $\boldsymbol{\Gamma}$ with the EM algorithm, and used it in obtaining posterior estimates of $\boldsymbol{\eta}$ and $\{\boldsymbol{\beta}_j\}$. Daniels and Gatsonis (1999) used such modeling to analyze geographic and temporal trends with clustered longitudinal binary data.

Although these days logistic regression is more popular than probit regression, for Bayesian inference the probit case has computational simplicities due to connections with an underlying normal regression model. Albert and Chib (1993) studied probit regression modeling, with extensions to ordered multinomial responses. They assumed the presence of normal latent variables Z_i (such that the corresponding binary $y_i = 1$ if $Z_i > 0$ and $y_i = 0$ if $Z_i \leq 0$) which, given the binary data, followed a truncated normal distribution. The normal assumption for $\mathbf{Z} = (Z_1, \ldots, Z_n)$ allowed Albert and Chib to use a hierarchical prior structure similar to that of Lindley and Smith (1972). If the parameter vector $\boldsymbol{\beta}$ of the linear predictor has dimension k, one can model $\boldsymbol{\beta}$ as lying on a linear subspace $\mathbf{A}\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ has dimension $p < k$. This leads to the hierarchical prior

$$\mathbf{Z} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}), \;\; \boldsymbol{\beta} \sim N(\mathbf{A}\boldsymbol{\beta}_0, \sigma^2\mathbf{I}), \;\; (\boldsymbol{\beta}_0, \sigma^2) \sim \pi(\boldsymbol{\beta}_0, \sigma^2),$$

where $\boldsymbol{\beta}_0$ and σ^2 were assumed independent and given noninformative priors.

Bedrick, Christensen, and Johnson (1996, 1997) took a somewhat different approach to prior specification. They elicited beta priors on the success probabilities at several suitably selected values of the covariates. These induce a prior on the model parameters by a one-to-one transformation. They argued, following Tsutakawa and Lin (1986), that it is easier to formulate priors for success probabilities than for regression coefficients. In particular, those priors can be applied to different link functions, whereas prior specification for regression coefficients would depend on the link function. Bedrick et al. (1997) gave an example of modeling the probability of a trauma patient surviving as a function of four predictors and an interaction, using priors specified at six combinations of values of the predictors and using Bayes factors to compare possible link functions.

Item response models are binary regression models that describe the probability that a subject makes a correct response to a question on an exam. The simplest models, such as the Rasch model, model the logit or probit link of that probability in terms of additive effects of the difficulty of the question and the ability of the subject. For Bayesian analyses of such models, see Tsutakawa and Johnson (1990), Kim et al. (1994), Johnson and Albert (1999, Ch. 6), Albert and Ghosh (2000), Ghosh et al. (2000b), and the references therein. For instance, Ghosh et al. (2000b) considered necessary and conditions for posterior distributions to be proper when priors are improper.

Another important application of logistic regression is in modeling trend, such as in developmental toxicity experiments. Dominici and Parmigiani (2001) proposed a Bayesian semiparametric analysis that combines parametric dose-response relationships with a flexible nonparametric specification of the distribution of the response, obtained using a Dirichlet process mixture approach. The degree to which the distribution of the response adapts nonparametrically to the observations is driven by the data, and the marginal posterior distribution of the parameters of interest has closed form. Special cases include ordinary logistic regression, the beta-binomial model, and finite mixture models. Dempster, Selwyn, and Weeks (1983) and Ibrahim, Ryan, and Chen (1998) discussed the use of historical controls to adjust for covariates in trend tests for binary data. Extreme versions include logistic regression either completely pooling or completely ignoring historical controls.

Greenland (2001) argued that for Bayesian implementation of logistic and Poisson models with large samples, both the prior and the likelihood can be approximated with multivariate normals, but with sparse data, such approximations may be inadequate. For sparse data, he recommended exact conjugate analysis. Giving conjugate priors for the coefficient vector in logistic and Poisson models, he introduced a computationally feasible method of augmenting the data with binomial "pseudo-data" having an appropriate prior mean and variance. Greenland also discussed the advantages conjugate priors have over noninformative priors in epidemiological studies, showing that flat priors on regression coefficients often imply ridiculous assumptions about the effects of the clinical variables.

Piegorsch and Casella (1996) discussed empirical Bayesian methods for logistic regression and the wider class of GLMs, through a hierarchical approach. Let $h[E(Y_i)] = \mathbf{x}_i'\boldsymbol{\beta}$ for a link function h. When considering a hierarchical logistic model, they placed a beta(A_i, B_i) prior on each response probability. Following Albert (1988), they let $A_i = \lambda h^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$ and $B_i = \lambda[1 - h^{-1}(\mathbf{x}_i' \boldsymbol{\beta})]$, creating a hyperparameter λ that reflects prior belief in the correctness of the functional form of the GLM. The classical GLM is the limit as $\lambda \to \infty$. They also suggested an extension of the link function through the inclusion of another hyperparameter. All the hyperparameters were estimated via marginal maximum likelihood.

Here is a summary of other literature involving Bayesian binary regression modeling. Hsu and Leonard (1997) proposed a hierarchical approach that smoothes the data in the direction of a particular logistic regression model but does not require estimates to perfectly satisfy that model. Chen, Ibrahim, and Yiannoutsos (1999) considered prior elicitation and variable selection in logistic regression. Chaloner and Larntz (1989) considered determination of optimal design for experiments using logistic regression. Zocchi and Atkinson (1999) considered design for multinomial logistic models. Dey, Ghosh, and Mallick (2000) edited a collection of articles that provided Bayesian analyses for GLMs. In that volume Gelfand and Ghosh (2000) surveyed the subject and Chib (2000) modeled correlated binary data.

5.2 Multi-category responses

For frequentist inference with a multinomial response variable, popular models include logit and probit models for cumulative probabilities when the response is ordinal (such as $\text{logit}[P(y_i \leq j)] = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta})$, and multinomial logit and probit models when the response is nominal (such as $\log[P(y_i = j)/P(y_i = c)] = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}_j$). The ordinal models can be motivated by underlying logistic or normal latent variables. Johnson and Albert (1999) focused on ordinal models. Specification of priors is not simple, and they used an approach that specifies beta prior distributions for the cumulative probabilities at several values of the explanatory variables (e.g., see p. 133). They fitted the model using a hybrid Metropolis-Hastings/Gibbs sampler that recognizes an ordering constraint on the $\{\alpha_j\}$. Among special cases, they considered an ordinal extension of the item response model.

Chipman and Hamada (1996) used the cumulative probit model but with a normal prior defined directly on $\boldsymbol{\beta}$ and a truncated ordered normal prior for the $\{\alpha_j\}$, implementing it with the Gibbs sampler. They illustrated with two industrial data sets. For binary and ordinal regression, Lang (1999) used a parametric link function based on smooth mixtures of two extreme value distributions and a logistic distribution. His model used a flat, non-informative prior for the regression parameters, and was designed for applications in which there is some prior information about the appropriate link function.

Bayesian ordinal models have been used for various applications. For instance, Johnson (1996) proposed a Bayesian model for agreement in which several judges provide ordinal ratings of items, a particular application being test grading. Johnson assumed that for a given item, a normal latent variable underlies the categorical rating. The model is used to regress the latent variables for the items on covariates in order to compare the performance of raters. Broemeling (2001) employed a multinomial-Dirichlet setup to model agreement among multiple raters. For other Bayesian analyses with ordinal data, see Cowles, Carlin and Connett (1996), Bradlow and Zaslavsky (1999), Ishwaran and Gatsonis (2000), Ishwaran (2000), and Rossi, Gilula, and Allenby (2001).

For nominal responses, Daniels and Gatsonis (1997) used multinomial logit models to analyze variations in the utilization of alternative cardiac procedures in a study of Medicare patients who had suffered myocardial infarction. They examined how the rates of cardiac procedures depend on patient-level characteristics and considered whether regional differences exists in the use of the procedures. Their model generalized the Wong and Mason (1985) hierarchical approach. They used a multivariate t distribution for the regression parameters, with vague proper priors for the scale matrix and degrees of freedom.

In the econometrics literature, many have preferred the multinomial probit model to the multinomial logit model because it does not require an assumption of "independence from irrelevant alternatives." McCulloch, Polson, and Rossi (2000) discussed issues dealing with the fact that parameters in the basic model are not identified. They used a multivariate normal prior for the regression parameters and a Wishart distribution for the inverse covariance matrix for the underlying normal model, using Gibbs sampling to fit the model. See references therein for related approaches with that model. Imai and van Dyk (2004) considered a discrete-choice version of the model, fitted with MCMC.

5.3 Multivariate response extensions and other GLMs

For modeling multivariate correlated ordinal (or binary) responses, Chib and Greenberg (1998) considered a multivariate probit model. A multivariate normal latent random vector with cutpoints along the real line defines the categories of the observed discrete variables. The correlation among the categorical responses is induced through the covariance matrix for the underlying latent variables. See also Chib (2000). Webb and Forster (2004) parameterized the model in such a way that conditional posterior distributions are standard and easily simulated. They focused on model determination through comparing posterior marginal probabilities of the model given the data (integrating out the parameters). See also Chen and Shao (1999), who also briefly reviewed other Bayesian approaches to handling such data.

Logistic regression does not extend as easily to multivariate modeling, because of a lack of a simple analog of the multivariate normal. However, O'Brien and Dunson (2004) formulated a multivariate logistic distribution incorporating correlation parameters and having marginal logistic distibutions. They used this in a Bayesian analysis of marginal logistic regression models, showing that proper posterior distributions typically exist even when one uses an improper uniform prior for the regression parameters.

Zeger and Karim (1991) fitted generalized linear mixed models using a Bayesian framework with priors for fixed and random effects. The focus on distributions for random effects in GLMMs in articles such as this one led to the treatment of parameters in GLMs as random variables with a fully Bayesian approach. For any GLM, for instance, for the first stage of the prior specification one could take the model parameters to have a multivariate normal distribution. Alternatively, one can use a prior that has conjugate form for the exponential family (Bedrick et al. 1996). In either case, the posterior distribution is not tractable, because of the lack of closed form for the integral that determines the normalizing constant.

Recently Bayesian model averaging has received much attention. It accounts for uncertainty about the model by taking an average of the posterior distribution of a quantity of interest, weighted by the posterior probabilities of several potential models. Following the previously discussed work of Madigan and Raftery (1994), the idea of model averaging was developed further by Draper (1995) and Raftery, Madigan, and Hoeting (1997). In their review article, Hoeting et al. (1999) discussed model averaging in the context of GLMs. See also Giudici (1998) and Madigan and York (1995).

6 Bayesian Computation

Historically, a barrier for the Bayesian approach has been the difficulty of calculating the posterior distribution when the prior is not conjugate. See, for instance, Leonard, Hsu, and Tsui (1989), who considered Laplace approximations and related methods for approximating the marginal posterior density of summary measures of interest in contingency tables. Fortunately, for GLMs with canonical link function and normal or conjugate priors, the posterior joint and marginal distributions are log-concave (O'Hagan and Forster 2004, pp. 29-30). Hence numerical methods to find the mode usually converge quickly.

Computations of marginal posterior distributions and their moments are less problematic with modern ways of approximating posterior distributions by simulating samples from them. These include the importance sampling generalization of Monte Carlo simulation (Zellner and Rossi 1984) and Markov chain Monte Carlo methods such as Gibbs sampling (Gelfand and Smith 1990) and the Metropolis-Hastings algorithm (Tierney 1994). We touch only briefly on computational issues here, as they are reviewed in other sources (e.g., Andrieu, Doucet, and Robert (2004) and many recent books on Bayesian inference, such as O'Hagan and Forster (2004), Sections 12.42-46). For some standard analyses, such as inference about parameters in 2×2 tables, simple and long-established numerical algorithms are adequate and can be implemented with a wide variety of software. For instance, Agresti and Min (2004) provided functions using the software R for tail confidence intervals for association measures in 2×2 tables with independent beta priors.

For binary regression models, noting that analysis of the posterior density of $\boldsymbol{\beta}$ (in particular, the extraction of moments) was generally unwieldy, Zellner and Rossi (1984) discussed other options: asymptotic expansions, numerical integration, and Monte Carlo integration, for both diffuse and informative priors. Asymptotic expansions require a moderately large sample size n, and traditional numerical integration may be difficult for very high-dimensional integrals. When these options falter, Zellner and Rossi argued that Monte Carlo methods are reasonable, and they proposed an importance sampling method.

Importance sampling — as opposed to naive (uniform) Monte Carlo integration — is designed to be more efficient, requiring fewer sample draws to achieve a good approximation. To approximate the posterior expectation of a function $h(\boldsymbol{\beta})$, denoting the posterior kernel by $f(\boldsymbol{\beta}|\mathbf{y})$, Zellner and Rossi noted that

$$\begin{aligned} E[h(\boldsymbol{\beta})|\mathbf{y}] &= \int h(\boldsymbol{\beta}) f(\boldsymbol{\beta}|\mathbf{y})\, d\boldsymbol{\beta} \Big/ \int f(\boldsymbol{\beta}|\mathbf{y})\, d\boldsymbol{\beta} \\ &= \int h(\boldsymbol{\beta}) \frac{f(\boldsymbol{\beta}|\mathbf{y})}{I(\boldsymbol{\beta})} I(\boldsymbol{\beta})\, d\boldsymbol{\beta} \Big/ \int \frac{f(\boldsymbol{\beta}|\mathbf{y})}{I(\boldsymbol{\beta})} I(\boldsymbol{\beta})\, d\boldsymbol{\beta}. \end{aligned}$$

They approximated the numerator and denominator separately by simulating many values $\{\boldsymbol{\beta}_i\}$ from the *importance function* $I(\boldsymbol{\beta})$, which they chose to be multivariate t, and letting

$$E[h(\boldsymbol{\beta})|\mathbf{y}] \approx \sum_i h(\boldsymbol{\beta}_i) w_i \Big/ \sum w_i,$$

where $w_i = f(\boldsymbol{\beta}_i|\mathbf{y})/I(\boldsymbol{\beta}_i)$.

Gibbs sampling, a highly useful Markov chain Monte Carlo (MCMC) method to sample from multivariate distributions by successively sampling from simpler conditional distributions, became popular in Bayesian inference following the influential article by Gelfand and Smith (1990). They gave several examples of its suitability in Bayesian analysis, including a multinomial-Dirichlet model. Epstein and Fienberg (1991) employed Gibbs sampling to compute estimates of the entire posterior density of a set of cell probabilities (a finite mixture of Dirichlet densities), not simply the posterior mean. Forster and Skene (1994) applied Gibbs sampling with adaptive rejection sampling to the Knuiman and Speed (1988) formulation of multivariate normal priors for loglinear model parameters. Other examples include George and Robert (1992), Albert and Chib (1993), Forster (1994), Albert (1996), Chipman and Hamada (1996), Vounatsou and Smith (1996), Johnson and Albert (1999), and McCulloch, Polson, and Rossi (2000).

Often, the increased computational power of the modern era enables statisticians to make fewer assumptions and approximations in their analyses. For example, for multinomial data with a hierarchical Dirichlet prior, Leonard (1977c) made approximations when deriving the posterior to account for hyperparameter uncertainty. By contrast, Nandram (1998) used the Metropolis-Hastings algorithm to sample from the posterior distribution, rendering Leonard's approximations unnecessary.

7 Final Comments

Despite the advances summarized in this paper and the increasingly extensive literature, Bayesian inference does not seem to be commonly used yet in practice for basic categorical data analyses such as tests of independence and confidence intervals for association parameters. This may partly reflect the absence of Bayesian procedures in the primary software packages. Although it is straightforward for specialists to conduct analyses with Bayesian software such as BUGS, widespread use is unlikely to happen until the methods are simple to use in the software most commonly used by applied statisticians and methodologists. For multi-way contingency table analysis, another complication is the plethora of parameters for multinomial models, which necessitates substantial prior specification.

For many who are tentative users of the Bayesian approach, specification of prior distributions remains the stumbling block. The sensitivity of the results to changes in the prior specification when prior information is vague is troublesome to those who prefer an objective approach to data analysis but are attracted to other aspects of the Bayesian approach (e.g., making probabilistic conclusions about parameters). Many will find daunting the task of specifying and understanding prior distributions on GLM parameters in models with non-linear link functions, particularly for hierarchical models. In this regard, we find helpful the approach of eliciting prior distributions on the probability scale at selected values of covariates, as in Bedrick, Christensen and Johnson (1996, 1997). It is simpler to comprehend such priors and their implications than priors for parameters pertaining to a non-linear link function of the probabilities.

For the frequentist approach, the GLM provides a unifying approach for categorical data analysis. This model is a convenient starting point, as it yields many standard analyses as special cases and easily generalizes to more complex structures. Currently Bayesian approaches for categorical data seem to suffer from not having a natural starting point. Even if one starts with the GLM, there is a variety of possible approaches, depending on whether one specifies priors for the probabilities or for parameters in the model, depending on the distributions chosen for the priors, and depending on whether one specifies hyperparameters or uses a hierarchical approach or an empirical Bayesian approach for them.

It is unrealistic to expect all problems to fit into one framework, but nonetheless it would be helpful to data analysts if there were a standard default starting point for dealing with basic categorical data analyses such as estimating a proportion, comparing two proportions, and logistic regression modeling. However, it may be unrealistic to expect consensus about this, as even frequentists take increasingly diverse approaches for analyzing such data.

Historically, probably many frequentist statisticians of relatively senior age first saw the value of some Bayesian analyses upon learning of the advantages of shrinkage estimates, such as in the work of C. Stein. These days it is possible to obtain the same advantages in a frequentist context using random effects, such as in the generalized linear mixed model. In this sense, the lines between Bayesian and frequentist analysis have blurred somewhat. Nonetheless, there are still some analysis aspects for which the Bayesian approach is a more natural one, such as using model averaging to deal with the thorny issue of model uncertainty. In the future, it seems likely to us that statisticians will increasingly be tied less dogmatically to a single approach and will feel comfortable using both frequentist and Bayesian paradigms.

Acknowledgments

This research was partially supported by grants from NIH and NSF. The authors thank Jon Forster for many helpful comments and suggestions about an earlier draft and to him and Steve Fienberg and George Casella for suggesting relevant references.

References

Agresti, A. and Chuang, C. (1989) Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. *Computational Statistics and Data Analysis*, **7**, 245–258.

Agresti, A. and Min, Y. (2004) Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics (to appear).*

Aitchison, J. (1985) Practical Bayesian problems in simplex sample spaces. In *Bayesian Statistics 2*, 15–31. North-Holland/Elsevier (Amsterdam; New York).

Albert, J. H. (1984) Empirical Bayes estimation of a set of binomial probabilities. *Journal of Statistical Computation and Simulation*, **20**, 129–144.

— (1987a) Bayesian estimation of odds ratios under prior hypotheses of independence and exchangeability. *Journal of Statistical Computation and Simulation*, **27**, 251–268.

— (1987b) Empirical Bayes estimation in contingency tables. *Communications in Statistics, Part A – Theory and Methods*, **16**, 2459–2485.

— (1988) Bayesian estimation of Poisson means using a hierarchical log-linear model. In *Bayesian Statistics 3*, 519–531. Clarendon Press (Oxford).

— (1990) A Bayesian test for a two-way contingency table using independence priors. *The Canadian Journal of Statistics*, **18**, 347–363.

— (1996) Bayesian selection of log-linear models. *The Canadian Journal of Statistics*, **24**, 327–347.

— (1997) Bayesian testing and estimation of association in a two-way contingency table. *Journal of the American Statistical Association*, **92**, 685–693.

— (2004) Bayesian methods for contingency tables. *Encyclopedia of Biostatistics on-line version (to appear).*

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

Albert, J. H. and Ghosh, M. (2000) Item response modeling. In *Generalized Linear Models: A Bayesian Perspective* (eds. D. K. Dey, S. K. Ghosh and B. K. Mallick), 173–194. New York: Marcel Dekker, Inc.

Albert, J. H. and Gupta, A. K. (1982) Mixtures of Dirichlet distributions and estimation in contingency tables. *The Annals of Statistics*, **10**, 1261–1268.

— (1983a) Estimation in contingency tables using prior information. *Journal of the Royal Statistical Society, Series B, Methodological*, **45**, 60–69.

— (1983b) Bayesian estimation methods for 2×2 contingency tables using mixtures of Dirichlet distributions. *Journal of the American Statistical Association*, **78**, 708–717.

— (1985) Bayesian methods for binomial data with applications to a nonresponse problem. *Journal of the American Statistical Association*, **80**, 167–174.

Altham, P. M. E. (1969) Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test. *Journal of the Royal Statistical Society, Series B, Methodological*, **31**, 261–269.

— (1971a) The analysis of matched proportions. *Biometrika*, **58**, 561–576.

— (1971b) Exact Bayesian analysis of an intraclass 2×2 table. *Biometrika*, **58**, 679–680.

— (1975) Quasi-independent triangular contingency tables. *Biometrics*, **31**, 233–238.

Alvo, M. and Cabilio, P. (1982) Bayesian estimation of the difference between two proportions. *The Canadian Journal of Statistics*, **10**, 139–145.

Andrieu, C., Doucet, A. and Robert, C. P. (2004) Computational advances for and from bayesian analysis. *Statistical Science*, **19**, 118–127.

Antelman, G. R. (1972) Interrelated Bernoulli processes. *Journal of the American Statistical Association*, **67**, 831–841.

Basu, D. and de Braganca Pereira, C. A. (1982) On the Bayesian analysis of categorical data: The problem of nonresponse. *Journal of Statistical Planning and Inference*, **6**, 345–362.

Bayes, T. (1763) An essay toward solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, **53**, 370–418.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996) A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, **91**, 1450–1460.

— (1997) Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, **51**, 211–218.

Berger, J. O. and Pericchi, L. R. (1996) The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Bernardo, J. M. (1979) Reference posterior distributions for Bayesian inference (C/R p128-147). *Journal of the Royal Statistical Society, Series B, Methodological*, **41**, 113–128.

Bernardo, J. M. and Ramón, J. M. (1998) An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *The Statistician*, **47**, 101–135.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Wiley.

Bernstein, S. (1934) *Theory of Probability*.

Berry, D. A. (2004) Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science*, **19**, 175–187.

Berry, D. A. and Christensen, R. (1979) Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics*, **7**, 558–568.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analyses: Theory and Practice*. MIT Press.

Bloch, D. A. and Watson, G. S. (1967) A Bayesian study of the multinomial distribution. *The Annals of Mathematical Statistics*, **38**, 1423–1435.

Bradlow, E. T. and Zaslavsky, A. M. (1999) A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *Journal of the American Statistical Association*, **94**, 43–52.

Bratcher, T. L. and Bland, R. P. (1975) On comparing binomial probabilities from a Bayesian viewpoint. *Communications in Statistics*, **4**, 975–986.

Broemeling, L. D. (2001) A Bayesian analysis for inter-rater agreement. *Communications in Statistics, Part B – Simulation and Computation*, **30**, 437–446.

Brooks, R. J. (1987) Optimal allocation for Bayesian inference about an odds ratio. *Biometrika*, **74**, 196–199.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101–133.

— (2002) Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, **30**, 160–201.

Brown, L. D., Chow, M. and Fong, D. K. H. (1992) On the admissibility of the maximum-likelihood estimator of the binomial variance. *The Canadian Journal of Statistics*, **20**, 353–358.

Casella, G. (2001) Empirical Bayes Gibbs sampling. *Biostatistics*, **2**, 485–500.

Casella, G. and Moreno, E. (2002) Objective bayesian analysis of contingency tables. *Tech. Rep. 2002-023*, Department of Statistics, University of Florida.

Chaloner, K. and Larntz, K. (1989) Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, **21**, 191–208.

Chen, J. J. and Novick, M. R. (1984) Bayesian analysis for binomial models with generalized beta prior distributions. *Journal of Educational Statistics*, **9**, 163–175.

Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1999) Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society, Series B, Methodological*, **61**, 223–242.

Chen, M.-H. and Shao, Q.-M. (1999) Properties of prior and posterior distributions for multivariate categorical response data models. *Journal of Multivariate Analysis*, **71**, 277–296.

Chib, S. (2000) Bayesian methods for correlated binary data. In *Generalized Linear Models: A Bayesian Perspective* (eds. D. K. Dey, S. K. Ghosh and B. K. Mallick), 113–132. New York: Marcel Dekker, Inc.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Chipman, H. and Hamada, M. (1996) Bayesian analysis of ordered categorical data ¿from industrial experiments. *Technometrics*, **38**, 1–10.

Consonni, G. and Veronese, P. (1995) A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, **90**, 935–944.

Copas, J. B. (1972) Empirical Bayes methods and the repeated use of a standard. *Biometrika*, **59**, 349–360.

Cornfield, J. (1966) A Bayesian test of some classical hypotheses – With applications to sequential clinical trials. *Journal of the American Statistical Association*, **61**, 577–594.

Cowles, M. K., Carlin, B. P. and Connett, J. E. (1996) Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association*, **91**, 86–98.

Crook, J. F. and Good, I. J. (1980) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II (corr: V9 p1133). *The Annals of Statistics*, **8**, 1198–1218.

— (1982) The powers and strengths of tests for multinomials and contingency tables. *Journal of the American Statistical Association*, **77**, 793–802.

Crowder, M. and Sweeting, T. (1989) Bayesian inference for a bivariate binomial distribution. *Biometrika*, **76**, 599–603.

Daniels, M. J. and Gatsonis, C. (1997) Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine*, **16**, 2311–2325.

DasGupta, A. and Zhang, T. (2004) Inference for binomial and multinomial parameters: A review and some open problems. *Encyclopedia of Statistics (to appear)*.

David, H. A. (1998) First (?) occurrence of common terms in probability and statistics — A second list, with corrections. *The American Statistician*, **52**, 36–40.

Davidson, R. R. and Solomon, D. L. (1973) A Bayesian approach to paired comparison experimentation. *Biometrika*, **60**, 477–487.

Dawid, A. P. and Lauritzen, S. L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21**, 1272–1317.

De Waal, D. J. and Groenewald, P. C. N. (1988) Bayesian tests for hypothesis of the equality of multinomial probabilities. *South African Statistical Journal*, **22**, 87–104.

Delampady, M. and Berger, J. O. (1990) Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *The Annals of Statistics*, **18**, 1295–1316.

Dellaportas, P. and Forster, J. J. (1999) Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, 615–633.

Dempster, A. P., Selwyn, M. R. and Weeks, B. J. (1983) Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association*, **78**, 221–227.

Dey, D. e., Ghosh, S. K. e. and Mallick, B. K. e. (2000) *Generalized Linear Models: a Bayesian Perspective*. Marcel Dekker Inc.

Diaconis, P. and Freedman, D. (1990) On the uniform consistency of Bayes estimates for multinomial probabilities. *The Annals of Statistics*, **18**, 1317–1327.

Dickey, J. M. (1983) Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, **78**, 628–637.

Dickey, J. M., Jiang, J.-M. and Kadane, J. B. (1987) Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, **82**, 773–781.

Dobra, A. and Fienberg, S. E. (2001) How large is the world wide web? *Computing Science and Statistics*, **33**, 250–268.

Dominici, F. and Parmigiani, G. (2001) Bayesian semiparametric analysis of developmental toxicology data. *Biometrics*, **57**, 150–157.

Draper, D. (1995) Assessment and propagation of model uncertainty (Disc: p71-97). *Journal of the Royal Statistical Society, Series B, Methodological*, **57**, 45–70.

Draper, N. and Guttman, I. (1971) Bayesian estimation of the binomial parameter. *Technometrics*, **13**, 667–673.

Efron, B. (1996) Empirical Bayes methods for combining likelihoods (Disc: p551-565). *Journal of the American Statistical Association*, **91**, 538–550.

Efron, B. and Morris, C. (1975) Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311–319.

Epstein, L. D. and Fienberg, S. E. (1991) Using Gibbs sampling for Bayesian inference in multidimensional contingency tables. In *Computing Science and Statistics. Proceedings of the 23rd Symposium on the Interface*, 215–223. Interface Foundation of North America (Fairfax Station, VA).

— (1992) Bayesian estimation in multidimensional contingency tables. In *Bayesian Analysis in Statistics and Econometrics*, 27–41. Springer-Verlag (Berlin; New York).

Ericson, W. A. (1969) Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **31**, 195–233.

Evans, M., Gilula, Z. and Guttman, I. (1993) Computational issues in the Bayesian analysis of categorical data: Log-linear and Goodman's Rc model. *Statistica Sinica*, **3**, 391–406.

Evans, M. J., Gilula, Z. and Guttman, I. (1989) Latent class analysis of two-way contingency tables by Bayesian methods. *Biometrika*, **76**, 557–563.

Ferguson, T. S. (1967) On characterizing distributions by properties of order statistics. *Sankhya, Series A, Indian Journal of Statistics*, **29**, 265–278.

— (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Fienberg, S. E. and Holland, P. W. (1970) Methods for eliminating zero counts in contingency tables. In *Random Counts on Models and Structures* (ed. G. P. Patil), 233–260. University Park, PA: Pennsylvania State University Press.

— (1972) On the choice of flattening constants for estimating multinomial probabilities. *Journal of Multivariate Analysis*, **2**, 127–134.

— (1973) Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, **68**, 683–691.

Fienberg, S. E., Johnson, M. S. and Junker, B. W. (1999) Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A, General*, **162**, 383–405.

Fienberg, S. E. and Makov, U. E. (1998) Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385–397.

Forster, J. J. (1994) A Bayesian approach to the analysis of binary crossover data. *The Statistician*, **43**, 61–68.

— (2004a) Bayesian inference for poisson and multinomial log-linear models.

— (2004b) Bayesian inference for square contingency tables.

Forster, J. J. and Skene, A. M. (1994) Calculation of marginal densities for parameters of multinomial distributions. *Statistics and Computing*, **4**, 279–286.

Forster, J. J. and Smith, P. W. F. (1998) Model-based inference for categorical survey data subject to non-ignorable non-response (disc: P89-102). *Journal of the Royal Statistical Society, Series B, Methodological*, **60**, 57–70.

Franck, W. E., Hewett, J. E., Islam, M. Z., Wang, S. J. and Cooperstock, M. (1988) A Bayesian analysis suitable for classroom presentation. *The American Statistician*, **42**, 75–77.

Freedman, D. A. (1963) On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, **34**, 1386–1403.

Geisser, S. (1984) On prior distributions for binary trials (c/r: P247-251). *The American Statistician*, **38**, 244–247.

Gelfand, A. and Ghosh, M. (2000) Generalized linear models: A bayesian view. In *Generalized Linear Models: A Bayesian Perspective* (eds. D. K. Dey, S. K. Ghosh and B. K. Mallick), 1–22. New York: Marcel Dekker, Inc.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

George, E. I. and Robert, C. P. (1992) Capture-recapture estimation via Gibbs sampling. *Biometrika*, **79**, 677–683.

Ghosh, M. and Chen, M.-H. (2002) Bayesian inference for matched case-control studies. *Sankhya, Series B, Indian Journal of Statistics*, **64**, 107–127.

Ghosh, M., Chen, M.-H., Ghosh, A. and Agresti, A. (2000a) Hierarchical Bayesian analysis of binary matched pairs data. *Statistica Sinica*, **10**, 647–657.

Ghosh, M., Ghosh, A., Chen, M.-H. and Agresti, A. (2000b) Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference*, **88**, 99–115.

Giudici, P. (1998) Smoothing sparse contingency tables: A graphical Bayesian approach. *Metron*, **56**, 171–187.

Good, I. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

— (1956) On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society, Series B*, **18**, 113–124.

— (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Good, I. J. (1967) A Bayesian significance test for multinomial distributions (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **29**, 399–431.

— (1976) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *The Annals of Statistics*, **4**, 1159–1189.

— (1980) Some history of the hierarchical Bayesian methodology. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia*, 489–504. University of Valencia (Spain).

Good, I. J. and Crook, J. F. (1974) The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of the American Statistical Association*, **69**, 711–720.

— (1987) The robustness and sensitivity of the mixed-Dirichlet Bayesian test for "independence" in contingency tables. *The Annals of Statistics*, **15**, 670–693.

Goutis, C. (1993) Bayesian estimation methods for contingency tables. *Journal of the Italian Statistical Society*, **2**, 35–54.

Greenland, S. (2001) Putting background information about relative risks into conjugate prior distributions. *Biometrics*, **57**, 663–670.

Griffin, B. S. and Krutchkoff, R. G. (1971) An empirical Bayes estimator for p[success] in the binomial distribution. *Sankhyā, Series B, Indian Journal of Statistics*, **33**, 217–224.

Gunel, E. and Dickey, J. (1974) Bayes factors for independence in contingency tables. *Biometrika*, **61**, 545–557.

Hald, A. (1998) *A History of Mathematical Statistics From 1750 to 1930*. Wiley.

Haldane, J. B. S. (1948) The precision of observed values of small frequencies. *Biometrika*, **35**, 297–303.

Hashemi, L., Nandram, B. and Goldberg, R. (1997) Bayesian analysis for a single 22 table. *Statistics in Medicine*, **16**, 1311–1328.

Hoadley, B. (1969) The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association*, **64**, 216–229.

Hodges, A. (1983) *Alan Turing: The Enigma*. New York: Simon and Schuster.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial (Pkg: p382-417). *Statistical Science*, **14**, 382–401.

Howard, J. V. (1998) The 2×2 table: A discussion from a Bayesian viewpoint. *Statistical Science*, **13**, 351–367.

Hsu, J. S. J. and Leonard, T. (1997) Hierarchical Bayesian semiparametric procedures for logistic regression. *Biometrika*, **84**, 85–93.

Ibrahim, J. G. and Laud, P. W. (1991) On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, **86**, 981–986.

Ibrahim, J. G., Ryan, L. M. and Chen, M.-H. (1998) Using historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association*, **93**, 1282–1293.

Imai, K. and van Dyk, D. A. (2004) A Bayesian analysis of the multinomial probit model using the data augmentation. *Journal of Econometrics*, **45**, –.

Irony, T. Z. and de Braganca Pereira, C. A. (1986) Exact tests for equality of two proportions: Fisher v. Bayes. *Journal of Statistical Computation and Simulation*, **25**, 93–114.

Ishwaran, H. (2000) Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *The Canadian Journal of Statistics*, **28**, 715–730.

Ishwaran, H. and Gatsonis, C. A. (2000) A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *The Canadian Journal of Statistics*, **28**, 731–750.

Jansen, M. G. H. and Snijders, T. A. B. (1991) Comparisons of Bayesian estimation procedures for two-way contingency tables without interaction. *Statistica Neerlandica*, **45**, 51–65.

Jeffreys, H. (1948) *Theory of probability.* Oxford University Press.

Johnson, B. M. (1971) On the admissible estimators for certain fixed sample binomial problems. *The Annals of Mathematical Statistics*, **42**, 1579–1587.

Johnson, V. E. (1996) On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association*, **91**, 42–51.

Johnson, V. E. and Albert, J. H. (1999) *Ordinal data modeling.* Springer-Verlag Inc.

Johnson, W. E. (1932) Probability: Deductive and inductive problems (appendix). *Mind*, **41**, 421–423.

Jones, P. W. (1974) A note on the Bayesian sequential estimation of a binomial parameter. *Biometrika*, **61**, 642–644.

Jones, P. W. and Madhi, S. A. (1986) Bayesian sequential methods for choosing the better of two binomial populations. *Sequential Analysis*, **5**, 127–138.

Joseph, L., Wolfson, D. B. and Berger, R. d. (1995) Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician*, **44**, 143–154.

Kadane, J. B. (1985) Is victimization chronic? A Bayesian analysis of multinomial missing data. *Journal of Econometrics*, **29**, 47–67.

Kass, R. E. and Vaidyanathan, S. K. (1992) Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B, Methodological*, **54**, 129–144.

Kim, S.-H., Cohen, A. S., Baker, F. B., Subkoviak, M. J. and Leonard, T. (1994) An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika*, **59**, 405–421.

King, R. and Brooks, S. P. (2001a) On the Bayesian analysis of population size. *Biometrika*, **88**, 317–336.

— (2001b) Prior induction in log-linear models for general contingency table analysis. *The Annals of Statistics*, **29**, 715–747.

— (2002) Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika*, **89**, 785–806.

Knuiman, M. W. and Speed, T. P. (1988) Incorporating prior information into the analysis of contingency tables. *Biometrics*, **44**, 1061–1071.

Laird, N. M. (1978) Empirical Bayes methods for two-way contingency tables. *Biometrika*, **65**, 581–590.

Lang, J. B. (1999) Bayesian ordinal and binary regression models with a parametric family of mixture links. *Computational Statistics and Data Analysis*, **31**, 59–87.

Laplace, P. S. (1774) Mémoire sur la probabilité des causes par les événements. *Mém. Acad. R. Sci., Paris*, **6**, 621–656.

Lee, M.-L. T., Schoenfeld, D., Wang, X., Penfornis, A. and Faustman, D. (2001) Bayesian analysis of case control polygenic etiology studies with missing data. *Biostatistics*, **2**, 309–322.

Leighty, R. M. and Johnson, W. J. (1990) A Bayesian loglinear model analysis of categorical data. *Journal of Official Statistics*, **6**, 133–155.

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

— (1973) A Bayesian method for histograms. *Biometrika*, **60**, 297–308.

— (1975) Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B, Methodological*, **37**, 23–37.

— (1977a) An alternative Bayesian approach to the Bradley-Terry model for paired comparisons. *Biometrics*, **33**, 121–132.

— (1977b) A Bayesian approach to some multinomial estimation and pretesting problems. *Journal of the American Statistical Association*, **72**, 869–874.

— (1977c) Bayesian simultaneous estimation for several multinomial distributions. *Communications in Statistics, Part A – Theory and Methods*, **6**, 619–630.

Leonard, T. and Hsu, J. S. J. (1994) The Bayesian analysis of categorical data – A selective review. In *Aspects of Uncertainty. A Tribute to D. V. Lindley*, 283–310. Wiley (New York).

— (1999) *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers.* Cambridge University Press.

Leonard, T., Hsu, J. S. J. and Tsui, K.-W. (1989) Bayesian marginal inference. *Journal of the American Statistical Association*, **84**, 1051–1058.

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

Lindley, D. V. (1964) The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, **35**, 1622–1643.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **34**, 1–41.

Little, R. J. A. (1989) Testing the equality of two independent binomial proportions. *The American Statistician*, **43**, 283–288.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical analysis with missing data, 2nd edition.* John Wiley & Sons.

Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

Madigan, D. and York, J. C. (1997) Bayesian methods for estimation of the size of a closed population. *Biometrika*, **84**, 19–31.

Maritz, J. S. (1989) Empirical Bayes estimation of the log odds ratio in 2×2 contingency tables. *Communications in Statistics, Part A – Theory and Methods*, **18**, 3215–3233.

Martz, H. F. and Lian, M. G. (1974) Empirical Bayes estimation of the binomial parameter. *Biometrika*, **61**, 517–523.

Matthews, J. N. S. (1999) Effect of prior specification on Bayesian design for two-sample comparison of a binary outcome. *The American Statistician*, **53**, 254–256.

McCulloch, R. E., Polson, N. G. and Rossi, P. E. (2000) A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, **99**, 173–193.

Meeden, G., Geyer, C., Lang, J. and Funo, E. (1998) The admissibility of the maximum likelihood estimator for decomposable log-linear interaction models for contingency tables. *Communications in Statistics, Part A – Theory and Methods*, **27**, 473–493.

Meng, C. Y. K. and Dempster, A. P. (1987) A Bayesian approach to the multiplicity problem for significance testing with binomial data. *Biometrics*, **43**, 301–311.

von Mises, R. (1964) *Mathematical Theory of Probability and Statistics.* New York: Academic Press.

Müller, P. and Roeder, K. (1997) A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**, 523–537.

Nandram, B. (1998) A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, **61**, 97–126.

Nayak, T. K. and Naik, D. N. (1989) Estimating multinomial cell probabilities under quadratic loss. *The Statistician*, **38**, 3–10.

Nazaret, W. A. (1987) Bayesian log linear estimates for three-way contingency tables. *Biometrika*, **74**, 401–410.

Novick, M. R. (1969) Multiparameter Bayesian indifference procedure (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **31**, 29–64.

Novick, M. R. and Grizzle, J. E. (1965) A Bayesian approach to the analysis of data ¿from clinical trials. *Journal of the American Statistical Association*, **60**, 81–96.

Ntzoufras, I., Forster, J. J. and Dellaportas, P. (2000) Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation*, **68**, 23–37.

Nurminen, M. and Mutanen, P. (1987) Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics*, **14**, 67–77.

O'Brien, S. M. and Dunson, D. B. (2004) Bayesian multivariate logistic regression. *Biometrics*, **60**, 739–746.

O'Hagan, A. and Forster, J. (2004) *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Arnold.

Park, T. (1998) An approach to categorical data with nonignorable nonresponse. *Biometrics*, **54**, 1579–1590.

Park, T. and Brown, M. B. (1994) Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, **89**, 44–52.

Paulino, C. D. M. and Pereira, C. A. D. B. (1995) Bayesian methods for categorical data under informative general censoring. *Biometrika*, **82**, 439–446.

Perks, W. (1947) Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, **73**, 285–334.

Piegorsch, W. W. and Casella, G. (1996) Empirical Bayes estimation for logistic regression and extended parametric regression models. *Journal of Agricultural, Biological, and Environmental Statistics*, **1**, 231–249.

Poirier, D. (1994) Jeffrey's prior for logit models. *Journal of Econometrics*, **63**, 327–339.

Polson, N. and Wasserman, L. (1990) Prior distributions for the bivariate binomial. *Biometrika*, **77**, 901–904.

Prescott, G. J. and Garthwaite, P. H. (2002) A simple Bayesian analysis of misclassified binary data with a validation substudy. *Biometrics*, **58**, 454–458.

Quintana, F. A. (1998) Nonparametric Bayesian analysis for assessing homogeneity in $k \times l$ contingency tables with fixed right margin totals. *Journal of the American Statistical Association*, **93**, 1140–1149.

Raftery, A. E. (1986) A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, Series B, Methodological*, **48**, 249–250.

— (1996) Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Rossi, P. E., Gilula, Z. and Allenby, G. M. (2001) Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, **96**, 20–31.

Routledge, R. D. (1994) Practicing safe statistics with the mid-p^*. *The Canadian Journal of Statistics*, **22**, 103–110.

Rukhin, A. L. (1988) Estimating the loss of estimators of a binomial parameter. *Biometrika*, **75**, 153–155.

Seaman, S. R. and Richardson, S. (2001) Bayesian analysis of case-control studies with categorical covariates. *Biometrika*, **88**, 1073–1088.

— (2004) Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika*, **91**, 15–25.

Sedransk, J., Monahan, J. and Chiu, H. Y. (1985) Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions. *Journal of the Royal Statistical Society, Series B, Methodological*, **47**, 519–527.

Sharma, R. S. (1975) Bayes approach to interval estimation of a binomial parameter. *Annals of the Institute of Statistical Mathematics*, **27**, 259–268.

Sinha, S., Mukherjee, B. and Ghosh, M. (2004) Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics*, **60**, 41–49.

Sivaganesan, S. and Berger, J. (1993) Robust Bayesian analysis of the binomial empirical Bayes problem. *The Canadian Journal of Statistics*, **21**, 107–119.

Smith, P. J. (1991) Bayesian analyses for a multiple capture-recapture model. *Biometrika*, **78**, 399–407.

Soares, P. and Paulino, C. D. (2001) Incomplete categorical data analysis: A Bayesian perspective. *Journal of Statistical Computation and Simulation*, **69**, 157–170.

Sobel, M. J. (1993) Bayes and empirical Bayes procedures for comparing parameters. *Journal of the American Statistical Association*, **88**, 687–693.

Spiegelhalter, D. J. and Smith, A. F. M. (1982) Bayes factors for linear and loglinear models with vague prior information. *Journal of the Royal Statistical Society, Series B, Methodological*, **44**, 377–387.

Springer, M. D. and Thompson, W. E. (1966) Bayesian confidence limits for the product of n binomial parameters. *Biometrika*, **53**, 611–613.

Stigler, S. M. (1982) Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society, Series A, General*, **145**, 250–258.

— (1986) *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (c/r: P541-550). *Journal of the American Statistical Association*, **82**, 528–540.

Teather, D. (1984) The estimation of exchangeable binomial parameters. *Communications in Statistics, Part A – Theory and Methods*, **13**, 671–680.

Tierney, L. (1994) Markov chains for exploring posterior distributions (Disc: p1728-1762). *The Annals of Statistics*, **22**, 1701–1728.

Tsutakawa, R. K. and Johnson, J. C. (1990) The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, **55**, 371–390.

Tsutakawa, R. K. and Lin, H. Y. (1986) Bayesian estimation of item response curves. *Psychometrika*, **51**, 251–267.

Viana, M. A. G. (1994) Bayesian small-sample estimation of misclassified multinomial data. *Biometrics*, **50**, 237–243.

Vounatsou, P. and Smith, A. F. M. (1996) Bayesian analysis of contingency tables: A simulation and graphics-based approach. *Statistics and Computing*, **6**, 277–287.

Wakefield, J. (2004) Ecological inference for 2 x 2 tables (with discussion). *Journal of the Royal Statistical Society, Series A*, **167**, 385–445.

Walley, P. (1996) Inferences from multinomial data: Learning about a bag of marbles (disc: P35-57). *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 3–34.

Walley, P., Gurrin, L. and Burton, P. (1996) Analysis of clinical data using imprecise prior probabilities. *The Statistician*, **45**, 457–485.

Walters, D. E. (1985) An examination of the conservative nature of "classical" confidence limits for a proportion. *Biometrical Journal. Journal of Mathematical Methods in Biosciences.*, **27**, 851–861.

Warn, D. E., Thompson, S. G. and Spiegelhalter, D. J. (2002) Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Statistics in Medicine*, **21**, 1601–1623.

Webb, E. L. and Forster, J. J. (2004) Bayesian model determination for multivariate ordinal and binary data.

Weisberg, H. I. (1972) Bayesian comparison of two ordered multinomial populations. *Biometrics*, **28**, 859–867.

Wong, G. Y. and Mason, W. M. (1985) The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, **80**, 513–524.

Wypij, D. and Santner, T. J. (1992) Pseudotable methods for the analysis of 2×2 tables. *Computational Statistics and Data Analysis*, **13**, 173–190.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zelen, M. and Parker, R. A. (1986) Case-control studies and Bayesian inference. *Statistics in Medicine*, **5**, 261–269.

Zellner, A. and Rossi, P. E. (1984) Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, **25**, 365–393.

Zocchi, S. S. and Atkinson, A. C. (1999) Optimum experimental designs for multinomial logistic models. *Biometrics*, **55**, 437–444.