Bayesian Inference for Categorical Data Analysis: A Survey

Alan Agresti
Department of Statistics
University of Florida
Gainesville, Florida

University of Melbourne

June 27, 2006

OUTLINE

- Early history
- Prior distributions for proportions
- Estimating binomial and multinomial parameters
- Estimating cell probabilities in contingency tables
- Tests and confidence intervals in two-way tables
- Regression models for categorical responses

Seminar based on survey paper prepared with David Hitchcock,
Dept. of Statistics, Univ. of South Carolina

1 Early History of Bayesian Categorical Data Analysis

Bayes (1763) and Laplace (1774) estimate a binomial parameter using a uniform prior distribution.

Let y denote binomial random variable for n trials and parameter π, and let $p = y/n$ (ML). Laplace estimate (“law of succession”) is $(y + 1)/(n + 2)$.

De Morgan (1847) proposed uniform distribution over the simplex for multinomial probabilities, extension of Laplace estimate.

Perks (1947) considered Dirichlet prior, but actual implementation seems to be due to Lindley (1964) and Good (1965)

Early critic of Bayesian approach is R. A. Fisher (first to use term “Bayesian,” in 1950). In **Statistical Methods and Scientific Inference** (1956), challenges use of uniform prior, noting uniform priors on other scales lead to different results.

Early applications of Bayesian methods to contingency tables involved smoothing cell counts to improve estimation of cell probabilities with small samples (I. J. Good).

(ex.: What if the count in a cell is zero?)

- Good (1953) used uniform prior distribution over several categories in estimating 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,

  association factor = (probability of cell)/(probability assuming independence)

- Good (1965) monograph on estimating multinomial probabilities using Dirichlet prior. Also considered hierarchical and empirical Bayesian approaches with this model.
  (Main statistical assistant in 1941 to Alan Turing)

- Lindley (1964) focused on estimating summary association measures. Using Dirichlet prior (actually, improper limiting case) for multinomial probabilities, obtains posterior distribution of contrasts of log probabilities, such as log odds ratio.

2 Estimating Binomial and Multinomial Parameters

2.1 Prior distributions for a binomial parameter

Suppose y distributed binomial(n, π)

Conjugate prior density for π is beta(α, β) density,

$$g(\pi) \propto \ \pi^{\alpha-1}(1-\pi)^{\beta-1} \ \text{for} \ \alpha > 0, \beta > 0, \ 0 < \pi < 1$$

for which $E(\pi) = \alpha/(\alpha+\beta)$.

Posterior density $h(\pi|y) = f(y|\pi)g(\pi)/f(y)$

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

is beta($y+\alpha, n-y+\beta$)

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

where $w = n/(n+\alpha+\beta)$

Special cases:

- ML estimator $p = y/n$ results from $\alpha = \beta = 0$ (improper).

  Posterior distribution improper if $y = 0$ or n.

  Corresponds to uniform prior for log odds,

  $$\text{logit}(\pi) = \log[\pi/(1-\pi)].$$

  Haldane (1948) argued taking $\log(\pi)$ roughly uniform for π near 0

  "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}."

- Jeffreys prior (scale invariant), proportional to square root of determinant of Fisher information matrix for parameters of interest, is beta(.5, .5).

Logistic-normal prior:

Alternative two-parameter approach specifies $N(\mu, \sigma^2)$ prior for logit(π)

Cornfield (1966): clinical trial application

Strongly promoted by T. Leonard (1972 on, apparently instigated by D. Lindley)

With $N(0, \sigma^2)$ prior for logit(π), prior density function for π over $0 < \pi < 1$ is logistic normal,

$$f(\pi) = \frac{1}{\sqrt{2(3.14)\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\log\frac{\pi}{1-\pi}\right)^2\right\}\frac{1}{\pi(1-\pi)}.$$

On π scale, symmetric, unimodal when $\sigma^2 \leq 2$ and bimodal when $\sigma^2 > 2$, always tapering off toward 0 as π approaches 0 or 1.

2.2 Bayesian inference about a binomial parameter

Point Estimation:

Bayesian estimators of binomial (or multinomial) parameters not uniformly better than ML: If π = 0, $p = y/n$ = 0 with probability 1.

Johnson (1971): sample proportion admissible

With loss function $(T-\pi)^2/[\pi(1-\pi)]$, Bayes estimator for uniform prior distribution is $p = y/n$.
(For this loss function, risk function constant, so also minimax)

Freedman (1963) showed consistency of Bayes estimators, for sampling from discrete distributions, showed asymptotic normality of posterior assuming smoothness of prior.
(Extends binomial results – Bernstein 1934, von Mises 1964)

Diaconis and Freedman (1990): Inequalities for posterior probability falling close to p as n increases.

Draper and Guttman (1971): Estimating n based on r independent bin(n, π) observations. (Related capture-recapture literature, e.g. Madigan and York 1997, accounts for model uncertainty by placing prior distribution over set of models)

Interval Estimation and Hypothesis Testing:

Brown, Cai, and Das Gupta (2001, 2002) showed that posterior distribution with Jeffreys prior yields confidence interval for π with good frequentist performance. Approximates small-sample frequentist CI based on inverting two binomial one-sided tests when use mid-P value.

e.g., for 95% CI, .025 = $(\frac{1}{2})P(Y = y|\pi_L) + P(Y > y|\pi_L)$

.025 = $(\frac{1}{2})P(Y = y|\pi_U) + P(Y < y|\pi_U)$

For H_0: $\pi \geq \pi_0$, H_a: $\pi < \pi_0$,
Bayesian P-value = P($\pi \geq \pi_0|y)$.

With Jeffreys prior and $\pi_0 = 1/2$, Routledge (1994) showed

$$P(\pi \geq 1/2|y) \approx [(\frac{1}{2})P(Y = y|\pi = .5) + P(Y < y|\pi = .5)]$$

2.3 Bayesian estimation of multinomial parameters

With c categories, cell counts $\mathbf{y} = (y_1, \ldots, y_c)$ have multinomial dist. with $n = \sum y_i$, parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$, $\sum \pi_i = 1$.

$$f(\mathbf{y}|\boldsymbol{\pi}) \propto \ \prod_{i=1}^{c} \pi_i^{y_i}$$

Conjugate density is Dirichlet,

$$g(\boldsymbol{\pi}) \propto \ \prod_{i=1}^{c} \pi_i^{\alpha_i - 1} \quad \text{for } 0 < \pi_i < 1 \text{ all } i, \quad \sum_i \pi_i = 1,$$

with $\{\alpha_i > 0\}$. $E(\pi_i) = \alpha_i/K$, with $K = \sum \alpha_j$.

Posterior is Dirichlet($\{y_i + \alpha_i\}$), so

$$E(\pi_i|y_1, \ldots, y_c) = (y_i + \alpha_i)/(n + K).$$

Let $\gamma_i = E(\pi_i) = \alpha_i/K$, $p_i = y_i/n$.

$$E(\pi_i|\mathbf{y}) = [n/(n+K)]p_i + [K/(n+K)]\gamma_i$$

- Jeffreys prior sets all $\alpha_i = 0.5$.
- Lindley (1964) used improper limiting case $\{\alpha_i = 0\}$, also recommended by Novick (1969). Discussion of Novick (1969) shows lack of consensus about meaning of ‘noninformative’.
- Good (1965) used symmetric Dirichlet $\alpha_i = K/c$, for which $K = \sum \alpha_i$ called a “flattening constant.”

$$E(\pi_i|\mathbf{y}) = [n/(n+K)]p_i + [K/(n+K)](1/c)$$

Logistic-normal prior

Good noted Dirichlet restricted by having relatively few parameters.

Can specify means through choice of $\{\gamma_i\}$ and variances through the choice of K, but no freedom to alter correlations.

Multivariate normal prior for multinomial logits:

If $(X_1, \ldots, X_c)$ has multivariate normal distribution, then $\mathbf{P} = (P_1, \ldots, P_c)$, where $P_i = \exp(X_i)/\sum_{j=1}^{c} \exp(X_j)$, has logistic normal distribution.

Leonard (1973) proposed this prior for estimating a histogram.

(For ordered categories, natural for probabilities nearer each other to be more highly correlated, e.g., with autoregressive structure)

2.4 Hierarchical Bayesian and empirical Bayes estimates of multinomial parameters

Good (1965, 1967, 1976) adopts hierarchical approach, specifies second-stage prior for parameters $\{\alpha_i\}$ of Dirichlet prior.

Albert and Gupta (1982) let the prior parameters take certain structure natural for a contingency table (independence, symmetry)

These approaches gain greater generality at expense of giving up simple conjugate Dirichlet form for posterior.

Empirical Bayesian approach uses data to determine parameter values in prior – e.g., use prior density that maximizes marginal probability of observed data.

Good (1956) – first to use empirical Bayesian approach with categorical data, estimating parameters in gamma and log-normal priors for association factors.

Disadvantage: does not account for variability due to substituting estimates for prior parameters. Hierarchical approach increasingly preferred.

3 Estimating Cell Probabilities in Contingency Tables

3.1 Simultaneous estimation of several binomial parameters

$\pi_1, \pi_2, ...\pi_r$ (parameters for $r \times 2$ table)

Mostly with hierarchical approach (Leonard 1972)

- Stage 1: Given μ and σ, logit(π_i) independent from N(μ, σ^2).
- Stage 2: Improper uniform prior for μ over the real line, $\nu\lambda/\sigma^2$ independent of μ and is chi-squared with df = ν, with λ a prior estimate of σ^2 and ν a measure of sureness of prior beliefs.
- For simplicity, suggested limiting case in which log(σ^2) has improper uniform prior.
- Integrating out μ and σ^2, two-stage approach corresponds to multivariate t prior for logits.
- Resulting $E[\text{logit}(\pi_i|\mathbf{y})]$ is approximately weighted average of logit(p_i) and weighted average of $\{$logit$(p_j)\}$.

3.2 Empirical Bayesian approaches in two-way tables

Fienberg and Holland (1970, 1972, 1973):

For Dirichlet($K, \boldsymbol{\gamma}$) prior, estimator of multinomial π_{ij} is

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

- Minimum total mean squared error occurs when

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

- Optimal $K = K(\boldsymbol{\gamma}, \boldsymbol{\pi})$ depends on $\boldsymbol{\pi}$, so used $K(\boldsymbol{\gamma}, \mathbf{p})$ with sample proportion $\mathbf{p}$ replacing $\boldsymbol{\pi}$.
- Selected $\{\gamma_{ij}\}$ based on fit of simple model, such as $\{\gamma_{ij} = p_{i+}p_{+j}\}$ (e.g., rather than $\{\gamma_{ij}\}$ identical).

3.3 Estimating loglinear parameters in two-way tables

Rather than focusing on estimating probabilities (e.g., multinomial $\{\pi_{ij}\}$), instead focus on association parameters.

Lindley (1964):Used Dirichlet prior distribution (and limiting improper prior) for multinomial sampling.

Contrasts of log cell probabilities, such as log odds ratio, have approximate (large-sample) joint normal posterior distribution.

However, normally sensible to model cell probabilities, rather than treat as exchangeable.

First Bayesian approach for loglinear models focused on parameters of saturated model (Leonard 1975). For cell counts $\{y_{ij}\}$,

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

- Exchangeability within each set of loglinear parameters more sensible than exchangeability of multinomial probabilities (as with Dirichlet prior).
- Treated row effects, column effects, and interaction effects as a priori independent.
- For each, given μ, σ^2, first-stage prior takes them independent and normal with mean μ and variance σ^2.
- At the second stage, each normal mean is assumed to have improper uniform distribution over the real line, and σ^2 assumed to have an inverse chi-squared distribution.
- For computational convenience, parameters estimated by joint posterior modes rather than posterior means.
- Analysis shrinks log counts toward ML fit of independence model.

3.4 Extensions to multi-dimensional tables

Knuiman and Speed (1988) generalized Leonard (1975) by taking multivariate normal prior for all parameters collectively
(rather than univariate normals on individual parameters).

Forster (2004): Conditions for prior distributions such that marginal inferences equivalent for Poisson and multinomial models.

Parameter governing overall size of cell means (which disappears after the conditioning that yields the multinomial model) has improper prior.

Derives necessary and sufficient conditions for posterior to be proper, and relates to conditions for ML estimates to be finite.

3.5 Graphical models

Graphical models have conditional independence structure summarized by graph with vertices for variables and edges between vertices to represent conditional association
(no edge: conditional indep.).

- Joint cell probabilities expressed in terms of marginal and conditional probabilities.
- Independent Dirichlet prior distributions for them induce independent Dirichlet posterior distributions.
- Dawid and Lauritzen (1993) introduce probability distribution over set of such graphs. Special case includes **hyper Dirichlet** distribution that is conjugate for multinomial sampling and implies certain marginal probabilities have Dirichlet distribution.
- Madigan and Raftery (1994), Madigan and York (1995): Used this family for model comparison and averaging in constructing posterior distributions for summary measures
- Giudici (1998): Prior dist. over space of graphical models to smooth sparse contingency tables (Smoothing maintains association structure imposed by the graphical models)

4 Tests and Confidence Intervals in Two-Way Tables

4.1 Confidence intervals for association parameters

With 2×2 tables, parameters of usual interest are $\pi_1 - \pi_2$, relative risk π_1/π_2, and odds ratio $[\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)]$.

Most common to assume y_i has bin(n_i, π_i) distribution,
use independent beta($\alpha_i, \beta_i)$ prior for $\pi_i,\ i = 1, 2$.

Alternatively, could use correlated prior for (π_1, π_2), e.g., bivariate normal for [logit(π_1), logit(π_2)].

With independent beta priors, Nurminen and Mutanen (1987) gave integral expressions for posterior distributions of these measures.

Hashemi, Nandrum and Goldberg (1997) found Bayesian highest posterior density (HPD) confidence interval for these parameters.

HPD interval lacks invariance under parameter transformation.

If (L, U) a 95% HPD interval using posterior of odds ratio, then 95% HPD interval using posterior of inverse of odds ratio (relevant if we reverse identification of groups being compared) is not (1/U, 1/L).

Alternative 100$(1-\alpha)$% “tail interval” consists of values between $\alpha/2$ and $(1-\alpha/2)$ quantiles. (Longer, but invariant)

Agresti and Min (2005) evaluate Bayesian confidence intervals for association parameters.

- Coverage probabilities vary substantially according to choice of prior distribution, even for moderate sample sizes
- For good coverage performance (in frequentist sense) over entire parameter space, best to use quite diffuse priors (recommend Jeffreys priors)
- Provide R functions for tail confidence intervals for association measures with independent beta priors

  (www.stat.ufl.edu/$\sim$aa/cda/software.html)

4.2 Tests comparing two independent binomial samples

Novick and Grizzle (1965) used independent beta priors to find posterior $P(\pi_1 > \pi_2)$, with application to sequential clinical trials

Altham (1969) discussed Bayesian testing for 2×2 tables. For multinomial cell counts $\{y_{ij}\}$ and posterior Dirichlet distribution with $\{\alpha^{'}_{ij} = \alpha_{ij} + y_{ij}\}$,

$$P(\pi_{11}\pi_{22}/\pi_{12}\pi_{21} < 1|\{y_{ij}\}) =$$

$$\sum_{s=\max(\alpha^{'}_{21}-\alpha^{'}_{12},0)}^{\alpha^{'}_{21}-1} \binom{\alpha^{'}_{+1}-1}{s}\binom{\alpha^{'}_{+2}-1}{\alpha^{'}_{2+}-1-s}/\binom{\alpha^{'}_{++}-2}{\alpha^{'}_{1+}-1}$$

- Equals one-sided P-value for Fisher's exact test with $H_a : \pi_{11}\pi_{22}/\pi_{12}\pi_{21} > 1$, when $\alpha_{11} = \alpha_{22} = 0$ and $\alpha_{12} = \alpha_{21} = 1$ (improper)
- i.e., ordinary P-value for Fisher's exact test corresponds to Bayesian P-value with conservative prior distribution
- If $\{\alpha_{ij} = \gamma\}$, with $0 \leq \gamma \leq 1$, Bayesian P-value $<$ Fisher P-value, and difference between the two no greater than null probability of observed data

Howard (1998) shows that with Jeffreys beta priors, posterior $P(\pi_1 \leq \pi_2)$ approximates one-sided P-value for large-sample z test using pooled variance

5 Regression Models for Categorical Responses

5.1 Binary regression

For binary data $\{y_i\}$, link function $g(\cdot)$,

$g[P(y_i = 1)] = \mathbf{x}_i^{'}\boldsymbol{\beta}$, where $\{y_i, i = 1, \ldots, n\}$ independent,

Zellner and Rossi (1984) derive approximate posterior densities with prior on $\boldsymbol{\beta}$ improper uniform or multivariate normal.

Wong and Mason (1985): Hierarchical logistic regression modeling, in multilevel structure.

Probit regression: Computational simplicities in connecting to underlying normal regression model. See Albert and Chib (1993), with extensions to ordered multinomial responses.

Bedrick, Christensen, and Johnson (1996, 1997): Elicit beta priors on $P(y_i = 1)$ at selected values of covariates. These induce prior on model parameters by one-to-one transformation.

- Easier to formulate priors for $P(y_i = 1)$ than for $\boldsymbol{\beta}$
- Can apply priors to different link functions, whereas prior specification for $\boldsymbol{\beta}$ would depend on link function.

Dey, Ghosh, and Mallick (2000): Edited collection of articles that provide Bayesian analyses for GLMs.

Chaloner and Larntz (1989): Determining optimal design for experiments using logistic regression

Zocchi and Atkinson (1999): design for multinomial logistic models

5.2 Multi-category responses

Ordinal response: Logit and probit models for cumulative probabilities, such as

logit$[P(y_i \leq j)] = \alpha_j + \mathbf{x}_i^{'}\boldsymbol{\beta},\ j = 1, ..., c-1.$

Johnson and Albert (1999): Models motivated by underlying logistic or normal latent variables.

Nominal response: Multinomial logit and probit models, such as

$$\log[P(y_i = j)/P(y_i = c)] = \alpha_j + \mathbf{x}_i^{'}\boldsymbol{\beta}_j.$$

Daniels and Gatsonis (1997): Generalize Wong and Mason (1985) hierarchical approach to analyze variations in utilization of alternative cardiac procedures in study of Medicare patients who had suffered myocardial infarction.
Multivariate t prior for regression parameters.

5.3 Multivariate response extensions and other GLMs

For multivariate correlated ordinal (or binary) responses, Chib and Greenberg (1998) consider multivariate probit model.

- Multivariate normal latent random vector defines categories of observed discrete variables.
- Correlation among categorical responses induced through covariance matrix for underlying latent variables.

O’Brien and Dunson (2004) formulate multivariate logistic distribution incorporating correlation parameters and having marginal logistic distibutions.

Zeger and Karim (1991) fit generalized linear mixed models using Bayesian framework with priors for fixed and random effects.

6 Bayesian Computation

For GLMs with canonical link function and normal or conjugate priors, posterior joint and marginal distributions are log-concave (O’Hagan and Forster 2004).

Computational methods for approximating posterior distributions by simulating samples from them include:

- Importance sampling (Zellner and Rossi 1984)
- Markov chain Monte Carlo methods such as Gibbs sampling (Gelfand and Smith 1990) and Metropolis-Hastings algorithm (Tierney 1994).

e.g., Epstein and Fienberg (1991) employed Gibbs sampling to estimate posterior density of cell probabilities (a finite mixture of Dirichlet densities), not merely posterior means.

For reviews, see Andrieu, Doucet, and Robert (2004), text by O’Hagan and Forster (2004, Sections 12.42-46)

7 Final Comments

Now quite a large body of Bayesian literature for CDA.

But, Bayesian inference does not seem to be commonly used yet in practice for basic categorical data analyses.

- For multi-way contingency table analysis, plethora of parameters for multinomial models necessitates substantial prior specification.
- Need to specify and understand prior distributions on GLM parameters may be daunting, especially for hierarchical models.
- Approach of eliciting prior distributions on probability scale at selected values of covariates may be useful, as in Bedrick, Christensen and Johnson (1996, 1997).
- May partly reflect the absence of Bayesian procedures in the primary software packages.

Currently Bayesian approaches for categorical data seem to suffer from not having standard default starting point.