Bayesian Statistics

Multinomial model

Nan Lin
Department of Mathematics
Washington University in St. Louis

Multinomial distribution

- The multinomial model is a generalization of the binomial model for the case where the response variable can take on more than two values.
  - Polytomous responses from survey
  - Examples:
    - strongly agree, agree,..., strongly disagree
- Data consist in a $K \times 1$ vector of counts y. The jth element of y is the number of sample units for which the response variable was equal to the jth value of the outcome.
- Example: If out of 100 rural intersections we have that 50 experienced no crashes over a year, 30 exhibited one crash, 15 exhibited 2 crashes and 5 exhibited more than 2 crashes, our data would be a $K = 4 \times 1$ vector $Y = [50\ 30\ 15\ 5]'$.

Multinomial distribution

- $(\theta_1, \dots, \theta_K)$ are the probabilities associated with each of the K possible outcomes

$$f(y|\theta_1, \dots, \theta_K) = \frac{n!}{y_1! \cdots y_K!} \theta_1^{y_1} \cdots \theta_K^{y_K},$$

$$\sum_{i=1}^{K} \theta_i = 1, \sum_{i=1}^{K} y_i = n$$

Prior

- Traffic engineers know that in rural intersections, the chances of 0 or 1 crashes are higher than the chances of 3 or more crashes.
- It would be good to have a prior that permits assigning different prior probabilities to each of the K outcomes.
  - When $K = 2$, we may use the beta prior for the binomial model.
  - For $K > 2$, we need a multivariate prior for probabilities.
    - Dirichlet distribution

Dirichlet distribution: a conjugate prior

- $\theta \sim Dirichlet(\alpha_1, \alpha_2, \dots, \alpha_K)$
- A generalization of the beta distribution

$$f(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}, 0 < \theta_j < 1, \alpha_j > 0 \text{ for all j}$$

where $B(\alpha) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{K} \alpha_j)}$

  - α_j can be thought as "prior counts" associated with the jth outcome
  - $\alpha_0 = \sum_{j=1}^{K} \alpha_j$ is then a "prior sample size"

Conjugacy

- Posterior $\theta|y$

$$f(\theta|y) \propto f(y|\theta)f(\theta) \propto \prod_{j=1}^{K} \theta_j^{y_j} \times \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

$$= \prod_{j=1}^{K} \theta_j^{y_j + \alpha_j - 1}$$

- That is, $\theta|y \sim Dirichlet(y_1 + \alpha_1, \ldots, y_K + \alpha_K)$
- Posterior mean

$$\begin{aligned} E(\theta_j|y) &= \frac{\alpha_j + y_j}{\alpha_0 + n} \\ &= \frac{\text{“\#” of obs. of jth outcome}}{\text{“total” \# of obs}}. \end{aligned}$$

Example

- Example: If out of 100 rural intersections we have that 50 experienced no crashes over a year, 30 exhibited one crash, 15 exhibited 2 crashes and 5 exhibited more than 2 crashes, our data would be a $K = 4 \times 1$ vector $Y = [50\ 30\ 15\ 5]'$.
- Suppose that the traffic engineer thinks a prior that the probabilities associated with the four crash levels are 0.6, 0.3, 0.08 and 0.02
  - If he has high confidence on these guesses, he may use $\alpha_0 = 200$, which leads to $\alpha_1 = 120, \alpha_2 = 60, \alpha_3 = 16, \alpha_4 = 4$
  - If he has low confidence on these guesses, he may use $\alpha_0 = 20$, which leads to $\alpha_1 = 12, \alpha_2 = 6, \alpha_3 = 1.6, \alpha_4 = 0.4$

Jeffrey's prior

- Likelihood for the multinomial model

$$\log f(y|\theta) = const + \sum_{j=1}^{k} y_j \log \theta_j$$

  - $\frac{\partial}{\partial \theta_j} \log f(y|\theta) = \frac{y_j}{\theta_j}$
  - $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(y|\theta) = \begin{cases} -\frac{y_j}{\theta_i^2}, i = j \\ 0, i \neq j \end{cases}$

- Jeffrey's prior
  - $\pi(\theta) \propto \prod_j^K \theta_j^{-\frac{1}{2}}$
  - Dirichlet with all $\alpha_i = \frac{1}{2}$

How to sample from a Dirichlet distribution?

$$Z_i \sim \text{Gamma}(\alpha_i, \beta) \text{ independently,}$$

$$\downarrow$$

$$S = \sum_{i=1}^{K} Z_i \sim \text{Gamma}\left(\sum_{i=1}^{K} \alpha_i, \beta\right)$$

$$\downarrow$$

$$V = (V_1, \cdots, V_K) = (Z_1/S, \cdots, Z_K/S) \sim \text{Dir}(\alpha_1, \cdots, \alpha_K)$$

Marginal distribution of the Dirichlet distribution is $Beta(\alpha_i, \alpha_0 - \alpha_i)$

1. Draw $x_1, x_2, \ldots, x_K$ one from each independent gamma distributions with parameters δ and $\alpha_j + y_j$, for any common δ.
2. Set $\theta_j = x_j / \sum_{i=1}^{K} x_i$.

Properties of Dirchlet distribution

- Expectation

$$\mathrm{E}[\theta_i] = \frac{\alpha_i}{\sum_j \alpha_j}$$

- Variance

$$Var(\theta_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

- Covariance

$$Cov(\theta_i, \theta_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

- A uniform density is obtained by setting $\alpha_i = 1$ for all i. This distribution assigns equal probability to each θ_i.
- If setting $\alpha_i = 0$ for all i, it is equivalent to placing a uniform prior on $\ln \theta_j$
- ..\plotDirichlet.m

Application: 2×2 Contingency table

A model for a two by two contingency table:

	Intervention New	Intervention Control	
Death	$\theta_{1,1}$	$\theta_{1,2}$	
No death	$\theta_{2,1}$	$\theta_{2,2}$	
			N

Data model:

$$p(y|\theta) \propto \prod_{j=1}^{2} \prod_{i=1}^{2} \theta_{i,j}^{y_{i,j}}$$

Prior model:

$$p(\theta) \propto \prod_{j=1}^{2} \prod_{i=1}^{2} \theta_{i,j}^{a_{i,j}-1}$$

Posterior model:

$$p(\theta|y) \propto \prod_{j=1}^{2} \prod_{i=1}^{2} \theta_{i,j}^{y_{i,j}+a_{i,j}-1}$$

Example: GREAT Trial

- Aim of study: to compare a new drug treatment to be given at home as soon as possible after a myocardial infarction and placebo.
- Outcome measure: Thirty-day mortality rate under each treatment, with the benefit of the new treatment measured by the odds ratio, i.e., the ratio of the odds of death following the new treatment to the odds of death on the conventional:
  - If $OR < 1$, the new treatment is in favor.

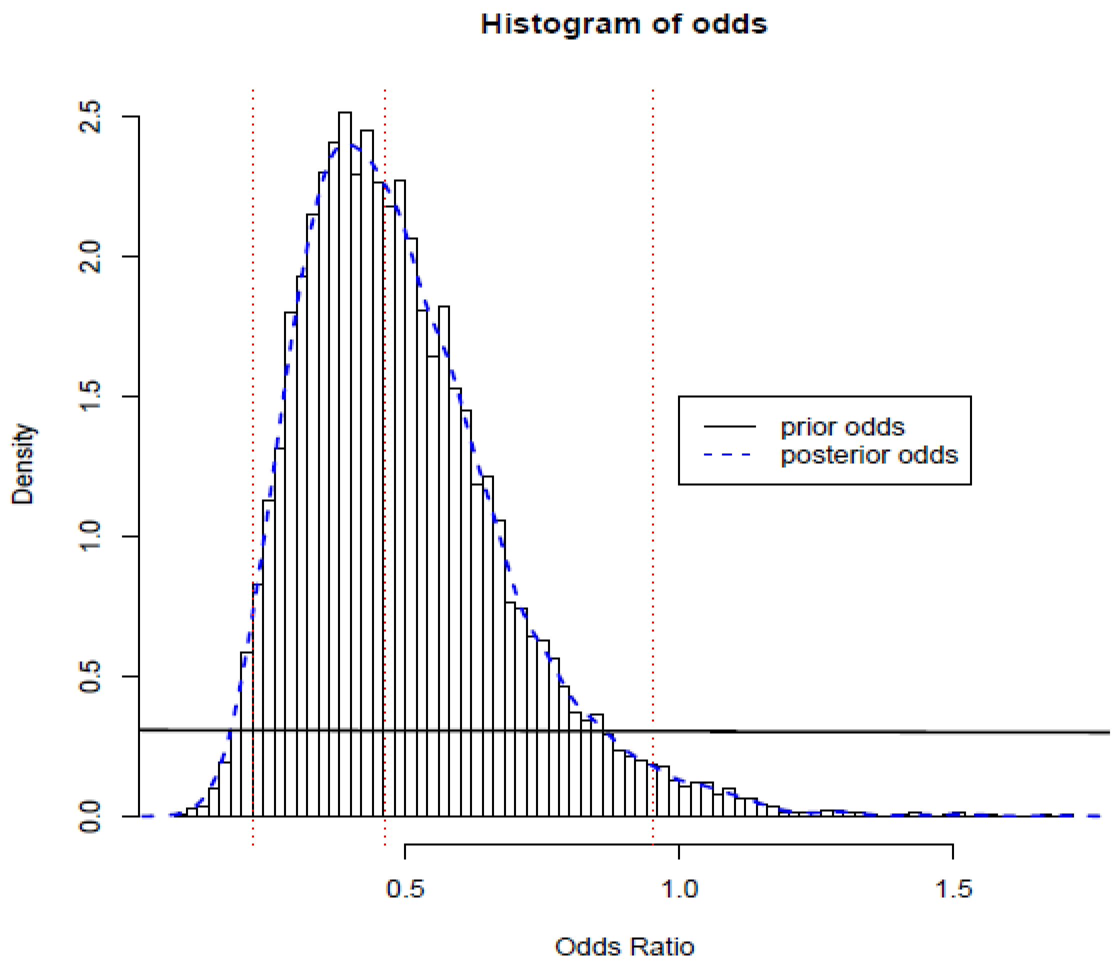
$$\Psi = \frac{\theta_{1,1}\theta_{2,2}}{\theta_{1,2}\theta_{2,1}}.$$

Example: GREAT Trial

	Intervention New	Control	
Death	13	23	36
No death	150	125	275
	163	148	311

- Uniform prior:
  - Dirichlet with parameters $a_{1,1} = a_{1,2} = a_{2,1} = a_{2,2} = 1$
- Posterior inference for the odds ratio

  - Simulate a large number of values for the vector θ from its posterio $\Psi = \dfrac{\theta_{1,1}\theta_{2,2}}{\theta_{1,2}\theta_{2,1}}$.
  - For each simulated value calculate Ψ^*.
  - Inference for Ψ is based on the histogram of Ψ^*.


Histogram of odds
Density
0.0
0.5
1.0
1.5
2.0
2.5
prior odds
posterior odds
0.5
1.0
1.5
Odds Ratio

Example: GREAT Trial

- Consider testing $H_0 : \Psi \geq 1$ vs. $H_1 : \Psi < 1$
- Frequentist: Fisher's exact test
  - p-value = 0.02817
- Bayesian: posterior probability $P(H_0|data)$

```
> sum(odds > 1)/10000
[1] 0.0175
```

- The two do not agree even though the prior is uniform

Example: GREAT Trial

- If we use a different Dirichlet prior with parameters

$$a_{1,1} = 2, a_{1,2} = 1, a_{2,1} = 1 \text{ and } a_{2,2} = 2$$

- that is,

$$p(\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2}) \propto \theta_{1,1}\theta_{2,2}$$

- posterior probability $P(H_0|data) \approx 0.0277$

- Fisher's exact test does not correspond to a 'non-informative' prior
  - Some weak information is implied