Chapter 10

Multinomial model

10.1 Introduction

The binomial distribution that we have seen earlier in the course allowed two possible outcomes for an observation, e.g., dead or alive; yes or no; works or doesn't work. The multinomial distribution generalizes the binomial to allow more than two possible outcomes. For example, we may be interested in characterizing people by their interest in the arts, with interest represented by: high, medium, low, nonexistent.

The multinomial distribution is used to describe data for which each observation is one of k multiple outcomes. If y is the vector of counts of the number of observations of each outcome, then

$$p(y \mid \theta) \propto \prod_{j=1}^{k} \theta_j^{y_i}$$

where the sum of the probabilities, $\sum_{j=1}^{k} \theta_j$ is 1. The distribution is typically thought of as implicitly conditioning on the number of observations, $\sum_{j=1}^{n} y_j = n$. The conjugate prior distribution is a multivariate generalization of the beta distribution known as the Dirichlet,

$$p(\theta \mid \alpha) \propto \theta_j^{\alpha_j - 1},$$

where the distribution is restricted to nonnegative $\theta_j's$ with $\sum_{j=1}^{k} \theta_j = 1$; the resulting posterior distribution for the

$$\theta_j \mid y, \alpha \sim Dirichlet(\alpha_j + y_j)$$

.

The prior distribution can be interpreted as containing equivalent information $\sum \alpha_j$ observations, with α_j observations of the *jth* outcome category. As in the binomial there are several plausible non-informative Dirichlet prior distributions. A uniform density is obtained by setting $\alpha_j = 1$ for all j; this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^{k} \theta_j = 1$. Setting $\alpha_j = 0$ for all j results in an improper prior distribution that is uniform in the $log(\theta_j)'s$. The resulting posterior distribution is proper if there is at least one observation in each of the k categories, so that each component of y is positive.

10.2 Example: Political poll

This example is from Gelman et al. (1995), p. 76.

Consider a sample survey question with three possible responses. In late October, 1988, a survey was conducted by CBS News of 1447 adults in the United States to find out their preferences in the upcoming Presidential election. Out of 1447 persons, $y_1 = 727$ supported George Bush, $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or expressed no opinion. Assuming no other information on the respondents, the 1447 observations

are exchangeable. If we also assume simple random sampling (that is, 1447 names drawn "out of a hat"), then the data (y_1, y_2, y_3) follow a multinomial distribution, with parameters $(\theta_1, \theta_2, \theta_3)$, the proportion of Bush supporters, Dukakis supporters and those with no opinion in the surveyed population. An estimand of interest is $\theta_1 - \theta_2$ the population difference in support for the two major candidates.

With a noninformative uniform prior distribution on θ, $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is Dirichlet(728, 584, 138). We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3)$ from the posterior Dirichlet distribution and compute $\theta_1 - \theta_2$ for each. We do this, using the following R code:

```
library(MCMCpack)   # Must have MCMCpack downloaded for this to work
 election.post=rdirichlet(1000,c(728,584,138))  # draw 1000 theta1, theta2, theta3
 election.post[1:3, ]    # view the first three rows.
          [,1]      [,2]       [,3]          # column1 is theta1, column 2 is theta2, etc
[1,] 0.5120700 0.4190138 0.06891619
[2,] 0.5115476 0.3791360 0.10931642
[3,] 0.5032504 0.3962764 0.10047312
 diffBush.Duk=election.post[,1]-election.post[,2]  # compute difference column 1-column 2
 diffBush.Duk[1:3]  # view: 0.09305628=0.5120700-0.4190138 from matrix above and so on
[1] 0.09305628 0.13241167 0.10697400    # first three
 hist(diffBush.Duk,main="Histogram of theta1-theta2 ")
 sum(diffBush.Duk>0)/1000  # post prob that Bush had more support than Dukakis
[1] 1      # The posterior probability of that event is 1
```

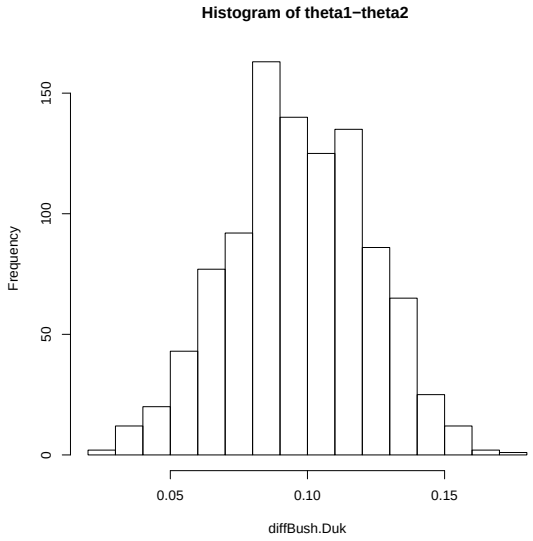

Figure 10.1: Histogram of values of $(\theta 1 - \theta_2)$ for 1000 simulations from the posterior distribution for the election polling example. Gelman et al. 1995, p.76

dirichlet.pdf}

10.3 Two by two tables

We also are interested often in the proportion of observations that satisfy two characteristics, for example: the proportion of victims that were exposed to lead poisoning and also died. As such, these proportions are interpreted as joint

probabilities of having the characteristic. Surveys also help us collect this type of data. When faced, for example, with the kind of data mentioned, we would have four possible joint proportions: the proportion of observations that were exposed to lead and did not die; the proportion of observations that were not exposed to lead and died; the proportion of observations that were not exposed to lead and did not die. Thus, we could interpret these data as multinomial data and analyze it as such.

10.3.1 Is Victimization Chronic?

How many times more likely it is that a household will be victimized the second time if they were victimized the first time than if they were not? Values of the odds ratio larger than 1 would indicate that "victimization is chronic." To set up a Bayesian scenario to answer this question, we will first set up a table that summarizes the information we got from the Crime Victimization Survey.

	2nd visit (1994)	
1st visit (1988)	Crime-free	Victims
Crime-free	4721	74
Victims	94	16

Following Kadane **??**, we suppose that the data in these four cells are multinomially distributed. The probabilities that we want to estimate are: θ_1 that a household is crime-free in both periods, θ_2 that it is crime-free in period 1 and victimized in period 2, θ_3 that it is victimized in period 1 and crime-free in period 2, and θ_4 that it is victimized in both periods. $\sum_{i=1}^{4} \theta_i = 1$ and $\theta_i > 0$. Let y_i be the number of observations corresponding to each of the categories $i = 1, 2, 3, 4$. $y = (4721, 74, 94, 16)$.

For prior, we assume an informative Dirichlet prior with parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. The prior distribution can be interpreted as containing equivalent information to $\sum_{i=1}^{4} \alpha_i$ observations with α_i observations of the *jth* outcome category. Kadane had this information when he did his study.

	2nd visit (1994)		
1st visit (1988)	Crime-free	Victims	Non-response
Crime-free	392	55 33	
Victims	76	38	9
Non-response	31	7	115

Thus, we set $\alpha_1 = 392, \alpha_2 = 55, \alpha_3 = 76, \alpha_4 = 38$.

The posterior distribution of the θ parameters is Dirichlet with parameters $\alpha_i + y_i$.

$$p(\theta \mid y) \propto \prod_{i=1}^{4} \theta_i^{y_i + \alpha_i - 1}$$

Thus, we will be drawing random numbers from a Dirichlet posterior distribution with parameters $(4721 + 392, 74 + 55, 94 + 76, 16 + 38) = (5113, 129, 170, 54)$. We sample $(\theta_1, \theta_2, \theta_3, \theta_4)$ from Dirichlet(5113,129,170,54). And we will be interested in the odds ratio or cross product $(\theta_1\theta_4)/(\theta_2\theta_3)$.

```
library(MCMCpack)   # Must have MCMCpack downloaded for this to work
 victim.post=rdirichlet(1000,c(5113,129,170,54))  # draw 1000 theta1, theta2, theta3, theta4
 victim.post[1:3, ]    # view the first three rows.
           [,1]       [,2]       [,3]        [,4]
[1,] 0.9366960 0.02248314 0.03122189 0.009598939
[2,] 0.9372178 0.02531207 0.03035691 0.007113238
[3,] 0.9434543 0.02324296 0.02564050 0.007662293
  odds.ratio=victim.post[,1]*victim.post[,4]/victim.post[,2]*victim.post[,3]
 odds.ratio[1:3]
 [1] 0.012486026 0.007995357 0.007974706
  hist(odds.ratio,main="Histogram of odds of being victimized twice ")
  summary(odds.ratio)
  sum(odds.ratio>1)
```

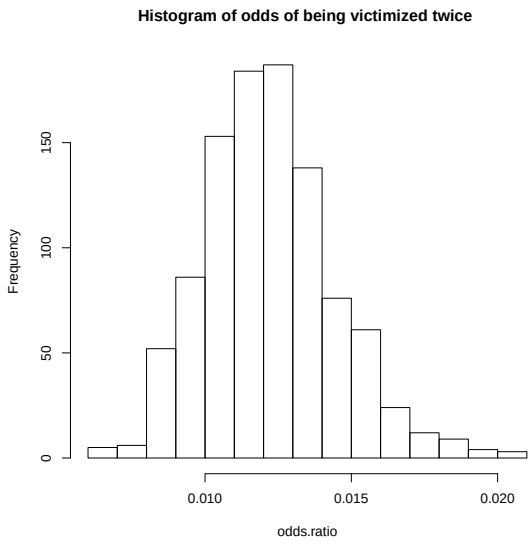

Figure 10.2: Histogram of odds ratio for victimization {fig:victimiza

We can see that none of the posterior odd ratios are larger than 1. Thus, we conclude that victimization is not chronic.

Exercise

Try to see whether the results are sensitive to the choice of prior distribution by altering the Dirichlet prior. For example, try a non-informative Dirichlet prior.

10.4 Many survey questions at once

In complicated problems - for example, analyzing the results of many survey questions simultaneously- the number of multinomial categories, and thus parameters, becomes so large that it is hard to usefully analyze a dataset of moderate size without additional structure in the model. Formally, additional information can enter the analysis through the prior distribution or the sampling model. An informative prior distribution might be used to improve inference in complicated problems, using the ideas of hierarchical models, which is the subject of our next and future lectures. Alternatively, log linear models can be used to impose structure on multinomial parameters that result from cross-classifying several survey questions. Section 14.7 of Gelman et al (1995), contains more information on this.