Statistics 520, Fall 2025
Assignment 7

1 Study Description

Epidemiologists are interested in the relations between disease status and the presence or absence of various risk factors. There are a number of study designs used to collect such data including what are known as *cohort* studies, *case-control* studies and *cross-sectional* studies. In a cross-sectional study a sample of N subjects are taken and each subject is observed for the presence or absence of both the risk factor and the disease under investigation. The simplest structure for data from this type of study can be represented with a 2×2 contingency table,

	Risk Factor		
Disease	R	R^c	
D	Y_{11}	Y_{01}	$Y_{11} + Y_{01}$
D^c	Y_{10}	Y_{00}	$Y_{10} + Y_{00}$
	$Y_{11} + Y_{10}$	$Y_{01} + Y_{00}$	N

Table 1: Protoype of a Cross-Sectional Study Design.

In Table 1, Y_{11}, Y_{10}, Y_{01} and Y_{00} are random variables giving counts for the cells of the table, and N is the sample size. For the random variables, the first index is risk factor and the second index is disease status.

2 Relative Risk

A number of quantities are used to reflect the association between risk and disease. In a cross-sectional study, one of the more useful of these is called *relative risk*. Let

θ_{ij} denote the probability an individual falls into cell ij in the contingency table, $i, j = 0, 1$. Relative risk is defined as,

$$RR = \frac{Pr(D|R)}{Pr(D|R^c)} = \frac{\theta_{11}(\theta_{01} + \theta_{00})}{\theta_{10}(\theta_{11} + \theta_{10})}. \tag{1}$$

Sample estimates of the cell probabilities are $\hat{\theta}_{ij} = y_{ij}/N$ and the sample version of RR is then,

$$\hat{RR} = \frac{y_{11}(y_{01} + y_{00})}{y_{01}(y_{11} + y_{10})}. \tag{2}$$

3 Probability Models

There are a number of probability models that can be assigned to a problem with the structure of Table 1. One of those is to take the random variables to follow a multinomial probability distribution. Let $\boldsymbol{y} = (y_{11}, y_{10}, y_{01}, y_{00})$ and $\boldsymbol{\theta} = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ and assume the probability mass function of $\boldsymbol{Y}$ is,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = K(\boldsymbol{y})\, \theta_{11}^{y_{11}}\ \theta_{01}^{y_{01}}\ \theta_{10}^{y_{10}}\ \theta_{00}^{y_{00}}, \tag{3}$$

subject to

$$Y_{11} + Y_{10} + Y_{01} + Y_{00} = N$$
$$\theta_{10} + \theta_{10} + \theta_{01} + \theta_{00} = 1.$$

This multinomial probability mass function has been written in what could be called degenerate form, as there are only 3 free random variables and parameters due to the constraints. The mass function (3) actually describes a three-dimensional distribution, but for our purposes the degenerate form will make notation easier. Because the θ_{ij} in this model are cell probabilities, RR based on the multinomial model is given by (1).

Another possibility is to treat N as a random variable as well as the cell counts. It can be shown that an equivalent model form to the multinomial is given by four independent Poisson random variables,

$$Y_{11} \sim Po(\lambda_{11})$$
$$Y_{10} \sim Po(\lambda_{10})$$
$$Y_{01} \sim Po(\lambda_{01})$$
$$Y_{00} \sim Po(\lambda_{00}),$$

and $N \sim Po(\lambda_N)$ as well, with $\lambda_N = \lambda_{11} + \lambda_{10} + \lambda_{01} + \lambda_{00}$. The multinomial formulation is the same as the Poisson formulation conditioned on the value of N. Here, the λ_{ij} are expected cell counts but the RR is analogous to that of the multinomial model, being

$$RR = \frac{\lambda_{11}(\lambda_{01} + \lambda_{00})}{\lambda_{01}(\lambda_{11} + \lambda_{10}}.$$

The expected probability for cell ij under the Poisson model is λ_{ij}/λ_N and computing RR with these cell probabilities gives the same result.

4 Bayesian Analysis

Our objective is to conduct a Bayesian analysis of a study with the structure of Table 1, under both multinomial and Poisson probability models. To this end, we will need prior distributions for the parameters θ_{11}, θ_{10}, θ_{01}, and θ_{00} in the multinomial model and λ_{11}, λ_{10}, λ_{01} and λ_{00} in the Poisson model. Assign the parameters of the multinomial model a Dirichlet distribution. Let $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{10}, \gamma_{01}, \gamma_{00})$ where $\gamma_{ij} > 0$ for $i, j = 0, 1$, and take,

$$\pi(\boldsymbol{\theta}|\boldsymbol{\gamma}) = C(\boldsymbol{\gamma})\, \theta_{11}^{\gamma_{11}-1}\, \theta_{10}^{\gamma_{10}-1}\, \theta_{01}^{\gamma_{01}-1}\, \theta_{00}^{\gamma_{00}-1}. \tag{4}$$

Note that (4) is still subject to $0 < \theta_{ij} < 1$ and $\sum_i \sum_j \theta_{ij} = 1$.

For the Poisson model, we will assign four independent gamma prior distributions.

Let $\alpha_{ij} > 0$, $\beta_{ij} > 0$ for $i, j = 01$ and take, for $\lambda_{ij} > 0$,

$$\pi(\lambda_{11}|\alpha_{11}, \beta_{11}) = \frac{\beta_{11}^{\alpha_{11}}}{\Gamma(\alpha_{11})} \lambda_{11}^{\alpha_{11}-1} \exp(-\beta_{11}\lambda_{11}),$$
$$\pi(\lambda_{10}|\alpha_{10}, \beta_{10}) = \frac{\beta_{10}^{\alpha_{10}}}{\Gamma(\alpha_{10})} \lambda_{10}^{\alpha_{10}-1} \exp(-\beta_{10}\lambda_{10}),$$
$$\pi(\lambda_{01}|\alpha_{01}, \beta_{01}) = \frac{\beta_{01}^{\alpha_{01}}}{\Gamma(\alpha_{01})} \lambda_{01}^{\alpha_{01}-1} \exp(-\beta_{01}\lambda_{01}),$$
$$\pi(\lambda_{00}|\alpha_{00}, \beta_{00}) = \frac{\beta_{00}^{\alpha_{00}}}{\Gamma(\alpha_{00})} \lambda_{00}^{\alpha_{00}-1} \exp(-\beta_{00}\lambda_{00}). \tag{5}$$

For prior parameters, in the multinomial model use $\gamma_{11} = 0.10$, $\gamma_{10} = 0.75$, $\gamma_{01} = 0.10$ and $\gamma_{00} = 0.75$. For the Poisson model use $\alpha_{11} = 0.5$, $\alpha_{10} = 3.0$, $\alpha_{01} = 0.5$, $\alpha_{00} = 3.0$, $\beta_{11} = 5.0$, $\beta_{10} = 4.0$, $\beta_{01} = 5.0$ and $\beta_{00} = 4.0$. These values were chosen so that joint cell probabilities have expected values $\theta_{11} = 0.06$, $\theta_{10} = 0.44$, $\theta_{01} = 0.06$ and $\theta_{00} = 0.44$. These values also imply that in the population $Pr(D|R) = Pr(D|R^c) = 0.12$ and relative risk is $RR = 1.0$, which would essentially correspond to a prior opinion that the risk factor does not influence disease status.

5 An Application

In late 2001 to early 2002 there was an outbreak of chickenpox in an elementary school in Oregon. One issue that was identified as possibly being involved was the number of *breakthrough* cases that occurred. Breakthrough cases are defined as incidents of the disease (here chickenpox) in individuals who have been vaccinated against the disease. To examine whether boosters for chickenpox are effective, a study was conducted from October 30 2001 to January 27 2002 (Tugwell et al. 2004). There were 422 students in the school. The number who had not already had chickenpox (and thus had not gained immunity) but who had been vaccinated was 211. The study then eliminated students who were in classrooms in which no cases had been observed, leaving 152 students who had been vaccinated, had not previously contracted chickenpox, but were in a classroom in which exposure to infected individuals had occurred. The risk factor examined was whether vaccination had occurred more than 5 years in the past, versus less than 5 years. The data for the 152 students in presented in Table 2.

	Time Since Vaccination	
Chickenpox	$>$ 5 years	$<$ 5 years
Yes	15	3
No	50	84

Table 2: Data on chickenpox and time since vaccination.

Reference:

Tugwell, B.D., Lee, L.E., Gillette, H., Lorber, E.M. and Cieslak, P.R. (2004), Chickenpox outbreak in a highly vaccinated school population. *Pediatrics***113**: 455-459.

6 The Assignment

Your assignment is to conduct a Bayesian analyis of relative risk for these data. You will do so using both the Multinomial probability model and the Poisson probability model.

1. (10 pts.) Using prior distributions and prior parameter values as given in Section 4, Bayesian Analysis, derive joint posterior distributions for $\boldsymbol{\theta}$ in the multinomial model and $\boldsymbol{\lambda}$ in the Poisson model.

2. (5 pts.) Using the data of Table 2 give posterior expected values for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, and the cell probabilities that would correspond to these means.

3. (5 pts.) Evaluate relative risk RR at the expected values of the posterior distributions you derived in exercise 1 and compare to the observed or sample-

based RR based only on the observed cell counts. Briefly explain why this is not the posterior expected value of RR (although we hope it is not far off).

4. (15 pts.) Find the posterior distributions of RR under both the multinomial and Poisson models. Give summary and 95% credible intervals. Produce some type of a graphical display of the posterior distributions. How do the results compare for the multinomial and Poisson models? How do the results compare to the initial sample version of RR in expression (2)? What would you conclude about the relation between time since vaccination and the chances of getting chickenpox in this population of school children? For example, what is your posterior probability that the risk factor of having more than 5 years since vaccination is positively related to the chance of contracting chickenpox in these students?