Markov chain Monte Carlo I
Introduction

Vladimir Minin, Kari Auranen, M. Elizabeth Halloran

SISMID 2021
University of Washington
Seattle, WA, USA

July 7, 2021

Introduction

- Bayesian inference
- Motivating examples
- Prior distributions

Transmission Probability

- Full probability model
- Varying data and prior information
- Prediction

Simple Gibbs sampler

- Chain binomial model
- Full conditionals

Notation

- Let θ denote unobservable vector quantities or population parameters of interest (such as the per-contact probability of transmission from an infected person to a susceptible person)
- Let y denote the observed data (such as the number of infecteds and uninfecteds who were exposed)
- Let $\tilde{y}$ or y_{n+1} denote unknown, but potentially observable quantities, such as the infection outcome of the next contact between an infected person and a susceptible person.

Bayesian Inference

- Bayesian statistical conclusions about a parameter θ, or unobserved data $\tilde{y}$, are made in terms of probability statements.
- These probability statements are conditional on the observed value of y.
- They can be written as $p(\theta|y)$ or $p(\tilde{y}|y)$
- We also implicitly condition on known values of any covariates.

Bayes' rule

- To make probability statements about θ given y, we start with a model providing the joint probability distribution for θ and y.
- The joint probability mass or density function can be written as a product of two densities that are often referred to as the prior distribution $p(\theta)$ and the sampling distribution (or data distribution) $p(y|\theta)$

$$p(\theta, y) = p(\theta)p(y|\theta)$$

- Simply conditioning on the known value of the data y, using the basic property of conditional probability known as Bayes' rule, yields the posterior density

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

Bayes' rule, cont'd

- Simply conditioning on the known value of the data y, using the basic property of conditional probability known as Bayes' rule, yields the posterior density

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

- where $p(y) = \sum_\theta p(\theta)p(y|\theta)$ and the sum is over all possible values of θ
- or $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous θ

Prior, likelihood, and posterior

- Let
  - $y = (y_1, \ldots, y_n)$: observed data
  - $f(y|\theta)$: model for the observed data, usually a probability distribution
  - θ: vector of unknown parameters, assumed a random quantity
  - $\pi(\theta)$: prior distribution of θ
- The posterior distribution for inference concerning θ is

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|u)\pi(u)du}.$$

Posterior and marginal density of y

- The integral $\int f(y|u)\pi(u)du$, the marginal density of the data y, does not depend on θ.
- When the data y are fixed, then the integral can be regarded as a normalizing constant C.
- In high dimensional problems, the integral can be very difficult to evaluate.
- Evaluation of the complex integral $\int f(y|u)\pi(u)du$ was a focus of much Bayesian computation.

Advent of MCMC Methods

- With the advent of the use of Markov chain Monte Carlo (MCMC) methods,
  $\longrightarrow$ one could avoid evaluating the integral, making use of the unnormalized posterior density.

$$f(\theta|y) \propto f(y|\theta)\pi(\theta).$$

- Equivalently, if we denote the likelihood function or sampling distribution by $L(\theta)$, then

$$\begin{aligned} f(\theta|y) &\propto L(\theta)\pi(\theta). \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \end{aligned}$$

- We will show how this works.

Other Uses of MCMC Methods

- Can simplify otherwise difficult computations.
- Sometimes a likelihood would be easy to evaluate if some data had been observed that was not observed or is unobservable.
- Examples:
  - infection times,
  - time of clearing infection,
  - when someone is infectious,
  - chains of infection.
- MCMC methods can be used to augment the observed data to make estimation simpler.

Likehood and Data Transforms Prior to Posterior

- Likelihood and data take prior to posterior:

$$\text{Prior} \xrightarrow[\text{–Likelihood}\ \text{–Data}]{\text{Transformation}} \text{Posterior}$$

- Bayesian data analysis is a study of the transformation.

Transmission probability

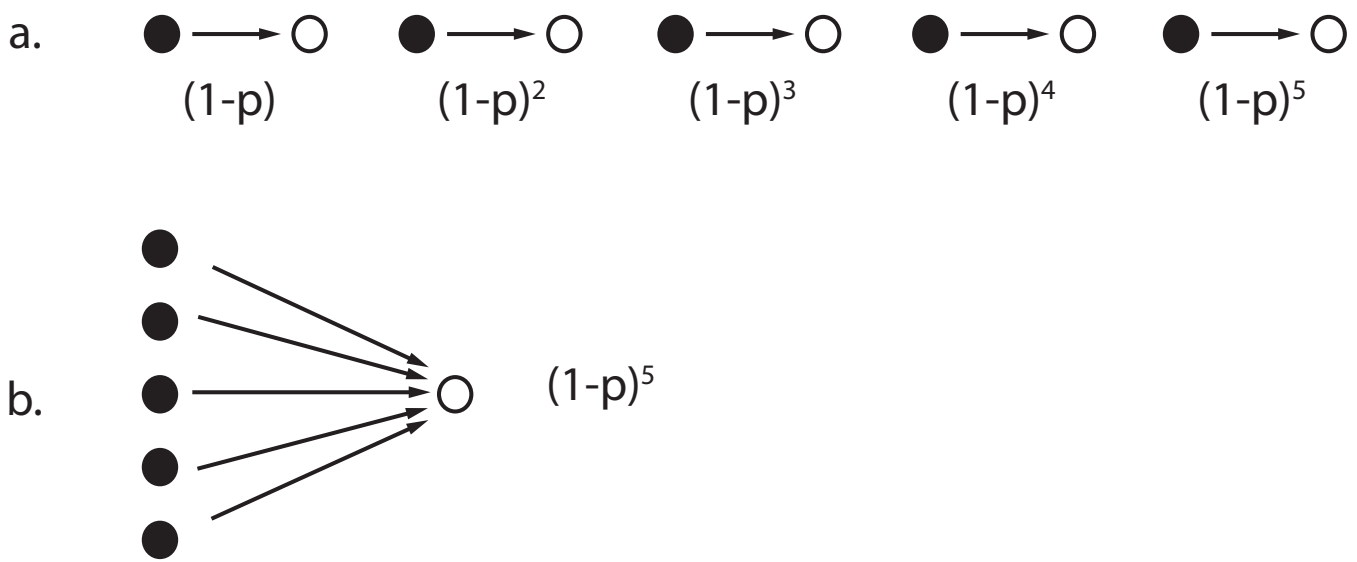
- p is the probability an infective infects a susceptible: transmission probability
- $q = 1 - p$ is the probability a susceptible escapes infection when exposed to an infective: escape probability
- Transmission versus escape ? which is the "success" and which the "failure" ?
- Given there are n exposures, and y infections, what is the estimate of the transmission probability?
- Given there are n exposures, and $n - y$ escapes, what is the estimate of the escape probability?

Chain-binomial model

- Assume independent households
- One person in each household introduces the infection into the household (index case).
- Infections occur within households in generations of infection (discrete time).
- p is the probability an infective infects a susceptible in a household in a generation
- $q = 1 - p$ is the probability a susceptible escapes infection when exposed to an infective in a household

Reed-Frost Chain Binomial Model

Figure : Independent exposures = independent Bernoulli trials



Chain Binomial Model

Table : Chain binomial probabilities in the Reed-Frost model in N households of size 3 with 1 initial infective and 2 susceptibles, $S_0 = 2, I_0 = 1$

Chain	Chain probability	Frequency	at p=0.4	at p=0.7	Final number infected
$1 \longrightarrow 0$	q^2	n_1	0.360	0.090	1
$1 \longrightarrow 1 \longrightarrow 0$	$2pq^2$	n_{11}	0.288	0.126	2
$1 \longrightarrow 1 \longrightarrow 1$	$2p^2q$	n_{111}	0.192	0.294	3
$1 \longrightarrow 2$	p^2	n_{12}	0.160	0.490	3
Total	1	N	1.00	1.00	

Chain binomial model

- Data: The observations are based on outbreaks of measles in Rhode Island 1929–1934.
- The analysis is restricted to $N = 334$ families with three susceptible individuals at the outset of the epidemic.
- Assume there is a single index case that introduces infection into the family.
- The actual chains are not observed, just how many are infected at the end of the epidemic.
- So the frequency of chains $1 \longrightarrow 1 \longrightarrow 1$ and $1 \longrightarrow 2$ are not observed.
- MCMC can be used to augment the missing data, and estimate the transmission probability p.

Chain Binomial Model

Table : Rhodes Island measles data: chain binomial probabilities in the Reed-Frost model in $N = 334$ households of size 3 with 1 initial infective and 2 susceptibles, $N_3 = n_{111} + n_{12} = 275$ is observed

Chain	Chain probability	Frequency	Observed frequency	Final number infected
$1 \longrightarrow 0$	q^2	n_1	34	1
$1 \longrightarrow 1 \longrightarrow 0$	$2pq^2$	n_{11}	25	2
$1 \longrightarrow 1 \longrightarrow 1$	$2p^2q$	n_{111}	not observed	3
$1 \longrightarrow 2$	p^2	n_{12}	not observed	3
Total	1	N	334	

General epidemic (SIR) model

- The population of N individuals
- Denote the numbers of susceptible, infective, and removed individuals at time t by $S(t)$, $I(t)$, and $R(t)$.
- The process can be represented by the compartmental diagram

$$S(t) \longrightarrow I(t) \longrightarrow R(t)$$

- Thus, $S(t) + I(t) + R(t) = N$ for all t.
- Initially, $(S(0), I(0), R(0)) = (N - 1, 1, 0)$

General epidemic model

- Each infectious individual remains so for a length of time $T_I \sim \exp(\gamma)$.
- During this time, infectious contacts occur with each susceptible according to a Poisson process of rate β/N
- Thus, the overall hazard of infection at time t is $\beta I(t)/N$
- The two model parameters of interest are β and γ

General epidemic model

- In a well-known smallpox data set, the removal times are observed. That is, when the people are no longer infectious for others.
- However, the infection times are not observed.
- Thus, estimating the two model parameters is difficult.
- The missing infection times are treated as latent variables.
- MCMC methods are used to augment the missing infection times and estimate the parameters β and γ.

Susceptible-infected-susceptible (SIS) model

- Background: Many infections are recurrent, occurring as an alternating series of presence and absence of infection
- Nasopharyngeal carriage of *Streptococcus pneumoniae* (Auranen et al 2000; Cauchemez et al 2006; Melegaro et al 2010)
- Nasopharyngeal carriage of *Neisseria meningitidis* (Trotter and Gay 2003)
- Malaria (Nagelkerke et al,)

Susceptible-infected-susceptible (SIS) model

- The population of N individuals
- Denote the numbers of susceptible and infected individuals at time t by $S(t)$ and $I(t)$.
- The process can be represented by the compartmental diagram

$$S(t) \leftrightarrow I(t)$$

- Thus, $S(t) + I(t) = N$ for all t.
- Acquisition and clearance times often remain unobserved
- Active sampling of the population to determine the current status of being infected or susceptible in individuals.

Susceptible-infected-susceptible (SIS) model

- Could be formulated as an infectious disease transmission process, as the general epidemic model.
- Too complicated for this introductory course
- We consider here the simple transition process, with rate parameters λ for acquisition and μ for clearance.
- The acquisition and clearance times are treated as latent variables.
- MCMC methods are used to augment the missing infection and clearance times, and estimate the parameters λ and μ.

Conjugate prior distributions

- Conjugacy: the property that the posterior distribution follows that same parametric form as the prior distribution.
- Beta prior distribution is conjugate family for binomial likelihood: posterior distribution is Beta
- Gamma prior distribution is conjugate family for Poisson likelihood: posterior distribution is Gamma

Conjugate prior distributions

- Simply put, conjugate prior distributions in tandem with the appropriate sampling distribution for the data have the same distribution as the posterior distribution.
- Conjugate prior distributions have computational convenience.
- They can also be interpreted as additional data.
- They have the disadvantage of constraining the form of the prior distribution.

More on prior distributions

- **Nonconjugate** prior distributions can be used when the shape of the prior knowledge or belief about the distribution of the parameters of interest does not correspond to the conjugate prior distribution.

- **Noninformative** prior distributions carry little population information and are generally supposed to play a minimal role in the posterior distribution.
  $\longrightarrow$They are also called diffuse, vague, or flat priors.

- Computationally nonconjugate distributions can be more demanding.

Data and Sampling Distribution

- Goal: Inference on the posterior distribution of the transmission probability

- Suppose that n people are exposed once to infection
  - y become infected ("successes")
  - $n - y$ escape infection ("failures")

- Let
  - $p =$ transmission probability
  - $1 - p = q =$ escape probability
- Binomial sampling distribution

$$L(y|p) = \text{Bin}(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} p^y q^{n-y}$$

Specify the Prior Distribution of p

- To perform Bayesian inference, we must specify a prior distribution for p.
- We specify a Beta prior distribution:

$$p \sim \text{Beta}(\alpha, \beta)$$

$$\text{Beta}(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \alpha > 0, \beta > 0.$$

- Mean: $E(p|\alpha, \beta) = \frac{\alpha}{\alpha+\beta}$
- Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{E(p|\alpha,\beta)[1-E(p|\alpha,\beta)]}{\alpha+\beta+1}$

Specify the Prior Distribution of p

- We specify a Beta prior distribution:

$$\begin{aligned} p &\sim \text{Beta}(\alpha, \beta) \\ \pi(p) &= \text{Beta}(p|\alpha, \beta) \\ \text{Beta} &\propto p^{\alpha-1}(1-p)^{\beta-1}. \end{aligned}$$

- Looks similar to binomial distribution
- $\alpha > 0$, $\beta > 0$, "prior sample sizes"

Posterior distribution of p

- The posterior distribution of the transmission probability p, $f(p|y)$:

$$\begin{aligned} f(p|y) &\propto p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1} \\ \text{posterior} &\quad \text{likelihood} \times \textit{prior} \\ &= p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \\ &\propto \text{Beta}(p|\alpha+y, \beta+n-y) \end{aligned}$$

- The role of α and β as prior sample sizes is clear.

Posterior mean of θ

- Posterior mean of p
  $\longrightarrow$ posterior probability of success (transmission) for a future draw from the population:

$$E(p|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

- posterior mean always lies between the prior mean $\alpha/(\alpha + \beta)$ and the sample mean y/n.
- Posterior variance of p:

$$\text{var}(p|y) = \frac{E(p|y)[1 - E(p|y)]}{\alpha + \beta + n + 1}$$

Uniform prior distribution

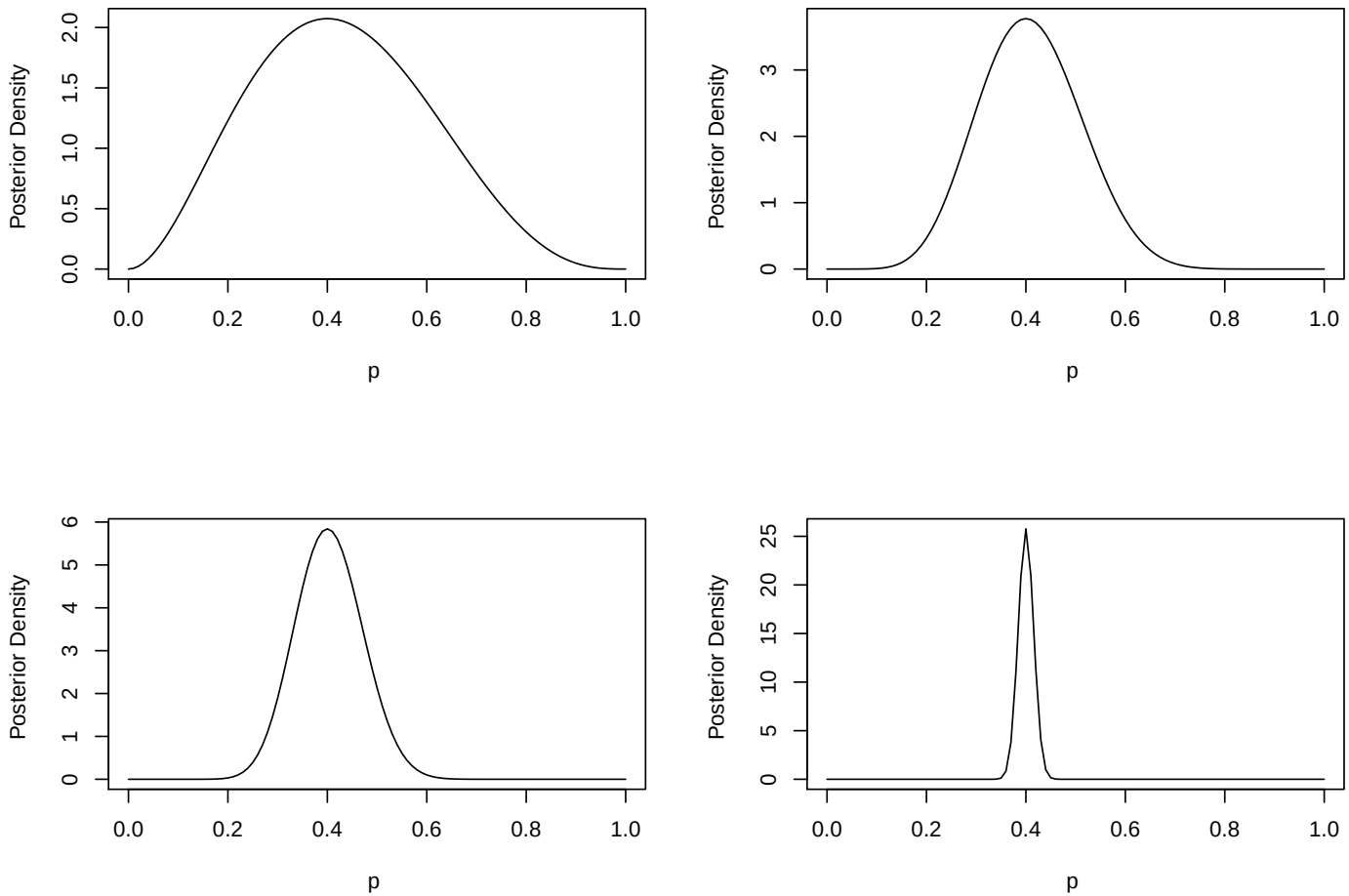
- The uniform prior distribution on [0,1] corresponds to $\alpha = 1$, $\beta = 1$. Essentially no prior information on p.

$$f(p|y) = \text{Beta}(p|y+1, n-y+1)$$

- Let's see how the posterior distribution of the transmission probability depends on the amount of data given a uniform prior distribution (Sample mean $y/n = 0.40$).

n, number exposed	y, number infected
5	2
20	8
50	20
1000	400

Figure : R program: Posterior distribution with differing amounts of data. Uniform Beta prior, Binomial sampling distribution.



Prediction

- After the data have been observed, we can predict a future unknown observable y_{n+1}.
- For example, we may observe n people who were exposed to infection, and whether they became infected.
- We may want to predict the probability that the next person to be observed would become infected.
- Posterior predictive distribution:
  $\longrightarrow$ posterior because conditional on the observed y
  $\longrightarrow$ predictive because it is a prediction for an observable y_{n+1}.

Prediction

- **Posterior predictive distribution of unknown observable** y_{n+1}:

$$\begin{aligned} f(y_{n+1}|y) &= \int f(y_{n+1}, p|y)dp \\ &= \int f(y_{n+1}|p, y)f(p|y)dp \\ &= \int f(y_{n+1}|p)f(p|y)dp \end{aligned}$$

- The last line follows because y and y_{n+1} are conditionally independent given p in this model.

- Useful in model checking.

References

- Gelman, A, Carlin, JB, Stern, HS, Dunson, DB, Vehtari, A, Rubin, DB. *Bayesian Data Analysis*, Chapman and Hall/CRC, third edition, 2014.
- Carlin, BP and Louis, TA. *Bayesian Methods for Data Analysis*, CRC Press, third edition, 2008.

Introductory Practical in R

- Before our module begins, do the exercises described in PracticalBayes12021.pdf
- The R code is available in bayesintro2021.R.
- This practical also has an exercise where you vary the amount of prior data keeping the amount of observed data constant to see how it affects the posterior distribution.
- If you already are familiar with R, this will be a simple exercise.

Chain Binomial Model

Table : Rhodes Island measles data: chain binomial probabilities in the Reed-Frost model in $N = 334$ households of size 3 with 1 initial infective and 2 susceptibles, $N_3 = n_{111} + n_{12} = 275$ is observed

Chain	Chain probability	Frequency	Observed frequency	Final number infected
$1 \longrightarrow 0$	q^2	n_1	34	1
$1 \longrightarrow 1 \longrightarrow 0$	$2pq^2$	n_{11}	25	2
$1 \longrightarrow 1 \longrightarrow 1$	$2p^2q$	n_{111}	not observed	3
$1 \longrightarrow 2$	p^2	n_{12}	not observed	3
Total	1	N	334	

Complete data likelihood for q

- The multinomial complete data likelihood for q:

$$f(n_1, n_{11}, N_3, n_{111}|q)$$

$$= \binom{334}{n_1, n_{11}, n_{111}, N_3 - n_{111}} (q^2)^{n_1} (2q^2p)^{n_{11}} (2qp^2)^{n_{111}} (p^2)^{N_3 - n_{111}}$$

- The observed data are (n_1, n_{11}, N_3), but we do not observe n_{111}.

- We could estimate q using a marginal model, but won't.

Gibbs sampler for chain binomial model

- The general idea of the Gibbs sampler is to sample the model unknowns from a sequence of full conditional distributions and to loop iteratively through the sequence.
- To sample one draw from each full conditional distribution at each iteration, it is assumed that all of the other model quantities are known at that iteration.
- In the theoretical lectures, it will be shown that the Gibbs sampler converges to the posterior distribution of the model unknowns.
- In the Rhode Island measles data, we are interested in augmenting the missing data n_{111} and estimating the posterior distribution of q, the escape probability.

Gibbs sampler for chain binomial model

- The joint distribution of the observations (n_1, n_{11}, N_3) and the model unknowns (n_{111}, q) is

$$f(n_1, n_{11}, N_3, n_{111}, q) = f(n_1, n_{11}, N_3, n_{111}|q) \times f(q)$$

complete data likelihood $\times$ prior

- We want to make inference about the joint posterior distribution of the model unknowns

$$f(n_{111}, q|n_1, n_{11}, N_3)$$

- This is possible by sampling from the full conditionals (Gibbs sampling): $f(q|n_1, n_{11}, N_3, n_{111})$ and $f(n_{111}|n_1, n_{11}, N_3, q)$

Algorithm for Gibbs sampler for chain binomial model

1. Start with some initial values $(q^{(1)}, n_{111}^{(1)})$
2. For $t = 1$ to M do
3. Sample $q^{(t+1)} \sim f(q|n_1, n_{11}, N_3, n_{111}^{(t)})$
4. Sample $n_{111}^{(t+1)} \sim f(n_{111}|n_1, n_{11}, N_3, q^{(t+1)})$
5. end for
6. How to get the two full conditionals in this model?

Full conditional of chain $1 \longrightarrow 1 \longrightarrow 1$

- Assume q is known
- Compute the conditional probability of chain $1 \to 1 \to 1$ when outbreak size is $N = 3$:

$$\Pr(1 \to 1 \to 1 | N = 3, q) = \frac{\Pr(N = 3, 1 \to 1 \to 1 | q)}{\Pr(N = 3 | q)}$$

$$= \frac{\Pr(N = 3 | 1 \to 1 \to 1, q) \Pr(1 \to 1 \to 1 | q)}{\Pr(N = 3 | 1 \to 1 \to 1, q) \Pr(1 \to 1 \to 1 | q) + \Pr(N = 3 | 1 \to 2, q) \Pr(1 \to 2 | q)}$$

$$= \frac{2p^2 q}{2p^2 q + p^2} = \frac{2q}{2q + 1}, \quad (0 \leq q < 1)$$

The full conditional of n_{111}

- We have found that

$$\Pr(1 \to 1 \to 1 | N = 3, q) = \frac{2q}{2q+1}$$

- So the full conditional distribution of n_{111} is

$$n_{111} | (n_1, n_{11}, N_3, q) \sim \text{Binomial}(275, 2q/(2q+1))$$

The full conditional of q

- Assume that n_{111} is known, that is, assume we know the complete data $(n_1, n_{11}, N_3, n_{111})$
- Assume a prior distribution for q: $q \sim \text{Beta}(\alpha, \beta)$,

$$f(q) \equiv f(q|\alpha, \beta) \propto q^{\alpha-1}(1-q)^{\beta-1}$$

- The full conditional distribution of q :

$$f(q|n_1, n_{11}, N_3, n_{111}, \alpha, \beta) \propto f(n_1, n_{11}, N_3, n_{111}|q, \alpha, \beta) f(q|\alpha, \beta)$$

$$\propto \underbrace{q^{2n_1+2n_{11}+n_{111}} p^{n_{11}+2N_3}}_{\text{complete data likelihood}} \times \underbrace{q^{\alpha-1}(1-q)^{\beta-1}}_{\times \text{ prior}}$$

The full conditional of q

- The full conditional distribution of q is thus a Beta distribution

$$q|\text{complete data}, \alpha, \beta \sim \text{Beta}(2n_1 + 2n_{11} + n_{111} + \alpha, n_{11} + 2N_3 + \beta)$$

- A uniform prior on q corresponds to $\alpha = 1$, $\beta = 1$.

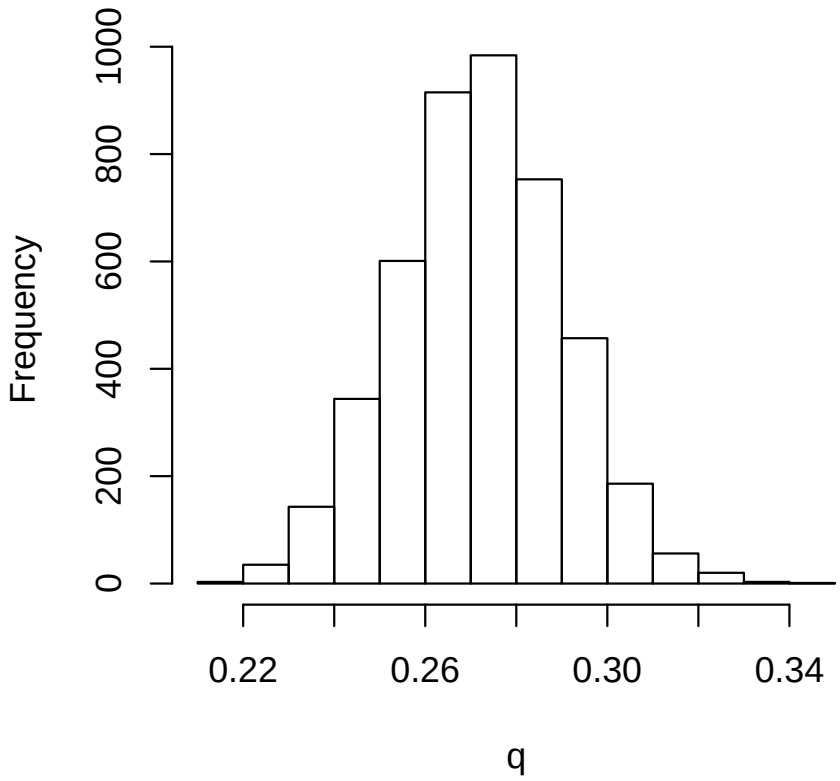
- With the complete data, a natural point estimate of the escape probability would be the mean of the Beta distribution, i.e., the proportion of "escapes" out of all exposures:

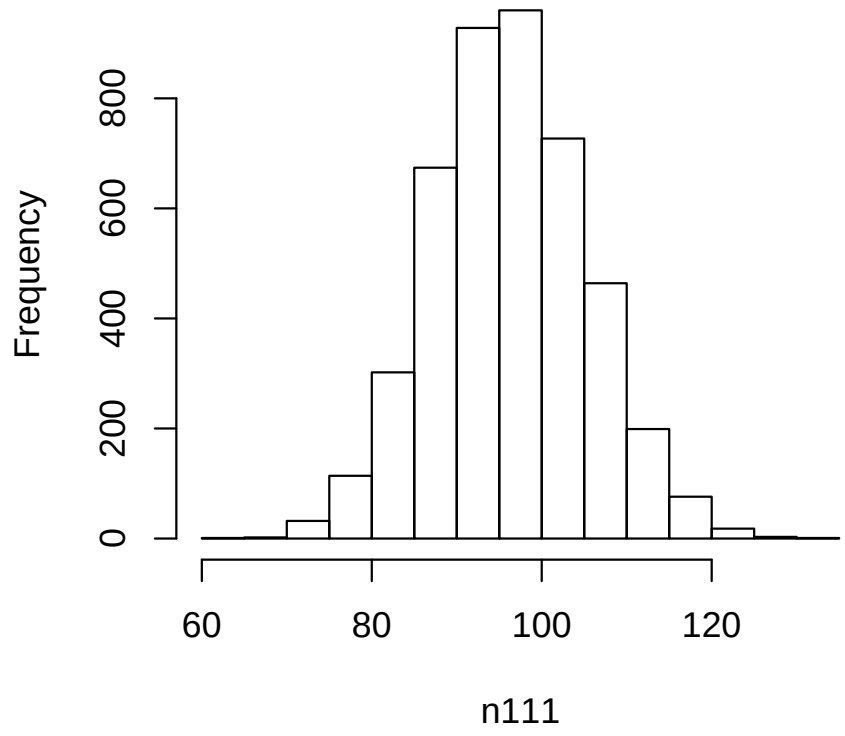
$$\frac{2n_1 + 2n_{11} + n_{111} + \alpha}{2n_1 + 3n_{11} + 3n_{111} + 2n_{12} + \alpha + \beta}$$

Algorithm for Gibbs sampler for chain binomial model

1. Start with some initial values $(q^{(1)}, n_{111}^{(1)})$
2. For $t = 1$ to M do
3. Sample $q^{(t+1)} \sim \text{Beta}(2n_1 + 2n_{11} + n_{111}^{(t)} + \alpha, n_{11} + 2N_3 + \beta)$
4. Sample $n_{111}^{(t+1)} \sim \text{Binomial}(275, 2q^{(t+1)}/(2q^{(t+1)} + 1))$
5. end for
6. Get summaries of the marginal posterior distributions.

Posterior distributions of q *and* n_{111}





Lab: First Gibbs Sampler

- Next will be the lab with first Gibbs sampler computational exercise.
- We will go over this in the module live.
- Before the module, you may try to do the Gibbs sampler if you like, but it is not necessary.