Model checking, hierarchical modeling and combined M-H and Gibbs

SISMID/July 13–15, 2015

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin

Outline

- The chain binomial model for household outbreaks of measles
  - Bayesian analysis of incompletely observed data, using data augmentation (cf. the earlier lecture and computer lab)
  - Checking the model fit through comparison of predictive data with the observed data of the final number infected
- Model extension by allowing heterogeneity across households $\rightarrow$ a hierarchical model
- Implementation of posterior sampling in the hierarchical model by a combined Gibbs and Metropolis algorithm

The observed outbreak sizes

Recall the observed data in the chain binomial model:

Chain	Chain probability	Frequency	Observed frequency	Final number infected
1	q_j^2	n_1	34	1
$1 \to 1$	$2q_j^2 p_j$	n_{11}	25	2
$1 \to 1 \to 1$	$2q_j p_j^2$	n_{111}	not observed	3
$1 \to 2$	p_j^2	n_{12}	not observed	3
Total	1	N	334	

- If the final number infected is 1 or 2, the actual chain is observed
- If the final number infected is 3, the actual chain data are not observed
  - We still know that $N_3 \equiv n_{111} + n_{12} = 275$
- In the previous analysis, we assumed $q_j = q$ for $j = 1, \ldots, 334$, i.e., for all 334 households

Prediction

- Recall that new (predictive) data y^{pred} can be generated by drawing from the posterior predictive distribution $f(y^{\text{pred}}|y)$
- Posterior predictive distribution because
  - conditioning on the observed data y
  - predicting a future observable y^{pred}
- Predictive data can be compared with the observed data to assess the fit of the model
- In this example, we compare the predictive and observed frequencies of chains 1 and $1 \to 1$

Posterior predictive distribution

- Denote the model parameters by θ. Then

$$f(y^{\text{pred}}|y) = \int f(y^{\text{pred}}, \theta|y)d\theta = \int f(y^{\text{pred}}|\theta, y)f(\theta|y)d\theta$$
$$= \int f(y^{\text{pred}}|\theta)f(\theta|y)d\theta$$

- Samples from the posterior predictive distribution can be realised as follows:
  - [1] Draw an MCMC sample θ_k from the posterior $f(\theta|y)$
  - [2] Draw a sample y_k^{pred} from $f(y^{\text{pred}}|\theta_k)$
  - [3] Repeat steps [1] and [2] K times $(k = 1, \ldots, K)$

Model checking

- The posterior predictive distribution of frequencies $(n_1, n_{11}, n_{111}, n_{12})$ is now

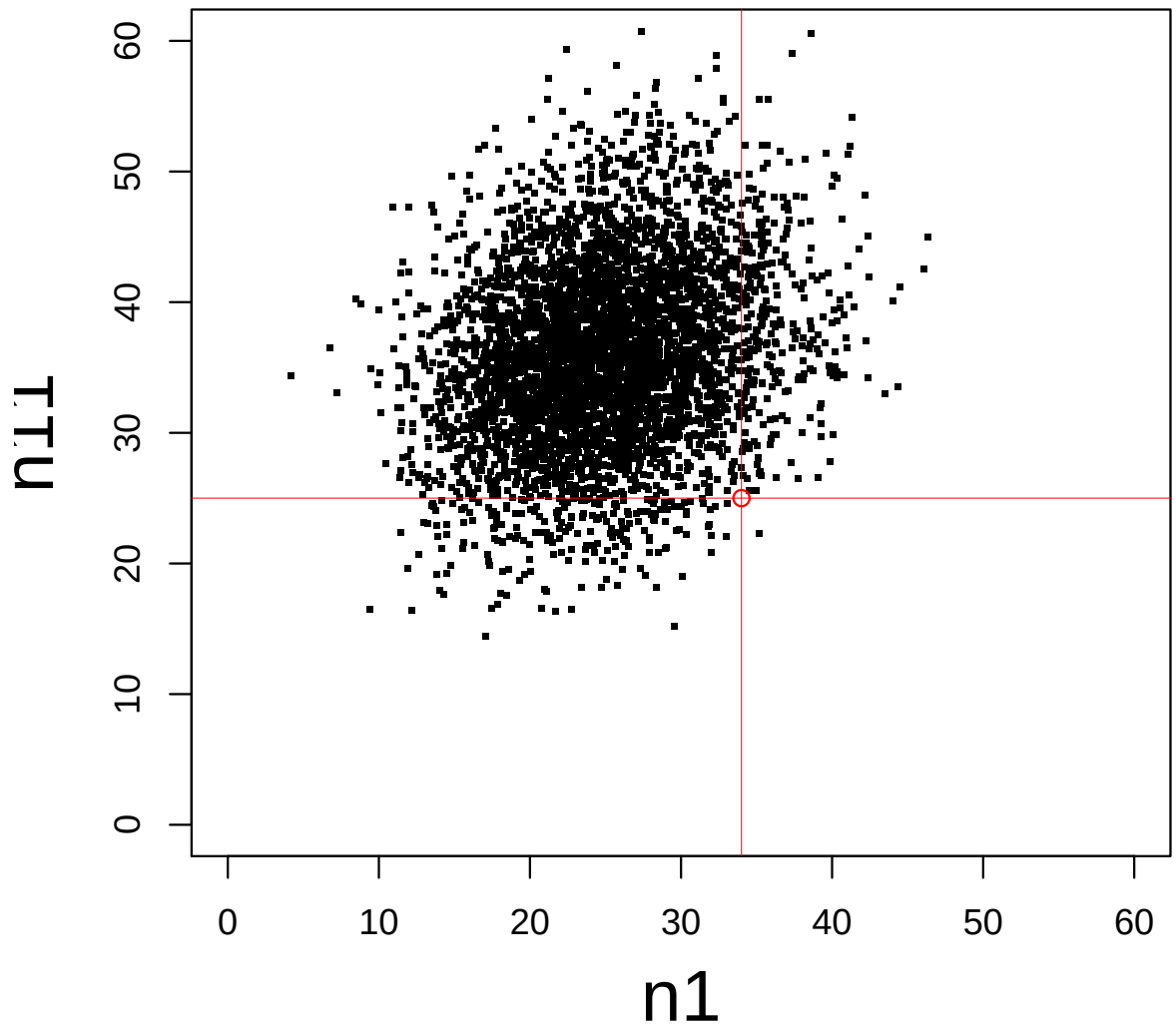
$$f(n_1^{\text{pred}}, n_{11}^{\text{pred}}, n_{111}^{\text{pred}}, n_{12}^{\text{pred}} | n_1, n_{11}, N_3)$$
$$= \int_0^1 f(n_1^{\text{pred}}, n_{11}^{\text{pred}}, n_{111}^{\text{pred}}, n_{12}^{\text{pred}} | q) f(q | n_1, n_{11}, N_3) dq$$

- Samples from the posterior predictive distribution:
  - [1] Draw an MCMC sample $q^{(k)}$ from the posterior $f(q|n_1, n_{11}, N_3)$
  - [2] Draw a sample $(n_1^{(k)}, n_{11}^{(k)}, n_{111}^{(k)}, n_{12}^{(k)})$ from Multinomial$(334, (q^{(k)}, 2(q^{(k)})^2 p^{(k)}, 2q^{(k)}(p^{(k)})^2, p^{(k)}))$
  - [3] Repeat steps [1] and [2] K times $(k = 1, \ldots, K)$

Model checking continues

- Comparison of a sample from the joint predictive posterior of $(n_1^{\text{pred}}, n_{11}^{\text{pred}})$ with the actually observed point (34,25) reveals a poor model fit (next page)
- The model did not take into account possible heterogeneity across households in the escape probability
- Therefore, we'll consider model extension through allowing such heterogeneity

Model checking continues

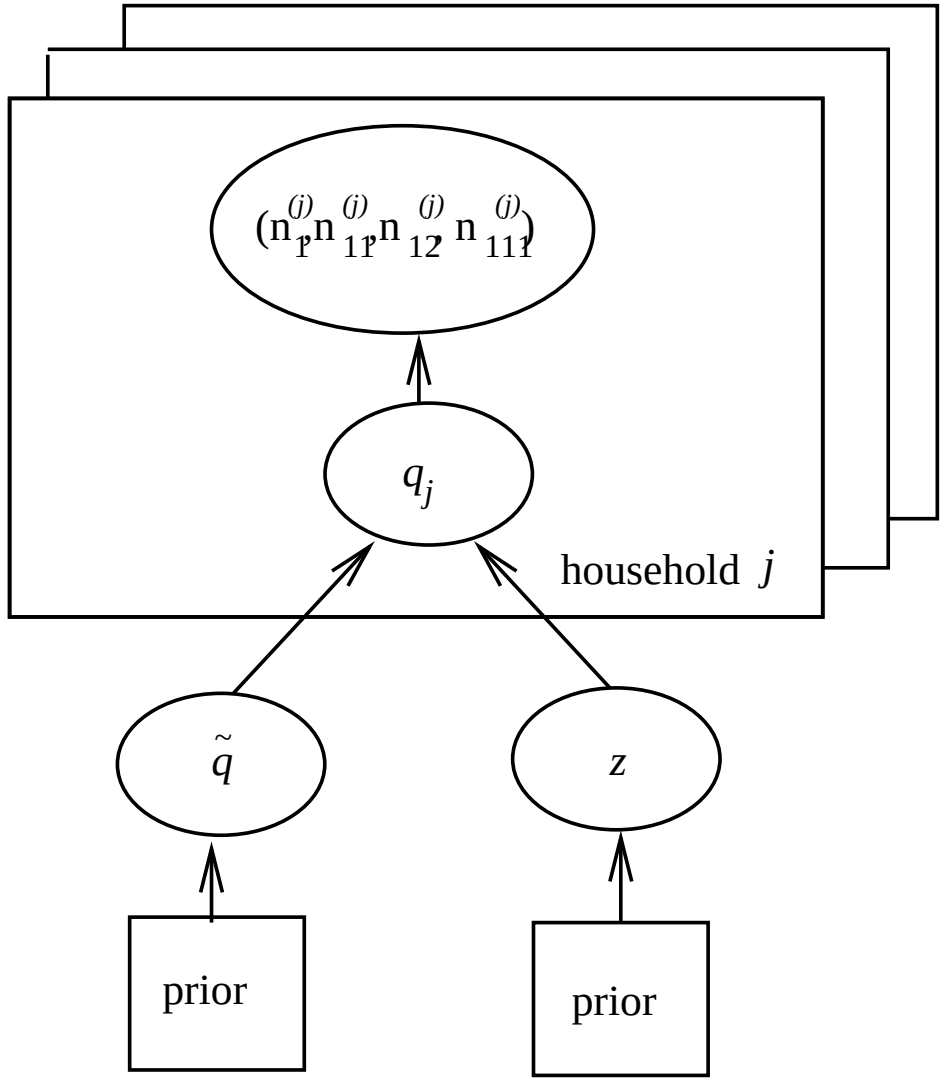


A hierarchical model

- In household j, the observation $(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)})$ follows a multinomial distribution with size 1 and probability vector $(q_j^2, 2q_j^2 p_j, 2q_j p_j^2, p_j^2)$, $j = 1, \ldots, 334$
- The *household-specific* escape probabilities q_j follow a $\text{Beta}(\tilde{q}/z, (1-\tilde{q})/z)$ distribution
- Assuming uniform and gamma priors for $\tilde{q}$ and z, respectively, the hierarchical model becomes fully defined:

$$
\begin{aligned}
(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)}) | q_j &\sim \text{Multinomial}(1, (q_j^2, 2q_j^2 p_j, 2q_j p_j^2, p_j^2)) \\
q_j | \tilde{q}, z &\sim \text{Beta}(\tilde{q}/z, (1-\tilde{q})/z)) \\
\tilde{q} &\sim \text{Uniform}(0, 1) \\
z &\sim \text{Gamma}(1.5, 1.5)
\end{aligned}
$$

A hierarchical model continues



The joint distribution

- The joint distribution of the parameters $\tilde{q}$ and z, the household-specific escape probabilities q_j $(j = 1, \ldots, 334)$, and the chain frequencies is

$$\prod_{j=1}^{334} \left(f(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)} | q_j) f(q_j | \tilde{q}, z) \right) f(\tilde{q}) f(z),$$

- The model unknowns are parameters $\tilde{q}$ and z, frequencies $n_{111}^{(j)}$ for all 275 household with outbreak size 3, as well as all 334 household-specific escape probabilities q_j

Sampling from the posterior

- Notation: $\alpha^{(k)} = \tilde{q}^{(k)}/z^{(k)}$, $\beta^{(k)} = (1-\tilde{q}^{(k)})/z^{(k)}$, k refers to iteration, j refers to household
- A sketch of the steps in kth iteration of the sampling algorithm:

$$q_j^{(k)}|\alpha^{(k-1)}, \beta^{(k-1)} \sim \text{Beta}(2+\alpha^{(k-1)}, \beta^{(k-1)}), \; j=1,\ldots,34$$

$$q_j^{(k)}|\alpha^{(k-1)}, \beta^{(k-1)} \sim \text{Beta}(2+\alpha^{(k-1)}, 1+\beta^{(k-1)}), \; j=35,\ldots,59$$

$$q_j^{(k)}|\alpha^{(k-1)}, \beta^{(k-1)}, n_{111}^{(j,k-1)} \sim \text{Beta}(n_{111}^{(j,k-1)}+\alpha^{(k-1)}, 2+\beta^{(k-1)}),$$

$$j=60,\ldots,334$$

$$n_{111}^{(j,k)}|q_j^{(k)} \sim \text{Binom}(1, 2q_j^{(k)}/(2q_j^{(k)}+1)), \; j=60,\ldots,334$$

$$\tilde{q}^{(k)}|z^{(k-1)}, q_1^{(k)}, \ldots, q_{334}^{(k)} \text{ using a Metropolis-Hastings step}$$

$$z^{(k)}|\tilde{q}^{(k)}, q_1^{(k)}, \ldots, q_{334}^{(k)} \text{ using a Metropolis-Hastings step}$$

Sampling from the posterior cont.

- In each household, the full conditional (Beta) distribution of $q_j^{(k)}$ depends on the current iterates of the numbers of escapes ($e_j^{(k-1)}$) and infections ($d_j^{(k-1)}$) *in that household* and the prior parameters $\alpha^{(k-1)}$ and $\beta^{(k-1)}$
- The numbers of escapes and infections: see Table
- So, $q_j^{(k)} \sim \text{Beta}(e_j^{(k-1)} + \alpha^{(k-1)}, d_j^{(k-1)} + \beta^{(k-1)})$

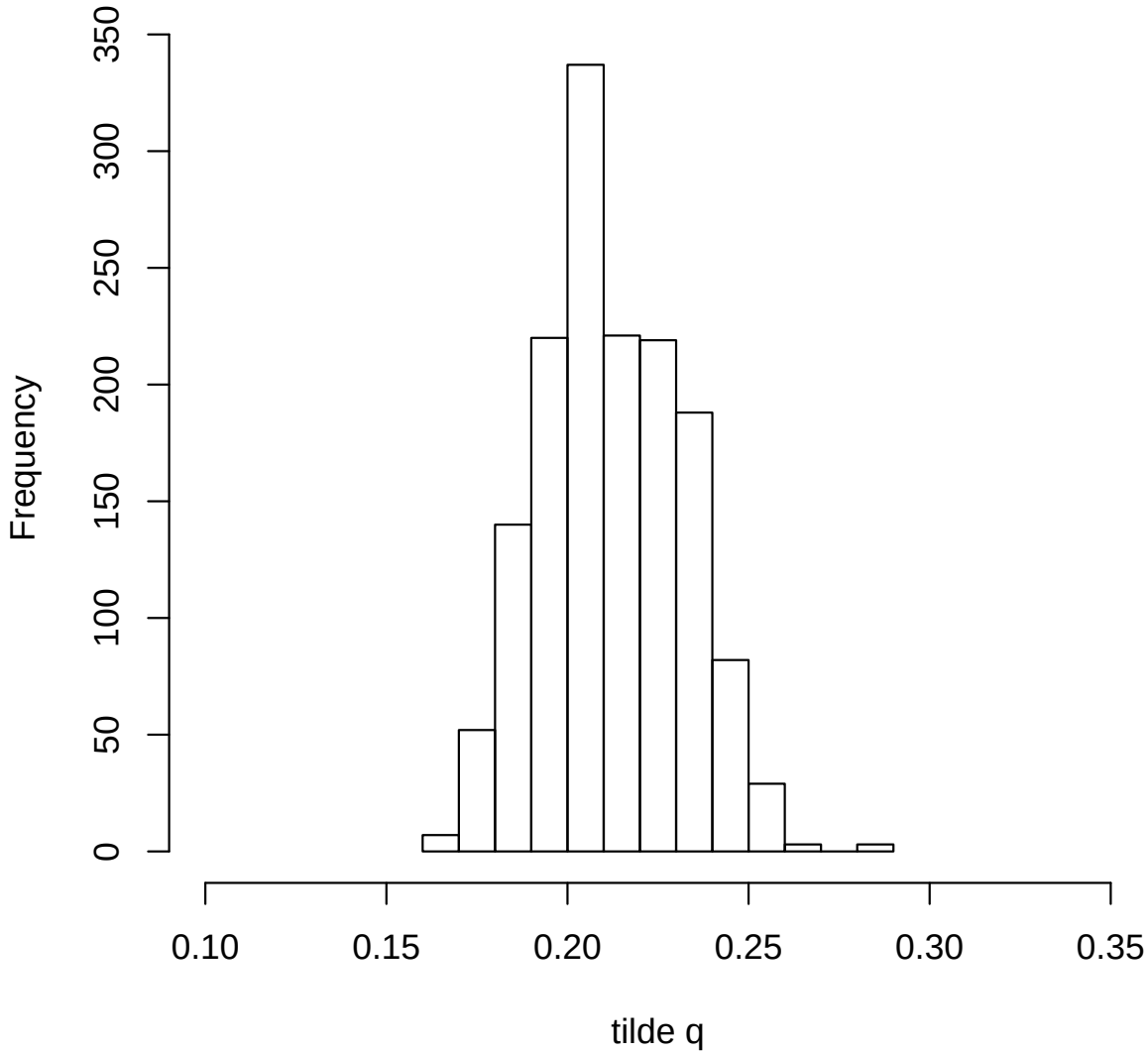
Chain	Number of escapes $e_j^{(k-1)}$	Number of infections $d_j^{(k-1)}$
1	2	0
1→1	2	1
1→1→1	$1 = n_{111}^{(j,k-1)}$	2
1→2	$0 = n_{111}^{(j,k-1)}$	2

Sampling from the posterior cont.

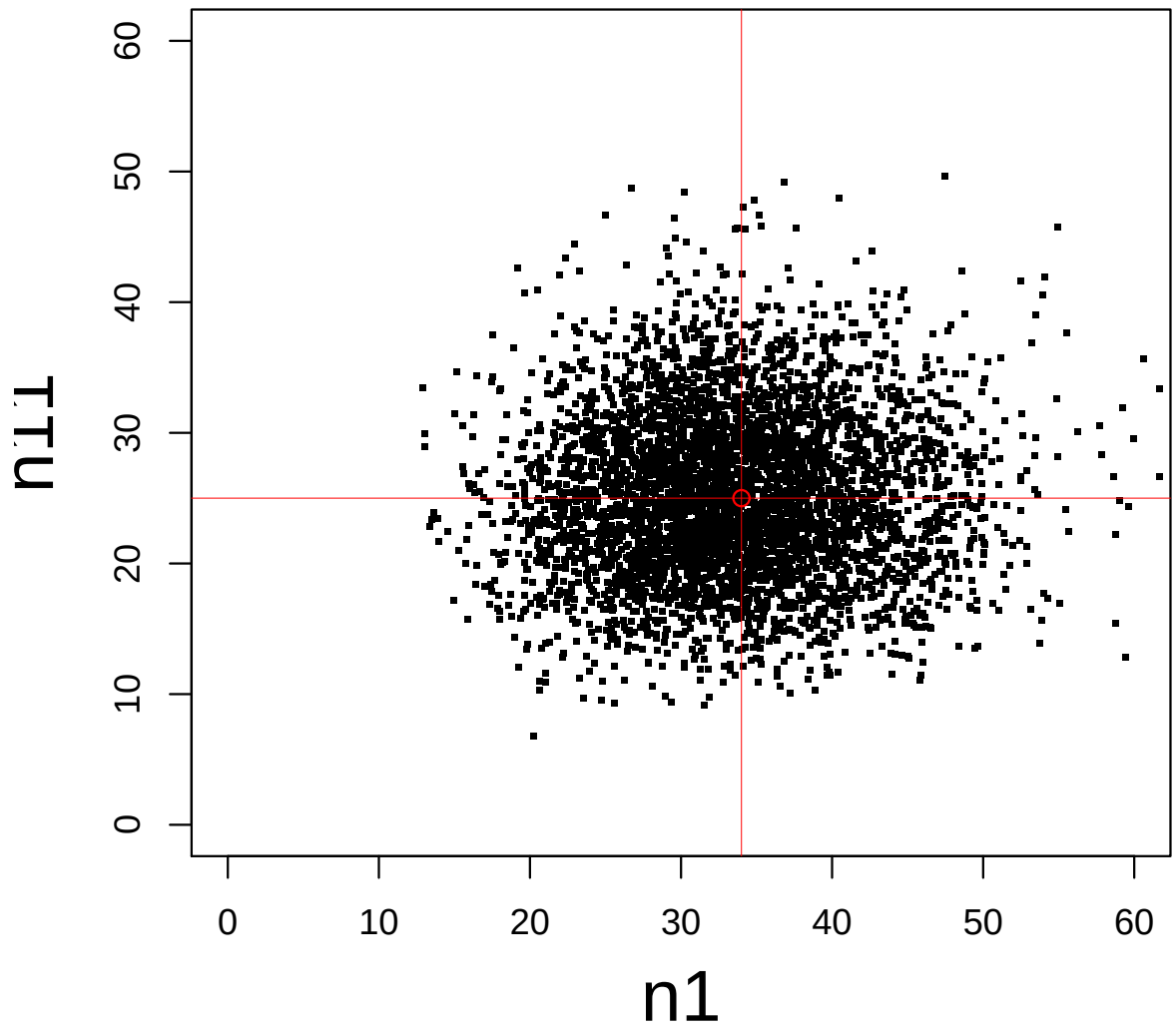
- Parameters $\tilde{q}$ and z require a Metropolis-Hastings step
- For $\tilde{q}$, if the current iterate is $\tilde{q}^{(k-1)}$, a new value $\bar{q}$ is first proposed (e.g.) uniformly about the current iterate (this is a symmetric proposal)
- The proposal is then accepted, i.e., $\tilde{q}^{(k)} := \bar{q}$, with probability

$$\min\{1, \frac{\prod_{j=1}^{334} f(q_j^{(k)}|\bar{q}, z^{(k-1)}) f(\bar{q})}{\prod_{j=1}^{334} f(q_j^{(k)}|\tilde{q}^{(k-1)}, z^{(k-1)}) f(\tilde{q}^{(k-1)})}\}$$

Posterior distribution of $\tilde{q}$



Checking the hierarchical model



An alternative approach

- In this example, it is possible to marginalise q_j over its prior distribution
- This means calculating the chain probabilities under as expectations of the respective probabilities in the previous table, with respect to Beta$(\tilde{q}/z, (1-\tilde{q})/z)$:

Chain	Chain probability	Frequency	Observed frequency	Final number infected
1	$\tilde{q}(\tilde{q}+z)/(1+z)$	n_1	34	1
1→1	$2\tilde{p}\tilde{q}(\tilde{q}+z)/((1+z)(1+2z))$	n_{11}	25	2
1→1→1	$2\tilde{p}\tilde{q}(\tilde{p}+z)/((1+z)(1+2z))$	n_{111}	missing	3
1→2	$\tilde{p}(\tilde{p}+z)/(1+z)$	n_{12}	missing	3

Alternative approach continues

- The following identity helps to calculate the expectations:

$$\mathrm{E}(p_j^u q_j^v) = \frac{\tilde{q}(\tilde{q}+z)\ldots(\tilde{q}+z(u-1))\tilde{p}(\tilde{p}+z)\ldots(\tilde{p}+z(v-1))}{(1+z)\ldots(1+z(u+v-1))}$$

- Using the probabilities as given in the table, it is straightforward to implement a Metropolis-Hastings algorithm to draw samples from the posterior of parameters $\tilde{q}$ and z

[1] Bailey T.J.N. The Mathematical Theory of Infectious Diseases. Charles Griffiths and Company, London 1975.

[2] O'Neill Ph. and Roberts G. Bayesian inference for partially observed stochastic processes. Journal of the Royal Statistical Society, Series A, 1999; 162: 121–129.

[3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

[4] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences 2002; 180:103-114.