

# Sampling Contingency Tables Preserving Observed Conditional Frequencies

Juyoun Lee  
Joint work with Dr. Slavković

Department of Statistics  
Pennsylvania State University

ENAR Fest  
February 26, 2009

# Contingency Tables in context of SDL

- Statistical disclosure limitation (SDL) in contingency tables
  - balance between disclosure risk and data utility
  - release of partial information: marginal total, **conditional rate**

## Example

**Table:** A 2x2x2 Table of results of clinical trial for the effectiveness of an analgesic drug

Statue	Treatment	Response		Total
		Poor	Good	
1	1	15	37	52
1	2	22	32	54
2	1	6	39	45
2	2	12	30	42
	Total	55	138	193

# Contingency Tables in context of SDL

- Statistical disclosure limitation (SDL) in contingency tables
  - balance between disclosure risk and data utility
  - release of partial information: marginal total, **conditional rate**

## Example

**Table:** [Treatment, Response] Marginal table of results of clinical trial for the effectiveness of an analgesic drug

Treatment	Response		Total
	Poor	Good	
1	21	76	97
2	34	62	96
Total	55	138	193

# Contingency Tables in context of SDL

- Statistical disclosure limitation (SDL) in contingency tables
  - balance between disclosure risk and data utility
  - release of partial information: marginal total, **conditional rate**

## Example

**Table:** [Response|Treatment] Table of conditional probabilities with [rounded probability]

Treatment	Response		Total
	Poor	Good	
1	$\frac{21}{97} = 0.2164948$ [0.2]	$\frac{76}{97} = 0.7835052$ [0.8]	97
2	$\frac{34}{96} = 0.3541667$ [0.4]	$\frac{62}{96} = 0.6458333$ [0.6]	96
Total	55	138	193

# Problem Statement

Cell counts of a table,  $\mathbf{n} = \{n_k\}$ , are conditionally independent Poisson with  $\lambda = \{\lambda_k\}$ . Let  $\mathcal{T} = \{\text{a set of observed conditional rates \& grand total } N\}$ , and  $\mathcal{F}_{\mathcal{T}}$  the set of all possible tables that preserve the given information  $\mathcal{T}$ .

**Goal:** Estimate 
$$P(\mathbf{n}|\mathcal{T}, \lambda) = \frac{P(\mathbf{n}|\lambda)I_{\mathbf{n} \in \mathcal{F}_{\mathcal{T}}}}{\sum_{\mathbf{n}' \in \mathcal{F}_{\mathcal{T}}} P(\mathbf{n}'|\lambda)},$$

where 
$$P(\mathbf{n}, \mathcal{T}|\lambda) = \begin{cases} P(\mathbf{n}|\lambda), & \mathbf{n} \in \mathcal{F}_{\mathcal{T}} \\ 0, & \textit{otherwise.} \end{cases}$$

- When  $\mathcal{T}$  is a set of marginal totals, a hypergeometric distribution is a special well-known case of  $P(\mathbf{n}|\mathcal{T}, \lambda) = P(\mathbf{n}|\mathcal{T})$ .

Generate samples from  $P(\mathbf{n}, \lambda | \mathcal{T})$  iteratively. An initial table  $\mathbf{n}^{(0)}$  is randomly chosen from  $\mathcal{F}_{\mathcal{T}}$ . Then at the  $(t + 1)^{th}$  step,

- 1 Sample  $\lambda^{(t+1)}$  from  $P(\lambda | \mathbf{n}^{(t)}, \mathcal{T}) \propto P(\mathbf{n}^{(t)} | \lambda) P(\lambda | \mathcal{T})$ .
- 2 Sample  $\mathbf{n}^{(t+1)}$  from  $P(\mathbf{n} | \mathcal{T}, \lambda^{(t+1)})$ .

Issues:

- Prior and posterior specification of  $\lambda$
- Generate a *synthetic table*.
  - Generating complete tables maintaining conditional rates
  - Utilize them as an alternative releasable data

# Prior and Posterior Specification

Conjugate prior,  $P(\lambda|\mathcal{T}) \sim \text{Gamma}(\alpha, \beta)$ , is considered where  $\alpha$  and  $\beta$  are induced by  $\mathcal{T}$ .

$$\begin{aligned} P(\lambda|\mathbf{n}^{(t)}, \mathcal{T}) &\propto P(\mathbf{n}_{i(t)}|\lambda)P(\lambda|\mathcal{T}) \\ &= \prod_k \text{Poisson}(n_k|\lambda_k) \text{Gamma}(\alpha_k, \beta_k) \\ &\propto \prod_k \text{Gamma}\left(n_k^{(t)} + \alpha_k, \left(1 + \frac{1}{\beta_k}\right)^{-1}\right). \end{aligned}$$

# Algebraic Algorithm for Generating Tables

The step 2 in the [MCMC Framework](#) is replaced by

- 1 Generate the candidate table  $\mathbf{n}^*$  from  $q(\mathbf{n}, \mathbf{n}^{(t)})$  induced by **Markov moves** [▶ details](#), then  $q(, )$  is symmetric.

That is, uniformly choose  $\mathbf{m} \in M$  and a sign  $\epsilon = \pm 1$  for the move with probability  $\frac{1}{2}$ , where  $M$  is a set of Markov moves.

- 2 Move to new state  $\mathbf{n}^* = \mathbf{n}^{(t)} + \epsilon \mathbf{m}$  with probability  $\alpha(\mathbf{n}^{(t)}, \mathbf{n}^*) = \min\{\rho, 1\}$  otherwise stay at  $\mathbf{n}^{(t)}$ .

$$\rho = \frac{P(\mathbf{n}^* | \lambda^{(t+1)})}{P(\mathbf{n}^{(t)} | \lambda^{(t+1)})}.$$

# Algebraic Algorithm for Generating Tables

## Tools from algebraic statistics

- The algebraic mapping is constructed by the constraints.
  - The Markov basis (move) is determined by the mapping.
  - $\mathbf{m}$  is a move, a table with integer entries.
  - Moves connect all tables maintaining the constraints.
  - Thus, moves allow to construct the connected Markov chain.
- 
- For the case given a set of marginal totals:  
Diaconis and Sturmfels (1998), Dobra et al (2006), Chen et al (2006)
  - For the case given a set of conditional rates:  
Slakovic (2004), Lee and Slakovic (2008)

# Example

**Table:** Collapsed 2x2 Table of results of clinical trial for the effectiveness of an analgesic drug and rounded conditionals  $P(\text{Response}|\text{Treatment})$

Treatment	Response [ $P(\text{Response} \text{Treatment})$ ]		Total
	Poor	Good	
Active	21[0.2]	76[0.8]	97
Placebo	34[0.4]	62[0.6]	96
Total	55	128	193

**Table:** Summaries of 10000 simulations of cell entries from the posterior distribution  $P(\mathbf{n}|\mathcal{T}, \lambda)$  with  $\lambda \sim \text{Gamma}(193, \frac{1}{4})$  and  $\mathcal{T} = \{P(\text{Response}|\text{Treatment}), N = 193\}$  vs.  $\mathcal{T} = \{n_{i+}, n_{+j}\}$ .

True Value	Posterior Mean	Posterior Median	Posterior Mode	Posterior Quantiles	MC St Error
	Con. Mar.	Con. Mar.	Con. Mar.	[ $Q_{0.025}, Q_{0.975}$ ] Con. Mar.	(Batch Mean) Con. Mar.
21	19.3 27.5	19 28	19 28	[16, 22] [21, 34]	0.0400 0.0689
76	72.1 69.5	72 69	72 69	[56, 80] [63, 76]	0.1601 0.0689
34	36.4 27.5	36 27	36 27	[32, 44] [21, 34]	0.0800 0.0689
62	65.2 68.5	65 69	65 69	[59, 77] [62, 75]	0.1201 0.0689

# Example

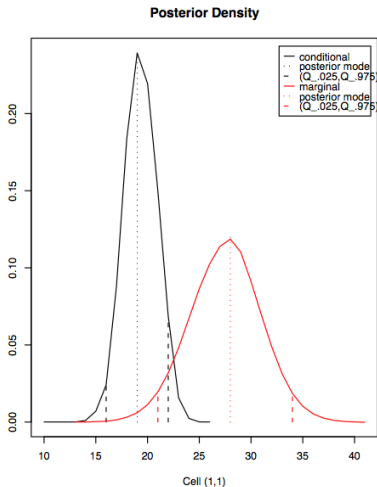


Figure: Posterior density of cell (1,1) given **two marginal totals** [Response][Treatment] and conditional rates [Response|Treatment]

## Conditional versus Marginal

- Posterior summaries given conditionals better estimate the original values than the posterior summaries given marginals.
- Posterior modes given conditionals preserves inferences about the odds-ratio while the posterior modes given marginals fail to provide consistent inferential results.

**Table:** Odds ratio ( $\theta$ ) and its 95% CI

	$\hat{\theta}$ (95%) CI on $\theta$	$\log \hat{\theta}$ (95%) CI on $\log \theta$
Original Table	0.5079 (0.2947, 0.8614)	-0.6854 (-1.2218, -0.1491)
Posterior Mode (Conditional)	0.5000 (0.2910, 0.8593)	-0.6931 (-1.2346, -0.1517)
Posterior Mode (Marginal)	1.0370 (0.6137, 1.7524)	0.0364 (-0.4883, 0.5610)

- **Better Data Utility but Worse Disclosure Risk**

# Considerations

- Extension to multi-way tables,
  - partial (small) conditionals
  - combination of marginals and conditionals
- Limitation of Markov moves
  - values of moves
  - additional computing
- Rounding issues with conditional rates
- Prior and Posterior specification on  $\lambda$
- Other sampling schemes
  - Importance sampling
  - Combination of multiple MCMC samplers



Chen, Y. and Dinwoodie, I.H. and Sullivant, S. (2006)

Sequential importance sampling for multiway tables. *Ann. Statist* **34**(1): 523–545.



Diaconis, P. and Sturmfels, B. (1998).

Algebraic algorithms for sampling from conditional distributions. *Ann. Statist* **26**(1): 363–397.



Dobra, A. and Tebaldi, C. and West, M. (2006).

Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference* **136**(2): 355–372.



Lee, J. and Slavković, A. (2008).

Synthetic tabular data preserving the observed conditional probabilities. *PSD 2008*.



Lee, J. and Slavković, A. (2008).

Posterior distributions for the unobserved cell counts in contingency tables. *ISBA 2008*.

# Acknowledges

- Joint work with Dr. Slavkovic
- NSF grant SES-0532407 to the Department of Statistics, Penn State University
- References: <http://www.stat.psu.edu/~sesa/privacy.html>

Thank you.



# Review of Sampling Algorithms for Contingency Tables

Our goal is to sample tables from the reference set preserving the observed conditional,  $\mathcal{F}_{R|T}$  in this example.

Markov Chain Monte Carlo(MCMC) with algebraic statistics  
: adapt the algebraic tools to generate a candidate table

- Markov bases are allowed to
  - perturb a table maintaining the marginal totals
  - calculate bounds on cell entries
  - have a connected Markov chain over a set of all possible tables preserving same marginal totals, referred as a *reference set*
  - specify the reference set
- A connected Markov chain based on Markov bases exists for the case given conditional probabilities.
  - Theoretically, MCMC with the algebraic flavor works.
  - Practically difficulties: relatively large Markov moves

# Reference Set for Marginal versus Conditional

Markov moves given the [rounded]  $[R|T]$  are

0	-1	-1	0	0	0	0	0	0	0
0	0	0	0	0	1	-1	0	0	0
0	1	1	0	0	0	0	0	1008 [1]	3648 [4]
0	0	0	0	0	-1	1	0	-1649 [-2]	-3007 [-3]

- Perturbation using first four moves are maintaining  $[TR]$ .
- Perturbation using the last moves are varying  $[TR]$ .
  - Last move based on rounded  $[R|T]$  allows more than one possible values for  $[TR]$ .
  - $\mathcal{F}_{[R|T]} = \mathcal{F}_{TR_1} \uplus \dots \uplus \mathcal{F}_{TR_i}$

Reference set given conditionals, for example  $[R|T]$ , is corresponding to either

- reference set given the corresponding marginal  $[TR]$
- disjoint union of finite reference sets given  $[TR]$

# Reference Set for Marginal versus Conditional

Markov moves given the [rounded]  $[R|T]$  are

0	-1	-1	0	0	0	0	0	0	0
0	0	0	0	0	1	-1	0	0	0
0	1	1	0	0	0	0	0	1008 [1]	3648 [4]
0	0	0	0	0	-1	1	0	-1649 [-2]	-3007 [-3]

- Perturbation using first four moves are maintaining  $[TR]$ .
- Perturbation using the last moves are varying  $[TR]$ .
  - Last move based on rounded  $[R|T]$  allows more than one possible values for  $[TR]$ .
  - $\mathcal{F}_{[R|T]} = \mathcal{F}_{TR_1} \uplus \cdots \uplus \mathcal{F}_{TR_i}$

Sampling tables given  $[R|T] =$  Combining samplings given  $[TR]$ 's

This framework is a variant of data augmentation in [3].

There are  $I$  possible values for marginal totals,  $\mathcal{T}_1, \dots, \mathcal{T}_I$ , for example,  $[TR]^1, \dots, [TR]^I$ .

- 1 For each  $i^{\text{th}}$  marginal  $\mathcal{T}_i = [TR]^i$ , an initial table,  $\mathbf{n}_{i0}$ , is randomly chosen from  $\mathcal{F}_{\mathcal{T}_i}$ . Then at  $(t+1)^{\text{th}}$  step,
  - 1 Sample  $\lambda^{(t+1)}$  from  $P(\lambda | \mathbf{n}_{i(t)}, \mathcal{T}) \propto P(\mathbf{n}_{i(t)} | \lambda) P(\lambda | \mathcal{T})$ .
  - 2 Sample  $\mathbf{n}_{i(t+1)}$  from  $P(\mathbf{n} | \mathcal{T}, \lambda_{i(t+1)})$ .
- 2 Combine  $\mathbf{n}_{11}, \dots, \mathbf{n}_{IJ}$ .

A contingency table  $\mathbf{n}$  is conditionally independent Poisson distribution with  $\lambda$ ,  $\mathcal{T}$  is the given information,  $\mathbf{n}_{ij}$  is  $j^{\text{th}}$  draw for a  $i^{\text{th}}$  marginal, and  $\mathcal{F}_{\mathcal{T}_i}$  is a set of all possible tables maintaining the given information  $\mathcal{T}_i$ .

# Prior and Posterior Specification in Step 1.1

Conjugate prior,  $P(\lambda|\mathcal{T}) \sim \text{Gamma}(\alpha, \beta)$ , is considered where  $\alpha$  and  $\beta$  are induced by  $\mathcal{T}$ .

$$\begin{aligned} P(\lambda|\mathbf{n}^{(t)}, \mathcal{T}) &\propto P(\mathbf{n}_{i(t)}|\lambda)P(\lambda|\mathcal{T}) \\ &= \prod_k \text{Poisson}(n_k|\lambda_k) \text{Gamma}(\alpha_k, \beta_k) \\ &\propto \prod_k \text{Gamma}\left(n_k^{(t)} + \alpha_k, \left(1 + \frac{1}{\beta_k}\right)^{-1}\right). \end{aligned}$$

# Generating Candidate Tables in Step 1.2

- Generate the candidate table  $\mathbf{n}^*$  from  $q(\mathbf{n}, \mathbf{n}^{(t+1)})$  induced by Markov moves, maintaining the given marginal totals, then  $q(\cdot, \cdot)$  is symmetric.
  - That is, randomly choose one move from the moves preserving the given marginal totals and its sign  $\pm$  with equal probabilities, then add the move to the previous table:  
 $\mathbf{n}^* = \mathbf{n}_{i(t)} + \mathbf{m}$ .
- Accept the candidate table,  $\mathbf{n}^*$  with  $\min\{1, \rho\}$ , where

$$\rho = \frac{P(\mathbf{n}^* | \lambda^{(t+1)})}{P(\mathbf{n}_{i(t+1)} | \lambda^{(t+1)})}$$

# Combining Multiple Samplings in Step 2

In the step 2 "Combine  $\mathbf{n}_{11}, \dots, \mathbf{n}_{IJ}$ " of the framework, we propose two ways:

- Equally-Averaging:  $\mathbf{n}_j = \sum_{i=1}^I \mathbf{n}_{ij}$
- Weighted-Averaging:  $\mathbf{n}_j = \sum_{i=1}^I w_i \mathbf{n}_{ij}$ ,  
where  $w_i = \frac{|\mathcal{M}_i|}{|\mathcal{M}|}$ ,  $\mathcal{M}_i$  is a set of Markov moves maintaining  $\mathcal{T}_i$   
and  $\mathcal{M} = \mathcal{M}_1 \uplus \dots \uplus \mathcal{M}_I$ .

# Example Revisited

A 2x2x2 [STR] Table given [R|T]

## Example

**Table:** A 2x2x2 Table of results of clinical trial for the effectiveness of an analgesic drug

Statue	Treatment	Response		Total
		Poor	Good	
1	1	15	37	52
1	2	22	32	54
2	1	6	39	45
2	2	12	30	42
Total		55	138	193

# Example Revisited

A 2x2x2 [STR] Table given [R|T]

## Example

**Table:** [Response|Treatment] Table of conditional probabilities 7 with [rounded probability]

Treatment	Response		Total
	Poor	Good	
1	$\frac{21}{97} = 0.2164948$ [0.2]	$\frac{76}{97} = 0.7835052$ [0.8]	97
2	$\frac{34}{96} = 0.3541667$ [0.4]	$\frac{62}{96} = 0.6458333$ [0.6]	96
Total	55	138	193

# Example Revisited

**Table:** Summaries for first eight cells in the table of 10000 simulations for each method of cell entries from the posterior distribution  $P(\mathbf{n}|\mathcal{I}, \lambda)$  with  $\lambda \sim \text{Gamma}(193, \frac{1}{8})$ . [Ordinary **Equally-Averaging** **Weighted-Averaging**]

Cell	True Value	Posterior Mean	Posterior Median	Posterior Quantiles [Q <sub>0.025</sub> , Q <sub>0.975</sub> ]
(1,1,1)	15	9.9 <b>10.6</b> 10.8	10 <b>10</b> 11	[5,15] <b>[5,16]</b> [6,15]
(1,1,2)	37	35.6 <b>37.9</b> 37.6	35 <b>37</b> 37	[24,50] <b>[24,53]</b> [27,54]
(1,2,1)	22	17.6 <b>17</b> 17.2	17 <b>17</b> 17	[11,26] <b>[9,26]</b> [10,24]
(1,2,2)	32	30.7 <b>30.7</b> 30.0	30 <b>30</b> 30	[21,39] <b>[20,43]</b> [22,40]

- Combining multiple samplings given marginal totals provide similar posterior statistics to sampling given conditionals but the point estimators based on combining multiple samples are slightly closer to the original cell entries.
- Combining sample based on weighted-averaging provides smaller intervals on cell entries than other two methods.

# Considerations

- Prior and posterior specification on  $\lambda$  in Step 1.1
- Different sampling schemes: Sequential importance sampling
- Difference combining schemes
- Extended to high dimensional tables
- The case given a set of marginal totals and conditional probabilities

# Alternative MCMC framework

$\mathcal{F}_{R|T} = \mathcal{F}_{TR_1} \uplus \dots \uplus \mathcal{F}_{TR_l}$ .  $\mathcal{M} = \mathcal{M}_1 \uplus \mathcal{M}_2$  is a set of Markov moves given  $[R|T]$ , where  $\mathcal{M}_1$  is a set of moves varying  $[TR]$  and  $\mathcal{M}_2$  is a set of moves maintaining  $[TR]$ . Define  $\mathcal{M}_{1_0} = \mathcal{M}_1 \cup \{\mathbf{0}\}$ .

$\mathbf{n}^{(0)}$  is randomly chosen from  $\mathcal{F}_{R|T}$ . Then, at  $(t+1)^{th}$  step

- 1 Sample  $\lambda^{(t+1)}$  from  $P(\lambda|\mathbf{n}^{(t)}, \mathcal{T}) \propto P(\mathbf{n}^{(t)}|\lambda)P(\lambda|\mathcal{T})$ .
- 2 Sample  $\mathbf{n}^{(t+1)}$  from  $P(\mathbf{n}|\mathcal{T}, \lambda^{(t+1)})$ .
  - 1 Generate the candidate table with two steps:
    - 1 Randomly choose one move,  $\mathbf{m}_1$  from  $\mathcal{M}_{1_0}$  and add it to the previous table:  $\mathbf{n}^{(t)} + \mathbf{m}_1$ . In this step, the chosen move either changes the marginal totals or preserves marginal totals, that is, the state either jumps to one of difference sub-reference sets or stays in the same sub-reference set.
    - 2 Randomly choose one move,  $\mathbf{m}_2$  from  $\mathcal{M}_2$  and add it to the table from the previous step:  $\mathbf{n}^{(*)} = \mathbf{n}^{(t)} + \mathbf{m}_1 + \mathbf{m}_2$ . In this step, the state travels within the chosen sub-reference set.
  - 2 Accept the candidate table,  $\mathbf{n}^*$  with  $\min\{1, \rho\}$ , where

$$\rho = \frac{P(\mathbf{n}^*|\lambda^{(t+1)})}{P(\mathbf{n}^{(t+1)}|\lambda^{(t+1)})}.$$