Sampling Importance Sampling for Contingency Tables

STAT 525

9/20/18

Motivating Example I: 82 descendants of Queen Victoria

Month of birth	Month of death												
	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
March	1	0	0	0	2	1	0	0	0	0	0	1	5
April	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
June	2	0	0	0	1	0	0	0	0	0	0	0	3
July	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sept	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

Exact Test of Independence

- Observations are from Multinomial$(n,\ p_{ij})$
- Null hypothesis: Independence between row and column variables.

$$H_0:\ p_{ij} = p_{i\cdot}p_{\cdot j}, \quad \text{for } i = 1, \ldots, k; j = 1, \ldots, l.$$

- Probability of observing table T

$$
\begin{aligned}
P(T) &= \binom{n}{n_{11}, \ldots, n_{kl}} \prod_{i=1}^{k} \prod_{j=1}^{l} p_{ij}^{n_{ij}} \\
&= \binom{n}{n_{11}, \ldots, n_{kl}} \prod_{i=1}^{k} \prod_{j=1}^{l} (p_{i\cdot} p_{\cdot j})^{n_{ij}} \\
&= \binom{n}{n_{11}, \ldots, n_{kl}} \left(\prod_{i=1}^{k} p_{i\cdot}^{n_{i\cdot}} \right) \left(\prod_{j=1}^{l} p_{\cdot j}^{n_{\cdot j}} \right)
\end{aligned}
$$

where $n_{i\cdot}$ and $n_{\cdot j}$ are sufficient statistics.

- If we fix $n_{i\cdot}$ and $n_{\cdot j}$, for $i = 1, \ldots, k$, $j = 1, \ldots, l$, the conditional distribution is

$$\pi(T) \propto \binom{n}{n_{11}, \ldots, n_{kl}} \propto \frac{1}{\prod_{i=1}^{k} \prod_{j=1}^{l} n_{ij}!} \quad \text{for } T \in \Omega$$

where Ω is the set of tables with given $n_{i\cdot}$ and $n_{\cdot j}$

- The exact p-value is

$$\mu = \sum_{T \in \Omega} 1_{\{\pi(T) \leq \pi(T_{obs})\}} \pi(T)$$

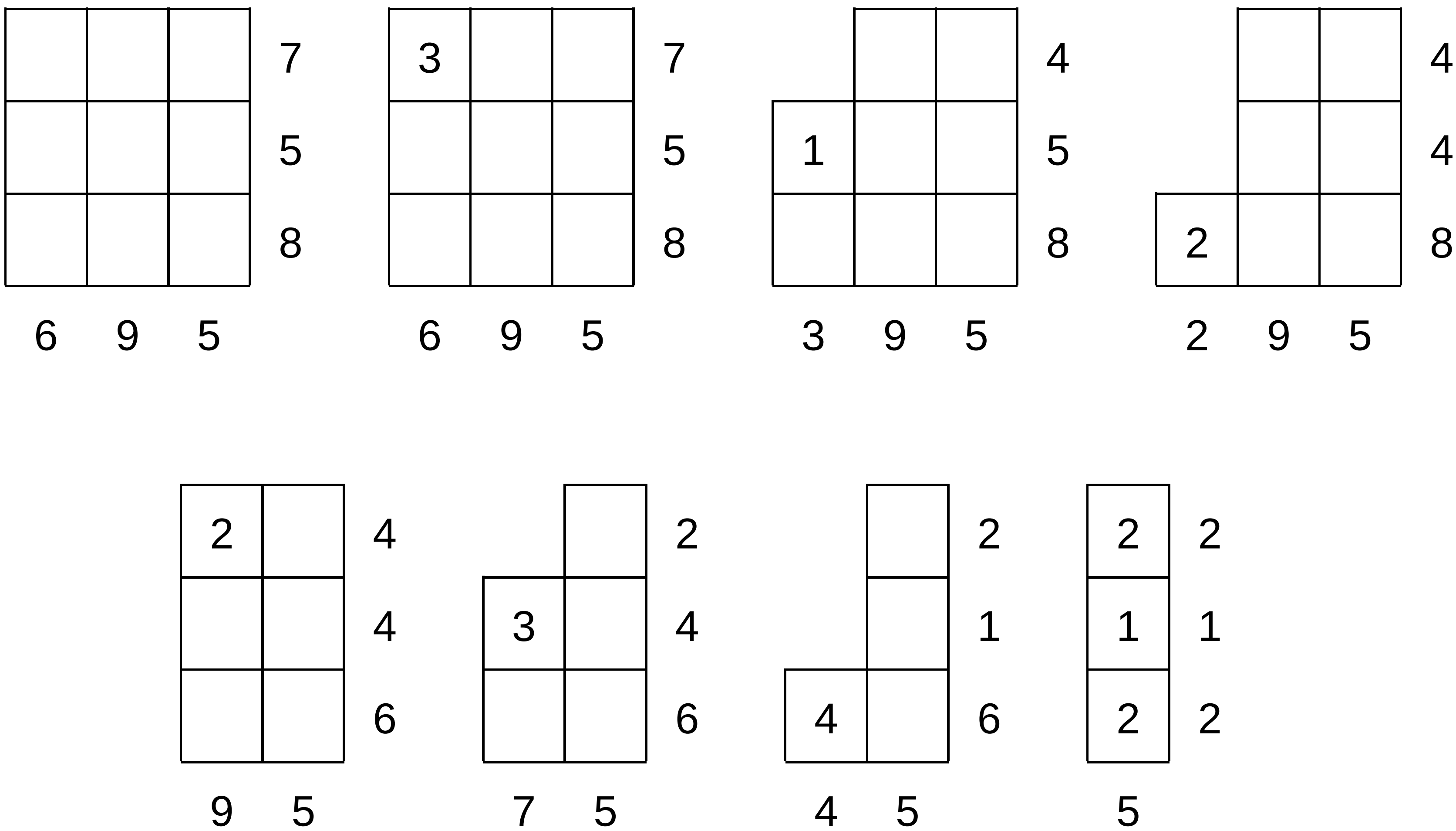
How to Compute the p-value ?

- The exact p-value is

$$\mu = \sum_{T \in \Omega} 1_{\{\pi(T) \leq \pi(T_{obs})\}} \pi(T)$$

- Consider sequential importance sampling (SIS)

Example:



Questions

- What is the support of the conditional distribution $t_i \big| (t_{i-1}, \ldots, t_1)$?
- How to sample from the support of the conditional distribution?

Fréchet Bounds for Two-Way Tables

			$\cdots$		r_1
			$\cdots$		r_2
			$\cdots$		r_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			$\cdots$		r_m
c_1	c_2	c_3	$\cdots$	c_n	

$$
\begin{aligned}
\max(0, c_1 - r_2 - \cdots - r_m) \leq \quad & t_{11} \quad \leq \min(r_1, c_1) \\
\max(0, c_1 - t_{11}^* - r_3 - \cdots - r_m) \leq \quad & t_{21} \quad \leq \min(r_2, c_1 - t_{11}^*) \\
& \vdots
\end{aligned}
$$

Sampling Distribution

- Difficult to obtain the true distribution of an entry conditional on the entries that have already been sampled.
- For a target uniform distribution, sample a cell value uniformly from the interval $[l, u]$.
- For a target hypergeometric distribution, sample a cell value from the hypergeometric distribution $p(x) = \binom{u}{x}\binom{u}{l+u-x}/\binom{2u}{l+u}$ on the interval $[l, u]$.

Counting Tables

- $\#P$ complete problem: How many tables satisfy the given constraints?
- Counting Tables by SIS
  - Note that $|\Omega| = \sum_{T \in \Omega} \frac{1}{q(T)} q(T)$.
  - Draw independent samples $T^{(1)}, \cdots, T^{(N)}$ from $q(T)$.
  - Estimate by

$$\widehat{|\Omega|} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{q(T^{(i)})}.$$

Eye Color versus Hair Color

	Black	Brunette	Red	Blonde	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

- Estimation: $(1.225 \pm 0.002) \times 10^{15}$.

  True: $1.225914276768514 \times 10^{15}$ (Diaconis and Gangolli, 1995).

References

- Chen, Y., Dinwoodie, I. H., and Sullivant, S. (2006). Sequential Importance Sampling for Multiway Tables. *The Annals of Statistics*, **34**, 523-545.
- Chen, Y., Diaconis, P., Holmes, S., and Liu, J.S. (2005). Sequential Monte Carlo Methods for Statistical Analysis of Tables. *Journal of the American Statistical Association*, **100**, 109-120.