Effection of Swapping on the Multinomial Example

Qiru Pan Yunhao Li Yicong Li

16:960:690: Ethical Statistical Learning

December 3rd, 2023

Data Swapping (Dalenius and Reiss 1982; Fienberg and McIntyre 2004)

1. Original Table:

Name	Sect	Role	#Attend	...
John Smith	01	Teacher	14	...
Ava Chen	02	Student	14	...
Zoe Kim	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

2. Partition Variables into $\mathbf{V}_{\text{Swap}}$ and $\mathbf{V}_{\text{Hold}}$:

Name	Sect	Role	#Attend	...
John Smith	01	Teacher	14	...
Ava Chen	02	Student	14	...
Zoe Kim	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

3. Define $\mathbf{V}_{\text{Match}} \subset \mathbf{V}_{\text{Hold}}$:

Name	Sect	Role	#Attend	...
John Smith	01	Teacher	14	...
Ava Chen	02	Student	14	...
Zoe Kim	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

4. Proceed **a swap step** on the records of $\mathbf{V}_{\text{Swap}}$:

Name	Sect	Role	#Attend	...
Zoe Kim	01	Teacher	14	...
Ava Chen	02	Student	14	...
John Smith	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

Data Swapping (Dalenius and Reiss 1982; Fienberg and McIntyre 2004)

1. Original Table:

Name	Sect	Role	#Attend	...
John Smith	01	Teacher	14	...
Ava Chen	02	Student	14	...
Zoe Kim	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

2. Partition Variables into $\mathbf{V}_{\text{Swap}}$ and $\mathbf{V}_{\text{Hold}}$:

Name	Sect	Role	#Attend	...
John Smith	01	Teacher	14	...
Ava Chen	02	Student	14	...
Zoe Kim	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

3. Define $\mathbf{V}_{\text{Match}} \subset \mathbf{V}_{\text{Hold}}$:

Name	Sect	Role	#Attend	...
John Smith	01	Teacher	14	...
Ava Chen	02	Student	14	...
Zoe Kim	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

4. Proceed **a swap step** on the records of $\mathbf{V}_{\text{Swap}}$:

Name	Sect	Role	#Attend	...
Zoe Kim	01	Teacher	14	...
Ava Chen	02	Student	14	...
John Smith	01	Observer	14	...
Leo Park	02	Student	13	...
⋮	⋮	⋮	⋮	⋱

Note that:

- Each record is independently selected with probability given by **swap rate** p.
- swapping is restricted to records which share the same values on $\mathbf{V}_{\text{Match}}$.

The Permutation Algorithm (Bailie, Gong and Meng, 2023)

Input: a dataset **X**.
Define **strata** as groups of records which match on the swap key $\mathbf{V}_{\text{Match}}$.
Within each stratum:

1. Select each record independently with probability p (the swap rate).
2. Derange swapping variable $\mathbf{V}_{\text{Swap}}$ of selected records, uniformly at random.

Output: the swapped dataset **Z**.

```
Input: Dataset X
 1: for j = 1, ..., 𝒥 do
 2:   if n_j = 0 or n_j = 1 then
 3:     continue
 4:   end if
 5:   for record i with category j do
 6:     Select i with probability p
 7:   end for
 8:   if 0 records selected then
 9:     continue
10:   else if exactly 1 record selected then
11:     go to line 5
12:   end if
13:   Sample uniformly at random a derangement σ of the selected records.
14:   /* Permute the swapping variable of the selected records according to σ: */
15:     Save copy X_0 ← X before permutation
16:     Let k^X(i) be the value of the swapping variable of record i in dataset X.
17:     for all selected records i do
18:       Set k^X(i) ← k^{X_0}(σ(i))
19:     end for
20: end for
21: Set Z ← X to be the swapped dataset.
22: return contingency table [n^Z_{jkl}]
```

Theorem (Bailie, Gong and Meng, 2023)

The Permutation Algorithm satisfies pure differential privacy with privacy loss budget

$$\epsilon = \ln(b+1) - \ln(o), \quad \text{for } 0 < p \leq 0.5,$$

conditioning on the invariants it induces, where $o = p/(1-p)$ and b is the largest stratum size.

Swapping Satisfies DP, Conditioning on its Invariants

Theorem (Bailie, Gong and Meng, 2023)

The Permutation Algorithm satisfies pure differential privacy with privacy loss budget

$$\epsilon = \ln(b+1) - \ln(o), \quad \text{for } 0 < p \leq 0.5,$$

conditioning on the invariants it induces, where $o = p/(1-p)$ and b is the largest stratum size.

Theorem (formal) (Bailie, Gong and Meng, 2023)

The Permutation Algorithm satisfies $(\mathscr{D}_{\mathbf{c}_{\text{Swap}}}, d^u_{\text{HamS}}, \text{Mult})$ differential privacy with privacy loss budget

$$\epsilon = \ln(b+1) - \ln(o), \quad \text{for } 0 < p \leq 0.5,$$

with $o = p/(1-p)$ and b is (roughly) the largest stratum size.

Swapping on the Multinomial Example

Suppose we have binary variables V_1, V_2 for Question 1 and Question 2, respectively, forming a $n \times 2$ data table of answer combinations.

Let $V_1 \in \mathbf{V}_{\text{Swap}}$, $V_2 \in \mathbf{V}_{\text{Hold}}$,

Swapping on the Multinomial Example

Suppose we have binary variables V_1, V_2 for Question 1 and Question 2, respectively, forming a $n \times 2$ data table of answer combinations.

Let $V_1 \in \mathbf{V}_{\text{Swap}}$, $V_2 \in \mathbf{V}_{\text{Hold}}$, say there exists $V_m \in \mathbf{V}_{\text{Match}}$ such that all records in $\mathbf{V}_{\text{Swap}}$ share the same values on V_m. Wolg, we omit V_m.

V_1	V_2
1	0
1	0
0	1
1	1
0	0
$\vdots$	$\vdots$

V_1	V_2
1 - 1	0
1	1
0+1	1
1	1
0	0
$\vdots$	$\vdots$

Swapping on the Multinomial Example

Suppose we have binary variables V_1, V_2 for Question 1 and Question 2, respectively, forming a $n \times 2$ data table of answer combinations.

Let $V_1 \in \mathbf{V}_{\text{Swap}}$, $V_2 \in \mathbf{V}_{\text{Hold}}$, say there exists $V_m \in \mathbf{V}_{\text{Match}}$ such that all records in $\mathbf{V}_{\text{Swap}}$ share the same values on V_m. Wolg, we omit V_m.

V_1	V_2
1	0
1	0
0	1
1	1
0	0
⋮	⋮

V_1	V_2
1 - 1	0
1	1
0+1	1
1	1
0	0
⋮	⋮

- Marginal Table

#(0, 0)	#(0, 1)
#(1, 0)	#(1, 1)

Swapping on the Multinomial Example

Suppose we have binary variables V_1, V_2 for Question 1 and Question 2, respectively, forming a $n \times 2$ data table of answer combinations.

Let $V_1 \in \mathbf{V}_{\text{Swap}}$, $V_2 \in \mathbf{V}_{\text{Hold}}$, say there exists $V_m \in \mathbf{V}_{\text{Match}}$ such that all records in $\mathbf{V}_{\text{Swap}}$ share the same values on V_m. Wolg, we omit V_m.

V_1	V_2
1	0
1	0
0	1
1	1
0	0
$\vdots$	$\vdots$

V_1	V_2
1 - 1	0
1	1
0+1	1
1	1
0	0
$\vdots$	$\vdots$

- Marginal Table

$\theta_1 = [\#(0,0)]/n$	$\theta_2 = [\#(0,1)]/n$
$\theta_3 = [\#(1,0)]/n$	$\theta_4 = [\#(1,1)]/n$

Swapping on the Multinomial Example

Suppose we have binary variables V_1, V_2 for Question 1 and Question 2, respectively, forming a $n \times 2$ data table of answer combinations.

Let $V_1 \in \mathbf{V}_{\text{Swap}}$, $V_2 \in \mathbf{V}_{\text{Hold}}$, say there exists $V_m \in \mathbf{V}_{\text{Match}}$ such that all records in $\mathbf{V}_{\text{Swap}}$ share the same values on V_m. Wolg, we omit V_m.

V_1	V_2
1	0
1	0
0	1
1	1
0	0
⋮	⋮

V_1	V_2
1 - 1	0
1	1
0+1	1
1	1
0	0
⋮	⋮

- Marginal Table

$\theta_1 = [\#(0,0)]/n$	$\theta_2 = [\#(0,1)]/n$
$\theta_3 = [\#(1,0)]/n$	$\theta_4 = [\#(1,1)]/n$

- Test Statistic

$$\phi = \frac{\theta_1\theta_4}{\theta_2\theta_3}$$

Interpretation of ϕ

	$V_1 = 0$	$V_2 = 0$
$V_1 = 0$	θ_1	θ_2
$V_1 = 1$	θ_3	θ_4

$$\phi = \frac{\theta_1 \theta_4}{\theta_2 \theta_3}$$

- $\phi = 1$: no association (independence) between V_1 and V_2, i.e. the occurrence of one variable does not affect the occurrence of the other.

  e.g. $\theta = (0.25, 0.25, 0.25, 0.25)$, $\phi = 1$

- $\phi > 1$: positive association. One variable positively influences the occurrence of the other.

  e.g. $\theta = (0.3, 0.2, 0.2, 0.3)$, $\phi > 1$

- $\phi < 1$: negative association (inverse relationship) between V_1 and V_2.

  e.g. $\theta = (0.2, 0.3, 0.3, 0.2)$, $\phi < 1$

Problem Formulation

- Primary Concern: Without altering any marginal totals, would a swap algorithm, satisfying DP, impact ϕ, on the $n \times 2$ dataset?

Problem Formulation

- Primary Concern: Without altering any marginal totals, would a swap algorithm, satisfying DP, impact ϕ, on the $n \times 2$ dataset?
- If yes, what causes this impact?
- As data analysts, assume the privatized data we get is trustworthy and analyze it using a naive analysis, what kind of results or errors might we encounter?
- How can we measure effect of swap algorithm so that we can effectively use Bayesian statistical methods to draw meaningful statistical inference from data that has been altered for privacy?

Generating Models

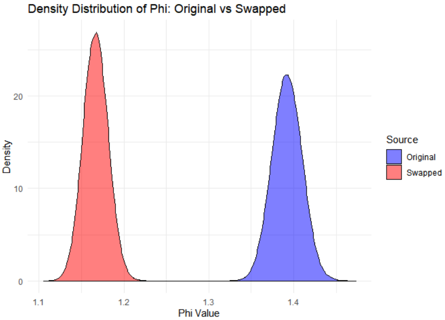
- Prior: $\pi(\theta) \sim \mathrm{Dir}(1, 1, 1, 1)$
- Likelihood: $\mathbf{y}|\theta \sim \mathrm{Mult}(n, \theta)$
- Posterior: $p(\theta|\mathbf{y}) \sim \mathrm{Dir}(1 + y_1, 1 + y_2, 1 + y_3, 1 + y_4)$

Observe Distribution of ϕ

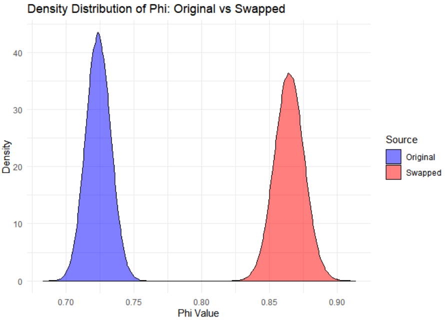
$$\phi = \frac{\theta_1\theta_4}{\theta_2\theta_3}$$

Findings

Fix swap rate $p = 0.5$, with different choices of θ:



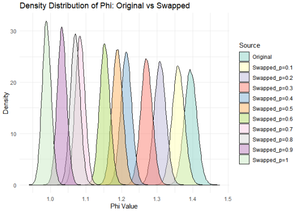
$\theta^* = (0.27, 0.23, 0.23, 0.27)$



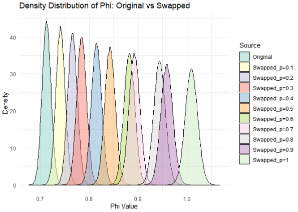
$\theta^* = (0.23, 0.27, 0.27, 0.23)$

Findings

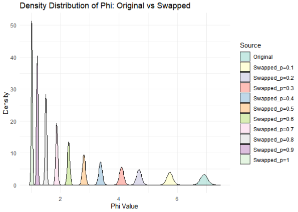
Apply different choices of p and θ:



$\theta^* = (0.27, 0.23, 0.23, 0.27)$



$\theta^* = (0.23, 0.27, 0.27, 0.23)$



$\theta^* = (0.7, 0.1, 0.1, 0.1)$

Note that by theorem, $0 < p \leq 0.5$.

Result

1. The influence of swapping algorithm on the data varies depending on specific values of θ^*, resulting in different trends in the ϕ distribution.
2. As swap rate p increases, swapping algorithm appears to consistently shift the value of ϕ closer to 1, likely because data becomes more uniform as a result of the swapping process.

Result

1. The influence of swapping algorithm on the data varies depending on specific values of θ^*, resulting in different trends in the ϕ distribution.
2. As swap rate p increases, swapping algorithm appears to consistently shift the value of ϕ closer to 1, likely because data becomes more uniform as a result of the swapping process.

- Data Privacy Implications: The swapping method seems to effectively increases data anonymity and randomness. It introduces additional randomness, which may disrupt the original structures or patterns in the data, enhancing privacy.
- Data Structure Considerations: Conversely, if the objective is to preserve the inherent structures or patterns of the original data, alternative approaches might be necessary. The swapping process, as noted, can disrupt some of the fundamental characteristics of the original data.

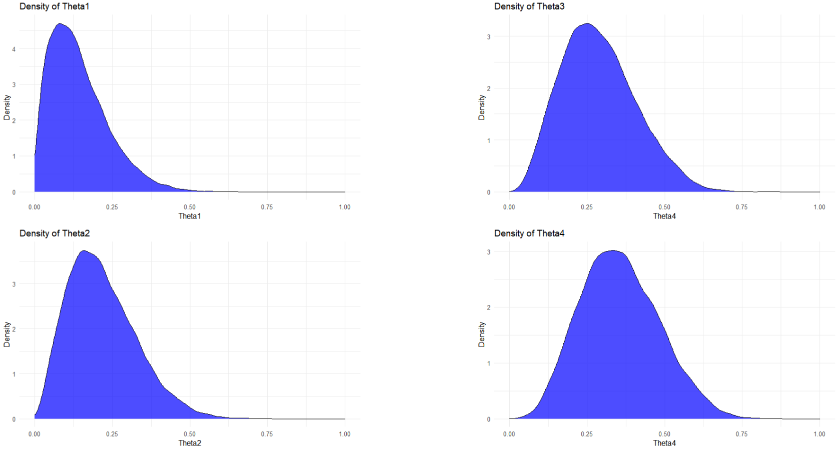
Statistical Inference with Approximate Bayesian Computation (ABC)

Methodology:

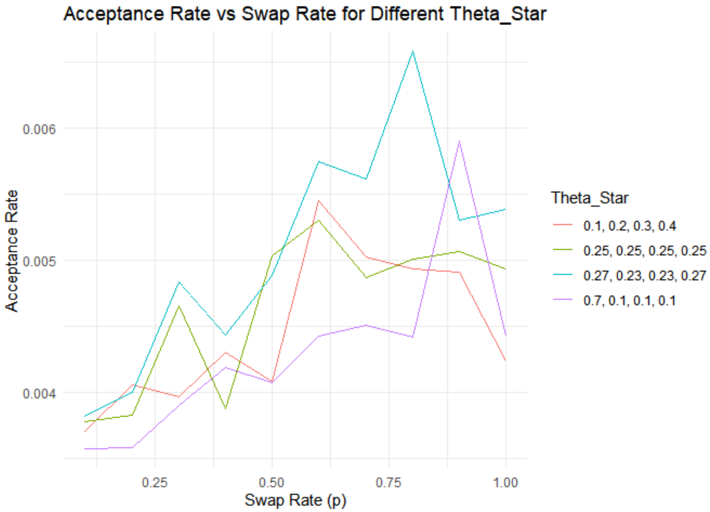
- Given dataset S_{obs}, assume Prior: $\theta^{(i)} \sim \text{Dir}(1, 1, 1, 1)$,
- For each iteration:
  1. generate $\theta^{(i)}$ from Prior.
  2. generate data $X^{(i)} | \theta^{(i)} \sim \text{Mult}(n, \theta)$.
  3. Apply swapping algorithm $\eta_p S | X$ with a swapping rate p
  4. Acceptance Criterion: compare the marginal table of S_{obs} with the marginal table of $S^{(i)}$.
     - If $S_{\text{obs}} = S^{(i)}$, retain $\theta^{(i)}$.
     - Otherwise, skip current iteration.
- Expect to obtain the distribution of θ concerning observed data S_{obs}

Findings

Let $p = 0.1$, marginal table of $S = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, with $5,000,000$ iterations:



Further Findings



Reference

1. Bailie, J., Gong, R., and Meng, X.-L. (2023). "Can Swapping be Differentially Private? A Refreshment Stirred, not Shaken." Retrieved from https://conference.nber.org/conf_papers/f178188.pdf.
2. Dalenius, T. and Reiss, S. P. (1982). "Data-Swapping: A Technique for Disclosure Control". Journal of Statistical Planning and Inference, 6(1), 73–85. doi: 10.1016/0378-3758(82)90058-1.
3. Fienberg, S. and McIntyre, J. (2004). "Data Swapping: Variations on a Theme by Dalenius and Reiss". In Privacy in Statistical Databases. doi: 10.1007/978-3-540-25955-8_2.
4. Gong, R. (2019). "Exact Inference with Approximate Computation for Differentially Private Data via Perturbations." arXiv:1909.12237.
5. Ju, N., Awan, J. A., Gong, R., and Rao, V. A. (2022). "Data Augmentation MCMC for Bayesian Inference from Privatized Data." arXiv:2206.00710.
6. Zhang, L. (2023). "STAT 16:960:690: Ethical Statistical Learning".