Synthetic Data Overview

Introduction to Statistical Data Privacy

Jordi Cortés Martínez

based on slides of Jingchen Monika Hu (Vassar College)

VIII RETREAT 2023, SITGES

Outline

1. Overview
2. Synthetic data generation
3. Utility
    - Global Utility
    - Specific-analysis utility
4. Disclosure risk

Overview

1. Examples
2. Synthetic data rationale
3. Typologies

Examples of data suitable of being protected

- Protecting **income information** in the National Health Interview Survey (NHIS)
- Protecting individual’s privacy (**political opinion**) in the AP VoteCast
- Protecting **price and available days information** of Airbnb listings in New York City
- Protecting individual’s privacy (**demographics**) in the STEM Labor Force survey
- Protecting patient's electronic health record (EHR) with **medical history**

What are synthetic data and how are they created?

- To provide **privacy protection** of individuals in datasets

- **Preserve data integrity** (e.g., important characteristics in the confidential data, such as means and correlations of variables)

- Usually created by **simulating variables** of records from statistical models or AI algorithms estimated on the confidential data

- If the statistical models and estimations are appropriate, we can capture **important characteristics in the confidential data**

- Moreover, they can provide some **levels of privacy protection**, as compared to releasing the confidential data

Types of synthetic data generation and evaluation

- Types of synthetic **data**:
    - Partial synthesis
    - Full synthesis
- Types of synthetic **data generation**:
    - Sequential synthesis
    - Joint synthesis
- Types of **evaluation**:
    - Utility
    - Disclosure risks

Data Generation

1. Partial vs. Full synthesis
2. Sequential vs. Joint synthesis
3. Evaluation
4. Datasets

Partial synthesis

- Proposed by **Little** (1993)

- **Some** variables in the collected dataset, such as **sensitive variables** and **key identifiers**, are synthesized

- The resulting synthetic data contain sensitive variables with synthesized values while **other variables remain unchanged**

Full synthesis

- Proposed by **Rubin** (1993)
- A **synthetic population** is first simulated
- Then a **synthetic sample** is selected from the synthetic population
- The resulting synthetic data have **every variable synthesized**, contain no records from the confidential data, and it may even have a **different sample size** than the confidential data if needed
- One can also create fully synthetic data following the **partial synthesis approach**, i.e., only working on the sample
    - This approach is actually more widely used when creating fully synthetic data

Synthesis approaches

- **Sequential synthesis**
  - Variable synthetized one by one
  - More commonly used and easier to estimate
- **Joint synthesis**
  - Variables jointly synthetized
  - Less commonly used and usually more challenging to estimate
  - The DPMPM model for multivariate nominal categorical data (Hu, Reiter, and Wang, 2014)

The CE sample

- Our sample is from the 1st quarter of 2019, containing **five** variables on **5133** CUs

Name	Variable information
UrbanRural	Binary; the urban/ rural status of CU: 1 = Urban, 2 = Rural
Income	Continuous; the amount of CU income before taxes in past 12 months (in *USD*)
Race	Categorical; the race category of the reference person: 1 = White, 2 = Black, 3 = Native American, 4 = Asian, 5 = Pacific Islander, 6 = Multi-race
Expenditure	Continuous; CU's total expenditures in last quarter (in *USD*)
KidsCount	Count; the number of CU members under age 16

The CE sample (3 rows)

```
## # A tibble: 3 x 5
##   UrbanRural Income  Race Expenditure KidsCount
##        <dbl>  <dbl> <dbl>       <dbl>     <dbl>
## 1          1  73720     1      27542.         3
## 2          1  12000     1       7416.         2
## 3          1  20000     1       8608.         0
```

The ACS sample

- Our sample is from 2012 ACS public use microdata, containing **10** (not all shown) categorical variables on **10,000** observations

Name	Variable information
SEX	1 = male, 2 = female
RACE	1 = White alone, 2 = Black or African American alone, 3 = American Indian alone, 4 = other, 5 = two or more races, 6 = Asian alone
MAR	1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married.
LANX	1 = speaks another language, 2 = speaks only English
WAOB	born in: 1 = US state, 2 = PR and US island areas, oceania and at sea, 3 = Latin America, 4 = Asia, 5 = Europe, 6 = Africa, 7 = Northern America.

The ACS sample (3 rows)

```
   SEX  RACE   MAR  LANX  WAOB   DIS HICOV   MIG   SCH  HISP
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
     2     1     1     2     1     2     1     3     1     1
     1     1     1     2     1     2     1     3     1     1
     2     6     5     1     4     2     2     2     1     1
```

Continuous synthetic variable: Transformations

- Suppose we want to use the **Income** variable to perform a partial synthesis for **Expenditure**
- Both Income and Expenditure are **highly skewed**, we apply the logarithm transformation for both in our model building, creating two new variables: **LogIncome** and **LogExpenditure**

A Bayesian simple linear regression: model specification

- Suppose we want to use the **Income** variable to perform a partial synthesis for **Expenditure**
- Y_i is **LogExpenditure** and X_i is **LogIncome** for CU
- A Bayesian simple linear regression model:

$$Y_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 X_i$$

- The expected **LogExpenditure** of CU i is μ_i, which is a linear function of **LogIncome** X_i through the intercept parameter β_0 and the slope parameter β_1
- A **priori distributions** should be defined for the parameters to estimate $(\beta_0, \beta_1, \sigma)$
- Both Income and Expenditure are **highly skewed**, we apply the logarithm transformation for both in our model building, creating two new variables: **LogIncome** and **LogExpenditure**

A Bayesian simple linear regression: model estimation

- We use the **`brms`** R package to estimate our chosen Bayesian simple linear regression model
- We will obtain pre-specified number of **posterior parameter** draws from the posterior distribution.
- These posterior parameter draws will be used for synthetic data generation through the **posterior predictive distribution**

A Bayesian simple linear regression: model estimation

- The key to applying Bayesian synthesis models is to **save posterior parameter draws** of estimated parameters
- These draws will be used to generate synthetic data given the **posterior predictive distribution**.
- We use the `posterior_samples` function to retrieve the posterior parameter draws.

```
##   b_Intercept b_LogIncome     sigma    lprior       lp
## 1    5.070943   0.3526935 0.7508692 -3.200975 -5799.107
## 2    5.147291   0.3460304 0.7521475 -3.201098 -5799.412
## 3    5.096195   0.3543455 0.7645371 -3.202682 -5802.832
```

A Bayesian simple linear regression: MCMC diagnostics

Don’t forget it

A Bayesian simple linear regression: synthesis

- To predict the **LogExpenditure**, Y_i^* for a CU given its **LogIncome**, X_i

and a set of parameter draws, denoted as $\{\beta_0^*, \beta_1^*, \sigma^*\}$:

$$Y_i^* | \beta_0^*, \beta_1^*, \sigma^* \sim Normal(\beta_0^*, +\beta_1^* X_i, \sigma^*)$$

- Now for each of the *n* CUs, we create a **predicted value**:

Compute $E[Y_1^*] = \beta_0^* + \beta_1^* X_1$ → Sample $Y_1^* \sim Normal(E[Y_1^*], \sigma^*)$

$\vdots$

Compute $E[Y_i^*] = \beta_0^* + \beta_1^* X_i$ → Sample $Y_i^* \sim Normal(E[Y_i^*], \sigma^*)$

$\vdots$

Compute $E[Y_n^*] = \beta_0^* + \beta_1^* X_n$ → Sample $Y_n^* \sim Normal(E[Y_n^*], \sigma^*)$

The joint distribution of synthesized variables

- This **sequential** versus **joint** classification is based on what strategy is used to estimate the joint distribution of the variables to be synthesized
- Variables to be **synthesized**: $\{y_1, y_2, y_3\}$
- **Insensitive** variables: $\{x_1, x_2\}$
- The **joint distribution** of synthesized variables:

$$f(y_1, y_2, y_3 | x_1, x_2)$$

Sequential synthesis

- This approach specifies a **sequence of univariate synthesis models** for the sensitive variables
- The **joint model** can be expressed as follows:

$$
\begin{aligned}
f(y_1, \cdots, y_{p_1} \mid x_1, \cdots, x_{p_2}) = \quad & f(y_1 \mid x_1, \cdots, x_{p_2}) \times \\
& f(y_2 \mid y_1, x_1, \cdots, x_{p_2}) \times \\
& \ldots \\
& f(y_{(p_1-1)} \mid y_1, \cdots, y_{(p_1-2)}, x_1, \cdots, x_{p_2}) \\
& f(y_{p_1} \mid y_1, \cdots, y_{(p_1-1)}, x_1, \cdots, x_{p_2})
\end{aligned}
$$

Sequential synthesis procedure

1. Specify a synthesis model for $y_1 | x_1, \dots, x_{p_2}$. Estimate this model on the confidential data, and generate synthetic y_1^* using $(x_1, \dots, x_{p_2})$.

2. Specify a synthesis model for $y_2 | y_1, x_1, \dots, x_{p_2}$. Estimate this model on the confidential data, and generate synthetic y_2^* using synthetic y_1^* from step 1 and confidential $(x_1, \dots, x_{p_2})$.

3. Repeat Step 2 for each of the variables of $\{y_3, \dots, y_{(p_1-1)}\}$.

4. Specify a synthesis model for $y_{p_1} | y_1, \dots, y_{(p_1-1)}, x_1, \dots, x_{p_2}$. Estimate this model on the confidential data, and generate synthetic $y_{p_1}^*$ using synthetic $\left(y_1^*, y_2^*, \dots, y_{(p_1-1)}^*\right)$ from previous steps and confidential $(x_1, \dots, x_{p_2})$.

Joint synthesis

- The **joint distribution**: $f(y_1, \ldots, y_{p_1} | x_1, \ldots, x_{p_2})$
- **Directly** specify a joint model for these sensitive variables
- For example, if $\{y_1, \ldots, y_{p_1}\}$ are all continuous (and marginally normal after transformation), we can use a **multivariate normal distribution**:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{p_1} \end{bmatrix} \sim \mathrm{MVN}_{p_1}\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{p_1} \end{bmatrix} \Sigma\right)$$

  - MVN_{p_1} stands for multivariate normal distribution of dimension p_1
  - $\mu_1, \ldots, \mu_{p_1}$ are the mean parameters (conditional on $x_1, \ldots, x_{p_2}$)
  - Σ is the covariance matrix

Joint synthesis cont'd

- If sensitive variables are all categorical, a well-researched model is the **Dirichlet Process mixture of products of multinomials** (DPMPM) (Hu, Reiter, and Wang, 2014);
- Joint synthesis model estimation is usually **more challenging** than sequential synthesis
- **Bayesian networks** are good approaches (Young, Graham, and Penny, 2009), Kaur et al. , 2021)

R Example 1: Sequential synthesis vs. Joint synthesis

- Generate **LogExpenditure** and **Kidcounts** using **sequential synthesis** with **CEData**
- Generate **DIS and HICOV** variables using **joint synthesis** with **ACSdata**

Utility

1. Global Utility
   - pMSE
   - eCDF
2. Specific-analysis utility
   - Interval overlap

Utility evaluation

Two general types:

1. **Global utility**: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
   - The key to all these global measures is in **discriminating** between the confidential and the synthetic

2. **Analysis-specific utility**: Evaluate whether synthetic data users can obtain statistical inferences on the synthetic data that are similar to those obtained on the confidential data
   - Utility measures are **tailored** to the analyses the data analyst is expected to perform on the synthetic data

Propensity scores

- Propensity scores measure the probability for individuals in a dataset being assigned to a specific treatment group given their information on other variables
- Consider two groups: group A receiving a treatment and group B not receiving it
- Given other measured covariates, we can estimate for each individual the probability that they were assigned to group A
- If the probability distributions differ for the two groups, this indicates that the individuals in the two groups differ with respect to these covariates

Propensity scores matching vs synthetic data

- **Propensity score matching**, a technique commonly employed in causal inference aimed at reducing bias due to confounding variables when estimating the effect of a treatment in an observational study
- To evaluate the **quality of synthetic datasets**, we use them to investigate whether the synthetic observations significantly differ from the original observations
- We follow Woo et al. (2009) and Snoke et al. (2018) and introduce the **pMSE** as the propensity scores utility measure

pMSE calculation

- In the synthetic data setting, the "*treatment*" is if the observation is part of the generated synthetic dataset
- Steps to compute ***pMSE*** metric:

1. Merge the confidential (n_c records) and the synthetic datasets (n_s records)
2. Add an additional variable, S:
   - $S_i = 0$ if it comes from the confidential dataset
   - $S_i = 1$ otherwise
3. For each record *i*, estimate the propensity score ($\hat{p}_i$) of it being in the synthetic dataset by fitting a model (i.e., a logistic regression) using available predictors

pMSE calculation

4. Compare the distributions of the $\hat{p}_i$ in the confidential and the synthetic datasets by computing the **propensity score mean-squared error**, known as **pMSE** , which is the mean-squared difference between the estimated propensity probabilities and the probability of a record being synthetic if original and synthetic observations were interchangeable, as:

$$pMSE = \frac{1}{n_c + n_s} \sum_{i=1}^{n_c+n_s} (\hat{p}_i - c)^2$$

where $c = n_s/(n_c + n_s)$ is the proportion of units with synthetic data

- When $n_c = n_s \rightarrow c = 1/2$
- When multiple synthetic datasets are generated, overall pMSE would be the average pMSE across all synthetic datasets

pMSE implications

- For a synthetic dataset of high utility, the classification model used in Step 3 will not be able to distinguish between the confidential and the synthetic datasets
- In such cases, the propensity scores will all be close to 0.5, leading to a pMSE of zero
- **Larger values of the *pMSE* indicate a lower level of utility**
- The largest possible value of pMSE when $n_c = n_s$ is $1/4$, when each $\hat{p}_i$ is equal to 0 or 1

Additional comments regarding *pMSE*

- We can use **any classification model** in Step 3, for example, logistic regression, classification tree, random forest, support vector machines (SVM)
- As pointed out by Snoke et al. (2018), if a logistic regression model is used, we can calculate the null distribution of the pMSE , which allows us to construct a **hypothesis test** (check out Nowok et al. (2020) for details)

Additional comments regarding *pMSE*

- The **choice of model** for classification will have a big impact on the pMSE:
    - Using a poor model will make it difficult to distinguish between the original and the synthetic dataset, artificially indicating good utility for the synthetic dataset
- The model used for classification should include at least **all of the variables which were synthesized**
- It is recommended to include **second-order and even third-order interactions** between the variables in the model if we want to test whether the relationships in the synthetic version of the confidential dataset are preserved (Woo et al. (2009), Snoke et al. (2018))

R Example 2: pMSE calculation for synthetic CE Expenditure

- We evaluate the **pMSE** global utility of synthetic CE expenditure example using a logistic regression with the main effects and first-order interactions of all 5 variables as a classification model in the CE sample
- Important: typically we evaluate utility of synthetic data of the **original scales**

eCDF definition

- An **empirical distribution** function, also commonly known as an empirical cumulative distribution function (empirical CDF), is the distribution function associated with the empirical measure of a sample

- It is a discrete distribution function which considers every observation in the sample to be an **equally likely outcome**:

Let $(y_1, \ldots, y_n)$ be the *n* sample observations. The eCDF is defined as:

$$\hat{F}_n(t) = \frac{number\ observations\ in\ the\ sample \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{y_i \leq t}$$

where $\mathbb{I}_{y_i \leq t}$ is the indicator of event $y_i \leq t$, and *n* the total sample size.

eCDF for synthetic data

- In the context of global utility of synthetic data, the **empirical CDF measures** rely on comparing the empirical distributions of the synthetic dataset with that of the confidential dataset
- Two similar samples should have **similar empirical CDFs**, and this can be used to obtain a global utility measure of synthetic data

eCDF calculation

- We follow Woo et al. (2009) and start with merging the two datasets (note that the merging step is not necessary to calculate the empirical CDFs):

  1. **Merge** the confidential and the synthetic datasets, resulting a merged dataset of dimension $(n_c + n_s)$ rows and (r) colums.
  2. Estimate the empirical CDF distribution of the **confidential dataset**, denoted as $\boldsymbol{ecdf^C}$, and that of the **synthetic dataset**, denoted by $\boldsymbol{ecdf^S}$.
  3. For record i $\left(i = 1, \ldots, (n_C + n_S)\right)$ in the merged dataset, estimate its **percentile** under the empirical CDF distribution of the confidential dataset $\boldsymbol{ecdf^C}$, denoted as $\boldsymbol{p^C}$, and its percentile under the empirical CDF distribution of the synthetic dataset $\boldsymbol{ecdf^S}$, denoted as $\boldsymbol{p^S}$

eCDF calculation

4. Use the following two measures:

$$U_m = \max_{1 \leq i \leq (n_c+n_s)} |p_i^C - p_i^S|,$$

$$U_a = \frac{1}{(n_c + n_s)} \sum_{i=1}^{(n_c+n_s)} (p_i^C - p_i^S)^2,$$

where U_m is the **maximum** absolute difference between the empirical CDFs, and U_a is the **average** squared differences between the empirical CDFs

eCDF implications

- The smaller the values of $\boldsymbol{U_m}$ and $\boldsymbol{U_a}$, the **higher the similarity** level between the confidential and the synthetic data, indicating **high utility**

- The larger the values of $\boldsymbol{U_m}$ and $\boldsymbol{U_a}$, the **lower the similarity** level between the confidential and the synthetic data, indicating **low utility**

R Example 3: eCDF calculation for synthetic CE Expenditure

- We evaluate the eCDF global utility of synthetic CE expenditure example

Additional comments regarding eCDF

- The empirical CDF and this global utility evaluation framework can be **generalized to more than one variable**

- However, in practice, it is often **challenging to estimate joint empirical CDFs for multiple variables** (no widely-used functions/ packages for this purpose, such as `ecdf` from the stats package we used for the univariate case)

Analysis-specific utility. Partially synthetic data: definition

- In a sample, only a subset of variables are deemed sensitive and then synthesized for privacy protection
- We follow Reiter (2003) and Drechsler (2011) Chapter 7.1.1. to describe the combining rules for valid inferences
- Combining rules refer to the valid inference methods for *m* synthetic datasets

Analysis-specific utility using partially synthetic data: inference methods

- Let $\boldsymbol{Q}$ be a **univariate parameter** of interest, such as a **population mean** of a univariate outcome or a univariate regression coefficient
- Let $\boldsymbol{q}$ and $\boldsymbol{v}$ be the **point estimate** and variance estimate of Q from the confidential data ($\boldsymbol{q}$ and $\boldsymbol{v}$ are not available unless one has access to the confidential data)
  - Note that $\boldsymbol{q}$ and $\boldsymbol{v}$ are estimates from a sample, for example, when Q is a population mean, $v = \sigma^2/n$ where σ is the population standard deviation if available, or $v = s^2/n$ where s is the sample standard deviation

Analysis-specific utility using partially synthetic data: inference methods

- Denote $Z = \left(Z^{(1)}, \dots, Z^{(m)}\right)$ the set of m partially synthetic datasets
- Let $\boldsymbol{q}^{(\boldsymbol{l})}$ and $\boldsymbol{v}^{(\boldsymbol{l})}$ be the values of q and v in the l-th synthetic dataset, $Z^{(l)}$,

that the data analyst is able to compute

- The analyst calculates:

$$
\begin{aligned}
\bar{q}_m &= \sum_{l=1}^{m} \frac{q^{(l)}}{m} \\
b_m &= \sum_{l=1}^{m} \frac{(q^{(l)} - \bar{q}_m)^2}{m-1} \\
\bar{v}_m &= \sum_{l=1}^{m} \frac{v^{(l)}}{m}
\end{aligned}
$$

Analysis-specific utility using partially synthetic data: inference methods

- The data analyst use $\bar{\boldsymbol{q}}_{\boldsymbol{m}}$ as the point estimate of Q, and

$$T_p = \frac{b_m}{m} + \bar{v}_m$$

as the variance estimate of $\bar{q}_m$

- Note that b_m is the variance of $(q_1, \dots, q_m)$

Discussion question: What are the effects of *m*? Does larger *m* produce larger T_p, and what does it imply for uncertainty? How do you think you would decide what *m* to use?

Analysis-specific utility using partially synthetic data: inference methods

- To make inferences for estimand Q, when the sample size of the synthetic data n is large, the data analyst can use a **t distribution** with degrees of freedom $v_p = (m-1)\left(1+\frac{\bar{v}_m}{b_m/m}\right)^2$

- The data analyst can obtain a **95% confidence interval for *Q*** as

$$\left(\bar{q}_m - t_{v_p}(0.975) \times \sqrt{\frac{b_m}{m} + \bar{v}_m},\ \bar{q}_m + t_{v_p}(0.975) \times \sqrt{\frac{b_m}{m} + \bar{v}_m}\right),$$

where $t_{v_p}(0.975)$ is the t score at 0.975 with the pertinent v_p degrees of freedom.

R Example 4: Inference for average expenditure in the synthetic CE

- Consider a population quantity of interest, *Q*, the average expenditure from the CE surveys. What is the 95% confidence interval for *Q* in the simulated synthetic data? How does it compare to that obtained from the confidential data?

Utility measures: Interval overlap

- Karr et al. (2006) first proposed the concept of using **interval overlap** as a utility measure
- We present two versions that are used in practice

Interval overlap utility measure definition 1

- The first version follows the description of Drechsler and Reiter (2009)
  - Let (L_S, U_S) denote the $(1-2\alpha)\%$ confidence interval for the estimand from *m* synthetic data, $Z = \left(Z^{(1)}, \dots, Z^{(m)}\right)$.
  - Let (L_C, U_C) denote the $(1-2\alpha)\%$ confidence interval for the estimand from the confidential data.
  - Compute the intersection of the two intervals, i.e. $(\max(L_S, L_C), \min(U_S, U_C))$, and denote it as (L_i, U_i)
- The utility measure of **interval overlap** is:

$$I = \frac{U_i - L_i}{2(U_C - L_C)} + \frac{U_i - L_i}{2(U_S - L_S)}$$

Interval overlap utility measure definition 2

- By design, the interval overlap measure **definition 1 returns 0 for any two non-overlapping** intervals
- Therefore, **it does not differentiate which one of two synthetic data confidence intervals is worse when both do not overlap** with that from the confidential data
- To make improvements, recent works such as Snoke et al. (2018) consider a second version of the interval overlap measure, where **non-overlapping would produce a negative interval overlap measure value**, which decreases as the distance between the two intervals increases
- We follow the description of Snoke et al. (2018)

Interval overlap utility measure definition 2

- Let (L_S, U_S) denote the $(1-2\alpha)\%$ confidence interval for the estimand from *m* synthetic data, $Z = \left(Z^{(1)}, \dots, Z^{(m)}\right)$.
- Let (L_C, U_C) denote the $(1-2\alpha)\%$ confidence interval for the estimand from the confidential data.

$$IO = \frac{1}{2}\left(\frac{\min(U_C, U_S) - \max(L_C, L_S)}{U_C - L_C} + \frac{\min(U_C, U_S) - \max(L_C, L_S)}{U_S - L_S}\right)$$

R Example 5: Interval overlap utility measure definitions for synthetic CE

- Calculate the interval overlap measure v1 for the **mean expenditure**
- Calculate the interval overlap measure v2 for the **mean expenditure**

Final comments

- The two versions of the interval overlap measure introduced above are designed for **frequentist inferences** of the confidential and the synthetic data

- They may still be applied to **credible intervals** obtained from Bayesian analyses, but in that case a **different interval overlap measure** has been proposed in Section 2.1 of Karr et al. (2006) which takes into account the additional information contained in the posterior distributions over these intervals

Disclosure Risk

1. Identification disclosurepMSE
   - Matching-based approaches
   - Record-linkage approaches
2. Attribute disclosure
   - The CAP risk measure
   - Classification-based risk measure

Disclosure risks evaluation

Assuming the intruder has access to external data, two common disclosure risks: identification and attribute (Hu, 2019):

1. **Identification disclosure**: The intruder correctly identifies records of interest in the released synthetic data
2. **Attribute disclosure**: The intruder correctly infers the true confidential values of the synthetic records using information from the released synthetic data

Overview of identification disclosure

- Identification disclosure:
  - The intruder correctly **identifies records of interest** in the released synthetic data
  - Only exist in **partially synthetic data**
- We will introduce two general approaches
  - **Matching-based** approaches
  - **Record-linkage** approaches

Discussion question: Suppose in the CE sample (five variables: `UrbanRural, Income, Race, Expenditure, KidsCount`), the synthetic data has `Expenditure` synthesized. Now suppose you know someone who's in the CE survey, and you know their `UrbanRural` and `Race`, how would you go about finding that person in the synthetic data?

Overview of attribute disclosure

- Attribute disclosure:
  - The intruder correctly infers the **true, confidential values** of the synthesized variables in the released synthetic data
  - Exist in **partially synthetic data** and **fully synthetic data**
- We will introduce two general approaches
  - The **CAP** risk measure
  - **Classification-based risk** measure

Overview of attribute disclosure

- In each method, it is assumed that an intruder knows certain characteristics for the targeted individual (one or multiple individuals), which we call the key variables
- It is further assumed that the intruder wishes to use the synthetic datasets to infer other characteristics for that individual, which we call the target variable(s)
- The approaches however differ with regards to assumptions about the process used by the intruder to infer the target variable(s), and the other information available

Overview: matching-based approaches

- In the matching-based approaches, we:

  1. **Make assumptions about the knowledge possessed by the intruder** for a confidential record *i*
  2. **Use the assumed knowledge** to investigate which records in the synthetic data are matched with record *i*
  3. Evaluate **whether the true match is among the matched records**, **how unique the true match is**, among other things

- We present a basic version of Reiter and Mitra (2009), which is a form of Bayesian probabilistic matching

Notation and setup

- $\boldsymbol{Y} = (\boldsymbol{Y}^{\boldsymbol{A}}, \boldsymbol{Y}^{\boldsymbol{U}})$ represents the confidential data sample: $\boldsymbol{Y}^{\boldsymbol{A}}$ denotes the variables available to the intruder from external databases and $\boldsymbol{Y}^{\boldsymbol{U}}$ denotes the variables unavailable to the intruder
- Similarly, we have $\mathbf{Z} = (\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{U}})$ for a partially synthetic dataset of $\boldsymbol{Y}$
- We can further split $\mathbf{Z}^{\mathbf{A}}$ into $\mathbf{Z}^{\mathbf{A_s}}$, the synthesized variables and $\boldsymbol{Z}^{\boldsymbol{A_{us}}}$,

the unsynthesized variables

- We assume here that the synthesized variables $\boldsymbol{Z}^{\boldsymbol{A_{us}}}$ are a **subset** of the variables which are available to the intruder
- Variables that are unavailable to the intruder,$\boldsymbol{Z}^{\boldsymbol{U}}$, could be synthesized or unsynthesized (they are not incorporated in the evaluation since the intruder has no access to them)
- The same split applies to $\mathbf{Y}^{\mathbf{A}}$ as in $\mathbf{Y}^{\mathbf{A}} = (\mathbf{Y}^{\mathbf{A_s}}, \mathbf{Y}^{\mathbf{A_{us}}})$

Notation and setup

- The intruder will have access to $\boldsymbol{y_i^A}$: available information that the intruder has access to through external databases, for target record *i*
- $\boldsymbol{y_i^A}$ contains the true values of the confidential variables
- In the synthetic dataset $\boldsymbol{Z}$, the corresponding available variables are in $\boldsymbol{Z^A}$
- The intruder will search for synthetic records in $\boldsymbol{Z}$ which can be matched to the target record $\boldsymbol{y_i}$
- Specifically, given the available variables, the intruder will try to match the target record $\boldsymbol{y_i^A}$ based on the available variables $\boldsymbol{Z^A}$
- Since some of these variables will be synthesized in the synthetic data, the **process of matching could create incorrect matches**. This is precisely how partially synthetic data are intended to protect from re-identification.

Notation and setup

- This suggests that the intruder will perform matching records in the synthetic data with a confidential, target record *i*, based on:
    - A set of available (known to the intruder), unsynthesized variables
    - The true, confidential values of the synthesized variables

Three risk summary measures

- Regardless of the type(s) of the synthetic variable(s), there are three **widely-used summaries of identification** disclosure risks based on matching:
  - The **expected match risk**
  - The **true match rate**
  - The **false match rate**

The expected match risk

- The **expected match risk** measures on average how likely it is to find the correct match for each record, and for the sample as a whole. It is defined as:

$$\sum_{i=1}^{n} \frac{T_i}{c_i}$$

where $T_i = 1$ if the true match is among the c_i units and $T_i = 0$ otherwise, and c_i is the number of records with the highest match probability for the target record *i*

Discussion question: When $T_i = 1$ and $c_i > 1$, what does $\frac{T_i}{c_i}$ represent? What happens to $\sum_{i=1}^{n} \frac{T_i}{c_i}$ when for record *i*, $T_i = 0$?

The expected match risk

- Each T_i/c_i is a **record-level probability** $\in [0,1]$
- The sum $\sum_{i=1}^{n}\frac{T_i}{c_i}$ is a **sample-level summary** of the expected risk, which is $\in [0,n]$
- The **higher the expected match risk**, the **higher the identification disclosure risk** for the sample, and vise versa

The true match rate

- The true match rate considers how large is a percentange of true unique matches that exist. It is defined as:

$$\sum_{i=1}^{n} \frac{K_i}{N}$$

  where $K_i = 1$ if the true match is the unique match (i.e., $c_i \cdot T_i = 1$) and $K_i = 0$ otherwise, and N is the total number of target records out of n total records.

- It is the **percentage of true unique matches** among the target records

Discussion question: The higher the true match rate, the lower or the higher the identification risk for the sample?

The false match rate

- The false match rate considers how large is a percentage of unique matches that are actually false matches. It is defined as:

$$\sum_{i=1}^{n} \frac{F_i}{s}$$

where $F_i = 1$ if there is a unique match but it is not the true match (i.e., $c_i(1 - T_i) = 1$) and $F_i = 0$, otherwise, and ***s*** is the number of uniquely matched records (i.e., $\sum_{i=1}^{n}(c_i = 1)$)

- It is the percentage of false matches among unique matches
- The **lower the false match rate**, the **higher the identification disclosure risk** for the sample, and vise versa

Summary and dicussion

- Higher expected match risk, higher true match rate, and lower false match rate indicate higher identification disclosure risk for the sample
- When m > 1 synthetic datasets are generated, we can calculate the three summaries on each synthetic dataset, and for each summary compute its averages across m samples
- Each measure is providing one aspect of the identification risk; we should consider them as a whole when comparing different synthetic datasets

R Example 6: ACS sample

- We use the matching-based approach to evaluate the identification disclosure risks of the synthetic ACS sample, where `DIS, HICOV` are synthesized and the other variables remain unsynthesized. The synthesis model is the DPMPM model with the NPBayesImputeCat R package. We assume that the intruder knows `SEX, RACE, MAR` of reach record.

Using the IdentificationRiskCalculation R package

- Check out the IdentificationRiskCalculation R package (Hornby and Hu (2020))
- Hornby and Hu (2021) provides examples
- Your results should match with the output from the package

When $m > 1$

- When dealing with m > 1, we calculate the three summaries in each synthetic dataset, and then take the **average to obtain three final summaries**
- The R script can be updated by specifying the c_vector, T_vector, K_vector, F_vector as matrices, and exp_match_risk, true_match_rate, false_match_rate as vectors

Record linkage methods: Overview

- **Record linkage methods** are developed mainly for the purpose of linking records from multiple databases
- Based on variables, called **keys**, a link between two records can be established. Therefore, record linkage approaches can be used as metrics of identification risks (William E. Winkler (2004))

Record linkage approaches for synthetic data

- For partially synthetic data, record linkage methods can be applied to **linking records in the synthetic dataset to the records in the confidential dataset**
- Among these linkages, we can evaluate identification risks in terms of true links (i.e., correct links) and false links (i.e., incorrect links)
    - **High percentage of true links** indicates **high identification disclosure risk**, and vice versa

Record linkage approaches for synthetic data

- We present the general procedure of performing probabilistic record linkage (Fellegi and Sunter, 1969) in partially synthetic data
- As with matching-based approaches
  - $\boldsymbol{Y} = (\boldsymbol{Y}^{\boldsymbol{A}}, \boldsymbol{Y}^{\boldsymbol{U}})$ to represent the confidential data sample containing *n* observations and *r* variables
  - $\boldsymbol{Y}^{\boldsymbol{A}}$ denotes the variables available to the intruder from external databases and $\boldsymbol{Y}^{\boldsymbol{U}}$ denotes the variables unavailable to the intruder
  - similarly, we have $\mathbf{Z} = (\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{U}})$ for a partially synthetic dataset of $\boldsymbol{Y}$
  - we can further split $\boldsymbol{Z}^{\boldsymbol{A}}$ into $\boldsymbol{Z}^{\boldsymbol{A}_{\boldsymbol{s}}}$, the synthesized variables and $\boldsymbol{Z}^{\boldsymbol{A}_{\boldsymbol{us}}}$, the unsynthesized variables

Procedure

1. Given $\boldsymbol{Y}^{\boldsymbol{A}}$,the set of variables available to the intruder, there are unsynthesized variables denoted by $\boldsymbol{Y}^{\boldsymbol{A_{us}}}$ and synthesized variables $\boldsymbol{Y}^{\boldsymbol{A_s}}$. We **generate pairs** between $\boldsymbol{Y}^{\boldsymbol{A}}$ and $\boldsymbol{Z}^{\boldsymbol{A}}$ based on $\boldsymbol{Y}^{\boldsymbol{A_{us}}}$ and $\boldsymbol{Z}^{\boldsymbol{A_{us}}}$
   - That is, a pair of record *i* from $\boldsymbol{Y}$ and record *j* from $\boldsymbol{Z}$ is generated only when $y_i^{A_{us}} = z_j^{A_{us}}$ (for the entire vector)
   - Call this collection of pairs as ***P***.
2. For each pair of records in ***P***, we, the data disseminators, compare the values of the synthesized variables
   - For example, if y_i and z_j is paired up in step 1, then this step **compares** $\boldsymbol{y_i^{A_s} = z_j^{A_s}}$
   - We can create a **set of similarity score** over all the synthesized variables, which will be used for scoring all pairs next.

Procedure

3. With calculated similarity score for each pair, we then score all the pairs
    - Think of this **as ranking all pairs**
    - **The higher the ranking** one pair is, the more likely a link will be established
    - This is the core of probabilistic record linkage as proposed by W. E. Winkler (2000), where we will use an **expectation-maximization algorithm** (**EM** algorithm)
    - In this approach, **a weight value will be estimated for each pair**, which will then be used next for determining links, also known as selecting pairs

Procedure

4. Each pair now comes with a weight from step 3. We then select one-to-one linkages between $\boldsymbol{Y^A}$ and $\boldsymbol{Z^A}$ records
   - That is, each record in the synthetic dataset Z **will be linked to at most one record** in the confidential Y, and vice versa
5. Among the selected pairs from step 4, we calculate the **percentages of true links** (i.e., the one-to-one links that are correct links) and of **false links** (i.e., the one-to-one links that are incorrect links)

R Example 6: ACS sample

- Now we will illustrate this record linkage approach to evaluating identification disclosure risks in the synthetic ACS sample

- The probabilistic record linkage algorithm is implemented by the `reclin` package (Laan (2018))

- We use the record linkage approach to evaluate the identification disclosure risks of a synthetic ACS sample, where `DIS, HICOV` are synthesized and the other variables remain unsynthesized. The synthesis model is the DPMPM model with the NPBayesImputeCat R package. We assume that the intruder knows `SEX, RACE, MAR, DIS, HICOV` of reach record.

Example of the ACS sample: calculate percentages of true links and false links

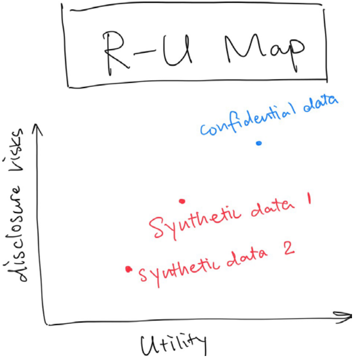
- Lastly, we need to evaluate among all the TRUE's in `greedy`, how many of them are true links and how many are false links
- A true link refers to a correct linkage, i.e., record *i* in `ACSdata_syn` is correctly linked to record *i* in `ACSdata`
- A false link refers to an incorrect linkage, i.e., record *i* in `ACSdata_syn` is incorrectly linked to record *j* in `ACSdata` where $i \neq j$
- Note that in our illustration SEX, RACE, MAR are unsynthesized, so our first step of generating pairs would have no errors
- It is possible that the intruder's knowledge of available variables includes some synthesized variables, which means the first step would generate incorrect pairs

Evaluation

- The choice depends on data disseminator’s protection goals Assuming a quality synthesis
    - Utility: higher in partially synthetic data
    - Disclosure risks: higher in partially synthetic data
- Utility-risk trade-off

Utility-risk trade-off

- Ideally, the released synthetic data have high utility and low disclosure risks
- However this is usually not the case, due to the utility-risk trade-off (Duncan, Keller-McNulty, and Stokes (2001))



Data generation References

References I

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes. 2001. “Disclosure Risk Vs Data Utility: The R-U Confidentiality Map.” National Institute of Statistical Sciences.

Hu, J. 2019. “Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data.” Transactions on Data Privacy 12: 61–89.

Hu, J., J. P. Reiter, and Q. Wang. 2014. “Disclosure Risk Evaluation for Fully Synthetic Categorical Data.” Privacy in Statistical Databases, 185–99.

Kaur, D., M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. 2021. “Application of Bayesian Networks to Generate Synthetic Health Data.” Journal of the American Medical Informatics Association 28: 801–11.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. 2011. “Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database.” International Statistical Review 79: 362–84.

References II

Little, R. J. A. 1993. “Statistical Analysis of Masked Data.” Journal of Official Statistics 9: 407–26.

Rubin, D. B. 1993. “Discussion Statistical Disclosure Limitation.” Journal of Official Statistics 9: 461–68.

Young, J., P. Graham, and R. Penny. 2009. “Using Bayesian Networks to Create Synthetic Data.” Journal of Official Statistics 25: 549–67.

Utility References

References I

Drechsler, J. 2011. Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics 201, Springer.

Drechsler, J., and J. P. Reiter. 2009. “Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB.” Journal of Official Statistics, 589–603.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. 2006. “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality.” The American Statistician 60: 224–32.

Nowok, B., G. M. Raab, C. Dibben, J. Snoke, and C. van Lissa. 2020. Synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control. https://cran.r-project.org/web/packages/synthpop/index.html.

Reiter, J. P. 2002. “Satisfying Disclosure Restrictions with Synthetic Data Sets.” Journal of Official Statistics 18: 531–44.

———. 2003. “Inference for Partially Synthetic, Public Use Microdata Sets.” Survey Methodology 29: 181–88.

References II

Ros, K., H. Olsson, and J. Hu. 2020. “Two-Phase Data Synthesis for Income: An Application to the NHIS.” Privacy in Statistical Databases (e-Proceedings).

Rubin, D. B. 1993. “Discussion Statistical Disclosure Limitation.” Journal of Official Statistics 9: 461–68.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. 2018. “General and Specific Utility Measures for Synthetic Data.” Journal of the Royal Statistical Society, Series A (Statistics in Society) 181: 663–88.

Woo, M. J., J. P. Reiter, A. Oganian, and A. F. Karr. 2009. “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation.” The Journal of Privacy and Confidentiality 1: 111–24.

Risk References

References I

Baillargeon, M., and A. Charest. 2020. "A Closer Look at the CAP Risk Measure for Synthetic Datasets." Privacy in Statistical Databases (e-Proceedings).

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks." In Proceedings of the 2nd Machine Learning for Healthcare Conference, edited by Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, 68:286–305. Proceedings of Machine Learning Research. Boston, Massachusetts: PMLR.

Elliot, M. 2014. "Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team." CMIST.

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." Journal of the American Statistical Association 64: 1183–1210.

References II

Hornby, R., and J. Hu. 2020. IdentificationRiskCalculation: Calculating the Identification Risk in Partially Synthetic Microdata. https://github.com/RyanHornby/IdentificationRiskCalculation.

— — — . 2021. “Identification Risks Evaluation of Partially Synthetic Data with the IdentificationRiskCalculation r Package.” Transactions of Data Privacy 14: 37–52.

James, Witten, G., and Tibshirani R. 2021. An Introduction to Statistical Learning with Applications in r, Second Edition. Springer.

Kaur, D., M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. 2021. “Application of Bayesian Networks to Generate Synthetic Health Data.” Journal of the American Medical Informatics Association 28: 801–11.

Laan, J. van der. 2018. Record Linkage Toolkit. R Package Version 0.1.1. Reiter, J. P., and R. Mitra. 2009. “Estimating Risks of Identification Disclosure in Partially Synthetic Data.” The Journal of Privacy and Confidentiality 1: 99–110.

References III

Taub, J., M. Elliot, M. Pampaka, and D. Smith. 2018. “Differential Correct Attribution Probability for Synthetic Data: An Exploration.” Privacy in Statistical Databases, 122–37.

Winkler, W. E. 2000. “Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage.” U.S. Bureau of the Census.

Winkler, William E. 2004. “Re-Identification Methods for Masked Microdata.” In Privacy for Statistical Databases, edited by J. Domingo-Ferrer and V. Torra, 216–30.