Methods for Utility Evalation part 2

Jingchen (Monika) Hu

Vassar College

Statistical Data Privacy

Outline

1. Introduction
2. Valid inferences for univariate estimands in partially synthetic data
3. Valid inferences for univariate estimands on fully synthetic data (rare)
4. Interval overlap utility measure (2 definitions)
5. Summary and References

Outline

Recap

- Lecture 6:
  - Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
  - Two measures: *pMSE* (propensity score) and ECDF

The CE sample

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

Variable 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.

Plan for this lecture

- Two general types of utility: global and analysis-specific (Lecture 1)

1. Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
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

Plan for this lecture

- Two general types of utility: global and analysis-specific (Lecture 1)

1. Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
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

- In this lecture, we focus on analysis-specific utility evaluation methods, with illustrations to the synthetic CE from Lectures 4 & 5

Discussion question: Given your readings and previous lectures, what are global utility evaluation methods you have seen?

Overview

- Analysis-specific utility measures are tailored to the analyses the data analyst is expected to perform on the synthetic data

Discussion question: What types of analyses do you think a data analyst might conduct on the synthetic CE data (suppose `Expenditure` is synthesized given `Income`)

Discussion question: What do we mean by synthetic data have high analysis-specific utility?

Outline

Partially synthetic data: definition

- In a sample, only a subset of variables are deemed sensitive and then synthesized for privacy protection
- In a sample, all variables are synthesized using synthesis models built for this sample

Partially synthetic data: inference methods

- 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

Partially synthetic data: inference methods

- Let Q be a univariate parameter of interest, such as a population mean of a univariate outcome or a univariate regression coefficient of a regression model

- Let q and v be the point estimate and variance estimate of Q from the confidential data (q and v are not available unless one has access to the confidential data)
    - Note that q and v are estimates from a sample, for example, when Q is a population mean, following the Central Limit Theorem, $v = \sigma^2/n$ where σ is the population standard deviation if available, or $v = s^2/n$ where s is the sample standard deviation

Partially synthetic data: inference methods

- Denote $\mathbf{Z} = (\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(m)})$ the set of m partially synthetic datasets
- Let $q^{(l)}$ and $v^{(l)}$ be the values of q and v in the l-th synthetic dataset, $\mathbf{Z}^{(l)}$, that the data analyst is able to compute
- The analyst calculates:

$$\bar{q}_m = \sum_{l=1}^{m} \frac{q^{(l)}}{m} \tag{1}$$

$$b_m = \sum_{l=1}^{m} \frac{(q^{(l)} - \bar{q}_m)^2}{m-1} \tag{2}$$

$$\bar{v}_m = \sum_{l=1}^{m} \frac{v^{(l)}}{m} \tag{3}$$

Partially synthetic data: inference methods

- The data analyst use $\bar{q}_m$ as the point estimate of Q, and
$$T_p = \frac{b_m}{m} + \bar{v}_m \tag{4}$$
as the variance estimate of $\bar{q}_m$

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

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?

Partially synthetic data: inference methods

- When the sample size of the synthetic data n is large, the data analyst can use a t distribution with degrees of freedom $v_p = (m-1)(1 + \bar{v}_m/(b_m/m))^2$ to make inferences for estimand Q
- The data analyst can obtain a 95% confidence interval for Q as $(\bar{q}_m - t_{v_p}(0.975) \times \sqrt{T_p},\ \bar{q}_m + t_{v_p}(0.975) \times \sqrt{T_p})$, that is:

$$\left(\bar{q}_m - t_{v_p}(0.975) \times \sqrt{\frac{b_m}{m} + \bar{v}_m},\ \bar{q}_m + t_{v_p}(0.975) \times \sqrt{\frac{b_m}{m} + \bar{v}_m}\right), \tag{5}$$

where $t_{v_p}(0.975)$ is the t score at 0.975 with degrees of freedom $v_p = (m-1)(1 + \bar{v}_m/(b_m/m))^2$

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?

Inference for average expenditure in the synthetic CE

- See hidden code for model fitting
- We can generate $m = 20$ synthetic datasets

```
n <- nrow(CEdata)
m <- 20
SLR_synthetic_m_partial <- vector("list", m)
for (l in 1:m){
  SLR_synthetic_one_partial <- SLR_synthesize(X = SLR_X,
                                              post_draws = post_SLR,
                                              index = 1980 + l,
                                              n = n,
                                              seed = m + l)
  names(SLR_synthetic_one_partial) <- c("Intercept", "LogIncome", "LogExpen
  SLR_synthetic_one_partial$Expenditure <- exp(SLR_synthetic_one_partial$Lo
  SLR_synthetic_one_partial$Income <- exp(SLR_synthetic_one_partial$LogInco
  SLR_synthetic_m_partial[[l]] <- SLR_synthetic_one_partial
}
```

Inference for average expenditure in the synthetic CE

- To obtain a valid point estimate and confidence interval of the unknown mean of `Expenditure` from $m = 20$ synthetic datasets
- First calculate the $q^{(l)}$ and $v^{(l)}$, the point estimate and the variance estimate of the mean of `Expenditure` in each of the $m = 20$ synthetic datasets, and $l = 1, \cdots, m$

```
q <- rep(NA, m)
v <- rep(NA, m)
for (l in 1:m){
  SLR_synthetic_one_partial <- SLR_synthetic_m_partial[[l]]
  q[l] <- mean(SLR_synthetic_one_partial$Expenditure)
  v[l] <- var(SLR_synthetic_one_partial$Expenditure)/n
}
```

Inference for average expenditure in the synthetic CE

- Next, we calculate $\bar{q}_m$, b_m, and $\bar{v}_m$

```
q_bar_m <- mean(q)
b_m <- var(q)
v_bar_m <- mean(v)
```

Inference for average expenditure in the synthetic CE

- Now, we can calculate $T_p = b_m/m + \bar{v}_m$ as the variance estimate of $\bar{q}_m$
- Furthermore, we need the degrees of freedom, $v_p = (m-1)(1 + \bar{v}_m/(b_m/m))^2$, of the t distribution for making inferences for the mean estimand Q

```
T_p <- b_m / m + v_bar_m
v_p <- (m - 1) * (1 + v_bar_m / (b_m / m))^2
```

Inference for average expenditure in the synthetic CE

- Finally, we have the point estimate for mean estimand Q as $\bar{q}_m =$ 10112.00\$, and the 95% confidence interval calculated as [9822.29\$, 10401.72\$]

```
q_bar_m
```

```
## [1] 10112
```

```
t_score_syn <- qt(p = 0.975, df = v_p)
c(q_bar_m - t_score_syn * sqrt(T_p),
  q_bar_m + t_score_syn * sqrt(T_p))
```

```
## [1]  9822.287 10401.715
```

Inference for average expenditure in the synthetic CE

- Given the confidential CE sample, we could obtain the point estimate of the unknown mean of `Income`, and its 95% confidence interval from the Central Limit Theorem and t score: 10197.38 and [9870.15, 10524.61].

```
mean_con <- mean(CEdata$Expenditure)
sd_con <- sd(CEdata$Expenditure)
t_score_con <- qt(p = 0.975, df = n - 1)
mean_con
```

```
## [1] 10197.38
```

```
c(mean_con - t_score_con * sd_con / sqrt(n),
  mean_con + t_score_con * sd_con / sqrt(n))
```

```
## [1]  9870.152 10524.612
```

Discussion question: What do you think of this particular analysis-specific utility?

Inference for a regression coefficient in the synthetic CE

Consider a simple linear regression model for inferences, presented below, where e_i is the error term. Suppose we are interested in the regression coefficient, the slope β_1, which is the population quantity of interest, Q.

$$Expenditure_i = \beta_0 + \beta_1 Income_i + e_i. \tag{6}$$

We compute the 95% confidence interval for β_1 from the simulated synthetic data and compare it to that obtained from the confidential data.

Inference for a regression coefficient in the synthetic CE

- Here we include relevant code for performing the inference procedure

```
ComputeBeta1 <- function(m, syndata){

Beta1_q <- rep(NA, m)
Beta1_v <- rep(NA, m)

for (l in 1:m){
  syndata_l <- syndata[[l]]
  syndata_l_lm <- stats::lm(formula = Expenditure ~ 1 + Income,
                            data = syndata_l)
  coef_output <- coef(summary(syndata_l_lm))
  Beta1_q[l] <- coef_output[2, 1]
  Beta1_v[l] <- coef_output[2, 2]^2
}

res <- list(Beta1_q, Beta1_v)
}
```

Inference for a regression coefficient in the synthetic CE

```
Beta1_qv <- ComputeBeta1(m = m,
                         syndata = SLR_synthetic_m_partial)
Beta1_q <- Beta1_qv[[1]]
Beta1_v <- Beta1_qv[[2]]
```

Inference for a regression coefficient in the synthetic CE

```
Beta1_q_bar_m <- mean(Beta1_q)
Beta1_b_m <- var(Beta1_q)
Beta1_v_bar_m <- mean(Beta1_v)

Beta1_T_p <- Beta1_b_m / m + Beta1_v_bar_m

Beta1_v_p <- (m - 1) * (1 + Beta1_v_bar_m / (Beta1_b_m / m))^2

Beta1_q_bar_m
```

```
## [1] 0.04115868
```

```
Beta1_t_score_syn <- qt(p = 0.975, df = Beta1_v_p)
c(Beta1_q_bar_m - Beta1_t_score_syn * sqrt(Beta1_T_p),
  Beta1_q_bar_m + Beta1_t_score_syn * sqrt(Beta1_T_p))
```

```
## [1] 0.03804092 0.04427644
```

Inference for a regression coefficient in the synthetic CE

- On the confidential data

```
confidata_lm <- stats::lm(formula = Expenditure ~ 1 + Income,
                 data = CEdata)
coef(summary(confidata_lm))
```

```
##                 Estimate   Std. Error  t value      Pr(>|t|)
## (Intercept) 5.538898e+03 1.971015e+02 28.10176 9.599075e-162
## Income      6.044031e-02 1.673219e-03 36.12218 8.797902e-255
```

```
confint(confidata_lm)
```

```
##                    2.5 %       97.5 %
## (Intercept) 5.152495e+03 5.925301e+03
## Income      5.716009e-02 6.372054e-02
```

Discussion question: What do you think of this particular analysis-specific utility? What reasons can you come up with?

Final comments

- Any synthesizer, even if developed carefully and is well estimated, can only preserve characteristics of the confidential data to a certain degree

- In fact, a synthesizer only captures what is specified by the synthesis model

- Any other characteristics beyond what the synthesis model captures could be lost

- It is therefore crucial, even though it may not be a trivial task, to develop comprehensive synthesizers that can capture most of the confidential data characteristics that can lead to correct population-level inference

Outline

Fully synthetic data: definition (Lecture 1)

- 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

Fully synthetic data: definition (Lecture 1)

- 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
- Much less common than partially synthetic data where every variable is synthesized

Fully synthetic data: inference methods

- We follow Drechsler (2011) Chapter 6.1.1.
- The notation for Q, q, v, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(m)})$, $q^{(i)}$, and $v^{(i)}$ are the same for the partial synthesis case mentioned previously
- Also, the quantities of $\bar{q}_m$, b_m, and $\bar{v}_m$ are calculated in the same way

Fully synthetic data: inference methods

- The data analyst use $\bar{q}_m$ as the point estimate of Q, and
$$T_f = \left(1 + \frac{1}{m}\right) b_m - \bar{v}_m \tag{7}$$
as the variance of $\bar{q}_m$

Fully synthetic data: inference methods

- When the sample size of the synthetic data n is large, the data analyst can use a t distribution with degrees of freedom $v_f = (m-1)(1 - \bar{v}_m/((1+1/m)b_m))^2$ to make inferences for estimand Q
- In other words, the data analyst can obtain a 95% confidence interval for Q as $(\bar{q}_m - t_{v_f}(0.975) \times \sqrt{T_f},\ \bar{q}_m + t_{v_f}(0.975) \times \sqrt{T_f})$

Fully synthetic data: inference methods

- Note the negative sign in front of $\bar{v}_m$: in some cases it is possible to obtain a negative value for T_f, which we then have to round up to zero

- Reiter (2002) suggests an alternative, non-negative variance estimator, which can replace T_f above
$$T_f^* = \max(0, T_f) + \delta\left(\frac{n_{syn}}{n}\bar{v}_m\right), \tag{8}$$
where $\delta = 1$ if $T_f < 0$ and $\delta = 0$ otherwise. n_{syn} is the number of observations in the released datasets sampled from the synthetic population

Outline

Overview

- 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, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(m)})$. 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 is

$$I = \frac{U_i - L_i}{2(U_c - L_c)} + \frac{U_i - L_i}{2(U_s - L_s)}. \tag{9}$$

Discussion question: What values of I indicate high utility?

Interval overlap utility measure definition 1 for synthetic CE

Calculate the interval overlap measure v1 for the mean expenditure and the regression coefficient for the synthetic CE data generated above.

Interval overlap utility measure definition 1 for synthetic CE

- We write a function to perform the calculation

```
CalculateIntervalOverlap_v1 <- function(confi_interval, syn_interval){

  L_i <- max(confi_interval[1], syn_interval[1])
  U_i <- min(confi_interval[2], syn_interval[2])

  if (L_i <= U_i){
    overlap <- (U_i - L_i) / (2 * (confi_interval[2] - confi_interval[1]))
    (U_i - L_i) / (2 * (syn_interval[2] - syn_interval[1]))
  }
  else
    overlap <- 0

  return(overlap)
}
```

Interval overlap utility measure definition 1 for synthetic CE

```r
I1 <- CalculateIntervalOverlap_v1(c(9870.15, 10524.61),
                                  c(9822.29, 10401.72))
I1
```

```
## [1] 0.8648142
```

```r
I2 <- CalculateIntervalOverlap_v1(c(0.057, 0.064),
                                  c(0.038, 0.044))
I2
```

```
## [1] 0
```

Discussion question: Does these I values indicate high utility? Why or why not?

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

Interval overlap utility measure definition 2

- To make improvements, recent works such as Snoke et al. (2018) and Ros, Olsson, and Hu (2020) 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, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(m)})$. Let (L_c, U_c) denote the $(1-2\alpha)\%$ confidence interval for the estimand from the confidential data. The interval overlap measure, IO, is calculated as

$$IO = \frac{1}{2}\left(\frac{\min(U_c, U_s) - \max(L_c, L_s)}{U_c - L_c} + \frac{\min(U_c, U_s) - \max(L_c, L_s)}{U_s - L_s}\right). \tag{10}$$

Interval overlap utility measure definition 2 for synthetic CE

Calculate the interval overlap measure v2 for the mean expenditure and the regression coefficient for the synthetic CE data generated above.

Interval overlap utility measure definition 2 for synthetic CE

- We write a function to perform the calculation

```
CalculateIntervalOverlap_v2 <- function(confi_interval, syn_interval){

  L_c <- confi_interval[1]
  U_c <- confi_interval[2]
  L_s <- syn_interval[1]
  U_s <- syn_interval[2]

  overlap <- 1 / 2 * ((min(U_c, U_s) - max(L_c, L_s)) / (U_c - L_c) +
                      (min(U_c, U_s) - max(L_c, L_s)) / (U_s - L_s))

  return(overlap)
}
```

Interval overlap utility measure definition 2 for synthetic CE

```
IO1 <- CalculateIntervalOverlap_v2(c(9870.15, 10524.61),
                                   c(9822.29, 10401.72))
IO1
```

```
## [1] 0.8648142
```

```
IO2 <- CalculateIntervalOverlap_v2(c(0.057, 0.064),
                                   c(0.038, 0.044))
IO2
```

```
## [1] -2.011905
```

Discussion question: Does these I values indicate high utility? Why or why not? How do they compared to the ones calculated using definition 1?

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

Outline

Summary

- Analysis-specific utility measures
  - Valid inferences for univariate estimands for partially synthetic data
  - Valid inferences for univariate estimands for fully synthetic data
  - Interval overlap measures (2 definitions)

Summary

- Analysis-specific utility measures
  - Valid inferences for univariate estimands for partially synthetic data
  - Valid inferences for univariate estimands for fully synthetic data
  - Interval overlap measures (2 definitions)

- No homework! But next week you will present global utility and analysis-specific utility results of your your simulated synthetic data for your project
- Lecture 8: Methods for risk evaluation part 1
  - Hu (2019), Section 4
  - Hornby and Hu (2021)

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.

Hornby, R., and J. Hu. 2021. "Identification Risks Evaluation of Partially Synthetic Data with the Identificationriskcalculation R Package." Transactions of Data Privacy 14 (1): 37–52.

Hu, J. 2019. "Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data." Transactions on Data Privacy 12 (1): 61–89.

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.

References II

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.

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 (3): 663–88.