Diagnostic Tools for Assessing Validity of Synthetic Data Inferences

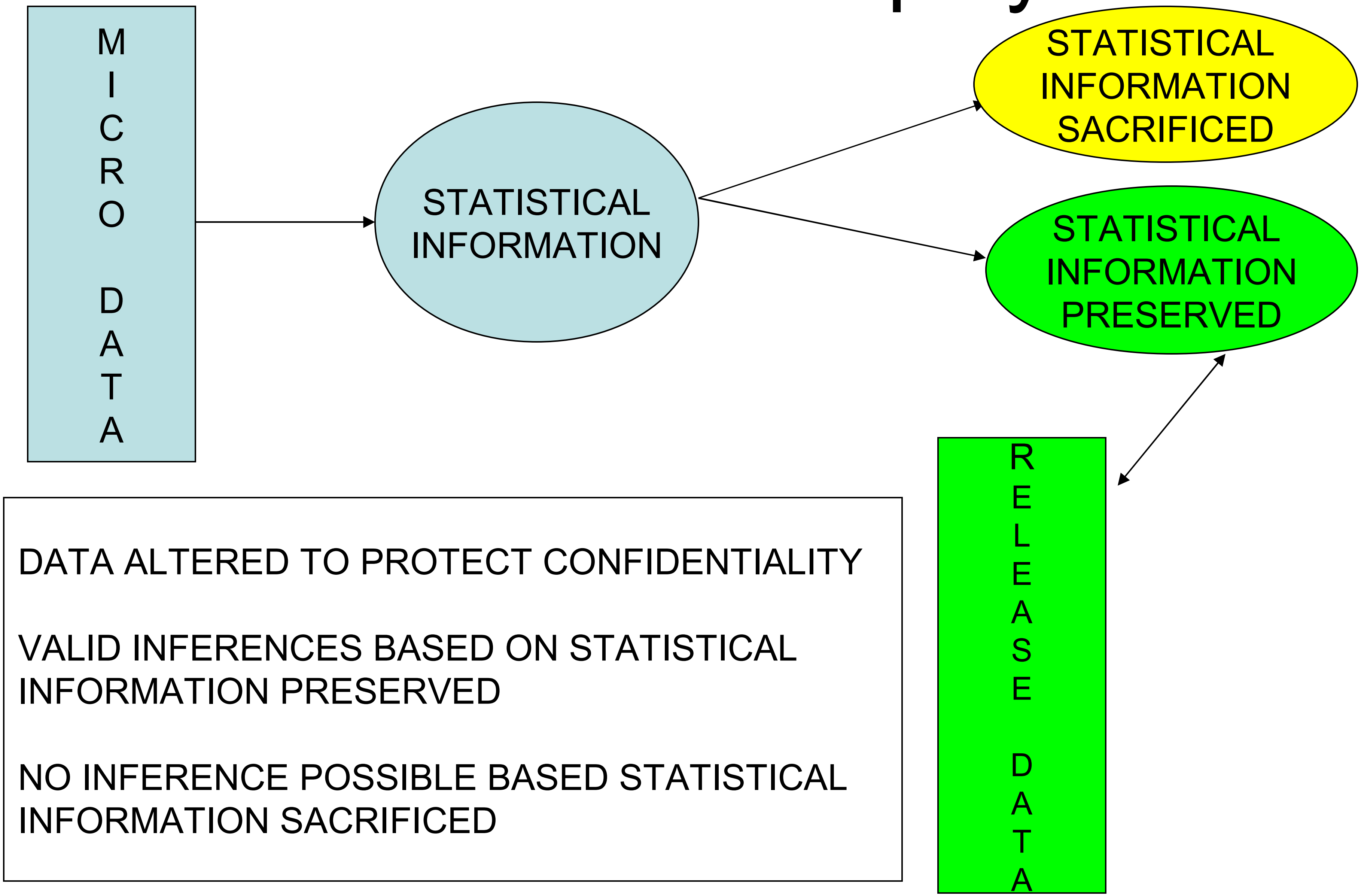
Trivellore Raghunathan (Raghu)
University of Michigan

U.S. Census Bureau, July 31, 2009

Disclosure Avoidance Procedures

- All approaches involve some form of alterations of data
- Many procedures are primarily aimed at reducing the risk of disclosure
- Marginal mean or proportion preserving alterations (such as variable specific data swapping)
- Survey data are used for variety of purposes other than descriptive statistics
  - Some alterations can change the “statistical information” and result in a biased estimates (Liu (2003), UM Ph.D. Dissertation)
- Overarching goal: Alterations that reduce risk of disclosure but preserve some “statistical information”

Schematic Display



MICRO DATA

STATISTICAL INFORMATION

STATISTICAL INFORMATION SACRIFICED

STATISTICAL INFORMATION PRESERVED

RELEASE DATA

DATA ALTERED TO PROTECT CONFIDENTIALITY

VALID INFERENCES BASED ON STATISTICAL INFORMATION PRESERVED

NO INFERENCE POSSIBLE BASED STATISTICAL INFORMATION SACRIFICED

EXAMPLE

- MICRO-DATA: P CONTINUOUS VARIABLES APPROXIMATELY NORMAL
- STATISTICAL INFORMATION TO BE PRESERVED: MEAN AND THE COVARIANCE MATRIX (HENCE, ALL LINEAR REGRESSION COEFFICIENTS)
- RELEASE DATA: ABILITY TO OBTAIN VALID INFERENCES BASED ON MEAN AND COVARIANCE MATRIX

$$\bar{X} = MEAN$$

$$S = COVARIANCE\ MATRIX$$

$$X_i^* \sim IND\ MVN(\bar{X}, S), i = 1, 2, \ldots, n$$

For valid inferences: Need to add uncertainty to sample mean and covariance matrix

Synthetic Data

For easy inference: Release Multiple independent copies

$$X_i^* \sim IND\ MVN(\bar{X}, S), i = 1, 2, \ldots, n$$

$$* = 1, 2, \ldots, M$$

EXAMPLE (CONTD.)

- Micro Data: P categorical variables
- Information to be preserved: Up to kth order interaction
- Release Data: Generate from a log-linear model where all interactions of order k+1 or higher are set to zero
- Multiple copies of the release data makes the analysis simpler through use of standard software with simple combining rules

General Association Preservation

- Sequential Regression approach

Variables : $Y_1, Y_2, ..., Y_P$

Draws from : $\Pr(Y_i \mid Y_1, Y_2, ..., Y_{i\text{-}1}, Y_{i+1}, ..., Y_P)$

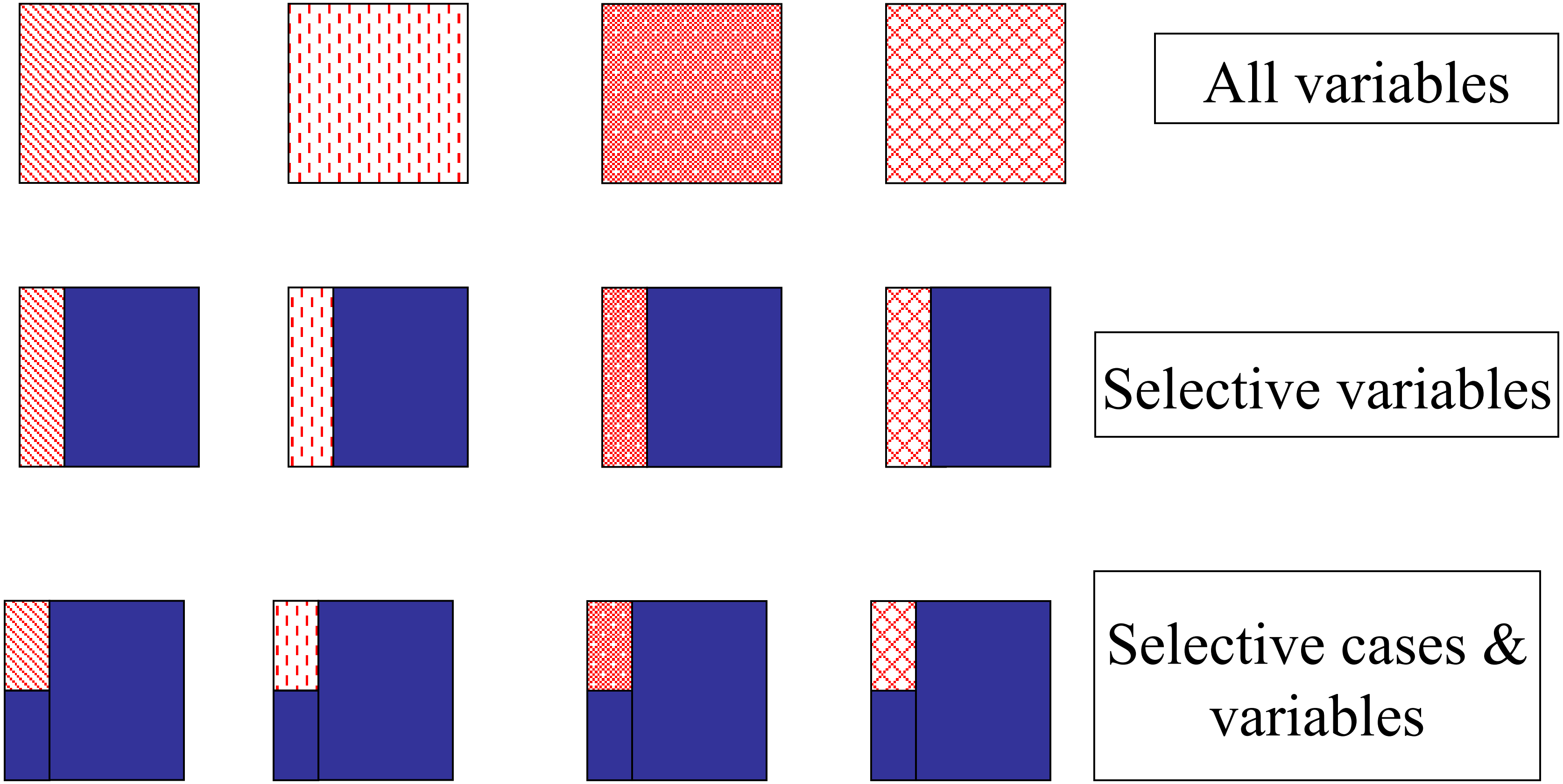
Parametric or Nonparametric regression models

Interactions could be added

Skip patterns and logical consistencies can be built - in

Flexible approach

Different versions of release samples



All variables

Selective variables

Selective cases & variables

Validity of the synthetic/altered data approach

- How to assess whether this approach is valid?
- Conceptualization of validity
  - Arithmetical or Numerical Validity
  - Inferential validity
    - Unconditional validity
    - Conditional validity

Arithmetical/Numerical validity

- Are the statistics computed (such means, proportions, regression coefficients, confidence intervals etc) from the actual data numerically close to those from the synthetic/altered data?
  - Given the stochastic nature of the way we create the release samples this is almost unattainable
    - Can imputations ever be real?
  - Can we get the same values from different samples?

Inferential Validity

- Unconditional Validity
  - Are the point and interval estimates have desirable properties from the repeated sampling point of view?
    - Bias: Across repeated sampling from the population are the estimates unbiased for the population quantity?
    - Confidence coverage: Across repeated sampling from the population do confidence intervals have the stated nominal level?
    - Type I and Type II error rates in the testing of hypothesis

Inferential Validity (Contd)

- Conditional Validity
  - Two stages or types of randomness
    - Stage 1: Original sample from the population
    - Stage 2: Creation of release/synthetic samples conditional on the original sample
  - Unconditional validity is marginal across both stages or types of randomness
  - Conditional validity is assessment with respect to Stage 2 and treats the original sample as fixed.
    - Whether the repeated release/synthetic sample statistics envelope the original data statistics tightly

Assessing Unconditional Validity

- Theoretical or analytical arguments
- Simulation studies
  - Create a population by assembling large data sets (e.g. concatenated CPS or SIPP dat)
  - Draw independent samples
  - Create synthetic data sets for each sample
  - Construct point and interval estimates from each sample and the corresponding synthetic data sets
  - Compare the properties such as bias, MSE and confidence coverage

Practical Approach to Assess Conditional Validity

- Suppose that a “release” consists of M synthetic samples
- The procedure is good if most “releases” are “close” to the original sample
- Create K=pM >> M synthetic samples
  - Draw a sample of M synthetic samples and compute the estimate of the target parameter
  - Repeat many times (Say, T times) and then assess the distribution of T estimates around the original sample estimate
  - Compare the spread among the T estimates with the standard error of the original sample estimates (could be problematic, if the original sample is “over powered” has negligible Standard error).

Remarks

- Computationally intensive much larger number of synthetic samples have to be created
- Statistics to be used in the evaluation needs to be determined and all these have to computed on each of the T potential release samples.
- Need to know what meaningful difference between the T estimates and the original sample estimate is alarming or unacceptable?
- Flavor of diagnostics

Bayesian Approach

- Use posterior predictive approach
- Model of interest is $f(y|\theta)$ and the prior is $\pi(\theta)$
- Compare $\pi(\theta \mid y_{act})$ and $\pi(\theta \mid y_{synth})$
- Identify several candidate analysis
- Draw values of the parameters using both data sets and then compare the distributional properties of the overlap of drawn values of the parameters

Alternative Approach

- Are the distributions of variables “balanced” across the original and each synthetic sample?
- Use the propensity score idea

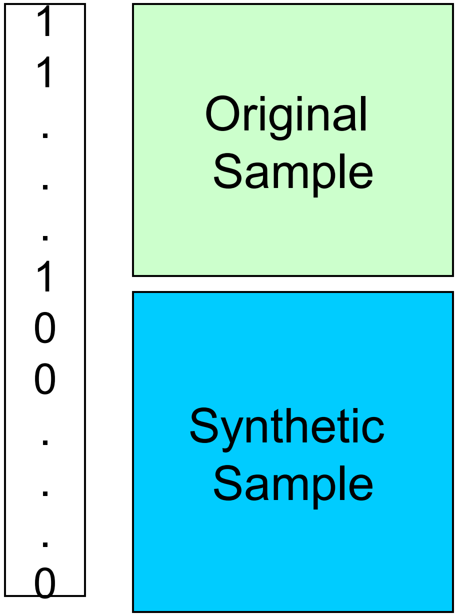
Run a regression of the indicator variable on the concatenated original/synthetic samples.

Include main effects and interaction effects

Estimate the propensity score and create classes

Ideally, the proportion of original sample cases in each class should be the same.

Check for deviations



1
1
.
.
.
1
0
0
.
.
.
0

Original Sample

Synthetic Sample

Remarks

- Idea is to balance across the M synthetic samples
  - Do the propensity score classification for each sample and then check whether the proportion of original sample cases averaged across the M samples are constant
- It is good to balance each synthetic sample but requiring balance for each sample is more a stringent requirement and may not be necessary

Conclusion and Discussion

- Focus should be on “information preserving” alterations and some information will have to be sacrificed
- It is important to assess the inferential validity rather than numerical validity
- Conditional validity is more meaningful from the practical point of view
  - Useful to refine the model to improve the conditional validity
- Unconditional validity is sort of a minimum requirement
  - Carefully planned simulation studies needed