Disclosure Limitation in Microdata with Multiple Imputation

Jerry Reiter
Institute of Statistics and Decision Sciences
Duke University

General setting

- Agency seeks to release data on individuals.
- Risk of re-identifications from matching to external databases.
- Statistical disclosure limitation applied to data before release.

Standard approaches to disclosure limitation

- Suppress data
- Add random noise
- Recode variables
- Swap data

Another approach: Partially synthetic data

Release multiple, partially synthetic datasets so that:

- Released data comprise mix of observed and synthetic values.
- Released data look like actual data.
- Statistical procedures valid for original data are valid for released data.

Little (1993, *JOS*), Reiter (2003, 2004 *Surv. Meth*)

Existing applications

- Replace sensitive values for selected units:

  Kennickel (1997, *Record Linkage Techniques*).

- Replace values of identifiers for selected units:
  Liu and Little (2002, *JSM Proceedings*),
  Current research with Sam and Rolando.

- Replace all values of sensitive variables:

  Abowd and Woodcock (2001, *Confid., Discl., and Data Access*),
  Survey of Income and Program Participation.
  Longitudinal Business Database.

Advantages of partially synthetic data

- Confidentiality protected since risky identifiers or sensitive values not genuine.
- Replacements come from realistic models, so associations are preserved (as long as model is good).
- Varying amounts of synthesis can be done, depending on risk/utility tradeoff.
- Provide information in tails of distributions and release finer geographic detail.
- Agency can describe imputation model, so that analysts have a sense how their results are affected.

Handling missing and synthetic data simultaneously

Reiter (2004, *Survey Methodology*)

- Create m completed datasets using MI for missing data.
- For each completed dataset, create r replacement datasets using MI for partially synthetic data.
- Release $M = mr$ datasets to public.

Inference with missing and partially synthetic data

Reiter (2004, *Survey Methodology*)

- Estimand: $Q = Q(X, Y)$
- In each synthetic dataset $d_k^{(i)}$

$$q_k^{(i)} = Q(d_k^{(i)}) \qquad u_k^{(i)} = U(d_k^{(i)})$$

Quantities needed for inference

$$\bar{q}_M = \sum_{i=1}^{m}\sum_{k=1}^{r} q_k^{(i)} / rm = \sum_{i=1}^{m} \bar{q}^{(i)} / m$$

$$b_M = \sum_{i=1}^{m} (\bar{q}^{(i)} - \bar{q}_M)^2 / (m-1)$$

$$\bar{w}_M = \sum_{i=1}^{m}\sum_{k=1}^{r} (q_k^{(i)} - \bar{q}^{(i)})^2 / m(r-1) = \sum_{i=1}^{m} w^{(i)} / m$$

$$\bar{u}_M = \sum_{i=1}^{m}\sum_{k=1}^{r} u_k^{(i)} / rm$$

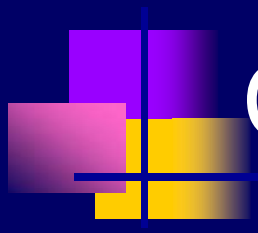
Inference with missing and partially synthetic data

- Estimate of Q: $\bar{q}_M$
- Estimate of variance is

$$T_M = \bar{u}_M + (1+1/m)b_M - \bar{w}_M / r$$

- For large n, m, and r, use normal based inference for Q:

$$\bar{q}_M \pm 1.96\sqrt{T_M}$$



Ongoing research

- Semi-parametric and non-parametric data generation methods.
- Risk/usefulness profile on genuine data in production setting.
- Packaged synthesizers.