Data Confidentiality and the Common Good.

Kyle Fennell
Julia Lane

Overview

- The issues
- Data access modalities
- Statistical Disclosure Limitation Techniques
- The Role of Archiving

The Issues

Data protection paramount

- Ethical requirement (promise to respondents)
- Legal requirement (often legislation)
- Practical (maintain response rates)

Dissemination paramount

- Ethical requirement (promise to respondents)
- Legal requirement (often legislation)
- Practical (maintain funding)

Dissemination Approaches

- Tables
  - Broad use
  - Answers predefined questions
  - Statistical validity of current approaches challenged
- Microdata
  - Narrower use
  - Marginal vs average effects
  - Data quality

Access Modalities

Research Data Centers

Remote Access

Licensing

Public Use files

Research Data Centers

Who uses this approach?

- Begun by Census Bureau in U.S.; now also in many other countries (CA, UK, NZ, NL…)

What they are

- Researchers physically go to access data on a site controlled by NSI
- Monitored by NSI Employees
- Supported by NSI, host institution, foundations

Research Data Centers

- Benefits
  - Access to “gold standard” datasets
  - Perceptions
- Costs
  - Length of review process
  - Cost in terms of time
  - Cost in terms of money
  - Disparate use

Licensing

Who uses this approach?

Wide variety of federal agencies

Licensing: Signed agreements that allow external researchers to access semi anonymized datafiles: typical protocols are

- Data Security Plan that defines location, security arrangements and access protocols
- Confidentiality pledges
- Institutional concurrence,
- Onsite security inspections

Licensing: Evaluation

Benefits

Higher quality data than public use files

Flexible use by researchers in their offices

Costs

Only works for individual, not business datasets

- Outliers removed
- Some evidence of violations

Remote Access

- Who uses this approach?
  - Oldest example is Luxembourg Income Study
  - Statistics Denmark, Netherlands,
  - NORC
- What it is varies
  - Buffered remote access: Users send in code; output examined and returned.
  - Web interface with custom tailored (commercial) software
  - True remote access

Remote Access: Evaluation

- Benefits
  - High quality data
  - Very low cost
  - Collaboratories possible
- Costs
  - Perceptions
  - Technological advances => higher risk
  - Buffered Remote Access
    - Slow
    - Outliers suppressed
    - Rigid framework

Public Use Files

- Who uses this approach?
  - – Pioneered by U.S. Census Bureau
  - – Used by almost every NSI
- What it is
  - - Microdata files anonymized so that there is “low disclosure risk” (FCSM working paper 22)

More detail on techniques

Reduce Information (recoding and

- variable deletion)
- recoding categorical variables into larger categories
- recoding continuous variables into categories
- rounding continuous variables
- using top and bottom code
- using local suppression and enlarging geographic areas

More detail on techniques

Perturb information

- noise addition
- Data swapping
- blanking and imputation
- micro-aggregation
- multiple imputation/modeling to generate synthetic data

Public Use Files: Evaluation

- Benefits
  - Broad use
  - Important training for graduate and even undergraduate students
- Costs
  - Decreasing quality, particularly wrt outliers
  - Vulnerability to admin data on web and technological advances in matching software

Data Access: Archiving

Access provides opportunity to engage researcher community in data documentation…but

- Research shows major reasons for not documenting are economic ☺
- Lack of incentives (lack of academic credit)
- Time cost of documentation
- Lack of funding
- Lack of standards

New approaches

- Develop tools to reduce costs to researchers
    - E.g. microdata documentation toolkit
- Add benefits:
    - Researcher incentives for metadata documentation Contributions indexed and attributed. Citations required, and posted
        - Monetary contributions – reduction in fees
        - Collabatory
    - Develop metadata system with feedback loop on data quality

Conclusion

- Fundamental tension in data dissemination
- Many access modalities; no “silver bullet”
- Each modality provides some opportunity for archivists to engage researcher community