Disclosure Avoidance at Statistics Canada

INFO747 Session on Confidentiality Protection April 19, 2007

Jean-Louis Tambay, Statistics Canada

jean-louis.tambay@statcan.ca

Outline

- Statistics Canada’s context
- Public Use Microdata Files
- Research Data Centres
    - Disclosure vetting at RDCs
- Remote Access
- References

Providing access to microdata

- The *Statistics Act*
  - Sections 11 & 12 data sharing agreements
  - Discretionary release
  - Use of “deemed employees”
  - Public Use Microdata Files

Public Use Microdata Files

- *Anonymized* microdata files for a *sample* of units – mostly household survey data
- Microdata Release Policy & Guidelines
- Need approval of Microdata Release Committee to release a PUMF
- Submissions must include data distributions, geographic level of detail, description of the weighting procedure and the methods to evaluate and decide on data to be presented

Preparation for PUMFs

- Suppress identifying variables
- Limit design & related information
    - Clusters (& households), strata, Bootstrap weights
- Consider level of geographic detail
- Examine distribution of weights (low weights, geographical information implied by weights)
- Special analyses (relationships, multiplicity, Data Intrusion Simulation, linkages, …)
- Data suppression and perturbation
- Longitudinal PUMFs have rarely been released!

Special analyses

- Multiplicity
  - Given a set of n indirect identifiers (ii), generate all 3-way tables involving 3 ii's at a time
  - Multiplicity = # tables in which unit is unique
  - Analysis can be by sub-group (e.g., province)
- Data Intrusion Simulation (Elliot)
  - Probability a unique match to a microdata record is a true match

    P(cm|um) ≈ #uniques / [#uniques + 2*#pairs*(weight-1)]
  - Expanded to Poisson sampling by Skinner & Carter

Research Data Centres

- Initially created to provide researcher access to longitudinal surveys – now housing population & housing survey data
- Around 20 centres provide access to researchers in a secure university setting
- Always staffed by STC employees
- Accessible only to researchers with approved projects who have been sworn in as “deemed employees” under the *Statistics Act*
- All outputs are vetted before being released

Disclosure vetting at RDCs

- Two types of risks:
  - Produce results for identifiable respondents
  - Compromise confidentiality of PUMF data
- Since results are from sample surveys and are aggregated, risks are low BUT
- ... many surveys release PUMFs – we do not want to risk compromising disclosure control methods used to protect PUMF data
- General rules implemented for all surveys – some surveys have additional rules

Disclosure vetting at RDCs

- Potential problems associated with availability of PUMF data
  - Statistics based on few observations could be linked to individual respondents – risks increase if survey weights can help in linking (note: survey results based on few respondents are not reliable)
  - Some distributional results provide information about extreme values (top-coded on PUMFs)
  - Approximate location of sample units can be revealed – this affects more than one survey as many have sample in the same clusters

Disclosure vetting at RDCs

- Key aspects:
  - Results should use survey weights (justify need for unweighted other than sample size indications)
  - No unit-level results: apply 5-respondent minimum for frequencies & statistics (some surveys use 10) – use higher threshold if releasing weighted and unweighted tabular results
  - Intermediary outputs increase the risk of residual disclosure and should be avoided
  - Analytical and model outputs entail less risks than tabular ouputs

Disclosure vetting at RDCs

- Other rules:
  - Careful about tables with full cells (i.e., only one nonzero cell in a row/column)
  - 5-respondent min. applies to descriptive statistics; for medians & percentiles need at least 5 units at or above & at or below value
  - No ranges, min. or max. for quantitative variables
  - Model outputs are generally safe but:
    - saturated models with categorical covariates should be vetted as if tabular results
    - covariances/correlations involving dichotomous variables are releasable if results by value of dichotomous variable are releasable
    - no unit-level results (e.g., residuals, scatterplots)

Disclosure vetting at RDCs

- Special rules for detailed geographical results:
  - Do not reveal sensitive information about the location of the sample or of sample units on a map, table, list or otherwise
  - Round weighted frequencies to base 50
  - Detailed geographical outputs for visible identifying characteristics, e.g., race or disability should only be released if they do not pose a risk (full cell problem)
  - Researchers who wish to release geographical contextual information must indicate how those relate to geographical areas – if some areas are clearly identified from the contextual information the vetting rules should be applied at the level of those areas

Disclosure vetting at RDCs

- Rules apply to household survey data at RDCs
- Plans to put census data and some admin data at RDCs
- Census rules will apply for census data. Additionally, geographical detail will stop at the census tract (or equivalent) level and intermediary outputs will not be allowed.
- Admin data put in feasibility study mode – rules to be developed

Rules for census data

- Random rounding for counts (usually base 5)
- Population thresholds for “standard” & custom geographies (40 & 100)
- Population & household thresholds for income characteristics (250 & 40)
- # same-sex common-law couples available for areas over 5,000 people
- For place of work data size limits are applied to the employed labour force
- Suppression of statistics if: $ values of units in cell are in a narrow range; <4 records used in calculation; sum of weights <10; or presence of outliers
- Otherwise totals for quantitative statistics obtained by multiplying average with rounded weighted frequency

Remote Access

- Provide indirect access since the 1990s
- Researchers obtain survey & datafile documentation and “dummy” test data
  - Note: Test files created from survey data need approval of Microdata Release Committee
- SAS/SPSS/Stata programs submitted by e-mail, results e-mailed back after manual vetting for confidentiality
- Popular for some surveys (e.g., health)
- Disclosure issues similar to RDCs

References

Elliot, M.J. (2000). Data Intrusion Simulation: Advances and a Vision for the Future of Disclosure Control. Presented at the *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*. Skopje, March 14-16, 2001.

Mayda, J.E., Mohl, C. and Tambay, J.L. (1996). Variance Estimation and Confidentiality: They Are Related! *Proceedings of the Survey Methods Section, SSC Annual Meeting*. June, 1996.

Skinner, C.J. and Carter, R.G. (2003). Estimation of a Measure of Disclosure Risk for Survey Microdata under Unequal Probability Sampling. *Survey Methodology*. 29, 177-180.

Statistics Canada (2005). Guide for Researchers under Agreement with Statistics Canada. October, 2005. http://www.statcan.ca/english/rdc/pdf/researchers_guide.pdf

Tambay, J.L., Goldmann, G. and White, P. (2001). Providing Greater Access to Survey Data for Analysis at Statistics Canada. *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.