Reporting the Quality of Output Data From the Integration of Multiple Data Sources

Linda J. Young
USDA National Agricultural Statistics Service
April 18, 2018

Levels of Transparency

- High Transparency
  - Information needed to validate findings
  - Known limitations in the methodology or supporting data
  - Underpins other levels of transparency
  - Audience: academics, agency specialists, subject-matter experts
- Moderate Transparency
  - Key, high-value transparency information
  - Close details are not immediately displayed
  - Users should be able to reveal high transparency information
  - Audience: policy makers, professional journalists, students

Levels of Transparency

- Low Transparency
  - Limited detailed transparency information
  - Findings easy to navigate and understand
  - Users should be able to reveal moderate and high transparency information
  - Audience: general public

Output Data

- Integrated estimates
  - Outcome statistics
  - Supporting statistics
- Micro-data files
  - Record-linked data files
  - Variables or other content on data files

Breaks in Series

Langton: NCVS

- National Crime Victimization Survey (NCVS)
  - 2016 Design change—Break or blip in series?
  - Decision: Release unadjusted estimates

Breaks in Series

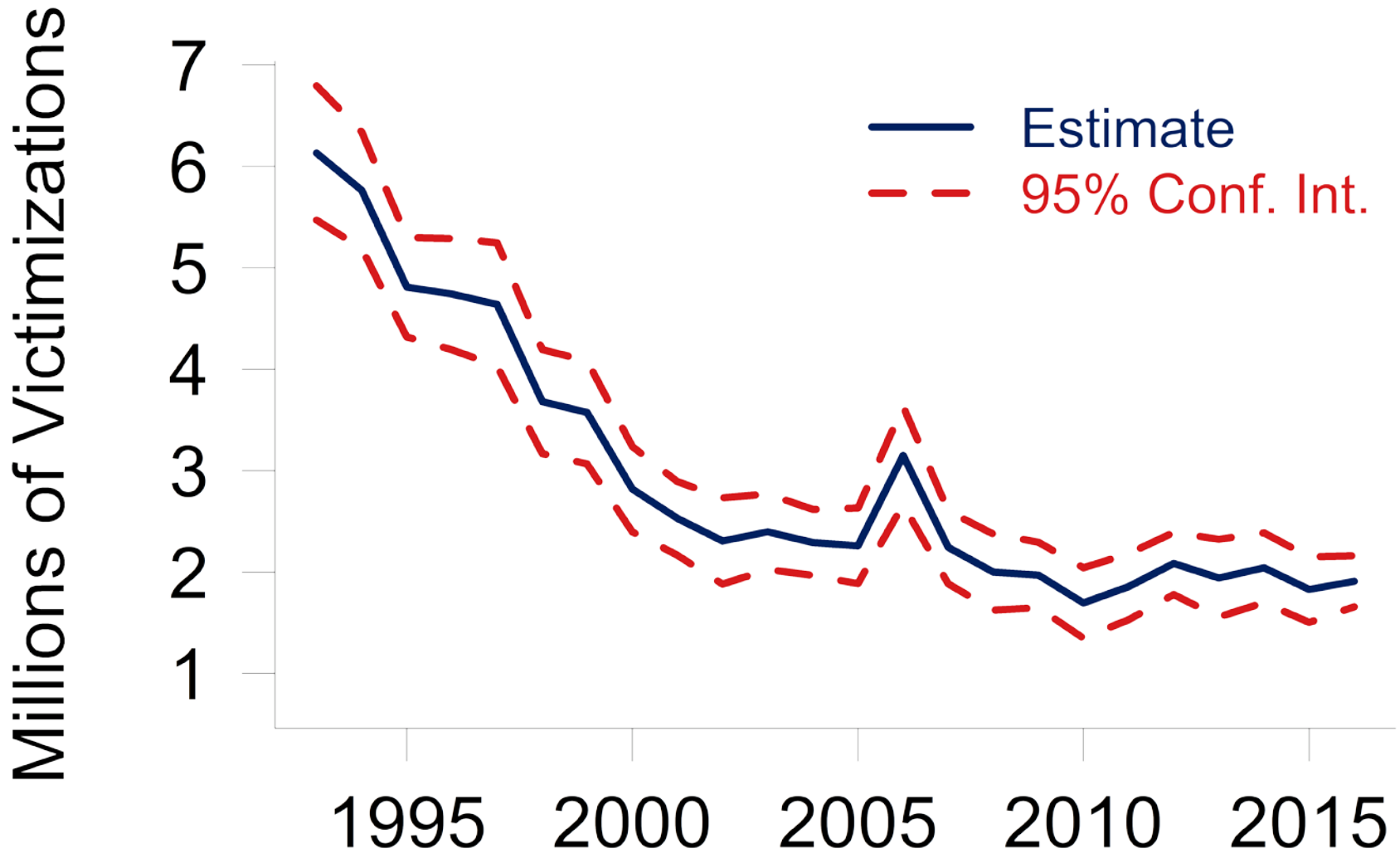
Eltinge: Discussant

- Not specific to integrated data
- Resulting decision processes apply
- Particularly challenging if key purpose is to monitor change
- Mitigate impact
  - Statistical adjustments
  - Planning
- Trade-offs: Incremental changes vs big changes
- Key: Communication to stakeholders

Is It Really a Break in Series?

- Breaks in series may be reported when simple changes are made
  - Change in question ordering
  - Change in design
  - Change in level of interviewer training
  - Change in data sources
  - Change in data collection procedures
  - Change in methodology used in analysis
- Are these breaks or a failure to fully reflect the uncertainty in the estimates?

NCVS Rape, robbery, agg. assault



Courtesy of Sharon Lohr

Combining Information from Multiple Data Sources

Ragunathan: Large study

- Estimate prevalence rates and trends for multiple disease outcomes
  - Attribute costs to these outcomes
  - Determine how much change in overall cost over time is due to (1) changes in prevalence or (2) changes in treatment costs
- 7 survey data sources and 5 non-survey data sources
- Propensity and imputation methods used to combine information from each source

Combining Information from Multiple Data Sources

- Issues
  - Types of respondents and sources of information differ
  - Differences in surveys: question wording, survey designs, coverage, mode effects, response error properties
  - Measurement error issues across data sources
- Opportunities:
  - Use big data
  - Improve non-probability information using probability sample data
- "**It is dangerous to think that we do not need high quality probability surveys anymore.**"

Combining Information from Multiple Data Sources

Bell: Connected Raghu's approach to Small Area Estimation

- Assumptions needed for success
  - Relationships between *Y* and *X*
  - Good estimates of sampling error are available and used
  - External standard can be used to assess error if it is unbiased or biases are negligible
- Assessment of estimates are optimistic as they assume models are true
- If improvements to estimates ae modest, effort may not be worth the risks of model failure

Administrative Data and Uncertainty

- Integrating Disparate Data
  - Survey data
  - Administrative data
- Quantifying Uncertainty
  - Uncertainty in the survey
  - Uncertainty in the administrative data
  - Uncertainty introduced in the analysis

Sensitivity Analyses

- Standards
- How should the results impact reported measures of uncertainty?

Frameworks for Assessing Data Quality

Biemer and Czajka: Total Survey Error (TSE) Framework

- Statistics Norway (Zhang) extended for integrated data
  - Renamed concepts to accommodate administrative data and integration
  - Phase 1: sources of error for input data
  - Phase 2: sources of error from integration and harmonization processes
- Extended by Statistics New Zealand (Reid)
  - Quality indicators for Phase 1 and Phase 2
  - Phase 3: sources of error for assessing estimates from Phase 2 products

Summary

- Focus on the quality of the target estimate instead of the datasets
- Think about proxies; no data are perfect
- Strengthen collaborations
  - Exchange knowledge
  - Build on data combining efforts
  - Share burdens (costs)
- Think differently about what we are doing