Noisy Data from the Noisy Census

Gary King

Institute for Quantitative Social Science
Harvard University

Center for Discrete Mathematics and Theoretical Computer Science, Rutgers University,
4/28/2022

Inference from Census Population Counts

Inference from Census Population Counts

- True counts, unobserved

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
    - 1990-2010: swapping (top coding, imputation) released publicly
    - 2020: Transparent Two-Step Process
        - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
        - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
    - Target for academics: the true counts

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
  - Target for academics: the true counts
  - Target for public policy & law: ?

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
  - Target for academics: the true counts
  - Target for public policy & law: ?
  - Everyone needs estimators with: (1) known statistical properties and (2) accurate uncertainty estimates

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
  - Target for academics: the true counts
  - Target for public policy & law: ?
  - Everyone needs estimators with: (1) known statistical properties and (2) accurate uncertainty estimates
- Feasibility of developing these estimators

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
  - Target for academics: the true counts
  - Target for public policy & law: ?
  - Everyone needs estimators with: (1) known statistical properties and (2) accurate uncertainty estimates
- Feasibility of developing these estimators
  - Swapping: open question, never studied

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
  - 1990-2010: swapping (top coding, imputation) released publicly
  - 2020: Transparent Two-Step Process
    - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
    - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
  - Target for academics: the true counts
  - Target for public policy & law: ?
  - Everyone needs estimators with: (1) known statistical properties and (2) accurate uncertainty estimates
- Feasibility of developing these estimators
  - Swapping: open question, never studied
  - NMF: doable, but every statistical method changes! lots to do.

Inference from Census Population Counts

- True counts, unobserved
- Enumerated counts, measured with error, not released publicly
- Privacy Protected Counts
    - 1990-2010: swapping (top coding, imputation) released publicly
    - 2020: Transparent Two-Step Process
        - Noisy Measurement File: Counts+noise (obviously not counts), not yet released
        - TopDown Post-processing: NMF squished, censored, adjusted (definitely not counts, but look that way), released 8/12/2021
- Inferential Goals
    - Target for academics: the true counts
    - Target for public policy & law: ?
    - Everyone needs estimators with: (1) known statistical properties and (2) accurate uncertainty estimates
- Feasibility of developing these estimators
    - Swapping: open question, never studied
    - NMF: doable, but every statistical method changes! lots to do.
    - TopDown: open question, very difficult or impossible

The Census DAS is not Differential Privacy!

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection
  - Chosen DP parameters are enormous

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information
  - $\leadsto$ Census Bureau has not released it

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information
  - $\rightsquigarrow$ Census Bureau has not released it
- TopDown post-processed file

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information
  - ⤳ Census Bureau has not released it
- TopDown post-processed file
  - Formal: privacy bounds from NMF are unchanged

The Census DAS is not Differential Privacy!

- Noisy measurements File: DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information
  - $\rightsquigarrow$ Census Bureau has not released it
- TopDown post-processed file
  - Formal: privacy bounds from NMF are unchanged
  - Informal: Census Bureau acts as if squishing, censoring, & adjusting provides additional privacy protection

The Census DAS is not Differential Privacy!

- **Noisy measurements File:** DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information
  - $\rightsquigarrow$ Census Bureau has not released it
- **TopDown post-processed file**
  - **Formal:** privacy bounds from NMF are unchanged
  - **Informal:** Census Bureau acts as if squishing, censoring, & adjusting provides additional privacy protection
  - **Privacy protections:** maybe better but no formal guarantees!

The Census DAS is not Differential Privacy!

- **Noisy measurements File:** DP provides little protection
  - Chosen DP parameters are enormous
  - Known attacks could reveal individual information
  - $\rightsquigarrow$ Census Bureau has not released it
- **TopDown post-processed file**
  - **Formal:** privacy bounds from NMF are unchanged
  - **Informal:** Census Bureau acts as if squishing, censoring, & adjusting provides additional privacy protection
  - **Privacy protections:** maybe better but no formal guarantees!
  - **Advantages of DP in this Census:** mostly of theoretical mathematical interest & establishing transparency principles

This session

This session

- Our speakers:

This session

- Our speakers:
  - Sam Hirsch

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen
  - Jeff Gill

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen
  - Jeff Gill
  - Jeff Lewis

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen
  - Jeff Gill
  - Jeff Lewis
- They'll help us understand:

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen
  - Jeff Gill
  - Jeff Lewis
- They'll help us understand:
  - Target dataset: true, enumerated, or privacy protected data

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen
  - Jeff Gill
  - Jeff Lewis
- They'll help us understand:
  - Target dataset: true, enumerated, or privacy protected data
  - Target quantities of interest, necessary statistical inferences

This session

- Our speakers:
  - Sam Hirsch
  - Clark Bensen
  - Jeff Gill
  - Jeff Lewis
- They'll help us understand:
  - Target dataset: true, enumerated, or privacy protected data
  - Target quantities of interest, necessary statistical inferences
  - And implications for democracy