Scientific Inferences From Privatized Census Data

Gary King[1]

Institute for Quantitative Social Science
Harvard University

Panel Discussion on the Future Direction of the Census Bureau's American Community Survey, for Federal Reserve Staff, 11/30/2021

[1] GaryKing.org

Requirements for Scientific Measurement

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects
2. Measure with known statistical properties

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects
2. Measure with known statistical properties
    - E.g.: If we apply this rule to data we have lots of times, on average we'll get the right answer ("unbiasedness")

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects
2. Measure with known statistical properties
    - E.g.: If we apply this rule to data we have lots of times, on average we'll get the right answer ("unbiasedness")
    - E.g.2: The more data, the closer we'll likely get to the right answer ("consistency")

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects
2. Measure with known statistical properties
    - E.g.: If we apply this rule to data we have lots of times, on average we'll get the right answer ("unbiasedness")
    - E.g.2: The more data, the closer we'll likely get to the right answer ("consistency")
3. Accurate uncertainty estimates

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects
2. Measure with known statistical properties
    - E.g.: If we apply this rule to data we have lots of times, on average we'll get the right answer ("unbiasedness")
    - E.g.2: The more data, the closer we'll likely get to the right answer ("consistency")
3. Accurate uncertainty estimates
    - E.g.: Margins of error (CIs), SEs, hypothesis tests, etc.

Requirements for Scientific Measurement

1. Quantity of interest defined separately from any measure
    - E.g.: Forecasts, descriptions, causal effects
2. Measure with known statistical properties
    - E.g.: If we apply this rule to data we have lots of times, on average we'll get the right answer ("unbiasedness")
    - E.g.2: The more data, the closer we'll likely get to the right answer ("consistency")
3. Accurate uncertainty estimates
    - E.g.: Margins of error (CIs), SEs, hypothesis tests, etc.
    - A scientific statement: not one that is necessarily correct, but one that comes with accurate uncertainty estimates

The Role of Differential Privacy: Statistics vs. CS

The Role of Differential Privacy: Statistics vs. CS

Population
⋮
Margaret
Robert
Bhashkar
Gary
Jerome
Sandra
Amanda
Ari
Hannah
Indraneel
$48

Mean income:

Quantity of Interest

The Role of Differential Privacy: Statistics vs. CS

Population	Sample
⋮	✗
Margaret	✓
Robert	✓
Bhashkar	✓
Gary	✓
Jerome	✓
Sandra	✓
Amanda	✓
Ari	✓
Hannah	✓
Indraneel	✓
$48	

Mean income:

Quantity of Interest

The Role of Differential Privacy: Statistics vs. CS

Population	Sample	$
⋮	✗	?
Margaret	✓	122
Robert	✓	76
Bhashkar	✓	145
Gary	✓	96
Jerome	✓	86
Sandra	✓	127
Amanda	✓	72
Ari	✓	132
Hannah	✓	95
Indraneel	✓	134
$48		$108

Mean income:

Classical Inference

Quantity of Interest

Usually no direct relevance

The Role of Differential Privacy: Statistics vs. CS

Population	Sample	$
⋮	✗	?
Margaret	✓	122
Robert	✓	76
Bhashkar	✓	145
Gary	✓	96
Jerome	✓	86
Sandra	✓	127
Amanda	✓	72
Ari	✓	132
Hannah	✓	95
Indraneel	✓	134
$48		$108

Mean income:

Classical Inference

Quantity of Interest

Usually no direct relevance

The Role of Differential Privacy: Statistics vs. CS

Population	Sample	$	+Privacy
⋮	✗	?	
Margaret	✓	122	
Robert	✓	76	
Bhashkar	✓	145	Noise & Censoring
Gary	✓	96	
Jerome	✓	86	
Sandra	✓	127	
Amanda	✓	72	
Ari	✓	132	
Hannah	✓	95	
Indraneel	✓	134	
$48		$108	

Mean income:

Classical Inference

Quantity of Interest

Usually no direct relevance

The Role of Differential Privacy: Statistics vs. CS

Population	Sample	$	+Privacy	=dp$
⋮	✗	?		
Margaret	✓	122		85
Robert	✓	76		103
Bhashkar	✓	145	Noise & Censoring	75
Gary	✓	96		113
Jerome	✓	86		125
Sandra	✓	127		97
Amanda	✓	72		101
Ari	✓	132		128
Hannah	✓	95		83
Indraneel	✓	134		201
Mean income:	$48	$108		$111

Classical Inference

Query-Response

Quantity of Interest

Usually no direct relevance

No direct relevance

The Role of Differential Privacy: Statistics vs. CS

Population	Sample	$	+Privacy	=dp$
⋮	✗	?		
Margaret	✓	122		85
Robert	✓	76		103
Bhashkar	✓	145		75
Gary	✓	96		113
Jerome	✓	86		125
Sandra	✓	127		97
Amanda	✓	72		101
Ari	✓	132		128
Hannah	✓	95		83
Indraneel	✓	134		201
Mean income:		$48	$108	$111

Noise & Censoring

Classical Inference

Query-Response

Statistically Valid Inferences from Privacy Protected Data

Analyzing Differentially Private Data (Data + Noise)

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP:

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP: larger CIs

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP: larger CIs
- The (only) valid objections to DP

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP: larger CIs
- The (only) valid objections to DP
  - I don't wanna learn new statistical methods!

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP: larger CIs
- The (only) valid objections to DP
  - ~~I don't wanna learn new statistical methods!~~

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP: larger CIs
- The (only) valid objections to DP
  - ~~I don't wanna learn new statistical methods!~~
  - Added privacy protections: not necessary

Analyzing Differentially Private Data (Data + Noise)

- Statistical methods: must change!
- Consequence of ignoring DP noise
  - Bias: any direction, any magnitude
- Proper analysis of DP data (with corrected methods)
  - estimates with known statistical properties (as with raw data)
  - accurate uncertainty estimates (as with raw data)
  - the only change with DP: larger CIs
- The (only) valid objections to DP
  - ~~I don't wanna learn new statistical methods!~~
  - Added privacy protections: not necessary
  - The larger CIs: too large for my QOI

US Census Privatization Strategies

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

2020: Public Privatization

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

2020: Public Privatization

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)

2020: Public Privatization

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified

2020: Public Privatization

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"
  - Valid inferences: easy, but *data not (yet) released!*

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"
  - Valid inferences: easy, but *data not (yet) released!*
- Post-Processed data released: "TopDown Algorithm"

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"
  - Valid inferences: easy, but *data not (yet) released!*
- Post-Processed data released: "TopDown Algorithm"
  - Motivation: CB's legacy code, users' statistical confusion

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"
  - Valid inferences: easy, but *data not (yet) released!*
- Post-Processed data released: "TopDown Algorithm"
  - Motivation: CB's legacy code, users' statistical confusion
  - Valid inferences: (most are) extremely difficult

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"
  - Valid inferences: easy, but *data not (yet) released!*
- Post-Processed data released: "TopDown Algorithm"
  - Motivation: CB's legacy code, users' statistical confusion
  - Valid inferences: (most are) extremely difficult
  - Proper statistical methods: not developed yet

What can the Fed do?

US Census Privatization Strategies

1990-2010: Secret (failed) Privatization

- Methods: Swapping, top coding, cell suppression (no details)
- Privatization fails: most people can be reidentified
- Valid inferences: impossible

2020: Public Privatization

- Method: Add DP noise to census block counts (public DGP)
- Privatized "Noisy Measurements File"
  - Valid inferences: easy, but *data not (yet) released!*
- Post-Processed data released: "TopDown Algorithm"
  - Motivation: CB's legacy code, users' statistical confusion
  - Valid inferences: (most are) extremely difficult
  - Proper statistical methods: not developed yet

What can the Fed do?

Push Census Bureau to release the Noisy Measurements File