Statistical Privacy in the 21st Century: Census and Consensus

An Overview of the 2020 Census Disclosure Avoidance System and Noisy Measurement Files

Joint Statistical Meetings
Toronto, Canada
August 7, 2023

Michael Hawes
Senior Statistician for Scientific Communication
Research and Methodology

United States® Census Bureau

Disclosure Avoidance for the 2020 Census

The 2020 Census improves on the noise injection methods of the 1990-2010 Censuses by employing a risk assessment framework based on Differential Privacy (DP) to assess and quantify disclosure risk and confidentiality protection.

Every individual that is reflected in a particular statistic contributes towards that statistic's value.

Every statistic that you publish "leaks" a small amount of private information.

DP as a framework allows you to assess each individual's contribution to the statistic, and to measure (and thus, limit) how much information about them will leak.

United States® Census Bureau

The 2020 Census Disclosure Avoidance System (DAS)

TopDown Algorithm (TDA)

Produces privacy-protected microdata (Microdata Detail File) that is ingested by Decennial tabulation system

- Redistricting Data (P.L. 94-171) Summary File
- Demographic Profile
- Demographic and Housing Characteristics File (DHC)
- Congressional District Summary Files

SafeTab PHSafe

Produce privacy-protected tabulations

- Detailed DHC-A
- Detailed DHC-B
- Supplemental DHC

United States® Census Bureau

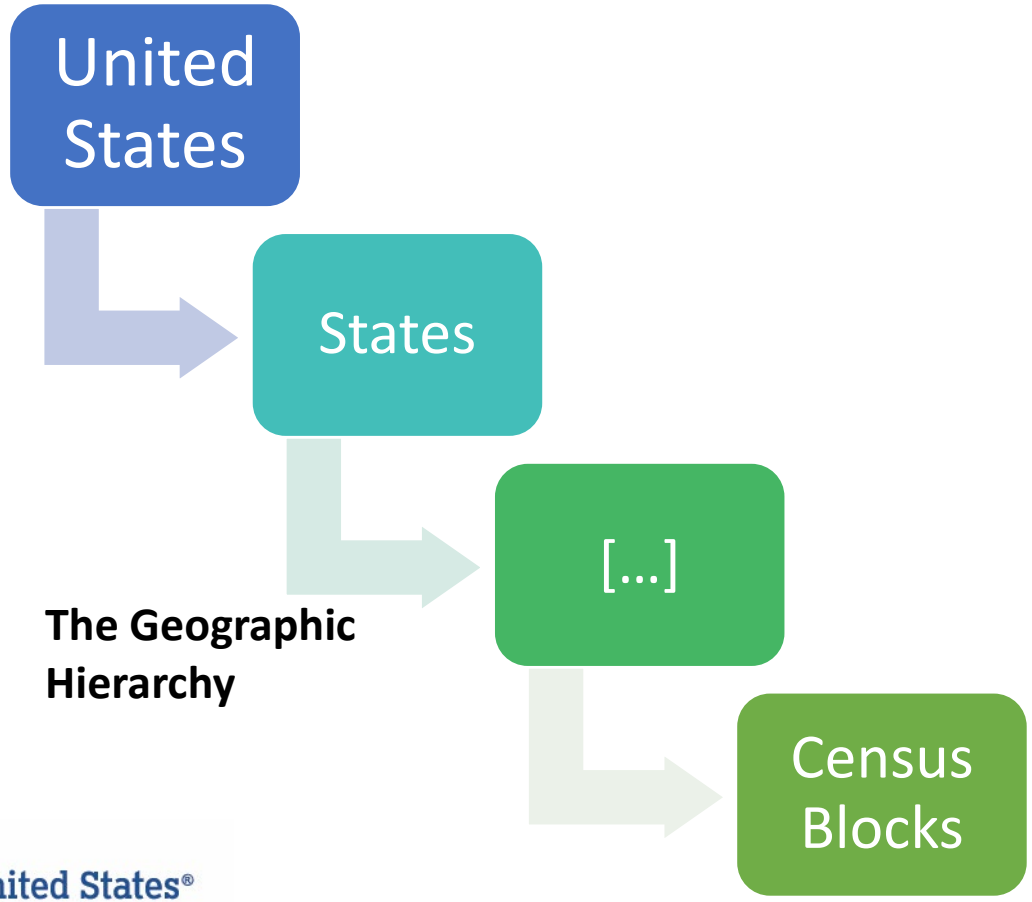
The TopDown Algorithm

*A histogram, in this context, is a tabular representation of the microdata with counts of records for each possible combination of values for each attribute in the microdata.

For complete details see: Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*. (June) https://doi.org/10.1162/99608f92.529e3cb9

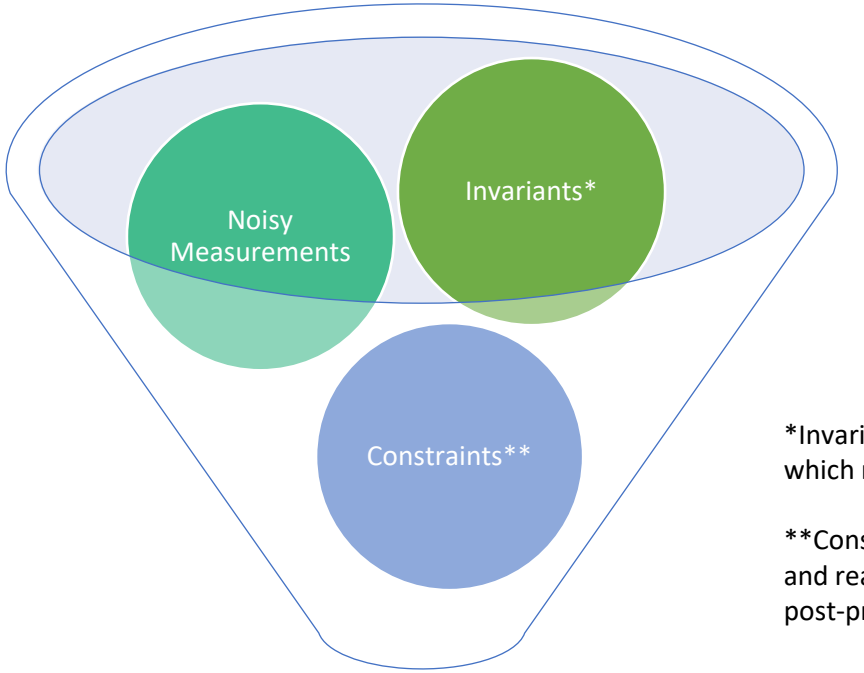
United States® Census Bureau

The TopDown Algorithm



The Geographic Hierarchy

At each geographic level:



Internally consistent histogram

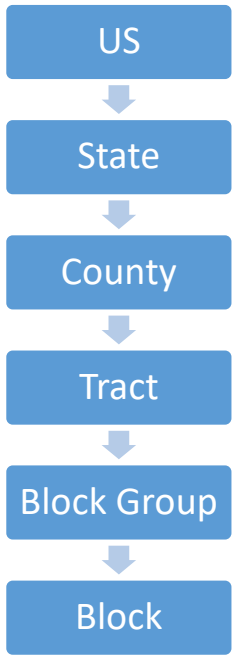
*Invariants are counts to which no noise is added.

**Constraints are consistency and reasonableness rules the post-processing must impose.

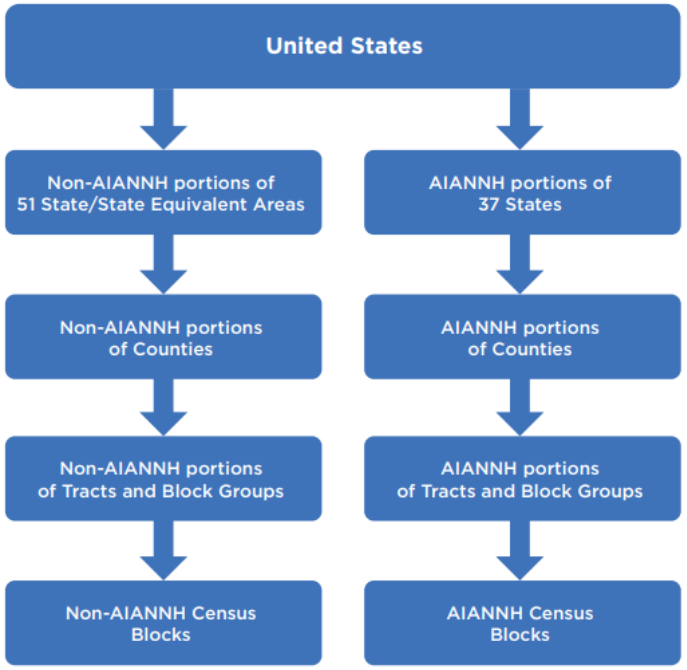
United States® Census Bureau

The Geographic Hierarchy ("Spine")

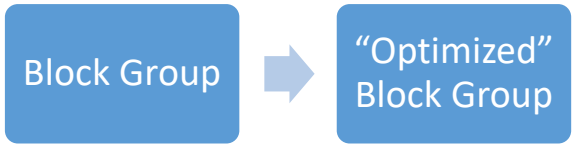
Standard Spine Tabulation Hierarchy



TDA's American Indian/Alaska Native/Native Hawaiian (AIANNH) Spine for Redistricting



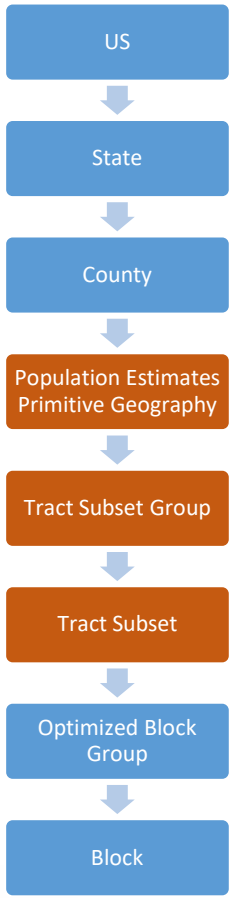
Geographic Optimization
(for Redistricting Data)



Reconfigured TDA's definition of block groups to optimize accuracy in statistics for certain types of geographies, including Minor Civil Divisions, Places, and individual AIANNH areas. Optimized block groups are used inside TDA. Tabulation block groups from the standard hierarchy are used for all published data tables.

United States® Census Bureau

Population Estimates Primitive Geographies



Population Estimates Primitive Geographies are the most granular geographic areas that are required in order to derive tables for every geography for which official Population Estimates are produced.

The Population Estimates Primitive Geographies form a complete, mutually exclusive partition of the U.S.

Tract Subsets are defined as the intersection of Population Estimates Primitive Geographies with census tabulation tracts.

Tract Subset Groups are defined as the union of multiple tract subsets that are all within the same Population Estimates primitive geography.

For the DHC, the TDA Geographic Hierarchy was further modified to include “Population Estimates Primitive Geographies” on the spine.

United States® Census Bureau

Queries and Privacy-loss Budget Allocation

Global *rho*	2.56
Global *epsilon*	17.90
delta	10^{-10}

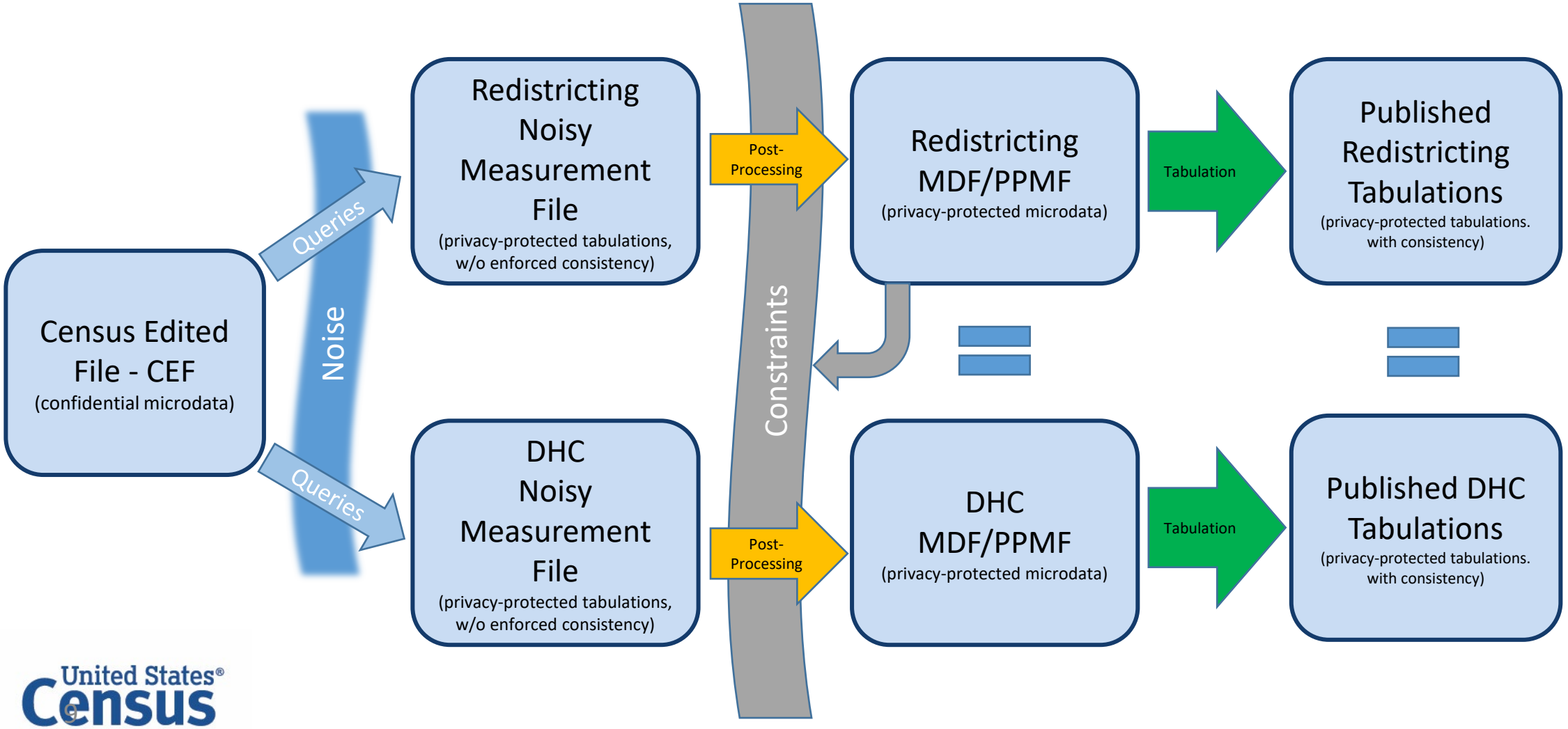
	***rho* Allocation by Geographic Level**
US	2.54%
State	35.13%
County	10.91%
Tract	16.76%
Optimized Block Group*	30.64%
Block	4.03%

Production settings for the 2020 Census Redistricting Data (P.L. 94-171) Summary File (Persons tables P1-P5)

Query	**Per Query *rho* Allocation by Geographic Level**					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		32.35%	8.32%	6.40%	12.75%	0.00%
CENRACE (63 cells)	0.03%	0.05%	0.03%	0.03%	0.02%	0.01%
HISPANIC (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHINSTLEVELS (3 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHGQ (8 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HISPANIC*CENRACE (126 cells)	0.08%	0.10%	0.07%	7.90%	7.89%	0.02%
VOTINGAGE*CENRACE (126 cells)	0.08%	0.10%	0.07%	0.08%	0.07%	0.02%
VOTINGAGE*HISPANIC (4 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE*HISPANIC*CENRACE (252 cells)	0.27%	0.29%	0.27%	0.27%	0.18%	0.07%
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	1.99%	1.97%	2.01%	1.97%	9.63%	3.88%

United States® Census Bureau

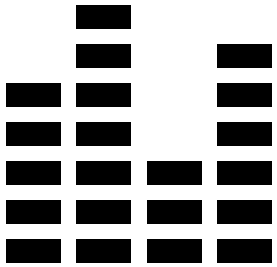
Noisy Measurement Files (NMFs), Privacy-Protected Microdata Files (PPMFs), Published Tabulations



Should I Use the NMF, the PPMF, or the Tabulations?

- There are two sources of error in the published statistics (PPMF and Tabulations):

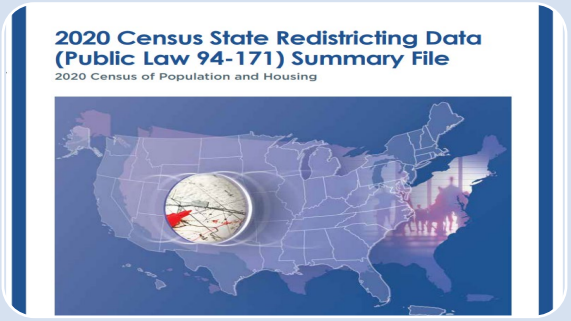
Differentially private noise
- Unbiased
- Known distribution
- Reflected in the noisy measurements



Post-processing
- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

United States® Census Bureau

Should I Use the NMF, the PPMF, or the Tabulations?



2020 Census Redistricting and DHC Tabulations

- Official 2020 Census Statistics
- Higher Accuracy (feature of TDA)
- Does include bias due to post-processing

2020 Census PPMF

- 100% microdata file
- Consistent with published tabulations
- Useful for special tabulations and microdata analysis

2020 Census NMF

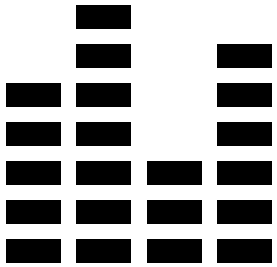
- Can be used to produce unbiased estimates and confidence intervals
- Can be used to evaluate alternate post-processing mechanisms
- Research product

United States® Census Bureau

Should I Use the NMF, the PPMF, or the Tabulations?

- There are two sources of error in the published statistics (PPMF and Tabulations):

Differentially private noise
- Unbiased
- Known distribution
- Reflected in the noisy measurements



Post-processing
- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

United States® Census Bureau

How to Access the NMFs

There are approximately 25 trillion numbers associated with the combined redistricting data and DHC Noisy Measurement Files, consuming 33 terabytes of space in a compressed format. Given their large size, the files are housed and accessible via off-site locations.

Use the links below to access the files:

2010 Demonstration Data	Release Date	Download Locations	
Redistricting Data (P.L. 94-171)	April 3, 2023	Inter-university Consortium for Political and Social Research (ICPSR)*	Harvard Dataverse
Demographic and Housing Characteristics File (DHC)	June 30, 2023	ICPSR*	N/A
2020 Research Data	**Release Date**	**Download Locations**	
Redistricting Data (P.L. 94-171)	June 15, 2023	ICPSR*	Harvard Dataverse
Demographic and Housing Characteristics File (DHC)	Fall 2023		

* Requires registering for a free Globus account to access and download.

United States® Census Bureau

Content of the NMF

	geocode	query_name	hhgq	votingage	hispanic	cenrace	query_shape	value	variance	plb
0	1100890045	total_dpq	*	*	*	*	[1 1 1 1]	[2980]	6.110105	[52707577526 322048826043]
1	1100890045	cenrace_dpq	*	*	*	cenrace	[1 1 1 63]	[1975 837 22 6 3 51 9 -9 26 7 -20 -40 72 47 -55 -9 44 -52 -44 44 55 -49 11 -39 -41 -1 -70 -20 12 20 22 -17 51 45 34 76 23 20 0 -27 5 48 6 22 -66 -24 -41 20 -6 -4 -15 -12 -7 -72 -42 1 -26 -14 16 -30 -14 26 -48]	1196.8168	[269087824 322048826043]
2	1100890045	hispanic_dpq	*	*	hispanic	*	[1 1 2 1]	[2890 123]	1914.9069	[168179890 322048826043]
3	1100890045	votingage_dpq	*	votingage	*	*	[1 2 1 1]	[379 2645]	1914.9069	[168179890 322048826043]
4	1100890045	hhinstlevels_dpq	hhinstlevels	*	*	*	[3 1 1 1]	[1781 1127 29]	1914.9069	[168179890 322048826043]
5	1100890045	hhgq_dpq	hhgq	*	*	*	[8 1 1 1]	[1754 1017 59 120 6 -68 -84 -10]	1914.9069	[168179890 322048826043]
6	1100890045	hispanic * cenrace_dpq	*	*	hispanic	cenrace	[1 1 2 63]	[1933 856 24 18 2 2 11 15 3 1 1 7 1 -2 -4 -1 2 -1 5 3 2 -1 -3 -3 1 -2 2 0 2 0 -1 0 -1 0 1 2 -6 -3 -1 3 1 0 -1 -1 1 -2 0 0 -3 -5 1 2 1 1 -2 -1 0 7 -3 -4 0 1 2 67 8 10 -2 -1 13 0 4 0 0 3 -1 0 -2 3 1 1 7 6 2 0 2 0 4 -2 -1 -1 3 0 2 -2 3 3 -2 4 3 4 -1 1 -3 -5 2 -2 -2 1 2 0 -1 4 0 1 -1 -1 4 3 2 -2 4 2 0 0 3 -1]	4.9531994	[65018345474 322048826043]

United States® Census Bureau

Using the NMFs to Generate Unbiased Confidence Intervals

- Recorded webinar on how to use the NMFs available here: Noisy Measurement Files for the Redistricting and DHC Data Products (census.gov)

- Sample code, examples, and useful tips: Computing Confidence Intervals Using the 2010 Census Redistricting (P.L. 94-171) Demonstration Data Noisy Measurement File (2023-04-03) by Ryan Cumings-Menon, Michael Hawes, and Matthew Spence (April 2023)

United States® Census Bureau

Generating Confidence Intervals

Generating a more precise county-level estimate and CI by leveraging multiple noisy measurements

Bullock County, AL (00110011):	Tract 1 (001100110001):	Tract 2 (001100110002):	Tract 3 (001100110003):
total_dpq: 10911 variance: 4.70162740	total_dpq: 1433 variance: 6.11010487	total_dpq: 7105 variance: 6.11010487	total_dpq: 2376 variance: 6.11010487

In [6]:
```
# Constants in this example:
coef_mat = np.vstack((np.ones((1, 3)), np.eye(3)))
var_county = 4.70162740
var_tracts = 6.11010487
sigma_sqrds = np.array([var_county] + [var_tracts] * 3)
dp_answers = np.array([10911, 1433, 7105, 2376])

weights = np.diag(1 / sigma_sqrds)
point_estimate, ci_half_width = simulation_based_ci(dp_answers, coef_mat, weights, sigma_sqrds, prng,
num_sims, alpha)
print(f"The confidence interval is: {point_estimate} +/- {ci_half_width}")
```

```
The confidence interval is: 10911.612405249798 +/- 3.193797375100286
```

Tip: Even when the statistic of interest is measured directly in the NMFs, it can be advantageous to incorporate additional information (e.g., noisy measurements for the child geounits) in order to obtain more precise estimates and smaller CIs.

United States® Census Bureau

How do these estimates compare?

2010 Total Population	SF1	2010 PPMF	2010 NMF
Bullock County, AL	10,914	10,912	10,911.61 +/- 3.19
Menominee, WI	4,232	4,230	4,232 +/- 5
Redfield, IA	835	835	835 +/- 4

Simulated Error in Population in Households for Counties (Excluding Puerto Rico) From Nonsampling Variability

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**117.27**	**-248**	**+230**
Counties with housing unit population between 0-999	37	10.03	-10	+27
Counties with housing unit population between 1,000-9,999	691	28.23	-38	+71
Counties with housing unit population between 10,000-99,999	1,849	74.45	-131	+177
Counties with housing unit population between 100,000-999,999	527	292.21	-784	+545
Counties with housing unit population at or above 1,000,000	39	1,463.12	-3,659	+1,351

Source: U.S. Census Bureau, Research and Methodology Directorate, simulations based on 2010 Census Coverage Measurement research and 2010 Census housing unit population.

Simulated Error in Population in Households for Counties (Excluding Puerto Rico) From Census Coverage Error

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**964.00**	**-1,841**	**+2,048**
Counties with housing unit population between 0-999	37	23.00	-22	+54
Counties with housing unit population between 1,000-9,999	691	121.00	-146	+284
Counties with housing unit population between 10,000-99,999	1,849	446.00	-832	+1,053
Counties with housing unit population between 100,000-999,999	527	2,930.00	-7,222	+6,278
Counties with housing unit population at or above 1,000,000	39	14,848.00	-44,833	+20,007

Source: U.S. Census Bureau, Research and Methodology Directorate, simulations based on 2010 Census Coverage Measurement research and 2010 Census housing unit population.

Source: Understanding Disclosure Avoidance-Related Variability 2020 Census

United States® Census Bureau

What additional resources would you like to see?