Disclosure Avoidance and the 2020 Census Demographic and Housing Characteristics (DHC) File

Michael Hawes
Senior Survey Statistician for Scientific Communication
Research and Methodology Directorate
U.S. Census Bureau

May 24, 2023

United States Census Bureau

Any viewpoints or opinions expressed in this presentation are entirely the author's own and do not represent the viewpoints or opinions of the U.S. Census Bureau

2020 Census Data Products

Released

Apportionment
April 26, 2021

Redistricting File
(Public Law 94-171)
August 12, 2021
September 16, 2021

Demographic Profile

Demographic and Housing Characteristics File (DHC)
May 25, 2023

Congressional District Summary Files
Planned August 2023

Detailed DHC-A
Planned September 2023

Detailed DHC-B
Release Date TBD

Supplemental DHC (S-DHC)
Release Date TBD

Future Effort

Public Use Microdata File

Special Tabulations

More information about the products is available on the About 2020 Census Data Products webpage.

United States Census Bureau

Apportionment Release

- Apportionment is the process of dividing the 435 memberships, or seats, in the U.S. House of Representatives among the 50 states. At the conclusion of each decennial census, the results are used to calculate the number of seats to which each state is entitled.

- **Results were released on April 26, 2021**
- **Subjects include:**
  - Resident population
  - Overseas population
  - Apportionment population
- **Geography:** 50 states, the District of Columbia (DC), and Puerto Rico
- **Disclosure avoidance:** Results do <u>not</u> undergo disclosure avoidance

United States® Census Bureau

Redistricting File (Public Law 94-171)

- Public Law 94-171 directs the Census Bureau to provide data to the governors and legislative leadership in each of the 50 states for redistricting purposes. This product is the first file released that includes demographic and housing characteristics.

- **Results were released on August 12, 2021 (Summary Files) and September 16, 2021 (data.census.gov)**
- **Subjects include:**
  - Voting age
  - Race
  - Hispanic or Latino origin
  - Housing occupancy
  - Group quarters (GQ) population by major GQ type
- **Lowest level of geography:** Census Block
- **Disclosure avoidance:** Differentially private TopDown Algorithm (TDA)

United States Census Bureau

Demographic Profile

- This product will provide select demographic and housing characteristics about local communities in a streamlined, easy to use format.

- **Expected release date:** May 25, 2023
- **Subjects include:**
  - Sex by 5-year age groups
  - Median age by sex
  - Race
  - Hispanic or Latino origin
  - Relationship to householder
  - GQ population
  - Household type
  - Housing occupancy
  - Housing tenure
- **Lowest level of geography:** Tract
- **Disclosure avoidance:** Differentially private TDA

United States® Census Bureau

Demographic and Housing Characteristics File (DHC)

- The DHC will include many of the demographic and housing tables previously included in 2010 Summary File 1 (2010 SF1). Some tables are repeated by race and ethnicity.

- **Expected release date:** May 25, 2023
- **Subjects include:**
  - Sex by single year-of-age
  - Hispanic or Latino origin of householder by race of householder
  - GQ population by sex by age
  - Relationship by age for population under 18 years
  - Household type by relationship and presence of people of specific ages
  - Multigenerational households
  - Family type by presence of children
  - Tenure by household size
  - Tenure by household type by age of householder
  - Vacancy Status
- **Lowest level of geography:** Varies with many tables at Census Block
- **Disclosure avoidance:** Differentially private TDA

United States Census Bureau

Detailed Demographic and Housing Characteristics File A (Detailed DHC-A)

- Detailed DHC-A includes population counts repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribal and village population groups
- **Expected release date:** Sept 2023
- **Subjects are repeated by detailed racial and ethnic groups:**
  - Total population
  - Sex by Age for Selected Age Categories
- **Proposed levels of geography:** Nation, State, County, Tract, Place, AIANNH areas
- **Disclosure avoidance:** Differentially private SafeTab-P algorithm

United States® Census Bureau

Detailed Demographic and Housing Characteristics File B (Detailed DHC-B)

- Detailed DHC-B includes household counts repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribal and village population groups

- **Expected release date:** TBD
- **Subjects are repeated by detailed racial and ethnic groups:**
  - Household Type
  - Tenure
- **Proposed levels of geography:** Nation, State, County, Tract, Place, AIANNH areas
- **Disclosure avoidance:** Differentially private SafeTab-H algorithm

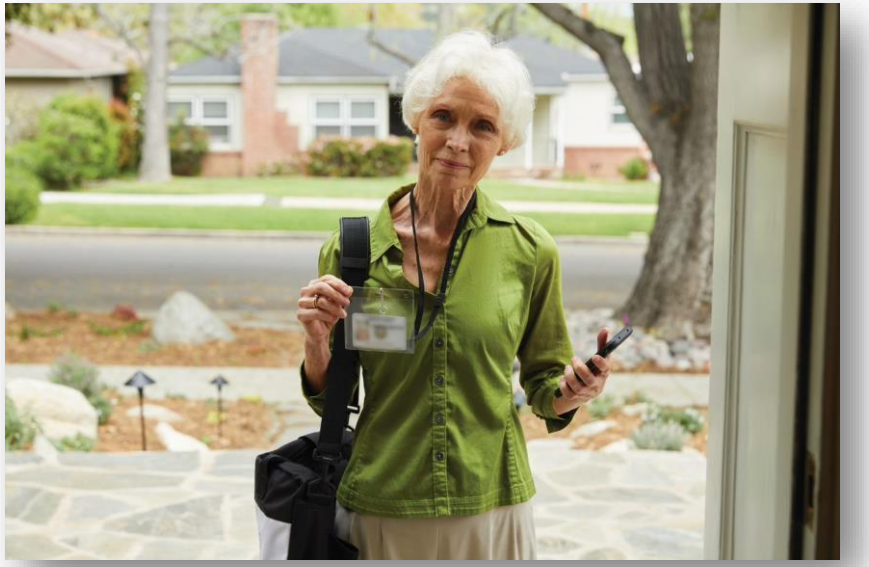
United States® Census Bureau

Background on Confidentiality Protections for the 2020 Census Data Products

United States Census Bureau

Keeping the Public's Trust: Title 13

"To stimulate public cooperation necessary for an accurate census…Congress has provided assurances that information furnished by individuals is to be treated as confidential. ***Title 13 U.S.C. §§ 8(b) and 9(a)*** *explicitly provide for nondisclosure of certain census data, and* ***no discretion is provided to the Census Bureau on whether or not to disclose such data****…"* (U.S. Supreme Court, Baldrige v. Shapiro, 1982)



To safeguard the public's confidential census responses, the Census Bureau has long employed a variety of statistical techniques to mitigate disclosure risk in our published data products.

United States® Census Bureau

Disclosure Avoidance for Past Censuses

1970-1980 Censuses

	528	
		794
	581	
137	941	189
931		
	250	
		590

SUPPRESSION

1990-2010 Censuses

668	178	779
91	8	159
809	112	811
518	424	955
989	352	765
237	411	686
77	820	590

SWAPPING

United States Census Bureau

The Ever-rising Risk of Disclosure

- Any data release carries some risk of disclosure.
- Improvements in computing power and the explosion of third-party data mean that disclosure risk has increased significantly.
- Protecting confidentiality means adapting and responding to these increasing threats

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 mathematical framework known as 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 inject a precise amount of noise into the statistic to 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

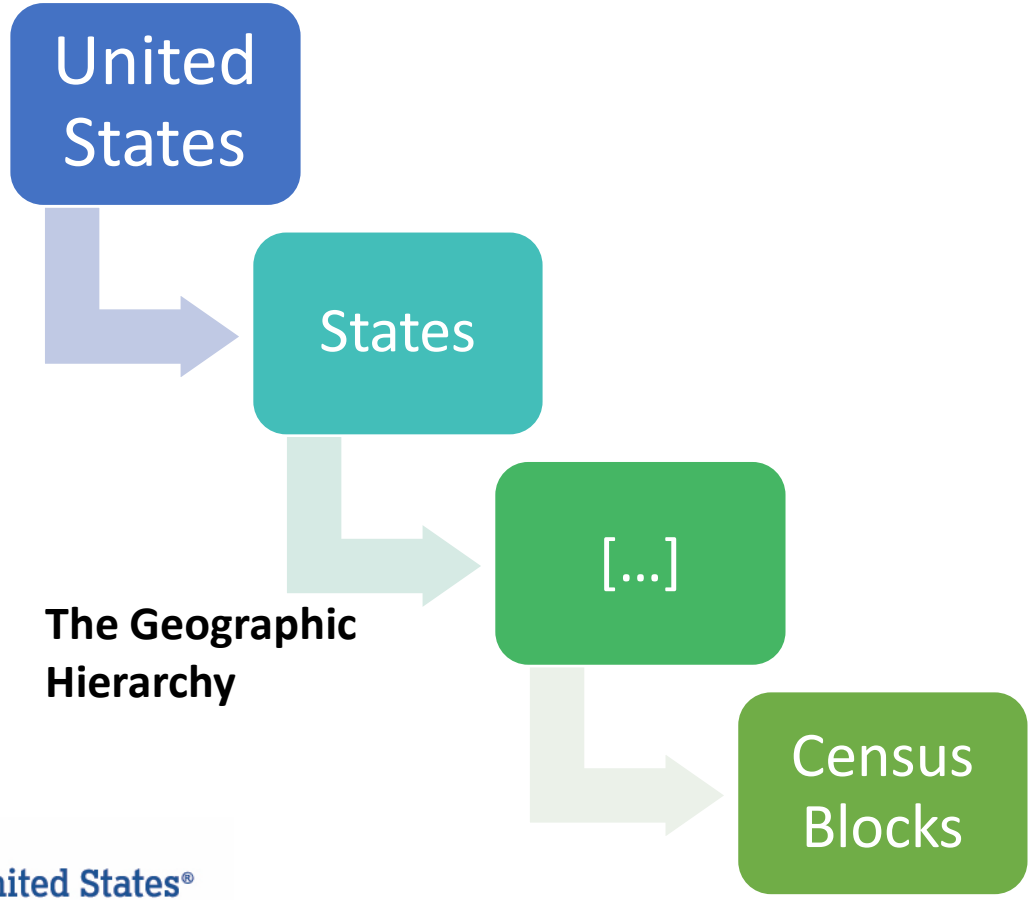
The TopDown Algorithm

*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

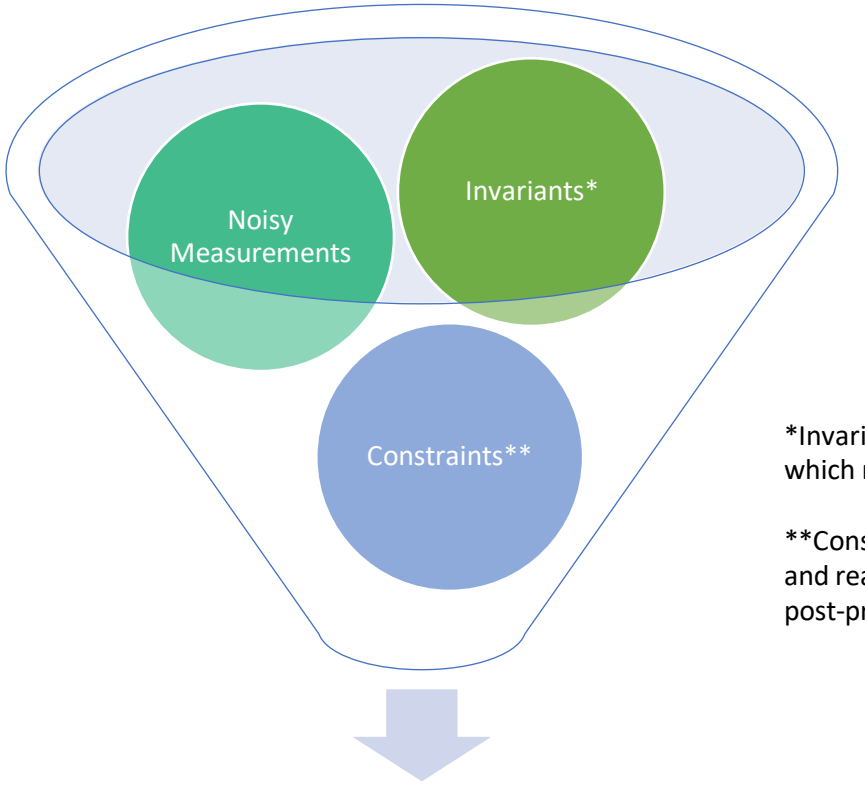
United States Census Bureau

The TopDown Algorithm



The Geographic Hierarchy

At each geographic level:



*Invariants are counts to which no noise is added.

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

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

2010 Demonstration Data

United States® Census Bureau

Components of the 2010 Demonstration Data Products Suite – Redistricting and Demographic and Housing Characteristics File – Production Settings (2023-04-03) **(2010 DDPS)**

(2010 Census data processed through the 2020 DAS at production settings)

- 2010 DDPS Fact Sheet
- Detailed Summary Metrics (and Metrics Overview)
- Privacy-Protected Microdata File (PPMF)
- DHC Tabulations (via IPUMS)
- Privacy-loss Budget (PLB) Allocations
- Noisy Measurement File (NMF)

United States Census Bureau

Fitness-for-Use: Detailed Summary Metrics

- One way to understand the anticipated fitness of use – Detailed Summary Metrics, which compare tabulations from 2010 Census data run through Disclosure Avoidance System (DAS) to published 2010 data.
- These comparisons (e.g., absolute value of difference) are averaged across geographies in a geographic level (e.g., counties) to create estimates of accuracy (e.g., Mean Absolute Error or MAE) or bias or to count large outliers.
- We can use these metrics to think about the fitness-for-use of 2020 DHC data, since the same software and settings are used to produce those.

United States® Census Bureau

Fitness-for-Use: Detailed Summary Metrics

- Have released **9** versions of Detailed Summary Metrics that show how the DAS and its parameter settings have evolved over time due to internal and external feedback on use cases and data user needs.
- This release showcases improvements that were made between the last demonstration data release (August 2022) and the final code settings for the production run in November 2022.
- Based on these 2010 evaluations, there are acuracy improvements for:
  - Householder race
  - Presence and age of own children
  - Relationship to householder
  - Single year of age for children for counties and school districts
  - Age for persons in group quarters
  - Same-sex married and unmarried partners

United States® Census Bureau

Highlights from the Detailed Summary Metrics

- Looking beyond average error (MAE/MAPE):
  - 95th Percentile Absolute Error
  - 95th Percentile Absolute Percent Error

  } 95% of geographies are off by this much (or less).
- Comparing to average populations to provide context

Tenure Numbers are: 2022-08-25 / 2023-04-03 Universe: Occupied Housing Units. Geography: Census Tract

	Mean Population	Mean Absolute Error (MAE)	95th Percentile Absolute Error (95th AE)	Mean Absolute Percent Error (MAPE)	95th Percentile Absolute Percent Error (95th APE)
Owned with a mortgage	725	1.65 / 2.04	4.00 / 5.00	1% / 1%	1% / 1%
Owned free and clear	315	1.65 / 2.05	4.00 / 5.00	2% / 2%	4% / 5%
Renter-occupied	528	1.66 / 2.04	4.00 / 5.00	1% / 1%	3% / 3%

- Acceptable errors even for the 95th percentile case

United States Census Bureau

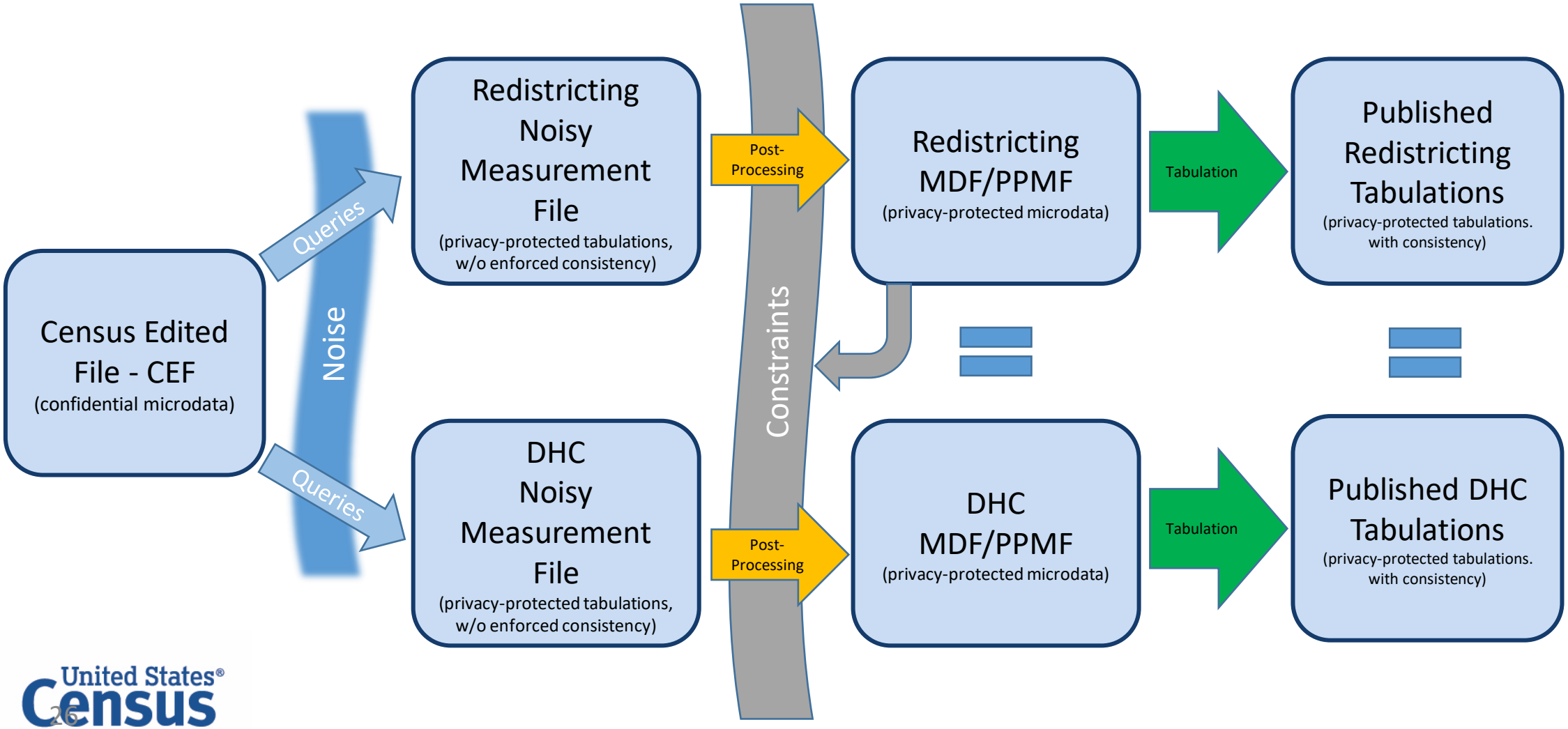
Updated Metrics on 2010 DDPS Accuracy

Coupled Household Type	Numbers are: 2022-08-25 / 2023-04-03			Universe: Households. Geography: County	
	Mean Population	MAE	95th AE	MAPE	95th APE
Opposite-sex married couple household	17,980	5.83/ 4.58	15 / 12	0% / 0%	1% / 1%
Same-sex married couple household	111	3.62/ 2.38	9 / 6	21% / 15%	80% / 60%
Opposite-sex unmarried partner household	2,177	5.12 / 3.68	13 / 10	2% / 1%	8% / 6%
Same-sex unmarried partner household	176	3.61 / 2.29	10 / 6	33% / 24%	200% / 100%

Tenure By Race of Householder	2022-08-25 / 2023-04-03			Universe: Occupied Housing Units. Geography: County	
	Mean Population	MAE	95th AE	MAPE	95th APE
Owner occupied					
White alone	20,187	83.61 / 5.54	258 / 15	2% / 0%	5% / 1%
Black or African American alone	1,992	48.09 / 4.10	157 / 12	238% / 32%	1,200% / 200%
American Indian and Alaska Native alone	162	18.08 / 3.45	60 / 10	81% / 23%	300% / 100%
Asian alone	856	31.60 / 2.96	112 / 9	227% / 35%	948% / 200%
Native Hawaiian and Other Pacific Islander alone	20	5.10 / 1.51	18 / 5	128% / 65%	400% / 200%
Some Other Race alone	629	15.29 / 3.13	53 / 8	66% / 26%	270% / 117%
Two or More Races	332	19.37 / 3.38	63 / 9	47% / 12%	188% / 44%
Renter occupied					
White alone	8,370	57.34 / 4.59	179 / 12	3% / 0%	11% / 2%
Black or African American alone	2,504	33.81 / 3.43	102 / 9	167% / 28%	800% / 200%
American Indian and Alaska Native alone	137	14.67 / 3.06	54 / 8	90% / 29%	375% / 140%
Asian alone	618	17.62 / 2.41	62 / 7	164% / 38%	700% / 200%
Native Hawaiian and Other Pacific Islander alone	26	6.61 / 1.53	27 / 5	146% / 66%	400% / 200%
Some Other Race alone	936	12.75 / 3.04	44 / 8	46% / 21%	200% / 100%
Two or More Races	368	16.26 / 3.08	54 / 8	55% / 16%	214% / 62%

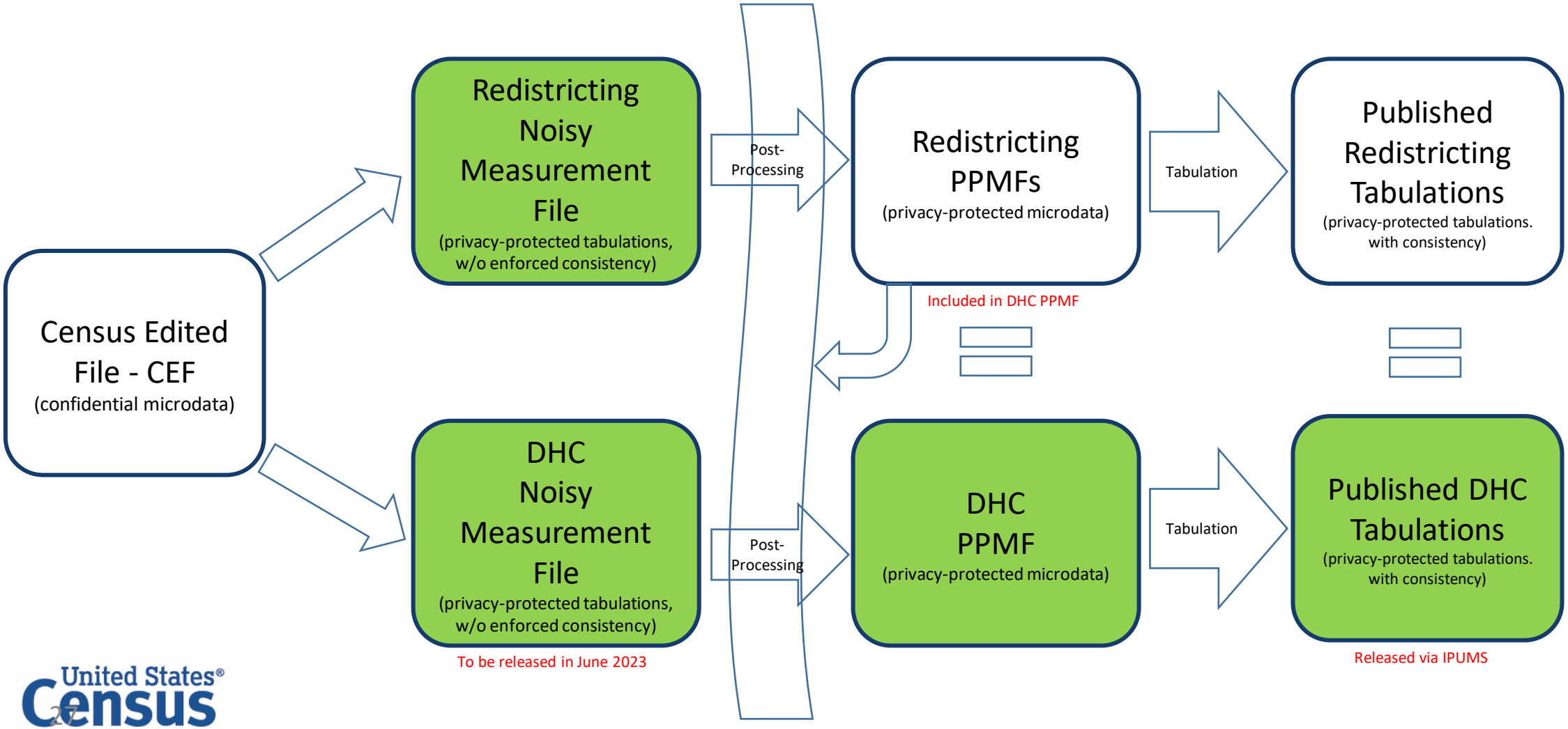
Single Year of Age for Ages 0-15	2022-08-25 / 2023-04-03				Universe: Persons. Geography: County
	Mean Population	MAE	95th AE	MAPE	95th APE
Under 1 year old	1,255	14.62 / 5.58	42 / 14	6% / 3%	25% / 14%
1 year old	1,266	14.93 / 5.47	44 / 14	6% / 4%	22% / 14%
2 years old	1,304	14.56 / 5.47	42 / 13	6% / 3%	22% / 13%
3 years old	1,311	13.15 / 5.96	38 / 14	6% / 4%	20% / 14%
4 years old	1,293	13.20 / 6.15	37 / 15	6% / 4%	21% / 15%
5 years old	1,291	16.50 / 5.93	50 / 14	7% / 4%	23% / 14%
6 years old	1,294	16.47 / 5.96	51 / 15	7% / 4%	25% / 14%
7 years old	1,282	16.42 / 6.10	47 / 15	7% / 4%	24% / 15%
8 years old	1,287	17.07 / 5.93	48 / 15	7% / 4%	24% / 14%
9 years old	1,320	17.01 / 5.98	49 / 15	7% / 4%	24% / 14%
10 years old	1,328	16.69 / 6.06	49 / 16	7% / 3%	24% / 13%
11 years old	1,309	16.55 / 6.02	46 / 16	7% / 3%	23% / 13%
12 years old	1,306	16.88 / 5.92	50 / 15	7% / 3%	25% / 13%
13 years old	1,310	16.82 / 6.02	48 / 15	7% / 4%	24% / 13%
14 years old	1,325	16.87 / 6.11	49 / 16	7% / 4%	23% / 13%
15 years old	1,350	5.11 / 3.98	13 / 11	3% / 2%	11% / 8%

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



United States Census Bureau

2010 DDPS NMFs, PPMF, Tabulations

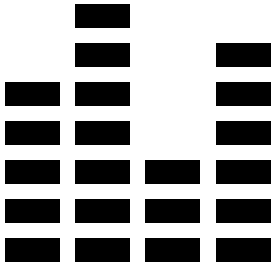


Should I Use the NMF, the PPMF, or the Tabulations?

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

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements

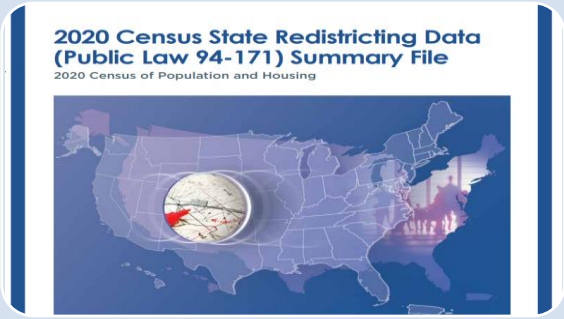


Post-processing

- 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

New Resources for Data Users

United States® Census Bureau

Reader-Friendly Disclosure Avoidance Briefs

- Disclosure Avoidance and the 2020 Redistricting Data
- Why the Census Bureau Chose Differential Privacy
- Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works

More resources are in development, as well as additional specific guidance and training for using the 2020 Census data.

United States Census Bureau

Coming Soon

- Guidance and examples on how to use the NMF to calculate unbiased estimates and confidence intervals.
- Subscribe to our newsletter to receive the announcement and related webinar info.

United States Census Bureau

Questions?

Or send them to 2020DAS@census.gov

United States Census Bureau

Supplemental Slides

United States® Census Bureau

About the Detailed DHC-A

- Subjects repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribes and villages:
  - Total population
  - Sex by age for selected age categories

- Geographic levels included:
  - Nation
  - State
  - County
  - Tracts
  - Places
  - American Indian/Alaska Native/Hawaiian Home Land (AIANNH) areas

- Planned for release in September 2023

United States® Census Bureau

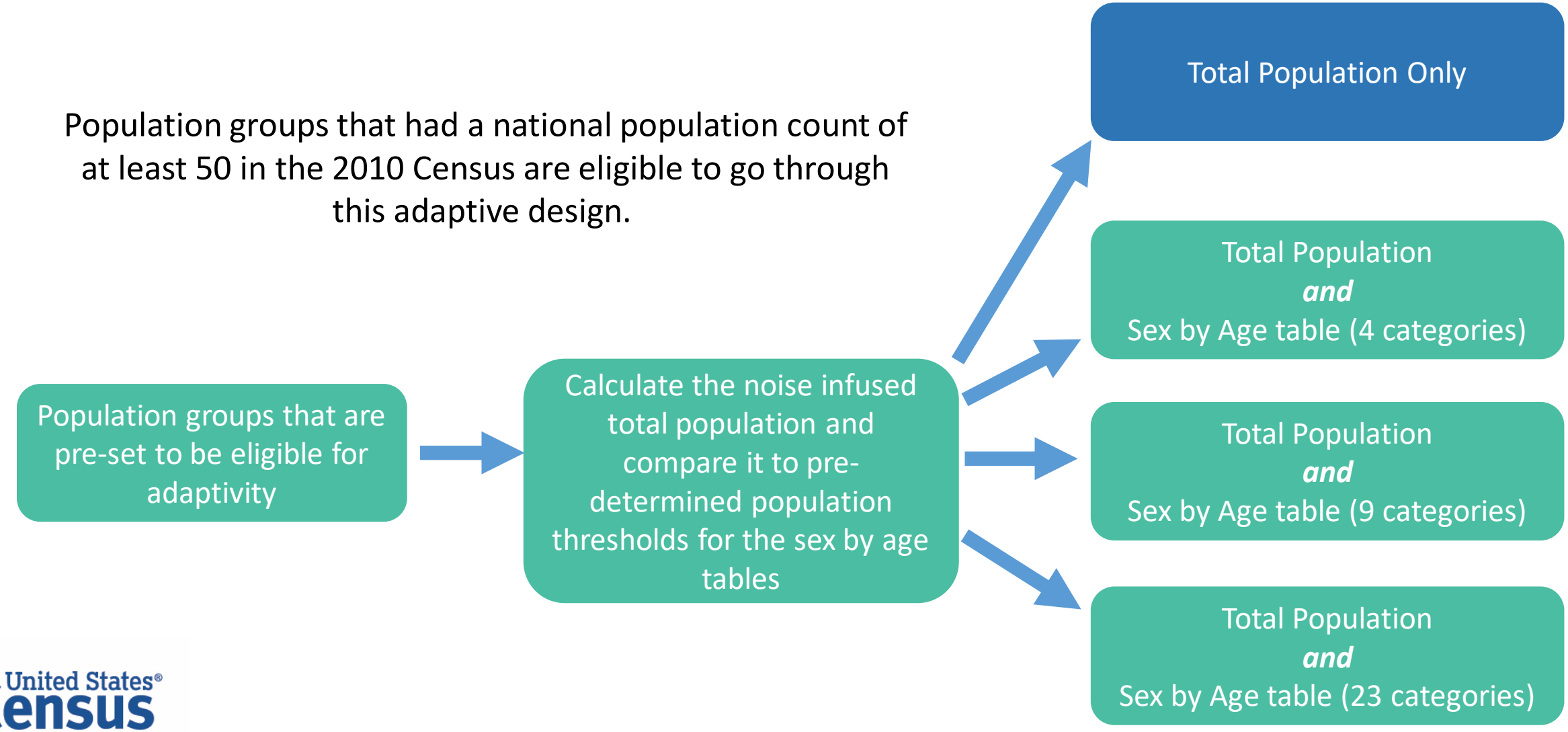
Using Adaptive Design to Produce the Detailed DHC-A

Detailed groups with a national population count less than 50 in the 2010 Census are pre-set to receive total population only in the Detailed DHC-A.

United States Census Bureau

Using Adaptive Design to Produce the Detailed DHC-A

Population groups that had a national population count of at least 50 in the 2010 Census are eligible to go through this adaptive design.



United States Census Bureau

Sex by Age
23 categories

Sex x Age(23)		
Age Group	**Male**	**Female**
Under 5 years	146	147
5 to 9 years	130	131
10 to 14 years	109	107
15 to 17 years	79	77
18 and 19 years	68	97
20 years	34	89
21 years	64	76
22 to 24 years	150	151
25 to 29 years	235	265
30 to 34 years	237	302
35 to 39 years	230	374
40 to 44 years	177	356
45 to 49 years	187	372
50 to 54 years	139	275
55 to 59 years	84	151
60 and 61 years	27	53
62 to 64 years	30	60
65 and 66 years	7	15
67 to 69 years	12	27
70 to 74 years	12	25
75 to 79 years	13	12
80 to 84 years	3	7
85 years and over	2	3

Sex by Age
9 categories

Sex x Age(9)		
Age Group	**Male**	**Female**
Under 5 years	146	147
5 to 17 years	318	315
18 to 24 years	316	413
25 to 34 years	472	567
35 to 44 years	407	730
45 to 54 years	326	647
55 to 64 years	141	264
65 to 74 years	31	67
75 years and over	18	22

Sex by Age
4 categories

Sex x Age(4)		
Age Group	**Male**	**Female**
Under 18 years	464	462
18 to 44 years	1,195	1,710
45 to 64 years	467	911
65 years and over	49	85

Note: These are fictitious data for illustration purposes.

United States® Census Bureau

Detailed DHC-A Minimum Noise Infused Population Counts and Margins of Error (MOE) by Geography

	Detailed groups		Regional groups	
Most comprehensive table type produced	**Nation & state (MOE=3)**	**Sub-state & AIANNH (MOE=11)**	**Nation & state (MOE=50)**	**Sub-state (MOE=50)**
Total count only	0-499	22-999	0-4,999	94-4,999
Sex by age – 4 categories	500-999	1,000-4,999	5,000-19,999	5,000-19,999
Sex by age – 9 categories	1,000-6,999	5,000-19,999	20,000-149,999	20,000-149,999
Sex by age – 23 categories	7,000+	20,000+	150,000+	150,000+

Note: The listed population thresholds are applied to the population counts after they have been processed by the approved differential privacy mechanism.
Note: MOE refers to the margin of error

United States® Census Bureau