# Differential Privacy and Census Data

Jonathan Buttle
California Department of Finance
Demographic Research Unit

# Why Differential Privacy?

- Title 13 specifies that “the Census Bureau shall not make any publication whereby the data furnished by any particular establishment or individual ... can be identified” (Title 13 U.S.C. § 9(a)(2), Public Law 87-813);
- Title 5 further prohibits “any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means” (Title 5 U.S.C. §502 (4), Public Law 107–347);

# Why Differential Privacy – Testing

- Based on simulations and testing, the Census Bureau determined that data protection techniques used in prior Censuses were no longer sufficient to meet statutory confidentially requirements.
- The Census Bureau performed a reconstruction experiment that correctly identified age, sex, race, and Hispanic ethnicity for an average of 50% of persons in each block;
- The Census Bureau then attempted to match the characteristics to an outside database and only a small number of re-identifications were correct;
- As a result, the Census Bureau concluded that the risk of re-identification is small (Abowd, 2018).

# What is Differential Privacy?

- Differential Privacy (DP) is a mathematical technique that allows for the formal quantification of the risk of data disclosure;
- Formally, DP is a property of algorithms for answering queries. An algorithm is considered differentially-private for a given epsilon ($\varepsilon$) if, for two databases that differ by one record, it satisfies:

$$\Pr[A(D) \in T] \leq \exp(\varepsilon) \Pr[A(D') \in T]$$

- If the algorithm satisfies this definition, the expression provides a bound on how much information can be inferred from adding or deleting a record in the database and prevents learning about a specific record by examining two datasets.

# What is Differential Privacy (con’t)

- As a result, DP allows for mathematically quantifying the risk of identifying a specific element in a dataset;

- Specifically, differentially private algorithms provide formal bounds as to how many queries can be made before the probability of learning specific information about a database increases beyond acceptable levels.

# The Components of Differential Privacy

- The privacy loss budget. The privacy loss budget is typically represented by epsilon ($\varepsilon$).
- When $\varepsilon = 0$, the resulting data would be random and essentially useless (perfect privacy).
- When $\varepsilon = \infty$, the resulting data would allow for full identification of survey participants (perfect accuracy).
- Values of epsilon between $0$ and $\infty$ represent a trade off between privacy and accuracy.

# The Privacy Budget

- An alternative interpretation of epsilon is that of a “privacy budget”.
- If only a single query on the data is expected to be performed, that query might use up the entirety of the budget;
- However, performing a series of queries on the data requires allocation of the budget over all the queries;
- There are two methods of allocating the privacy budget – sequential and parallel.

# Sequential Composition

- Sequential composition is where information from a database is released on an overlapping set of individuals;
- Example – a query to generate the population total for a county and a separate query generating the total by age group for that same county;
- In this case, the total privacy budget is the sum of the privacy budgets for the overlapping queries;
- In other words, the analyst must account for all the operations performed on the data to ensure the global privacy for the dataset.

# Parallel Composition

- Parallel composition is where a series of queries on a database release information on a disjoint set of individuals;
- Example – a query generates the number of persons in all counties in one county while another query returns the number of persons by age category who reside in a second county;
- The total privacy budget would be the maximum of the individual query budgets;

# The Privacy-Accuracy Tradeoff

This graph illustrates the privacy-accuracy trade off for a privacy mechanism with epsilon values between 1 and 6.

# The DP Mechanism

- The DP mechanism works by injecting statistically calibrated ”noise” into the data;
- The amount of noise injected is determined by epsilon and by sensitivity – sensitivity being the amount that one or more individuals (or records) can influence the output of the mechanism;
- Statistical “noise” is typically derived from two distributions:
  - ➢ The Laplace distribution, or the
  - ➢ The Geometric distribution;
- The geometric distribution has the advantage of returning integer values, while the Laplace distribution does not, and so the geometric mechanism has been employed in the Census Bureau’s DP engines.

# Post-Processing

- One important characteristic of DP is that once a dataset has been privatized through a DP algorithm, additional processing on the privatized dataset maintains the differential privacy;
- Therefore, additional data processing can address issues such as:
  - ➢ Counts less than zero;
  - ➢ Ensuring the sum of counts for lower geographies are equal to counts for higher geographies (for example, the sum of the counts for all counties in a state equal the total count for the state).

# Census Bureau and DP

- Early implementation
  - ➢ 2008 – OnTheMap/LEHD
- Post-Secondary Employment Outcomes
  - ➢ Earnings Distributions
- 2020 Census
- Note: the Census Bureau is not planning on implementing DP for the American Community Survey before 2025

# DP and the 2020 Census

- Original test implementation – 1940 Census Dataset
  - Employs top-down methodology;
  - Creates a histogram of demographic attributes (total population, voting age, race/ethnicity, group quarters type);
  - Assigns them iteratively to various geographies (nation, state, county, enumeration district);
  - Applies ‘noise’ to the attributes by adding results from random number generator to the attribute counts;
  - Post-processes the resulting noisy data subject to ‘invariants’ – total population at the state level and total housing unit and group quarters counts at the block level and lower and upper bounds based on housing and population counts.

# DP and 1940 Census Dataset Testing

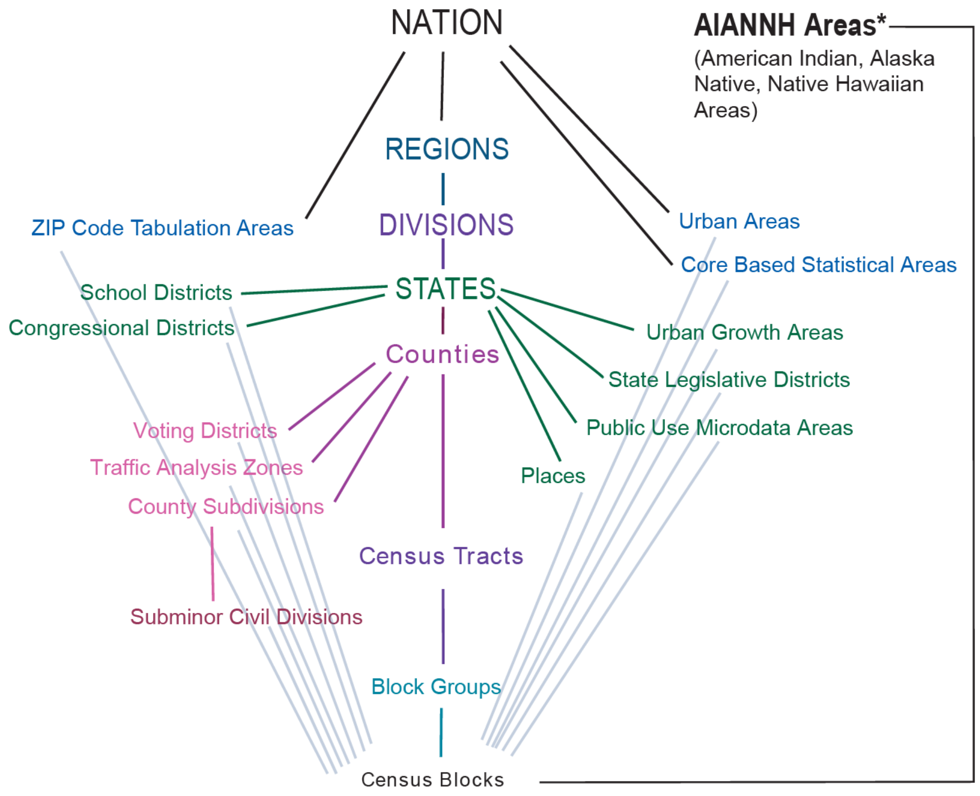
- 1940 Census Dataset
  - ➢ The Census Bureau released the source code (scripted in Python) and the 1940 Census dataset was made available through IPUMS;
  - ➢ The Census Bureau also released a series of DP runs for various epsilon levels (0.25, 0.5, 0.75, 1, 2, 4, and 6);
- Analysis of the results
  - ➢ Low privacy loss budget (epsilon) – 0.25 – resulted in significant distortions in smaller geographic areas and attributes such as race/ethnicity relative to original data;

# DP – 2010 DAS Release

- 2010 Demonstration Data Products Disclosure Avoidance System (DAS) release -
  - ➢ Updated DP applied to the Census Edited File used in the 2010 Census to generate person and housing tables from the PL94 and SF1;
  - ➢ DP process employed a global epsilon of 6.0 – 4.0 allocated to person tables and 2.0 allocated to housing tables;
  - ➢ Geographies expanded to include tract groups, tracts, block groups and blocks;
  - ➢ Tables expanded to include age by groupings by sex and households by race/ethnicity, sex, and presence of persons age 60 plus;

# DP - 2010 DAS Release - Analysis

- Analysis of the resulting tables by the Minnesota Population Center, National Conference of State Legislatures, and others found:
  - ➢ Transfer of population counts from larger geographic areas to smaller geographic areas as a result of invariants and post-processing error;
  - ➢ Significant distortions in demographic categories such as 5-year age groups;
  - ➢ Distortions in population counts for American Indian and Alaska Native Tribal areas, 'off-spline' geographic areas (geographic areas not included in the DAS geographic hierarchy), and small-population areas (such as census blocks);
  - ➢ Distortions in housing statistics (vacant and occupied housing units) and persons per household ratios.



# 2010 Demonstration Files - Issues

- The Census Bureau identified the following issues:
  - Measurement error due to DP noise;
  - Post-processing error from creating internally consistent, non-negative integer counts from noisy measurements;
  - Of those errors, post-processing errors tend to be larger than DP error;

# DAS Errors by County Population Count

- The scatter plots illustrate the error spread ('Noisy' Estimates – Original Estimates) by population size pre-post-processing;
- 'Noisy' estimates were generated using the geometric distribution engine from the 2010 DAS program;
- 'Original' estimates are county population counts drawn from the 2010-2014 American Community Survey (5-year estimates);
- Results – counties with smaller populations have a larger spread of errors then do counties with larger populations.

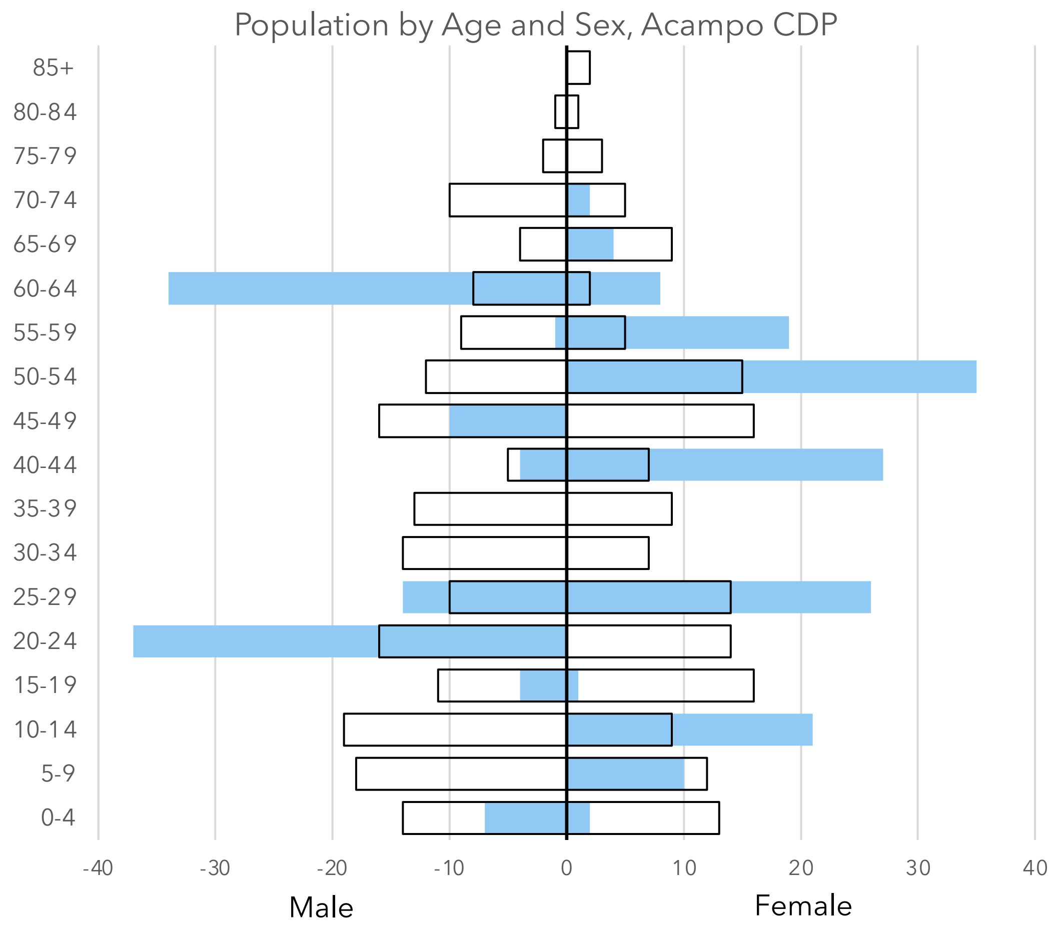
# A Tale of 3 Population Pyramids – Small Population

This pyramid compares the population distribution derived from the 2010 SF1 published data with data derived from the 2010 DAS for Acampo CDP.

2010 Census with DP

Published 2010 Census Data

2010 SF1 Population: 341




| | Absolute Error |
|---|---|
| Under 5 years | 18 |
| 5 to 9 years | 20 |
| 10 to 14 years | 7 |
| 15 to 17 years | 19 |
| 18 and 19 years | 3 |
| 20 years | 4 |
| 21 years | 1 |
| 22 to 24 years | 12 |
| 25 to 29 years | 16 |
| 30 to 34 years | 21 |
| 35 to 39 years | 22 |
| 40 to 44 years | 19 |
| 45 to 49 years | 22 |
| 50 to 54 years | 8 |
| 55 to 59 years | 6 |
| 60 and 61 years | 17 |
| 62 to 64 years | 15 |
| 65 and 66 years | 3 |
| 67 to 69 years | 6 |
| 70 to 74 years | 13 |
| 75 to 79 years | 5 |
| 80 to 84 years | 2 |
| 85 years and over | 2 |

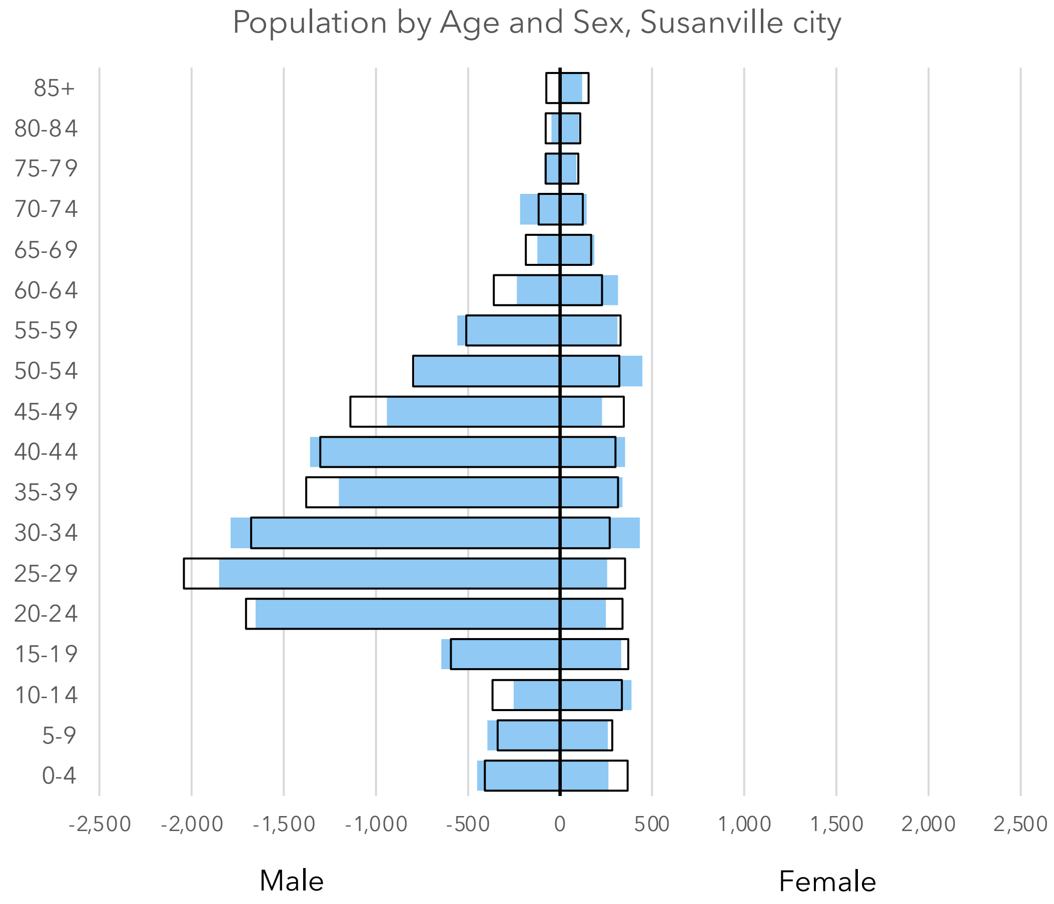
# A Tale of 3 Population Pyramids – Mid Population

This pyramid compares the population distribution derived from the 2010 SF1 published data with data derived from the 2010 DAS for Susanville city.

2010 Census with DP

Published 2010 Census Data

2010 SF1 Population: 17,947




| | Absolute Error |
|---|---|
| Under 5 years | 61 |
| 5 to 9 years | 31 |
| 10 to 14 years | 65 |
| 15 to 17 years | 4 |
| 18 and 19 years | 17 |
| 20 years | 32 |
| 21 years | 48 |
| 22 to 24 years | 125 |
| 25 to 29 years | 285 |
| 30 to 34 years | 276 |
| 35 to 39 years | 154 |
| 40 to 44 years | 108 |
| 45 to 49 years | 316 |
| 50 to 54 years | 120 |
| 55 to 59 years | 35 |
| 60 and 61 years | 66 |
| 62 to 64 years | 104 |
| 65 and 66 years | 28 |
| 67 to 69 years | 15 |
| 70 to 74 years | 121 |
| 75 to 79 years | 7 |
| 80 to 84 years | 38 |
| 85 years and over | 100 |

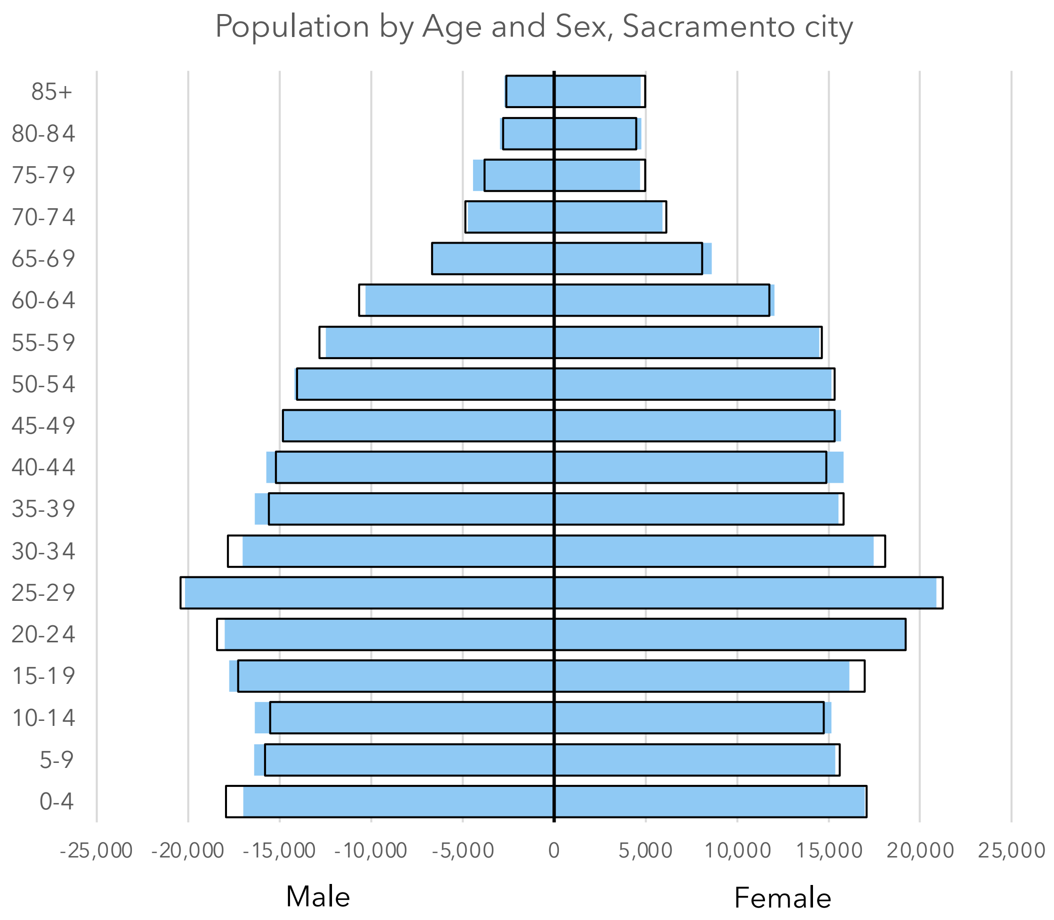
# A Tale of 3 Population Pyramids – Large Population

This pyramid compares the population distribution derived from the 2010 SF1 published data with data derived from the 2010 DAS for Sacramento city.

2010 Census with DP

Published 2010 Census Data

2010 SF1 Population: 466,488



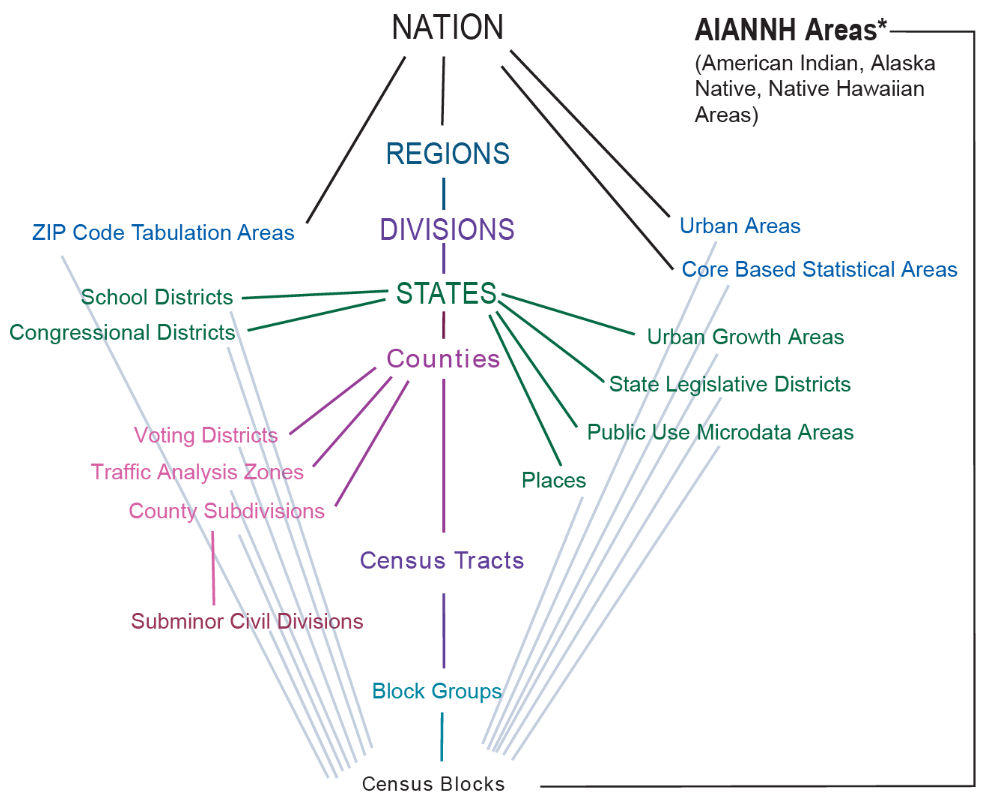

| | Absolute Error |
|---|---|
| Under 5 years | 1,049 |
| 5 to 9 years | 367 |
| 10 to 14 years | 1,259 |
| 15 to 17 years | 447 |
| 18 and 19 years | 115 |
| 20 years | 149 |
| 21 years | 49 |
| 22 to 24 years | 344 |
| 25 to 29 years | 592 |
| 30 to 34 years | 1,436 |
| 35 to 39 years | 487 |
| 40 to 44 years | 1,453 |
| 45 to 49 years | 430 |
| 50 to 54 years | 43 |
| 55 to 59 years | 479 |
| 60 and 61 years | 540 |
| 62 to 64 years | 627 |
| 65 and 66 years | 522 |
| 67 to 69 years | 26 |
| 70 to 74 years | 331 |
| 75 to 79 years | 360 |
| 80 to 84 years | 446 |
| 85 years and over | 140 |

# Census Plan to Improve Data Accuracy

- How Census plans to address these issues:
  - Select a level for epsilon that reduces measurement error while maintaining privacy;
  - Adopt a revised post-processing mechanism –
    - Multi-pass post-processing –
      - First pass: compute total population and GQ populations;
      - Second pass for redistricting file;
      - Third pass for population-estimates program; and
      - Fourth pass: rest of DHC-H and DHC-P.
  - Updated DAS development cycle consisting of 4-week development sprints followed by 2-week evaluation windows;
  - Revised accuracy metrics released to coincide with evaluation windows;

# Demonstration Products – Metrics Tables

- Starting in March 2020, Census began releasing updated metrics designed around use cases and stakeholder feedback;
- The purpose is to allow users/stakeholders to see improvements from changes to the DAS mechanism;
- The metrics will include measures of accuracy, bias, and outliers;
- Census plans to add AIAN and off-spline geographies, and to improve race metrics and outlier measures (see right).



# Demonstration Products – Metrics Tables - Accuracy

- Measures of accuracy.
  - ➢ Accuracy is measured by comparing the post-disclosure protected tabulations to the original, publicly available tabulations from the 2010 Census and the internal pre-disclosure avoidance microdata from the 2010 Census.
- Proposed accuracy measures include –
  - ➢ Mean/Median Absolute Error (MAE);
  - ➢ Mean/Median Numeric Error (ME) ;
  - ➢ Root Mean Squared Error (RMSE);
  - ➢ Mean/Median Absolute Percent Error (MAPE); and
  - ➢ Coefficient of Variation (CV)

# Demonstration Products – Metrics Tables - Bias

- Measures of bias.
  - ➢ Related to accuracy, but bias measures the direction of change and whether it varies with population size or some other characteristic.
- Proposed bias measures include –
  - ➢ Mean/Median Numeric Error (ME); and
  - ➢ Mean/Median Percent Error (MALPE)

# Demonstration Products – Metrics Tables – Examples – Accuracy

- Sample metrics table with measures of accuracy (5/27/2020 compared with the 3/25/2020 release):

| Table 1: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers | | | | | |
|---|---|---|---|---|---|
| Universe: Total population | | | | | |
| Geography: Summary Level 050 - State-County | | | | | |
| | Count of Units (N) | MAE | RMSE | MAPE (%) | CV |
| All counties | 3,143 | 15.95 | 21.15 | 0.14 | 0.02 |
| Counties with total population less than 1,000 | 35 | 13.51 | 17.19 | 2.72 | 2.50 |
| Counties with total population 1,000 to 4,999 | 268 | 14.40 | 19.42 | 0.52 | 0.64 |
| Counties with total population 5,000 to 9,999 | 395 | 15.51 | 20.72 | 0.21 | 0.28 |
| Counties with total population 10,000 to 49,999 | 1,469 | 14.75 | 19.58 | 0.07 | 0.08 |
| Counties with total population 50,000 to 99,999 | 398 | 17.05 | 22.22 | 0.02 | 0.03 |
| Counties with total population of 100,000 or more | 578 | 19.42 | 25.14 | 0.01 | 0.01 |

| Table 1: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers | | | | | |
|---|---|---|---|---|---|
| Universe: Total population | | | | | |
| Geography: Summary Level 050 - State-County | | | | | |
| | Count of Units (N) | MAE | RMSE | MAPE (%) | CV |
| All counties | 3,143 | 82.18 | 141.39 | 0.78 | 0.14 |
| Counties with total population less than 1,000 | 35 | 76.49 | 128.60 | 28.49 | 18.71 |
| Counties with total population 1,000 to 4,999 | 268 | 62.11 | 74.27 | 2.35 | 2.43 |
| Counties with total population 5,000 to 9,999 | 395 | 58.77 | 71.60 | 0.81 | 0.95 |
| Counties with total population 10,000 to 49,999 | 1,469 | 58.53 | 73.59 | 0.29 | 0.29 |
| Counties with total population 50,000 to 99,999 | 398 | 63.99 | 86.08 | 0.09 | 0.12 |
| Counties with total population of 100,000 or more | 578 | 180.45 | 287.70 | 0.07 | 0.07 |

# Demonstration Products – Metrics Tables – Example – Accuracy, Bias, Outliers

- Sample metrics table with measures of accuracy, bias, and outliers (5/27/2020 compared with the 3/25/2020 release):

| Table 1: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers | | | |
|---|---|---|---|
| Universe: Total population | | | |
| Geography: Summary Level 050 - State-County | | | |
| | MALPE (%) | Count of counties where the absolute percent difference is 5% to 10% | Count of counties where the absolute percent difference exceeds 10% |
| All counties | 0.02 | 2 | 2 |
| Counties with total population less than 1,000 | (0.03) | 2 | 2 |
| Counties with total population 1,000 to 4,999 | 0.14 | - | - |
| Counties with total population 5,000 to 9,999 | 0.07 | - | - |
| Counties with total population 10,000 to 49,999 | - | - | - |
| Counties with total population 50,000 to 99,999 | - | - | - |
| Counties with total population of 100,000 or more | - | - | - |

| Table 1: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers | | | |
|---|---|---|---|
| Universe: Total population | | | |
| Geography: Summary Level 050 - State-County | | | |
| | MALPE (%) | Count of counties where the absolute percent difference is 5% to 10% | Count of counties where the absolute percent difference exceeds 10% |
| All counties | 0.69 | 31 | 17 |
| Counties with total population less than 1,000 | 28.35 | 13 | 13 |
| Counties with total population 1,000 to 4,999 | 2.31 | 18 | 4 |
| Counties with total population 5,000 to 9,999 | 0.75 | - | - |
| Counties with total population 10,000 to 49,999 | 0.20 | - | - |
| Counties with total population 50,000 to 99,999 | (0.03) | - | - |
| Counties with total population of 100,000 or more | (0.06) | - | - |

# Questions/Discussion

# Resources – Census Bureau

- Basics of Differential Privacy –
  - Differential Privacy: An Introduction For Statistical Agencies - https://gss.civilservice.gov.uk/wp-content/uploads/2018/12/12-12-18_FINAL_Privitar_Kobbi_Nissim_article.pdf
  - Differential Privacy: A Primer for a Non-technical Audience - http://www.jetlaw.org/wp-content/uploads/2018/12/4_Wood_Final.pdf
- Census Bureau –
  - Disclosure Avoidance and the 2020 Census - https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html
  - 2020 Disclosure Avoidance System Updates - https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html
  - 2020 Census Data Products - https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products.html#par_textimage_153223444
  - 2010 Demonstration Products - https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html
- Github Python repositories -
  - DAS 2010 Demonstration Data Products Disclosure Avoidance System Release - https://github.com/uscensusbureau/census2020-das-2010ddp
  - DAS E2E Release - https://github.com/uscensusbureau/census2020-das-e2e
  - Disclosure Avoidance Repository - https://github.com/uscensusbureau/census-dp

# Resources – Outside Analysis and Data Products

- IPUMS –
  - Changes to Census Bureau Data Products - https://ipums.org/changes-to-census-bureau-data-products
  - Demonstration Data For U.S. Census Bureau Disclosure Avoidance System (1940 Full-Count Dataset) - https://usa.ipums.org/usa/1940CensusDASTestData.shtml
  - Differentially Private 2010 Census Data (2010 DAS data tables in wide and long format by various geographies) - https://www.nhgis.org/differentially-private-2010-census-data

- National Academy of Sciences Committee on National Statistics (CNSTAT) December 11-12 workshop on the 2010 Demonstration Data Products - https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518?#

# Contact Information

- Jonathan Buttle – jonathan.buttle@dof.ca.gov

- California Department of Finance
- Demographic Research Unit
- (916) 323-4086