Disclosure Avoidance for the Demographic and Housing Characteristics File (DHC) and Guidance for Data Users

2020 Census Briefs

By the Population Reference Bureau and
the U.S. Census Bureau's 2020 Census Data Products and Dissemination Team

C2020BR-08

August 2024

This is the sixth in a series of 2020 Census guidance briefs describing disclosure avoidance methods used to protect 2020 Census data products and the implications of those methods for data users. This brief is the first to focus on the 2020 Census Demographic and Housing Characteristics File (DHC) and Demographic Profile. The DHC, a successor to the 2010 Census Summary File 1 (SF1), provides demographic and housing characteristics down to the census block level. The Demographic Profile provides a snapshot of communities, based on selected demographic and housing characteristics, down to the census tract level.[1] This brief describes how disclosure avoidance works and illustrates the potential impacts of confidentiality protections on the 2020 DHC and Demographic Profile.[2]

At the Census Bureau, disclosure avoidance is a process used to protect the confidentiality of respondents' personal information. The Census Bureau has applied disclosure avoidance methods for decades to keep respondents' information confidential and maintain public trust in the data. For the 2020 Census, the Census Bureau employed new disclosure avoidance methods to protect against increasing disclosure risks.

Over time, the Census Bureau has published more detailed data, while advances in data science, more powerful computers, and externally accessible data have increased the risk of identifying individuals from published statistics. With ever-advancing technology, the threats to disclosure are expected to continue growing with time. To reduce this risk, the Census Bureau implemented new disclosure avoidance methods for the 2020 Census based on a framework known as "differential privacy." Previous methods included data swapping and data suppression. However, these methods were deemed to be insufficient given the growing threats to confidentiality protection.

This brief focuses on the disclosure avoidance implications for the DHC and potential use cases. If you're interested in understanding how disclosure avoidance was applied, refer to "Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works."

What Is Differential Privacy?

Differential privacy is a scientific framework for processing data to protect the identities and personal information of the people in the data. It works by adding statistical noise—small, random additions or subtractions—to every published statistic to reduce the likelihood that characteristics about a specific person or household can be accurately inferred using any combination of the published data.

Differential privacy forms the foundation of the Disclosure Avoidance System applied to the data to protect 2020 Census respondent confidentiality.

[1] The U.S. Census Bureau's About 2020 Census Data Products webpage provides a detailed description of all 2020 Census products.

[2] Although the DHC and Demographic Profile are two separate data products, the Demographic Profile is a subset of the DHC. Therefore, throughout this brief, we reference the DHC although the information also applies to the Demographic Profile.

United States® Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

WORKING WITH THE DHC DATA

The 2020 Census is the first census to be protected using differential privacy. This means there are new considerations for data users to understand before they begin analyses. These considerations vary based on the specific 2020 Census data product. For the DHC, the following are some "dos and don'ts" for using these statistics:

- **DON'T** use data for small counts. Instead, aggregate data into larger groups or geographic areas.
  - Census Bureau researchers found that for block groups, a minimum total population between 450 and 499 is sufficient to provide reliable total population and population counts for selected race and Hispanic origin groups. For places and minor civil divisions, total population between 200 and 249 provides reliable total population, race, and Hispanic origin counts. This is the same guidance the Census Bureau provided for the 2020 Census Redistricting Data.[3, 4]
  - Users may need to aggregate more data for certain characteristics that are based on narrower slices of the population such as single year of age, detailed vacancy, and homeownership rates by race of householder.
  - Refer to "Assessing Disclosure Avoidance Uncertainty in the 2020 Census: Determining Reliability Thresholds for Demographic and Housing Characteristics Data" for information about size thresholds and reliability for characteristics other than population by race and Hispanic origin.
- **DO** use demonstration data and summary metrics to evaluate the "fitness" of the DHC statistics for your work. For example, refer to "2020 Census Production Disclosure Avoidance System Detailed Summary Metrics" for information about the typical range of noise at various geographic levels.
- **DO** add and subtract across person tables to calculate counts for other groups. For example, you can subtract data in P18 (Group quarters population by sex by age by major group quarters type) from the data in P12 (Sex by age for selected age categories) to calculate the household population aged 65 and over. Similarly, you can add and subtract across housing tables to calculate other housing counts.
- **DO** use caution when combining data across person and housing tables, which may result in less accuracy than combining data within person or housing tables. In the DHC, as in redistricting data, noise was added to person and housing tables independently, so combining across person and housing tables may result in implausible results. This is by design.
  - **DO** use data from the Supplemental Demographic and Housing Characteristics File (S-DHC) where possible. That product, planned for release in September 2024, will provide official ratios, such as people per household, at the nation and state level.
  - **DO** use caution when calculating averages for small counts. Instead, aggregate data from smaller geographic units or groups into larger ones. Analyses of 2010 demonstration data showed that averages are more reliable for ratios where there are at least 100 households in the denominator.
  - **DO** refer to "Calculating and Interpreting Average Household Size Ratios in the Demographic and Housing Characteristics File" for more information about the various average calculations and the accuracy of different approaches.
- **DO** expect implausible and impossible results for small counts. When comparing data across person and housing tables, the data may not match (e.g., the count of householders or reference person in the person file may not match counts of occupied households in the housing unit file). Improbable results are also expected within the person or housing tables (e.g., blocks with only children under 18 years old and not adults). As data are aggregated to larger geographic areas with greater population counts, the implausible results are less frequent and accuracy increases.
- **DO** remember that some statistics, such as the total number of people in each state and the total number of housing units at the block level, are preserved "as enumerated" without any noise infusion. For example, the total number of people in Alabama is published as reported without noise infusion (i.e., 5,024,279).
  - For a full list of statistics without any noise infusion, reference the "Invariants" section of this brief.

[3] Tommy Wright and Kyle Irimata, "Empirical Study of Two Aspects of The TopDown Algorithm Output for Redistricting: Reliability & Variability (August 5, 2021 Update)," Working paper #2021-02, U.S. Census Bureau, Washington, DC, 2021, <www.census.gov/library/working-papers/2021/adrm/SSS2021-02.html>.

[4] Population Reference Bureau and U.S. Census Bureau, "Disclosure Avoidance and the 2020 Census Redistricting Data, *2020 Census Briefs*, C2020BR-02, Washington, DC, 2023, <www.census.gov/library/publications/2023/decennial/c2020br-02.html>.

- **DO** compare 2020 Census data with prior census data. However, use caution when drawing inferences based on changes observed for very small population groups or geographies, as they will tend to have a higher relative amount of noise compared to larger population groups and geographies. Refer to "Assessing Disclosure Avoidance Uncertainty in the 2020 Census: Determining Reliability Thresholds for Demographic and Housing Characteristics Data." As with every census, data users should also review guidance regarding methodology changes, such as changes to the questionnaires and geographic boundaries, when making comparisons.[5]
- **DO** compare 2020 Census data with estimates from the American Community Survey (ACS). However, data users should keep in mind the differences between the two sources. For example, the ACS is sent to a sample of U.S. addresses each year, while the decennial census attempts to count every person living in the United States**.**
  - For information about comparisons, refer to "Understanding and Using American Community Survey Data: What All Data Users Need to Know."

IMPLICATIONS OF DISCLOSURE AVOIDANCE

When working with any dataset, including data from the decennial census, it is important to consider various sources of error and their effects on the results. Errors may be introduced during data collection—such as erroneous enumerations, omissions, or counting people or households in the wrong block. Statistical errors may also possibly be introduced through data processing, including during the processes of imputing information for nonresponding households or adding statistical noise to protect confidentiality.

For the DHC, the disclosure avoidance system works by adding statistical noise—small, random additions or subtractions—to published statistics so that no one can reidentify a specific person or household with certainty using any combination of the published data. In addition, the system used for the DHC imposes certain consistencies (for example, ensuring that the population totals for counties within a state sum to the state's total population). By doing this, postprocessing also introduces bias into some counts. For more information on how the system works for the DHC, refer to "Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works."

[5] U.S. Census Bureau, "Guidance for Decennial Census of Population and Housing Data Users," <www.census.gov/programs-surveys/decennial-census/guidance.html>.

Noise and Bias

The DHC has more detailed information about population and households than other 2020 Census data products. These details may result in small counts, which have more disclosure avoidance-related error (noise and bias)—on average, relative to their size—than larger counts.

Because the DHC population totals are controlled to the redistricting data, levels of noise and bias for total population counts are the same in both data products. For these counts, Census Bureau researchers found that block-level error introduced by the Disclosure Avoidance System (DAS) are comparable in size to the uncertainty resulting from census operational, measurement, and coverage errors for counts of total population.[6] The average change between enumerated and published block-level total population was around ±5 people. For other DHC data, the amount of error for individual characteristics may vary more (or less) than ±5. This brief provides information on where to find information about the disclosure avoidance-related error in individual tables.

In addition, there is a small amount of bias introduced due to disclosure avoidance. Small counts have a slight positive bias, meaning their published counts are more likely to be the same or larger than the enumerated count. Larger counts can have a negative bias, meaning their published counts are more likely to be the same or smaller than enumerated. For more information on noise and bias, refer to "Disclosure Avoidance and the 2020 Census Redistricting Data" and to "2020 Census Production Disclosure Avoidance System Detailed Summary Metrics."

Implausible and Impossible Results

Similar to the redistricting data, disclosure avoidance results in some implausible and impossible results. Some implausible and impossible results mainly impact the smallest populations and geographies. These types of inconsistencies resolve with data aggregation. Examples include:

- Blocks with resident children under the age of 18 but no adults.
- Only one occupied housing unit but many people (implying an implausibly large household).

[6] U.S. Census Bureau, "Understanding Disclosure Avoidance-Related Variability in the 2020 Census Redistricting Data," 2022, <www.census.gov/library/fact-sheets/2022/variability.html> and William R. Bell and Joseph L. Schafer, "Block-Level Simulation of Non-Sampling Variability in Decennial Census Population Counts," U.S. Census Bureau, Washington, DC, 2021, <www.census.gov/library/working-papers/2021/adrm/CED-WP-2021-007.html>.

- People living in households in an area with only vacant housing units.

Other implausible and impossible results are present given the data structure. The redistricting data and the DHC have two underlying data files—the person file and the housing unit file. The person file contains individual-level characteristics (i.e., sex, age, Hispanic origin, race, relationship to householder, and group quarters), and the housing unit file contains household-level and householder-level characteristics (i.e., tenure, vacancy, household type, family type, and characteristics of the householder). In the 2020 Census, the person and housing unit files were protected by the disclosure avoidance system separately. As a result, implausible and impossible results may exist between the person and the housing unit characteristics. These types of inconsistencies do not resolve with data aggregation. For example, any geography may have data showing:

- The count of householders (i.e., reference person) in the person file may not match counts of occupied households in the housing unit file.
- The count of same-sex spouses in the person file may not equal the count of same-sex married households in the housing unit file.
- There may be more householders of a given race in the housing unit file than people of that race in the person file.
- The count of householders who are Hispanic in the person file may not equal the count of householders who are Hispanic in the housing unit file.
- A summation of the household size data from the housing unit file may be greater than the number of people in the person file.

Invariants

Redistricting data and the DHC also include certain "invariants"—data that are kept exactly as enumerated with no disclosure avoidance noise added. Invariant statistics include the following:

- Total number of people in each state, the District of Columbia, and Puerto Rico.
- Total number of housing units (but not population counts) in each census block and all other geographic levels.
- Number of occupied group quarters facilities (but not population counts) in each census block by major group quarters types:
  - Correctional facilities for adults.
  - Juvenile facilities.
  - Nursing facilities/skilled-nursing facilities.
  - Other institutional facilities.
  - College/university student housing.
  - Military quarters.
  - Other noninstitutional facilities.

TOOLS TO ASSESS FITNESS FOR USE

There are several tools available, and more under development, to help data users get a sense of how much disclosure avoidance-related error (noise and bias) impacts a given count.

2020 Detailed Summary Metrics

The Census Bureau created metrics comparing a subset of the 2020 enumerated data to the corresponding published 2020 statistics to help data users assess the fitness of the DHC data for their applications. The metrics are presented in a spreadsheet that includes a set of tables showing various measures of noise and bias for different characteristics and geographies—such as block groups, tracts, counties, incorporated places, and school districts.

In the summary metrics tables, a data user can look at their geography of interest, as well as the characteristics of the population, and determine how different the reported statistics are from the enumerated counts. These metrics provide guidance to data users about how to interpret their findings of interest.

Each table comes with two measures of error: "mean absolute error" and "mean error." Mean absolute error is the average noise—regardless of whether the noise is positive or negative—and reflects accuracy. A large mean absolute error means that the reported counts are far from the enumerated count. The other metric, mean error, reflects the impact of the bias introduced through the postprocessing. A negative mean error means that reported counts are, on average, lower than enumerated. A positive mean error means that reported counts tend to be higher than enumerated.

For more information, data users can refer to the "2020 Census Production Disclosure Avoidance System Detailed Summary Metrics."

Accuracy Factsheet

Data users provided feedback indicating a need for information to help understand the accuracy of published counts based on the size of the universe (such as the total population or the number of households). To respond to that feedback, the Census Bureau conducted a detailed analysis of 2010 demonstration

data for the DHC to determine how accurate selected counts were by geography and size.

The goal of the analysis was to determine the universe size necessary for the distribution of characteristics for two accuracy levels—to vary by less than 5 percentage points at least 90 percent of the time and to vary by less than 3 percentage points at least 90 percent of the time. The results can help data users determine when the data are large enough that the results are "fit for use."

Results are available in "Assessing Disclosure Avoidance Uncertainty in the 2020 Census: Determining Reliability Thresholds for Demographic and Housing Characteristics Data."

Average Household Size Factsheet

Data user feedback made it clear that people per household ratios—such as the average number of children under 18 years old per household—are important derived estimates for a wide variety of applications. The S-DHC will provide official counts for combined person and housing data, however, it is only available for the nation and states. Because data users may need to construct averages and ratios for lower-level geographies, the Census Bureau created this resource.

The 2010 Demonstration Data for the DHC was used to assess the accuracy of ratio calculation methods. The "Calculating and Interpreting Average Household Size Ratios in the Demographic and Housing Characteristics File Factsheet" summarizes the results of the analysis. It provides a reference for data users to determine the best method for deriving ratios.

In general, the analysis found the amount of error varies by table (i.e., average household size for total population and by race and ethnicity, voting age, and tenure). The analysis provides results for county, tract, place, and American Indian, Alaska Native, and Native Hawaiian areas. One takeaway was having 100 households or more in the universe improves the accuracy of the ratio.

2010 Demonstration Data Products Suite—Redistricting Data and the DHC

In addition to the smaller set of 2020 Detailed Summary Metrics, the Census Bureau created a more extensive Demonstration Data Products Suite (DDPS) based on 2010 Census data to help data users assess the broader expected fitness of the redistricting and the DHC data for their applications. There are two main components of the DDPS—metrics and privacy-protected microdata. Both resources were generated by running 2010 Census data through the same algorithm that was used to apply confidentiality protections for the 2020 Census Redistricting and the DHC data.

2010 Detailed Summary Metrics are presented in a spreadsheet with tables showing various measures of noise and bias for different characteristics and geographies—such as blocks, block groups, tracts, counties, incorporated places, and school districts. The summary metrics include a much longer list of measures than in the 2020 Detailed Summary Metrics, including information about the frequency of implausible results. Because they contain more detailed information than the 2020 metrics—including more measures of noise and bias and more geographic details—these metrics can provide useful guidance to data users about how to interpret their 2020 Census data of interest. For more information about these measures, refer to "Revised Data Metrics for 2020 Disclosure Avoidance."

2010 Privacy-Protected Microdata present the 2010 Census data as rows of individual records (called microdata) rather than the aggregated totals found in a table-based format and allow data users to create custom estimates and tables for comparison with the published microdata from the 2010 Census.

For more information and links to each product, data users can refer to the "Factsheet on Disclosure Avoidance for the 2010 Demonstration Data Products Suite – Redistricting and Demographic and Housing Characteristics File – Production Settings (2023-04-03)."

2020 Noisy Measurement File

Noise may lead to nonsensical results, such as negative population counts, so the Census Bureau added postprocessing steps to correct some of these results and to improve accuracy for lower-level geographic areas. The disclosure avoidance system for the DHC and redistricting data eliminates negative counts but must maintain certain totals, like the total population in a state, so that it equals the sum of the data for substate areas. This process introduces bias in the numbers (for more information, refer to the "Noise and Bias" section in this brief).

Because some researchers may want to work with noisy but unbiased counts, or analyze the uncertainty introduced by differential privacy, the Census Bureau published Noisy Measurement Files—the noisy data before the postprocessing step. While the Noisy Measurement Files are based on 2020 Census data,

they are not the official statistics, however, researchers may be interested in working with these data.

For more information on the Noisy Measurement Files, refer to "Working With Noisy Measurement Files for the Redistricting and DHC Data Products."

Variability Brief: Noise and Bias Estimates

Focusing on redistricting data, the Census Bureau provided a brief summary of noise and bias estimates, by population size category, for the following:

- Total population for blocks.
- White and Black or African American, non-Hispanic population for blocks.
- Total population for incorporated places and minor civil divisions.
- Total population for counties.

For example, among all counties with at least 3,250 people, the average variation in total population counts resulting from confidentiality protections is ±1.8 people, with an approximate 90 percent confidence interval of a loss of 4 people to a gain of 4 people.[7]

For a short overview of the level of noise and bias in some commonly used tables, refer to "Understanding Disclosure Avoidance-Related Variability in the 2020 Census Redistricting Data."

GUIDANCE FOR WORKING WITH THE DHC: EXAMPLES FROM USE CASES

The disclosure avoidance system was developed to protect respondents' information while ensuring the accuracy for the published DHC tables. The Census Bureau recognized that the data are used for a wide variety of purposes, which could not all be known in advance. To guide decisions about disclosure avoidance for the DHC, the Census Bureau considered use cases, input from Census Bureau subject matter experts, and feedback on the demonstration data products. Data users, of course, want to understand the implications of new disclosure avoidance methods.

The section below provides guidance for working with 2020 Census data for some common applications.

School Demographer: Analyzing the School District Population by Age

A school demographer at a larger elementary school district (more than 9,999 people under 18 years old) is preparing a presentation for the school board. They need to share information about the population expected to enroll in kindergarten over the next 2 years. Using the 2020 Detailed Summary Metrics Use Case Table 5.b: Single Years of Age for Population 0 to 17 Years Old for Elementary School District size categories—Mean Absolute Error and Mean Error, the demographer notices that the mean absolute error for 4-year-olds is 10.19 children, with a mean error of 0.26 children, suggesting a small inflation of this population (Table 1).[8] For 3-year-olds, the mean absolute error is 11.64 children, with a mean error of –1.21 children, suggesting that the published count is smaller than what was enumerated. The demographer also knows from Demographic Analysis that children under the age of 5 had a substantial undercount in the 2020 Census, unrelated to disclosure avoidance.[9]

[7] Ibid.

[8] The Excel file containing the 2020 Detailed Summary Metrics is available at <https://www2.census.gov/programs-surveys/decennial/2020/data/demographic-and-housing-characteristics-file/2020-Census-Disclosure-Avoidance-System-Detailed-Summary-Metrics.xlsx>.

[9] U.S. Census Bureau, "Census Bureau Expands Focus on Improving Data for Young Children," 2022, <www.census.gov/library/stories/2022/03/despite-efforts-census-undercount-of-young-children-persists.html>.

Table 1.
Elementary School Districts With More Than 9,999 People Under 18 Years Old: 2020

Age	Count of units	Mean absolute error	Mean error
Under 1 year old	130	10.95	–2.90
1 year old	130	12.84	–1.47
2 years old	130	12.41	–3.74
3 years old	130	11.64	–1.21
4 years old	130	10.19	0.26
5 years old	130	11.12	–0.49
6 years old	130	11.16	–2.58

Note: DHC Use Case Table 5.b: Single Years of Age for Population 0 to 17 Years Old for Elementary School District Size Categories—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

When presenting information about the district population, the demographer shares information about the mean absolute error, mean error, and young child undercount to help the district plan for kindergarten enrollment in the years ahead.

Emergency Response Staff: Identifying Households With Older Adults Living Alone

To identify where to focus evacuation efforts in advance of a hurricane, emergency response staff need to identify census tracts at high flood risk. They may also need to target populations that might be less mobile—such as older adults living alone.

The analysts utilize the 2020 Detailed Summary Metrics unit Table 8.d and find that, on average, the tract-level count of households with one or more people aged 65 and over living alone is not biased (mean error of –0.01 people) (Table 2).[10] However, these counts are noisy, with the average published count differing from the enumerated count by about 6 people (mean absolute error of 6.16). Based on this information, they use 2020 Census data to focus their efforts on tracts with the most households with older adults living alone. They also ensure services are available in all tracts—even those with a published count of zero households with older adults living alone—because, based on the mean absolute error data, they know a tract with a zero count could still have a small number of households with one or more people aged 65 and older living alone.

Emergency Response Staff: Mapping Wildfire Risk for People in Group Quarters

To identify neighborhoods that might require special evacuation during a wildfire, emergency response staff want to produce a map of group quarters populations in their county. While they collaborate closely with a college in their county and know the location and current population in student housing, they rely on decennial census data for locations and counts for smaller group quarters facilities such as nursing facilities, emergency and transitional shelters, and residential treatment centers.

The analysts turn to the 2020 Detailed Summary Metrics Table 18.d and find that, on average, the tract-level count of group quarters population is quite accurate, with nearly all major group quarters types having mean absolute errors below 1 (meaning that the published count is generally within 1 of the enumerated count) (Table 3).[11]

[10] Ibid.

[11] Ibid.

Table 2.
Presence of People 65 Years and Over Living Alone: 2020

Characteristic	Count of units	Mean absolute error	Mean error
Households with 1 or more people 65 years and over living alone	84,414	6.16	–0.01

Note: DHC Use Case Table 8.d: Presence of People 65 Years and Over Living Alone for Tracts—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

Table 3.
Group Quarters Population by Major Institutionalized and Noninstitutionalized Group Quarters Type: 2020

Characteristic	Count of units	Mean absolute error	Mean error
Institutionalized population	84,414	1.51	0.02
Correctional facilities for adults	84,414	0.35	Z
Juvenile facilities	84,414	0.25	Z
Nursing facilities/skilled-nursing facilities	84,414	0.99	Z
Other institutional facilities	84,414	0.11	0.01
Noninstitutionalized population	84,414	2.90	–0.01
College/university student housing	84,414	0.49	Z
Military quarters	84,414	0.06	Z
Other noninstitutional facilities	84,414	2.52	Z

Z Represents or rounds to zero.
Note: DHC Use Case Table 18.d: Group Quarters Population by Major GQ Type and Noninstitutionalized for Tracts—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

In addition, the team knows that group quarters locations were held invariant—meaning no group quarters population for a given major group quarters type (such as juvenile facilities or military quarters) would be assigned to a block that did not have a group quarters facility of that type. For these reasons, they use the data as published.

Emergency Response Staff: Getting a Housing Count

Housing counts are reported as enumerated and have no additional noise added. However, counts of characteristics—such as vacancy, tenure (owner/renter), and household size—have noise added.

For example, a local emergency response manager needs to get an accurate count of the number of houses that need to be notified of evacuation orders before a hurricane. The manager has the evacuation area and housing count data from the DHC. Using the invariants list, the manager knows the count of housing units is not noise-infused, so, the count reflects the number of doors their team needs to knock on to make sure evacuation orders are known.

State Demographer: Analyzing the County Population by Age Group

A demographer is working on producing population projections for the counties in their state. Using the 2020 Detailed Summary Metrics person file Table 16.a, the demographer notices that the mean absolute error for the population under 5 years of age is 9.52, with a mean error of –1.53 children (negative bias) (Table 4).[12] The population aged 5 to 9 is similarly noisy (mean absolute error of 10.64 children and mean error of –1.28 children [negative bias]). In addition, the demographer notes a mean error of 3.6 people (positive bias) for the population aged 15 to 19.

The demographer also knows from Demographic Analysis that—unrelated to disclosure avoidance—young children had a substantial undercount in the 2020 Census and that teens aged 15 to 19 had a net overcount.[13]

Based on the noise—both from disclosure avoidance and other sources of error—the demographer decides to augment their base year data for children and teens by incorporating local administrative data such as birth certificates and school enrollment records before producing county projections. They use counts for other ages as published.

[12] Ibid.
[13] U.S. Census Bureau, "Census Bureau Expands Focus on Improving Data for Young Children," 2022, <www.census.gov/library/stories/2022/03/despite-efforts-census-undercount-of-young-children-persists.html>.

Table 4.
Mean Absolute Error and Mean Error of 5-Year Age Groups: 2020

Characteristic	Count of units	Mean absolute error	Mean error
Under 5 years	3,143	9.52	–1.53
5 to 9 years	3,143	10.64	–1.28
10 to 14 years	3,143	10.63	–0.69
15 to 19 years	3,143	8.31	3.60
20 to 24 years	3,143	4.57	0.16
25 to 29 years	3,143	4.59	0.08
30 to 34 years	3,143	4.39	Z
35 to 39 years	3,143	6.56	–0.28
40 to 44 years	3,143	6.36	0.01
45 to 49 years	3,143	6.25	–0.09
50 to 54 years	3,143	6.22	–0.12
55 to 59 years	3,143	6.17	–0.01
60 to 64 years	3,143	8.08	0.07
65 to 69 years	3,143	7.90	–0.12
70 to 74 years	3,143	6.01	–0.05
75 to 79 years	3,143	5.87	–0.17
80 to 84 years	3,143	5.88	–0.04
85 to 89 years	3,143	5.39	–0.07
90 to 94 years	3,143	5.29	–0.10
95 to 99 years	3,143	4.77	–0.03
100 to 104 years	3,143	3.14	0.04
105 to 109 years	3,143	1.29	0.13
110 to 115 years	3,143	0.64	0.09

Z Represents or rounds to zero.
Note: DHC Use Case Table 16.a: Sex by 5 Year Age Groups for Counties—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

Regional Planner: Measuring Change in the Older Adult Population

Data users can compare 2020 Census data with prior census data. However, they should use caution when drawing inferences based on changes observed for very small population groups or geographies, as they tend to have a higher relative amount of noise compared to larger population groups and geographies—both from disclosure avoidance and from other sources of error. In addition, as with every census, data users should also review guidance regarding methodology changes, such as changes to the questionnaires and geographic boundaries, when making comparisons.[14]

For example, a regional planner is looking at the change in the older adult population in their metropolitan area from 2010 to 2020. Based on the size of the population aged 65 and older in their area, they are confident that the noise in the 2020 Census data is small relative to the count of people aged 65 and over (Table 4). Since the counties included in their metropolitan area have not changed, they are confident in using the data to show differences in the population aged 65 and over during the decade.

[14] U.S. Census Bureau, "Guidance for Decennial Census of Population and Housing Data Users," <www.census.gov/programs-surveys/decennial-census/guidance.html>.

Epidemiologist: Calculating Rates for a Population Group

An epidemiologist is trying to assess whether a public health program has been successful at reducing mortality on a reservation over the past decade. The epidemiologist needs to calculate mortality rates for American Indians and Alaska Natives living on the reservation but knows that there is both disclosure avoidance noise and other error in the 2020 Census data. Using the 2020 Detailed Summary Metrics person file Table 7.g, they find that on federal American Indian Reservations/Off-Reservation Trust Lands the mean absolute error is 5.94 for counts of people who are American Indian and Alaska Native alone or in combination (Table 5).[15]

The epidemiologist also knows that according to the results from the 2020 Post-Enumeration Survey that—unrelated to disclosure avoidance—American Indian and Alaska Native alone or in combination populations living on reservations show a statistically significant undercount rate of 5.64 percent.[16]

The epidemiologist chooses to use the mean error and undercount information to calculate upper and lower bounds on the mortality rate estimates.

[15] The Excel file containing the 2020 Detailed Summary Metrics is available at <https://www2.census.gov/programs-surveys/decennial/2020/data/demographic-and-housing-characteristics-file/2020-Census-Disclosure-Avoidance-System-Detailed-Summary-Metrics.xlsx>.

[16] U.S. Census Bureau, "Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census," 2022, <www.census.gov/newsroom/press-releases/2022/2020-census-estimates-of-undercount-and-overcount.html>.

Table 5.
Race Alone or in Combination With One or More Other Races for Federal American Indian Reservation/Off-Reservation Trust Lands: 2020

Characteristics	Count of units	Mean absolute error	Mean error
White alone or in combination with one or more other races	327	5.06	0.06
Black alone or in combination with one or more other races	327	4.67	2.20
American Indian or Alaska Native alone or in combination with one or more other races	327	5.94	–0.41
Asian alone or in combination with one or more other races	327	4.89	2.28
Native Hawaiian or Pacific Islander alone or in combination with one or more other races	327	4.00	2.66
Some Other Race alone or in combination with one or more other races	327	4.78	1.72

Note: DHC Use Case Table 7.g: Race Alone or in Combination With One or More Other Races for Federal American Indian Reservation/Off-Reservation Trust Lands—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

City Manager: Measuring Vacancy

Housing data, including occupancy and vacancy characteristics, are frequently used by local decision-makers who need accurate data to support plans and policies.

For example, a city manager has been asked to evaluate a new ordinance on short-term rentals. To understand the potential implications of the ordinance, the city manager needs an accurate count of housing units that are only used in certain seasons, on weekends, or for other occasional use throughout the year. They find a count of units for seasonal, recreational, or occasional use in their city in Table H5: Vacancy Status. To determine the accuracy of the count, they refer to "Assessing Disclosure Avoidance Uncertainty in the 2020 Census: Determining Reliability Thresholds for Demographic and Housing Characteristics Data" and find that for places, the share of vacant units that are vacant for seasonal, recreational, or occasional use differs by less than 3 percentage points at least 90 percent of the time when there are at least 225 to 249 vacant units (Table 6). Since their city has more than 1,000 vacant units, they find the data, as published to be reliable for their analysis.

City Planner: Measuring Homeownership

Housing characteristics, including tenure (whether housing units are owned or rented), are frequently used by planners who need accurate data for local areas to develop population, housing, and transportation forecasts, housing plans, and other analyses.

For example, a city planner is updating the housing element of their city's general plan. As part of that process, they need information on homeownership rates for demographic groups within their city. Using the 2020 Detailed Summary Metrics housing unit file Table 3.c, the planner finds that the mean absolute error for Hispanic or Latino owner-occupied households is 2.16 with a very slight negative bias (mean error –0.08 units) (Table 7).[17] The mean absolute error for Hispanic or Latino renter occupied households (needed in the denominator for calculating the rate) is 1.96 units with little bias (mean error –0.01 units). In any mid-sized city or larger, the noise would not be expected to change the ownership rate (rounded to the nearest tenth of a percentage point).

The planner decides to use the Hispanic or Latino homeownership rates as is.

The planner also reviews the data provided in "Assessing Disclosure Avoidance Uncertainty in the 2020 Census: Determining Reliability Thresholds for Demographic and Housing Characteristics Data" and

[17] The Excel file containing the 2020 Detailed Summary Metrics <https://www2.census.gov/programs-surveys/decennial/2020/data/demographic-and-housing-characteristics-file/2020-Census-Disclosure-Avoidance-System-Detailed-Summary-Metrics.xlsx>.

Table 6.
Size Categories at Which 90 Percent of 2010 Census Values Differ by 3 Percentage Points or Less for Selected Characteristic: 2020

Table ID	Numerator	Denominator	Block Groups	Places	Tracts
H5	Housing units for seasonal, recreational, or occasional use	Vacant housing units	450–474	225–249	125–149

Source: U.S. Census Bureau, 2020 Census.

Table 7.
Occupied Housing Units by Hispanic or Latino Origin of Householder for Incorporated Places: 2020

Characteristic	Count of units	Mean absolute error	Mean error
Owner-Occupied			
Householder who is Hispanic or Latino	19,519	2.16	–0.08
Householder who is not Hispanic or Latino	19,519	2.83	0.18
Renter-Occupied			
Householder who is Hispanic or Latino	19,519	1.96	–0.01
Householder who is not Hispanic or Latino	19,519	2.58	0.13

Note: DHC Use Case Table 3.c: Tenure by Hispanic or Latino Origin of Householder for Incorporated Places—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

notices that the share of households owned free and clear by a householder who is American Indian and Alaska Native alone, as a share of American Indian and Alaska Native households, is accurate for places with at least 575 to 599 American Indian and Alaska Native householders (Table 8).

Based on these results, a planner in a city with at least 575 American Indian and Alaska Native households—a threshold met by many large cities—might decide to use the data to calculate and report the ownership rate for that group.

A planner in a city with fewer than 575 American Indian and Alaska Native households would have several choices:

- Report the American Indian and Alaska Native homeownership rate along with a note about noise in the data.
- Group American Indian and Alaska Native households with another group to calculate a more accurate ownership rate, along with a note about why the groups are combined.
- Report the American Indian and Alaska Native homeownership rate for their county instead of their city, along with a note explaining the choice of geography.

They work with their tribal working group to identify the best approach.

Market Researcher: Classifying Counties by Household Size

A market researcher wants to develop a county-level model of purchasing behavior that includes a measure of household size. Using the 2020 Detailed Summary Metrics housing unit file Table 7.b, the researcher finds that for smaller counties (e.g., those with total population 5,000 to 9,999) there is substantial negative bias in the counts for 1-, 2-, and 3-person households (mean errors of –10.52, –17.54, and –8.43 units, respectively) and substantial positive bias in 5-, 6-, and 7-person households (mean errors of 9.82, 12.60, and 12.86 units, respectively) (Table 9).[18] In addition, they

[18] Ibid.

Table 8.
Size Categories at Which 90 Percent of 2010 Census Values Differ by 5 Percentage Points or Less for Selected Characteristic: 2020

Table ID	Numerator	Denominator	Block Groups	Places	Tracts
H4C	Households owned free and clear with an American Indian and Alaska Native householder	American Indian and Alaska Native households	225–249	575–599	200–224

Source: U.S. Census Bureau, 2020 Census.

Table 9.
Household Size for Occupied Housing Units by Size of County: 2020

Characteristic	Count of units	Mean absolute error	Mean error
Counties With Total Population 5,000 to 9,999			
1-person household	413	12.45	–10.52
2-person household	413	20.06	–17.54
3-person household	413	17.40	–8.43
4-person household	413	14.59	0.53
5-person household	413	15.74	9.82
6-person household	413	15.36	12.60
7-or-more-person household	413	15.15	12.86
Counties With Total Population of 100,000 or More			
1-person household	604	37.66	33.09
2-person household	604	87.70	66.43
3-person household	604	65.34	32.95
4-person household	604	45.79	4.73
5-person household	604	50.41	–29.96
6-person household	604	76.83	–53.20
7-or-more-person household	604	83.53	–53.77

Note: DHC Use Case Table 7.b: Household Size for County Size Categories—Mean Absolute Error and Mean Error.
Source: U.S. Census Bureau, Detailed Summary Metrics.

note that the pattern of bias is reversed for the largest counties—with positive bias for smaller household sizes and negative bias for larger household sizes.

The researcher decides that the noise and bias in the estimates for individual size categories are too large for the model they had in mind; so, they choose to use other household characteristics in their model instead.

District Planner: Deriving Data on Average People per Household or Average Household Size

A water district planner needs to calculate the average number of people per household (also referred to as average household size) for service areas—based on groups of tracts—in their district. Calculating average household size requires dividing the count of household population by the count of occupied housing units (also referred to as households).

The planner knows that noise was added independently to person and housing unit files, and dividing household population data from the person file by household data from the housing unit file to calculate average household size may result in implausible or impossible results. In addition, the planner knows that there are at least three ways to calculate the average household size, and that results may differ across methods. The three methods they consider include:

Person tables only: Total population (person) / Number of householders (person) such as using household population data from Table P15 (Population in Households by Age) divided by householder counts from Table P17 (Household Type [Including Living Alone] by Relationship).

Housing unit tables only: Household size (unit) / Total occupied housing units (unit) such as using data on household population in Table H9 (Household Size) divided by occupied unit counts (equivalent to household counts) from Table H3 (Occupancy Status).

A **combination** of person and housing unit tables: Total population (person) / Total occupied housing units (unit) such as using household population data from Table P15 (Population in Households by Age) divided by occupied unit counts from Table H3 (Occupancy Status).

The planner refers to "Calculating and Interpreting Average Household Size Ratios in the Demographic and Housing Characteristics File" and learns that ratios are most accurate when there are at least 100 households in the denominator, which there are in each of the service areas. They also see that for the average people per household ratio for the total population at the tract level, the calculation based on the person tables only is more accurate for more tracts than the unit file only or combination approaches (Table 10).

Using household population data from Table P15 (Population in households by age) and householder counts from Table P17 (Household Type [Including Living Alone] by Relationship) the planner calculates average household size for zones—based on groups of tracts—for their district. If they had wanted data by race, they might have chosen a different ratio calculation.

For more information on the reliability of ratios based on person/person, housing/housing, or person/housing combinations, refer to "Calculating and Interpreting Average Household Size Ratios in the Demographic and Housing Characteristics File."

Table 10.
Accuracy Analysis for the Number of Times Each Ratio Was the Most Accurate—Tracts: 2020

Characteristic	Number of tracts	Person file only	Unit file only	Combination of person and unit files
Total Population	**72,858**	**[1]51,225**	**10,074**	**49,273**
White alone	70,151	23,616	[1]35,042	28,057
Black or African American alone	28,883	9,062	7,885	[1]15,518
American Indian and Alaska Native alone	1,015	281	[1]390	[1]390
Asian alone	11,792	3,473	4,021	[1]5,285
Native Hawaiian and Other Pacific Islander alone	130	37	31	[1]69
Some Other Race alone	15,150	4,561	3,620	[1]7,990
Two or More Races	1,812	153	[1]1,407	283
Hispanic or Latino	30,174	8,333	[1]13,033	10,832
White alone, not Hispanic or Latino	66,997	16,341	[1]44,837	17,347

[1] Ratio type that is most accurate most often.
Source: U.S. Census Bureau, 2020 Census.

Housing Equity Advocate: Choosing Between Household and Householder Counts

A housing equity advocate needs city-level data on the number of households with a householder who is Black or African American alone. The analyst finds similar data in Tables H10 (Tenure by Race of Householder) and PCT17B (Household Type [Including Living Alone] by Relationship [Black or African American Alone]), but notices that the two do not match. The sum of owner and renter data from Table H10 differs from the householder count in Table PCT17B.

Because the advocate needs the number of households, they choose to use the data from Table H10 which provides the official count of households. If the advocate had needed other characteristics of the population living in households, they would have used Table PCT17B.

HOW HAS DATA USER FEEDBACK INFORMED THE PLANNING PROCESS?

The Census Bureau received invaluable feedback on disclosure avoidance from external stakeholders that informed our efforts and decision-making. These came via the 2020 DAS e-mail, <2020DAS@census.gov>, advisory meetings, tribal consultations, and comments provided during presentations at conferences and the Disclosure Avoidance Webinar Series.[19] The Census Bureau and external data users identified several issues with preliminary versions of the DAS that needed additional attention before it could be applied to the 2020 Census data, including:

- Situations where small populations tended to gain population, whereas larger populations tended to lose population such as rural counties with small populations gaining population and large urban counties losing population.
- Limitations of the noise-infused data for emergency planning operations.
- Issues for populations living on American Indian reservations, such as large changes in population counts.
- Problems with the accuracy of census data for certain geographic areas that do not follow the Census Bureau's standard geographic hierarchy, such as school districts.[20]
- Identification of extreme outliers, such as areas with children under the age of 18 but no adult population.
- Distortions in the data that effectively moved individuals from high- to low-density populations, such as from cities to rural areas or from larger race groups to smaller race groups.
- Concern about accuracy for specific demographic and housing characteristics, including children under 18 years old and tenure by race and ethnicity.
- Concern about accuracy for specific geographies, such as counties and school districts.
- Frequency of inconsistencies between the person and housing unit files.

The Census Bureau used this feedback to make improvements to the DAS and to adjust parameter settings to improve overall accuracy for geographic areas and other characteristics, but never to favor a particular demographic group over another. As a result of this work, the Census Bureau was able to greatly reduce, eliminate, or better document many of these limitations.

Data user feedback has also been incorporated in a series of demonstration products to test whether the noise-infused data were fit for use.[21] Advanced data users may download demonstration data that were generated by applying the 2020 DAS to the 2010 Census data.

While not all data user concerns were or could be addressed, the Census Bureau has continued gathering feedback to help inform future products. In addition, the Census Bureau is working on tools that will help data users interpret the accuracy of the data for their specific needs.

One of the challenges many data users have expressed is the difficulty of explaining and understanding disclosure avoidance methods. The Census Bureau is working to provide technical documentation, training, briefs, and tools to the data user community to provide guidance.

[19] To view any webinar in the series, visit <www.census.gov/data/academy/webinars/series/disclosure-avoidance.html>.

[20] Committee on National Statistics, workshop on "2020 Census Data Products: Data Needs and Privacy Considerations," <https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518>.

[21] U.S. Census Bureau, "Developing the DAS: Demonstration Data and Progress Metrics, Detailed Summary Metrics for Production Settings 2021-06-08," <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html>.

WHERE CAN I LEARN MORE?

- Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works
  <www.census.gov/library/publications/2023/decennial/c2020br-04.html>
- Disclosure Avoidance and the 2020 Census Redistricting Data: What Data Users Need to Know
  <www.census.gov/library/publications/2023/decennial/c2020br-02.html>
- Disclosure Avoidance for the 2020 Census: An Introduction
  <www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html>
- Disclosure Avoidance: Latest Frequently Asked Questions
  <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates/2020-das-faqs.html>
- 2020 Decennial Census: Processing the Count: Disclosure Avoidance Modernization
  <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html>
- Disclosure Avoidance Webinar Series
  <www.census.gov/data/academy/webinars/series/disclosure-avoidance.2021.List_882320526.html#list-tab-List_882320526>
- Protecting Privacy in Census Bureau Statistics
  <www.youtube.com/watch?v=1AaoaBcHoss>

You can also subscribe to the Census Bureau's 2020 Census Data Products Newsletter for timely updates and contact us at <2020DAS@census.gov> if you have questions.[22]

[22] U.S. Census Bureau, "Decennial Census: Data Products and Operational Updates," <https://public.govdelivery.com/accounts/USCENSUS/signup/15409>.