Disclosure Avoidance and the Supplemental Demographic and Housing Characteristics File (S-DHC): How PHSafe Works

2020 Census Briefs

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

C2020BR-12

September 2024

INTRODUCTION

This is the seventh in a series of briefs describing disclosure avoidance methods used to protect the confidentiality of respondent information in 2020 Census data products and the implications of those methods for data users. This brief describes how differential privacy works and how it is applied to the 2020 Census Supplemental Demographic and Housing Characteristics File (S-DHC). The methodology used to protect the S-DHC is different than the methodology used in other census data products. This brief also explains those differences and provides guidance for data users.

At the U.S. Census Bureau, disclosure avoidance is defined as a process used to protect the confidentiality of the information provided by our respondents. The Census Bureau has applied disclosure avoidance methods for decades to keep respondents' information confidential and maintain public trust in the data.

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 confidentiality protections 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.

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.

WHAT IS THE S-DHC?

The S-DHC tables provide counts of people living in households and the average household size, by different characteristics (such as age, family type, and tenure). The S-DHC includes eight tables, six of which are repeated (sometimes referred to as "iterated") by race and Hispanic origin. These tables were previously provided in the 2010 Census Summary File 1. The S-DHC tables include:

- PH1. Average Household Size by Age.*
- PH2. Household Type for the Population in Households.

United States® Census Bureau

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

- PH3. Population Under 18 Years by Relationship and Household Type.*
- PH4. Population in families by Age.*
- PH5. Average Family Size by Age.*
- PH6. Family Type and Age for Own Children Under 18 Years.
- PH7. Total Population in Occupied Housing Units by Tenure.*
- PH8. Average Household Size of Occupied Housing Units by Tenure.*

* Tables marked with an asterisk are repeated by the following race and Hispanic origin groups:

A. White alone.
B. Black or African American alone.
C. American Indian and Alaska Native alone.
D. Asian alone.
E. Native Hawaiian and Other Pacific Islander alone.
F. Some Other Race alone.
G. Two or More Races.
H. Hispanic or Latino.
I. White alone, not Hispanic or Latino.

A list of the tables and the data contained in each table is provided in the section "How Accurate Is the S-DHC?"

The S-DHC combines information about households and the people in them. This requires combining information from two data files: the person file and the housing unit file. The person file includes characteristics about individuals such as their age, race, ethnicity, and relationship to the householder. The housing unit file includes characteristics about the household such as household type (e.g., family/nonfamily) and tenure (owner/renter). When the person file and housing unit file are combined, it provides information about all people within the household, including the relationships between household members. For example, it provides the count of people living in married-couple households.

For some other 2020 Census data products, like the Demographic and Housing Characteristics File (DHC), the person and housing unit files are protected separately without enforcement of consistency between the two. The S-DHC tables, however, require information from both files, so they use a different algorithm that allows the person and housing unit data to be joined and protected together.

Combining person and household data has a higher disclosure risk than publishing population or household data alone. This is because information for each person in the household (especially the householder) influences the results for everyone else in the household. This interrelationship makes it much harder to obscure the effect that one person's record has on the others, which in turn makes it harder to guarantee that they are protected.

Geographies in the S-DHC

The S-DHC tables are available for the nation, 50 states, District of Columbia, and Puerto Rico. The S-DHC tables are not available for geographies below the state level. The S-DHC is provided for limited geographies because of concerns about meeting data quality and accuracy standards given confidentiality considerations.

For geographies below the state level, the DHC can be used to calculate average household size by certain characteristics without the associated "credible intervals" (a measure of accuracy described later in this brief). The Census Bureau published guidance for these calculations in the resource, Calculating and Interpreting Average Household Size Ratios in the Demographic and Housing Characteristics File.

WORKING WITH THE DATA

The S-DHC data product is different than some other 2020 Census data products, although the noise infusion is applied similarly to how it was applied for the Detailed DHC-A and Detailed DHC-B. Here are some "do's and don'ts" for using the S-DHC tables:

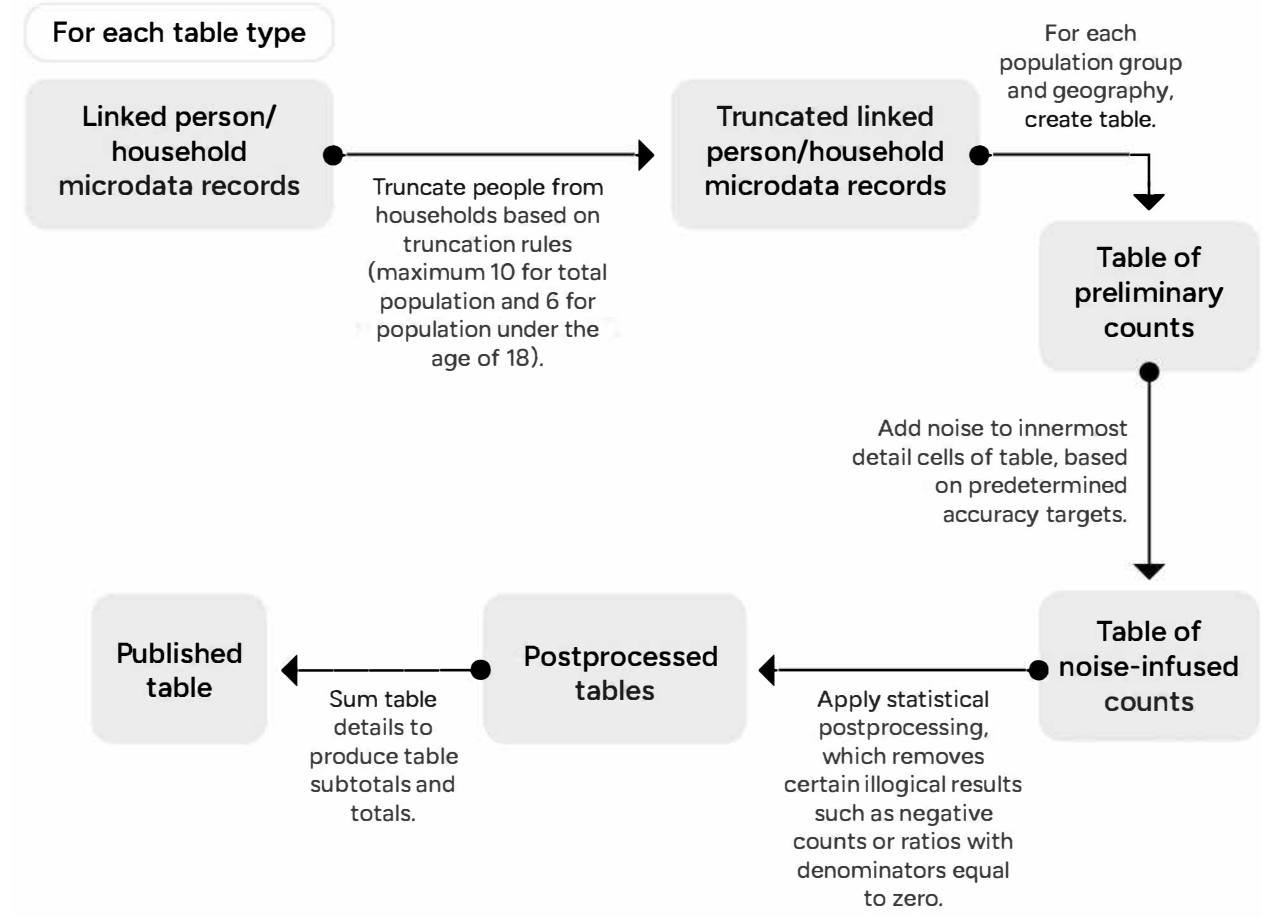
- **DON'T** expect consistency across the S-DHC tables.
  - "Total population in households" (table PH2) should match "total population in occupied housing units" (table PH7), but it often does not. Use the total from table PH7 when possible because it has less noise than the total from table PH2.
  - Average household size usually matches across tables PH1 and PH8 but occasionally may not. Use the average household size with the narrowest credible interval (refer to "What Is a Credible Interval?" text box).
  - The two previous patterns also apply to tables repeated by race and ethnicity, including tables PH2A through PH2I compared to PH7A through PH7I and tables PH1A through PH1I compared to PH8A through PH8I.
- **DON'T** expect consistency when aggregating the S-DHC data and comparing it to other S-DHC data. For example, summing data from table "PH2. Household Type for the Population in Households" across states may not match the national total. Summing race iterations A through G does not sum to the total household population count.

- **DON'T** expect consistency with other 2020 Census data products. For example, "total population in households" (table PH2) does not match "total population in households" from the DHC table "PH5. Population in Households by Age."
- **DON'T** interpret the credible interval (published with the S-DHC data on data.census.gov) as total error or as reflective of all sources of error, such as coverage error and truncation error (described below).
- **DO** understand that the underlying data are adjusted, through a process referred to as truncation, to enforce household size maximum thresholds. For more information on truncation, refer to the section on "How PHSafe and Postprocessing Methods Work."
- **DO** understand that the underlying data have noise added. Noise infusion may result in some implausibly large ratios, such as very large average household size, for some small population groups. For more information on both noise and bias, refer to the section on "Understanding the Accuracy of the S-DHC."
- **DO** use caution when aggregating published counts to produce statistics for custom groups or geographies. Aggregations of the S-DHC data get noisier in proportion to the number of counts that are included in the aggregation.

HOW PHSAFE AND POSTPROCESSING METHODS WORK

PHSafe, the algorithm that protects the S-DHC, works by calculating statistics using the joined confidential 2020 Census Edited File population and household data, applying truncation rules, and adding or subtracting noise to the tabulated results. Then, S-DHC undergoes statistical postprocessing which ensures the data meet a series of logical constraints, such as eliminating negative values that were introduced by noise infusion (Figure 1). The full list of logical constraints is provided in the section "Step 3: Postprocessing."

Figure 1.
Steps in the PHSafe and Statistical Postprocessing Method



Source: U.S. Census Bureau.

Step 1: Household Size Truncation

Because linked person and household data have higher disclosure risk than either person or household data alone, PHSafe adds a disclosure avoidance step that is not applied to other 2020 Census data products, to protect the confidentiality of households' characteristics.

When the count of people in an enumerated household exceeds a set threshold based on the number of people, individuals are removed or "truncated," at random until the household meets that household size threshold. Those thresholds are as follows:

- The threshold is ten people for the following table universes: "population in households" (table PH2), "population in families" (table PH4), and "population in occupied housing units" (table PH7).
- The threshold is six people for tables with the following universes: "population under 18 years in households" (table PH3) and "own children under 18 years" (table PH6).

The truncation process follows a series of steps based on whether the table is for the total household population or for people under 18 years of age. First, each member of the household is assigned an index number at random, and the records are sorted by that index number. For tables based on total household population, each person ranked above ten is removed.[1] For tables based on the population under 18 years of age, people aged 18 or older are excluded from the universe. Then, the remaining records are ranked from lowest to highest based on their index number, and any person ranked above six is removed. If the remaining number of people in the household is less than or equal to the threshold (ten for total household population, six for children under 18 years of age), no people are dropped.

Table 1 illustrates truncation for the threshold of ten people per household. Household A includes three people, which is below the threshold, so no change is made. Household B has 12 people, which exceeds the threshold of ten. In Household B, two people are removed at random.

[1] The ranking includes "Person 1" who is also referred to as "householder." For truncation, that person is ranked independently of their status as householder and may be removed from the data.

Table 1.
Illustration of Truncation for Household Maximum of Ten People

Initial household	Person number in the enumerated household	Action	Person number in the truncated household
Household A	1	No change.	1
Household A	2	No change.	2
Household A	3	No change.	3
Household B	1	No change.	1
Household B	2	No change.	2
Household B	3	Selected at random for removal.	Not included.
Household B	4	No change.	3
Household B	5	No change.	4
Household B	6	No change.	5
Household B	7	Selected at random for removal.	Not included.
Household B	8	No change.	6
Household B	9	No change.	7
Household B	10	No change.	8
Household B	11	No change.	9
Household B	12	No change.	10

Source: U.S. Census Bureau.

Table 2 illustrates truncation for tables where the universe is children under the age of 18, and the threshold of six is applied. Household A includes only one person under the age of 18, so no truncation is applied. Household B has 12 people of any age, nine of whom are under the age of 18. That exceeds the threshold of six. In Household B, three people under the age of 18 are removed at random. Household C also has 12 people of any age, but only two of those people are under the age of 18, so no truncation is applied. However, Households B and C would both be truncated when used to produce total household population tables, since they each exceed the population threshold of ten (not displayed in Table 2).

Table 2.
Illustration of Truncation for Household Maximum of Six People Under the Age of 18

Initial household	Person number in the enumerated household	Age	In universe?	Action	Person number in the truncated household
Household A	1	51	Not in universe.	Exclude from universe.	Not in universe.
Household A	2	49	Not in universe.	Exclude from universe.	Not in universe.
Household A	3	17	In universe.	No change.	1
Household B	1	63	Not in universe.	Exclude from universe.	Not in universe.
Household B	2	40	Not in universe.	Exclude from universe.	Not in universe.
Household B	3	22	Not in universe.	Exclude from universe.	Not in universe.
Household B	4	2	In universe.	No change.	1
Household B	5	5	In universe.	No change.	2
Household B	6	1	In universe.	No change.	3
Household B	7	2	In universe.	Selected at random for removal.	Not included.
Household B	8	5	In universe.	Selected at random for removal.	Not included.
Household B	9	6	In universe.	No change.	4
Household B	10	6	In universe.	Selected at random for removal.	Not included.
Household B	11	9	In universe.	No change.	5
Household B	12	17	In universe.	No change.	6
Household C	1	26	Not in universe.	Exclude from universe.	Not in universe.
Household C	2	22	Not in universe.	Exclude from universe.	Not in universe.
Household C	3	24	Not in universe.	Exclude from universe.	Not in universe.
Household C	4	23	Not in universe.	Exclude from universe.	Not in universe.
Household C	5	21	Not in universe.	Exclude from universe.	Not in universe.
Household C	6	21	Not in universe.	Exclude from universe.	Not in universe.
Household C	7	19	Not in universe.	Exclude from universe.	Not in universe.
Household C	8	19	Not in universe.	Exclude from universe.	Not in universe.
Household C	9	19	Not in universe.	Exclude from universe.	Not in universe.
Household C	10	18	Not in universe.	Exclude from universe.	Not in universe.
Household C	11	17	In universe.	No change.	1
Household C	12	17	In universe.	No change.	2

Source: U.S. Census Bureau.

How Many Households Are Affected by Truncation?

Although truncation removes many people from the data, the relative magnitude of the effect on the published data varies by geography and population group. We cannot provide the exact number of people truncated at the national or state level without incurring additional disclosure risk. As an approximation, however, Tables 3, 4, 5, and 6 provide information that helps to understand the magnitude of truncation.

For total population at the national level, 126,263 (or 0.1 percent) households are truncated for total household population. However, the impact of truncation varies by demographic group. For example, nationwide, 53,264 or 0.3 percent of households with a Hispanic or Latino householder had more than ten people (Table 3).

The impact of truncation also varies by location. For example, in Virginia, 2,202 or 0.1 percent of households had more than ten people, while in Hawaii there were 3,543 or 0.7 percent truncated (Table 4).

To understand the difference between the truncated household population in the S-DHC and the total household population, data users can compare the truncated counts in the S-DHC to untruncated results published in the DHC. This comparison provides an approximation of the number of people truncated because both sets of data underwent noise infusion prior to the comparison.

Table 3.

Number and Percentage of Households That Exceed Truncation Threshold by Race and Hispanic Origin: United States

Race and Hispanic origin	Threshold			
	10 people		6 children	
	Number	Percent	Number	Percent
Total	**126,263**	**0.1**	**111,072**	**0.1**
White alone	43,001	0.1	58,666	0.1
Black or African American alone	15,012	0.1	19,531	0.1
American Indian and Alaska Native alone	4,676	0.4	3,293	0.3
Asian alone	13,079	0.2	3,340	0.1
Native Hawaiian and Other Pacific Islander alone	2,488	1.4	1,345	0.7
Some Other Race alone	30,653	0.4	14,285	0.2
Two or More Races	17,354	0.2	10,612	0.1
Hispanic or Latino	53,264	0.3	26,570	0.2
White alone, not Hispanic or Latino	35,830	Z	53,654	0.1

Z Represents or rounds to zero.

Note: The denominators are derived from the 2020 Census Demographic and Housing Characteristics File (DHC). Public release of the Supplemental Demographic and Housing Characteristics File (S-DHC) counts was authorized by the Census Bureau's Disclosure Review Board (clearance number CBDRB-FY24-0380).

Source: U.S. Census Bureau.

Table 4.

Households That Exceed Truncation Threshold by State

State	Threshold			
	10 people		6 children	
	Number	Percent	Number	Percent
United States[1]	**126,425**	**0.1**	**111,326**	**0.1**
Alabama	980	Z	1,386	0.1
Alaska	591	0.2	680	0.3
Arizona	3,353	0.1	3,543	0.1
Arkansas	726	0.1	953	0.1
California	34,989	0.3	12,365	0.1
Colorado	1,434	0.1	1,388	0.1
Connecticut	451	Z	402	Z
Delaware	236	0.1	227	0.1
District of Columbia	250	0.1	160	0.1
Florida	4,757	0.1	4,203	Z
Georgia	3,042	0.1	3,338	0.1
Hawaii	3,543	0.7	992	0.2
Idaho	587	0.1	1,051	0.2
Illinois	3,655	0.1	3,017	0.1
Indiana	1,653	0.1	2,949	0.1
Iowa	727	0.1	1,281	0.1
Kansas	618	0.1	1,182	0.1
Kentucky	1,063	0.1	1,817	0.1
Louisiana	869	Z	1,421	0.1
Maine	215	Z	310	0.1
Maryland	2,644	0.1	1,554	0.1
Massachusetts	1,170	Z	759	Z
Michigan	2,715	0.1	3,631	0.1
Minnesota	2,239	0.1	3,117	0.1
Mississippi	637	0.1	856	0.1
Missouri	1,529	0.1	2,502	0.1
Montana	372	0.1	545	0.1
Nebraska	489	0.1	822	0.1
Nevada	1,178	0.1	994	0.1
New Hampshire	195	Z	188	Z
New Jersey	3,889	0.1	3,687	0.1
New Mexico	741	0.1	693	0.1
New York	12,760	0.2	11,125	0.1
North Carolina	1,463	Z	2,009	Z
North Dakota	173	0.1	302	0.1
Ohio	2,797	0.1	4,990	0.1
Oklahoma	870	0.1	1,355	0.1
Oregon	1,312	0.1	1,092	0.1
Pennsylvania	3,107	0.1	4,353	0.1
Rhode Island	140	Z	156	Z
South Carolina	817	Z	1,187	0.1
South Dakota	489	0.1	629	0.2
Tennessee	1,589	0.1	2,248	0.1
Texas	10,026	0.1	9,266	0.1
Utah	2,149	0.2	2,622	0.2
Vermont	102	Z	109	Z
Virginia	2,202	0.1	1,929	0.1
Washington	2,954	0.1	2,700	0.1
West Virginia	256	Z	400	0.1
Wisconsin	1,522	0.1	2,586	0.1
Wyoming	160	0.1	255	0.1
Puerto Rico	326	Z	103	Z

Z Represents or rounds to zero.

[1] United States includes the 50 states plus the District of Columbia.

Note: The denominators are derived from the 2020 Census Demographic and Housing Characteristics File (DHC). Public release of the Supplemental Demographic and Housing Characteristics File (S-DHC) counts has been authorized by the Census Bureau's Disclosure Review Board (clearance number CBDRB-FY24-0380).

Source: U.S. Census Bureau.

For example, in the DHC, the U.S. household population under the age of 18 is 72,903,622, while in the S-DHC the number is 72,710,247, a difference of 193,375 (Table 5). While the numeric difference may appear large, the proportional difference is typically less than 1 percent (not provided). Variations are proportionally larger for some population groups, such as Native Hawaiian and Other Pacific Islander alone, where the truncated S-DHC count is 0.6 percent lower than the published DHC count.

Results also vary by state. For example, in the DHC the Florida household population is 21,073,604, while in the S-DHC the number is 21,060,426, a difference of 13,178 (or 01. percent) (Table 6). For South Dakota, the difference is 3,533 (or 0.4 percent).

The steps following truncation use the truncated population as the baseline. In other words, any references to accuracy, such as predetermined accuracy targets and credible intervals—both of which are explained next—apply only to the truncated population, which is smaller than the enumerated population.

Step 2: Noise Infusion

Once the microdata are run through the truncation routine, PHSafe produces preliminary tabulations of characteristics and then adds or subtracts noise to those results to produce a set of preliminary counts referred to as noisy measurements. The amount of noise added or subtracted is determined by a "privacy-loss budget." That budget was set to meet predetermined accuracy targets (referred to as "margins of error") based on the truncated (not enumerated) population in Step 1. The noisy measurements meet those targets at least 90 percent of the time. (The initial target margins of error are listed in Technical Appendix A.)

Table 5.
Difference Between Truncated S-DHC Household Population Count and DHC Household Population Count by Race and Hispanic Origin

Race and Hispanic Origin	Household population (all ages)			Household population (under the age of 18)		
	S-DHC	DHC	Difference	S-DHC	DHC	Difference
Total	**322,873,821**	**323,210,265**	**–336,444**	**72,710,247**	**72,903,622**	**–193,375**
White alone	X	X	X	38,543,300	38,643,656	–100,356
Black or African American alone	X	X	X	10,059,755	10,092,637	–32,882
American Indian and Alaska Native alone	X	X	X	996,771	1,004,125	–7,354
Asian alone	X	X	X	3,994,485	4,000,538	–6,053
Native Hawaiian and Other Pacific Islander alone	X	X	X	192,205	193,367	–1,162
Some Other Race alone	X	X	X	7,943,330	7,968,979	–25,649
Two or More Races	X	X	X	10,977,355	11,000,320	–22,965
Hispanic or Latino	X	X	X	18,661,050	18,709,543	–48,493
White alone, Not Hispanic or Latino	X	X	X	34,418,907	34,508,363	–89,456

X Not applicable. The DHC does not include the count of people living in households with a householder of a given race and ethnicity, so we were unable to provide comparisons to the S-DHC.

Note: Public release of the Supplemental Demographic and Housing Characteristics File (S-DHC) counts was authorized by the Census Bureau's Disclosure Review Board (clearance number CBDRB-FY24-DSEP-003).

Source: U.S. Census Bureau, 2020 Census Demographic and Housing Characteristics File (DHC).

Table 6.

Difference Between Truncated S-DHC Household Population Count and DHC Household Population Count by State

State	Household population (all ages)			Household population (under the age of 18)		
	S-DHC (PH7)	DHC	Difference	S-DHC (PH3)	DHC	Difference
United States[1]	**322,873,821**	**323,210,265**	**–336,444**	**72,710,247**	**72,903,622**	**–193,375**
Alabama	4,893,672	4,896,345	–2,673	1,102,494	1,104,713	–2,219
Alaska	701,275	703,100	–1,825	177,554	178,669	–1,115
Arizona	6,982,914	6,991,233	–8,319	1,598,329	1,605,489	–7,160
Arkansas	2,926,649	2,929,006	–2,357	695,452	697,263	–1,811
California	38,528,800	38,620,291	–91,491	8,662,304	8,683,082	–20,778
Colorado	5,643,231	5,646,866	–3,635	1,259,641	1,261,796	–2,155
Connecticut	3,496,277	3,497,942	–1,665	734,030	734,640	–610
Delaware	966,840	967,203	–363	205,806	206,203	–397
District of Columbia	648,043	648,863	–820	112,672	112,651	21
Florida	21,060,426	21,073,604	–13,178	4,183,787	4,190,081	–6,294
Georgia	10,448,933	10,456,730	–7,797	2,481,641	2,486,762	–5,121
Hawaii	1,403,056	1,414,530	–11,474	295,927	298,405	–2,478
Idaho	1,788,302	1,789,377	–1,075	460,332	461,799	–1,467
Illinois	12,527,574	12,537,143	–9,569	2,803,945	2,809,337	–5,392
Indiana	6,603,292	6,607,154	–3,862	1,584,889	1,589,704	–4,815
Iowa	3,089,229	3,091,811	–2,582	735,632	738,397	–2,765
Kansas	2,849,593	2,850,940	–1,347	704,851	706,817	–1,966
Kentucky	4,377,655	4,380,927	–3,272	1,015,969	1,019,504	–3,535
Louisiana	4,530,674	4,533,634	–2,960	1,084,304	1,085,600	–1,296
Maine	1,325,265	1,325,905	–640	251,246	251,273	–27
Maryland	6,045,148	6,051,719	–6,571	1,356,941	1,359,495	–2,554
Massachusetts	6,780,585	6,784,399	–3,814	1,360,478	1,361,054	–576
Michigan	9,848,681	9,855,615	–6,934	2,150,102	2,156,456	–6,354
Minnesota	5,562,524	5,568,201	–5,677	1,307,781	1,313,665	–5,884
Mississippi	2,866,127	2,867,608	–1,481	680,682	682,039	–1,357
Missouri	5,981,743	5,985,899	–4,156	1,371,054	1,375,625	–4,571
Montana	1,053,419	1,054,921	–1,502	230,756	231,844	–1,088
Nebraska	1,910,792	1,912,002	–1,210	483,166	484,139	–973
Nevada	3,064,590	3,067,660	–3,070	688,889	690,285	–1,396
New Hampshire	1,334,321	1,334,794	–473	256,011	256,014	–3
New Jersey	9,099,303	9,108,619	–9,316	1,997,939	2,005,129	–7,190
New Mexico	2,072,748	2,074,672	–1,924	476,044	477,123	–1,079
New York	19,560,565	19,592,466	–31,901	4,058,641	4,078,290	–19,649
North Carolina	10,148,981	10,153,866	–4,885	2,276,343	2,279,654	–3,311
North Dakota	752,514	752,844	–330	181,485	182,298	–813
Ohio	11,491,913	11,499,820	–7,907	2,578,943	2,587,745	–8,802
Oklahoma	3,839,891	3,842,049	–2,158	943,762	945,914	–2,152
Oregon	4,136,430	4,140,326	–3,896	861,752	863,993	–2,241
Pennsylvania	12,589,984	12,598,604	–8,620	2,634,560	2,642,106	–7,546
Rhode Island	1,050,878	1,051,624	–746	208,629	208,581	48
South Carolina	4,977,626	4,979,822	–2,196	1,097,960	1,100,087	–2,127
South Dakota	850,981	854,514	–3,533	213,352	215,672	–2,320
Tennessee	6,749,537	6,754,418	–4,881	1,519,790	1,523,518	–3,728
Texas	28,514,649	28,539,460	–24,811	7,250,989	7,266,786	–15,797
Utah	3,210,707	3,216,151	–5,444	939,237	944,134	–4,897
Vermont	617,673	618,176	–503	117,832	117,879	–47
Virginia	8,389,042	8,394,747	–5,705	1,878,086	1,881,235	–3,149
Washington	7,536,246	7,545,074	–8,828	1,671,674	1,676,728	–5,054
West Virginia	1,740,874	1,741,649	–775	359,145	359,771	–626
Wisconsin	5,738,264	5,742,419	–4,155	1,274,726	1,279,336	–4,610
Wyoming	563,175	563,523	–348	134,151	134,842	–691
Puerto Rico	3,246,770	3,248,365	–1,595	559,714	560,164	–450

[1] United States includes the 50 states plus the District of Columbia.

Note: Public release of the Supplemental Demographic and Household Characteristics File (S-DHC) counts has been authorized by the Census Bureau's Disclosure Review Board (clearance number CBDRB-FY24-DSEP-0003).

Source: U.S. Census Bureau, 2020 Census Demographic and Housing Characteristics File (DHC).

Noise is added to each of the most detailed characteristics in a table, sometimes referred to as the "innermost cells," in a table individually. The subtotals and totals within each table are not output or aggregated during this step in the processing. For example, in table "PH3. Population Under 18 Years by Relationship and Household Type," noise is added individually to seven "innermost cells," and those are later aggregated to two subtotals ("own child" and "other relatives") and one "total" count (Table 7).[2]

Step 3: Postprocessing

To reduce implausible results, improve accuracy, and provide measures of disclosure avoidance-related uncertainty, the noise-infused results from Step 2 are then run through a series of statistical postprocessing steps.

Statistical postprocessing does not impact confidentiality protections. Once data are protected, the data can be altered in any way without increasing the risk of disclosure—even if that means the counts are closer to the enumerated population.

[2] To reduce the amount of noise introduced into the S-DHC data and to reduce the privacy-loss budget, tables reuse existing noise-infused data where possible. There are two instances where this is possible. First, noise-infused data from table "PH4. Population in Families by Age" provide the numerators for averages in table "PH5. Average Family Size by Age." Second, noise-infused data from table "PH7. Total Population in Occupied Housing Units by Tenure" provide the numerators for averages in table "PH8. Average Household Size of Occupied Housing Units by Tenure."

The postprocessing steps ensure that the following criteria are met:

- Values may not be negative.
- Average household sizes (for ratios that are not age-limited) are at least one, since at least one person must be present in a household.
- Average family sizes (for ratios that are not age-limited) are at least two, since at least two people must be present in a family.
- The published statistic reflects that no more than ten people (six children) could be contributed per household.

The constraints introduced by statistical postprocessing also reduce, but may not fully eliminate, demographically unreasonable averages (such as very large averages).

The statistical postprocessing model starts with the noisy measurements. The model, which uses a Bayesian approach, uses the noisy measurements and the logical constraints (listed above) that would be present in the enumerated data to generate a new set of estimates. The statistical postprocessing is applied independently for the national and state-level tables. Within each table, the postprocessing is applied to the innermost cells, which are then aggregated to produce the corresponding table's subtotals and totals. The new statistically postprocessed estimates are typically more precise (relative to the truncated, enumerated

Table 7.
Noise Infusion Occurs at Innermost Cells Only; Subtotals and Totals Are Calculated Later by Aggregating Across the Detailed Cells

Table element	Detail level
Total:	Total:
Householder, spouse, unmarried partner, or nonrelative	Innermost cell (noise added here)
Own child:	Subtotal:
In married couple family	Innermost cell (noise added here)
In cohabiting couple family	Innermost cell (noise added here)
In male householder, no spouse or partner present family	Innermost cell (noise added here)
In female householder, no spouse or partner present family	Innermost cell (noise added here)
Other relatives:	Subtotal:
Grandchild	Innermost cell (noise added here)
Other relatives	Innermost cell (noise added here)

Source: U.S. Census Bureau.

value) than the noisy measurements themselves. In addition to correcting some of the implausible data, the postprocessing method also generates a measure of accuracy known as a "credible interval."

UNDERSTANDING THE ACCURACY OF THE S-DHC

The Census Bureau set data accuracy and protection targets for the statistics included in the S-DHC. Those targets are reflected in the privacy-loss budget for the S-DHC, which ensures that the noise-infused counts are within (plus or minus) established margins of error (MOEs) of their truncated, enumerated count about 90 percent of the time. The noise-infused counts were then statistically postprocessed to create the S-DHC statistics. These MOEs do not account for the impact of truncation, nor do they reflect the data quality improvements and uncertainty quantification from the statistical postprocessing. Accuracy of the S-DHC statistics, excluding the effects of truncation, is reflected in the published credible intervals.[3]

Setting Accuracy Targets in the Noise Infusion Step

As shown in Technical Appendix A, the MOEs used in the noise infusion step for S-DHC differ by geography, race and ethnicity iteration, and table line. For example, at the state level, in table "PH7H. Total Population in Occupied Housing Units by Tenure (Hispanic or Latino)," the initial noise-infused count of people in renter-occupied households is within ±200 of the truncated, enumerated count at least 90 percent of the time.

In practical terms, this means that the initial noise-infused count is within the target MOE of the enumerated count 90 percent of the time *if all households have ten or fewer people* (or six or fewer children for tables based on population under the age of 18). However, we know that some households exceed the truncation thresholds (refer to the section "How Many Households Are Affected by Truncation").

[3] The Disclosure Avoidance System is not the only source of uncertainty in 2020 Census data. Noise introduced by disclosure avoidance may compound underlying errors or may offset those errors. Examples of these types of errors are available in the 2020 Census Post-Enumeration Survey.

What Is a Credible Interval?

Similar to a margin of error, the credible interval represents a range of values that contain the truncated, enumerated value with 90 percent probability. Credible intervals reflect the noise infused by disclosure avoidance and the impact of statistical postprocessing; they do not provide estimates of expected error introduced by truncation or non-disclosure-avoidance error (such as coverage error). The credible intervals are displayed on data.census.gov alongside the estimates.

For example, for table "PH2. Household Type for the Population in Households" shows that Utah had 2,255,097 people in opposite-sex, married-couple households in 2020, with a credible interval of 2,254,897 to 2,255,298. This means that there is a 90 percent probability that truncated, enumerated value for Utah had between 2,254,897 and 2,255,298 people living in opposite-sex, married-couple households in the 2020 Census.

Reducing Implausible Results and Calculating the Credible Interval

As described above, after the noise-infusion step, the S-DHC data are statistically postprocessed. In addition to correcting some implausible results, the postprocessing method also generates the credible interval for each statistic.

Credible intervals, which are displayed on data.census.gov, represent a range of values that contain the truncated, confidential value with 90 percent probability.

For information about finding S-DHC data including their credible intervals, refer to Chapter 2 of the 2020 Census Supplemental Demographic and Housing Characteristics File (S-DHC) Technical Documentation.

Accuracy of Subtotals and Totals

It is important to remember that totals and subtotals (such as married-couple household, which is the sum of opposite-sex, married couple and same-sex, married couple) represent the sum of the characteristics data. As a result, noise is aggregated from characteristics to subtotals and then to the total household counts. In other words, the sums are noisier than the parts.

For this reason, the total household count from table "PH7. Total Population in Occupied Housing Units by Tenure" is more accurate than the total household count from table "PH2. Household Type for the Population in Households."

Is There Bias in S-DHC?

Although truncation improves confidentiality protection, it introduces bias by removing individuals from larger households. Information about the impact of truncation is described in the section "How Many Households Are Affected by Truncation?" Postprocessing also introduces small upward bias by enforcing non-negativity. The noise-infusion step for S-DHC does not introduce any additional bias.

HOW DOES DISCLOSURE AVOIDANCE FOR S-DHC COMPARE WITH OTHER 2020 CENSUS PRODUCTS?

The 2020 Census Redistricting Data (P.L. 94-171) Summary File and the DHC data are protected through a disclosure avoidance system called the TopDown Algorithm. The Detailed DHC-A and Detailed DHC-B are protected by algorithms called SafeTab-P and SafeTab-H, respectively. However, the data requirements for the S-DHC are different than those for other data products. In particular, S-DHC needed to incorporate person-household joins that were not present in other data products. These differences led the Census Bureau to create a new differential privacy algorithm that was tailored to produce S-DHC data PHSafe. The primary frameworks are compared in Table 8.

One advantage of PHSafe is that it allows person-household joins by providing the count of people in certain types of households. However, the S-DHC is only available for a limited set of geographies, S-DHC tables are not consistent with each other, and the S-DHC is inconsistent with other data products.

For more information on the other algorithms, refer to the following:

- Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works.
- Disclosure Avoidance and the 2020 Census: How SafeTab-P Works.
- Disclosure Avoidance and the 2020 Census: How SafeTab-H Works.

HOW HAS DATA USER FEEDBACK INFORMED THE PLANNING PROCESS?

The Census Bureau received invaluable feedback on disclosure avoidance from external stakeholders that informed our 2020 Census 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.

The Census Bureau also acknowledges data user concerns about the relevancy of the S-DHC, given the late release and limited geographies available. This feedback helps inform planning for the 2030 Census.

WHERE CAN I LEARN MORE?

- Calculating and Interpreting Average Household Size Ratios in the Demographic and Housing Characteristics File

  www.census.gov/library/fact-sheets/2024/dec/calculating-interpreting-avg-hh-size.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.html

For timely updates and to contact us, subscribe to the Census Bureau's 2020 Census Data Products Newsletter at <2020DAS@census.gov> or on the Decennial Census: Data Products and Operational Updates email sign-up page.

For key terms, refer to the "Glossary" in Disclosure Avoidance for the 2020 Census: An Introduction.

Table 8.

Comparison of TopDown Algorithm With the Safe Suite (SafeTab-P, SafeTab-H, and PHSafe)

TopDown	SafeTab-P, SafeTab-H	PHSafe
Data product(s)		
Redistricting Data (P.L. 94-171), Demographic and Housing Characteristics File (DHC), Demographic Profile	Detailed DHC-A, Detailed DHC-B	Supplemental Demographic and Housing Characteristics File (S-DHC)
Methodology		
Algorithm produces privacy-protected microdata.	Algorithm directly produces privacy-protected tabulations (but not microdata).	Algorithm directly produces privacy-protected tabulations (but not microdata).
Does not use adaptive design.[1]	Uses adaptive design to determine the amount of data published.[1]	Does not use adaptive design.[1]
No link between person and housing unit files.	No link between person and housing unit files.	Person and housing unit files are joined.
No truncation of data (but reported household sizes are topcoded at 7+).	No truncation of data.	Household size is truncated.
Postprocessing ensures no negative values.	Negative values are suppressed.	Postprocessing ensures no negative values.
Aggregating data		
All geographies aggregate as expected.	Geographies do not aggregate as expected.	Geographies do not aggregate as expected.
When aggregating data, disclosure avoidance noise generally cancels out and the statistics become more accurate.	When aggregating data, it generally becomes noisier.	When aggregating data, it generally becomes noisier.
Data accuracy		
Results are consistent within data products.	Results are inconsistent within data products.	Results are inconsistent within data products.
Invariant (no noise added) data included 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.	Invariant (no noise added) data included the following: • Total number of housing units (but not population counts) in each census block.	There are no invariant counts included in the published S-DHC statistics. S-DHC is only published for areas with at least one housing unit.
Overall accuracy can be targeted, but the exact levels of accuracy cannot be known in advance.	All margins of error are determined in advance and met 95 percent of the time.	All margins of error for the noisy measurements are determined in advance and met 90 percent of the time for the truncated population.
No credible interval information.	No credible interval information.	Credible intervals are reported for the truncated population.

[1] For information on adaptive design, refer to Disclosure Avoidance Methods for the Detailed Demographic and Housing Characteristics File A (Detailed DHC-A): How SafeTab-P Works and Disclsoure Avoidance and the 2020 Census: How SafeTab-H Works.

Source: U.S. Census Bureau.

TECHNICAL APPENDIX A

As shown in the appendix table, margins of error (MOEs) used in the noise infusion step for S-DHC differ by geography, race and ethnicity iteration, and table line. For example, at the state level, in table "PH7H. Total Population in Occupied Housing Units by Tenure (Hispanic or Latino)," the initial noise-infused count of people in renter-occupied households is within ±200 of the truncated, enumerated count at least 90 percent of the time.

For average tables, the MOEs reflect the accuracy of the numerator and denominator used in calculating the average. For example, in table "PH1. Average Household Size by Age," the numerator and denominator used to calculate the average number of children per household (children under 18 years of age divided by households) for a state are each within ±200 of the truncated, enumerated count at least 90 percent of the time.

While these MOEs are not reflective of the final accuracy of the published estimates, they impact the final accuracy of the data used to create the estimates, and reflect key inputs into the disclosure avoidance framework for S-DHC.

Appendix Table.

Margins of Error (±) by Level of Geography and Characteristic Detail

Table and line	Nation total	State		
		Total	Iterations H-I[1]	Iterations A-G[1]
PH1. Average Household Size by Age				
Margins of error are for numerator (count of people in households by age)				
Total	**705**	**282**	**282**	**96**
Under 18 years	500	200	200	68
18 years and over	500	200	200	68
PH2. Household Type for the Population in Households				
Total	**1,410**	**564**	**X**	**X**
In married-couple household	705	282	X	X
Opposite-sex married couple	500	200	X	X
Same-sex married couple	500	200	X	X
In cohabiting-couple household	705	282	X	X
Opposite-sex cohabiting couple	500	200	X	X
Same-sex cohabiting couple	500	200	X	X
Male householder, no spouse or partner present	705	282	X	X
Living alone	500	200	X	X
Living with others	500	200	X	X
Female householder, no spouse or partner present	705	282	X	X
Living alone	500	200	X	X
Living with others	500	200	X	X
PH3. Population Under 18 Years by Relationship and Household Type				
Total	**1,319**	**528**	**528**	**53**
Householder, spouse, unmarried partner, or nonrelative	500	200	200	20
Own child	1,000	400	400	40
In married-couple family	500	200	200	20
In cohabiting-couple family	500	200	200	20
In male householder, no spouse or partner present family	500	200	200	20
In female householder, no spouse or partner present family	500	200	200	20
Other relatives	705	282	282	29
Grandchild	500	200	200	20
Other relatives	500	200	200	20
PH4. Population in Families by Age				
Total	**705**	**282**	**282**	**96**
Under 18 years	500	200	200	68
18 years and over	500	200	200	68
PH5. Average Family Size by Age				
Margins of error are for numerator (duplicated from PH4)				
Total	**705**	**282**	**282**	**96**
Under 18 years	500	200	200	68
18 years and over	500	200	200	68
PH6. Family Type and Age for Own Children Under 18 Years				
Total	**1,994**	**798**	**X**	**X**
In married-couple families	1,000	400	X	X
Under 4 years	500	200	X	X
4 and 5 years	500	200	X	X
6 to 11 years	500	200	X	X
12 to 17 years	500	200	X	X
In cohabiting-couple families	1,000	400	X	X
Under 4 years	500	200	X	X
4 and 5 years	500	200	X	X
6 to 11 years	500	200	X	X
12 to 17 years	500	200	X	X

Footnotes available at the end of table.

Table 9.
Margins of Error (±) by Level of Geography and Characteristic Detail—Con.

Table and line	Nation total	State		
		Total	Iterations H-I[1]	Iterations A-G[1]
In male householder, no spouse or partner present family	1,000	400	X	X
Under 4 years	500	200	X	X
4 and 5 years	500	200	X	X
6 to 11 years	500	200	X	X
12 to 17 years	500	200	X	X
In female householder, no spouse or partner present family	1,000	400	X	X
Under 4 years	500	200	X	X
4 and 5 years	500	200	X	X
6 to 11 years	500	200	X	X
12 to 17 years	500	200	X	X
PH7. Total Population in Occupied Housing Units by Tenure				
Total	**864**	**346**	**346**	**118**
Owned with a mortgage or loan	500	200	200	68
Owned free and clear	500	200	200	68
Renter-occupied	500	200	200	68
PH8. Average Household Size of Occupied Housing Units by Tenure Margins of error are for denominator (total count of occupied housing units by tenure)				
Total	**705**	**282**	**282**	**96**
Owner-occupied	500	200	200	68
Renter-occupied	500	200	200	68

X Not applicable. Tables PH2 and PH6 are not iterated by major race and ethnicity group.

[1] "Total" represents total household population, "A-G" represents race iterations, and "H-I" represents Hispanic or Latino origin iterations.

Note: Margins of error are rounded. For tables PH1 and PH5, the denominator margins of error are equal to the detailed characteristic line's margins of error in their respective tables. For table PH8, the numerators are derived from table PH7 and do not have independent margins of error.

Source: U.S. Census Bureau.