Assessing Disclosure Avoidance Uncertainty in the 2020 Census: Determining Reliability Thresholds for Demographic and Housing Characteristics Data

This factsheet provides new information regarding expected disclosure avoidance-related uncertainty for the 2020 Census Demographic and Housing Characteristics File (DHC) data by selected characteristics and geographies. This guidance is based on a detailed analysis of the 2010 Demonstration Data Products Suite (2023-04-03), which ran 2010 Census data through the same disclosure avoidance system, with the same settings and parameters, as the production 2020 Census. This demonstration data allows users to assess the performance of the disclosure avoidance system.

MOTIVATION

The U.S. Census Bureau modernized its disclosure avoidance methods for the 2020 Census using a framework based on the principles of differential privacy. The adoption of this new disclosure avoidance system created a new need for data user guidance. For example, to illustrate the changes in methods, to explain the need for these changes, and to give users detailed advice on how to work with data products protected using the new system.

In addition to webinars, blog posts, and public outreach, the Census Bureau in partnership with the Population Reference Bureau developed several 2020 Census briefs specifically about the new disclosure avoidance methods. From the very first brief, data users were encouraged to aggregate data, especially small counts, across geographies or across demographic groups to increase accuracy. For the redistricting data and DHC, aggregation reduces the relative effect of noise from disclosure avoidance on tabulated values and increases the reliability of these counts.

The first brief included specific thresholds to achieve reliability, but this guidance was focused on race and ethnicity characteristics based on total population thresholds for the redistricting (P.L. 94-171) data release.

Census Bureau researchers found that for block groups, "a minimum total population between 450 and 499 is sufficient to provide reliable characteristics of various demographic groups, whereas a minimum total population between 200 and 249 provides reliable characteristics for places and minor civil divisions." Refer to Disclosure Avoidance for the 2020 Census: An Introduction.

A subsequent factsheet provided additional detail on the reliability (or variability) of total population data for census blocks, places/minor civil divisions, and counties, and it directly compared the noise added by the disclosure avoidance system to a simulated total amount of variability resulting from other census operations (coverage error, operational error, etc.). However, this factsheet was limited to analyzing total population variability.

With the release of additional statistics and geographies in the DHC, data users indicated a need for guidance on how to assess the reliability of counts for additional characteristics and geographies. In response to that feedback, the Census Bureau conducted a detailed analysis of 2010 demonstration data for DHC to estimate the anticipated impact of disclosure avoidance uncertainty on the accuracy of the published 2020 Census counts by selected characteristics, geographies, and size categories. The goal of this analysis was to provide specific thresholds when aggregating data to achieve reliability—effectively answering the question "How large must a particular group be in order to draw reliable conclusions about it from the 2020 DHC?"

METHODS

The original research on reliability, documented in the first brief, focused on the redistricting data, particularly on race and ethnicity variables. Its guidance on aggregations considered the reliability of a count by showing how much its relative share shifted with the addition of noise; in the context of the redistricting and Voting Rights Act use cases, a share that changed by more than 5 percentage points was deemed "unreliable." Specifically, the researchers identified an area's largest race and ethnicity group and calculated how much the share of that group changed relative to the area's total

population. For example, Table 2 of their paper shows the block group for Loving County, Texas, having 60 non-Hispanic White people out of a total population of 82 (73.2 percent) in the swapped 2010 data, but 58 out of 85 (68.3 percent) for the noise-infused 2010 demonstration data. This absolute difference of 4.9 percentage points was just under the threshold of 5 percentage points and so the noise-infused demonstration data would be deemed "reliable" for the stated use cases.

The researchers also organized their results by geographic areas and total population size categories and found that more populous areas tended to have smaller changes in relative share for the largest demographic group. To support user interpretation, the size categories were ranked to determine a threshold above which at least 95 percent of areas had a change of 5 percentage points or less. Those population thresholds were 450–499 for block groups and 200–249 for places and minor civil divisions.

The current analysis extends the original research to assess reliability for a variety of characteristics published in the DHC. It goes beyond calculating counts as a share of total population and instead considers a count as a share of a related population called a "universe." Almost all DHC tables have a universe count on their first row. Examples of universes are shown in the "denominator" column of Table 1.

This analysis assesses the universe size necessary for the distribution of characteristics to vary by less than 5 percentage points at least 90 percent of the time. It also assesses the universe size necessary for the share to vary by less than 3 percentage points at least 90 percent of the time. Although the results for shares varying by less than 3 percentage points are more accurate, they require a larger universe size (or more aggregation of geographies and/or populations).

The analysis was conducted for selected characteristics listed in Table 1. Characteristics were chosen to provide an illustrative sample across content, geographies, and the population sizes of numerators and denominators (size of the universes). Motivated data users could replicate this analysis for any particular characteristic of interest using the 2010 Demonstration Data Product

Table 1.
DHC Characteristics and Tables Analyzed

Table ID	Numerator	Denominator
H4C	Housing units owned free and clear with an American Indian and Alaska Native householder.	American Indian and Alaska Native households.
H5	Housing units for seasonal, recreational, or occasional use.	Vacant housing units.
HCT2	Owner-occupied households with own children under age 6 only.	Occupied households.
P12	Males under age 5.	Male population.
P16	Male householder, no spouse present.	Occupied households.
P16	Married-couple households.	Occupied households.
P17	Same-sex unmarried partners.	People in households.
PCT11/custom universe denominator	Grandchildren under age 3.	Children under age 3.
PCT15/P16	Female-female married-couple households.	Occupied households.
PCT8/custom universe denominator	Own children of householder under age 3.	People in households.
PCT19	Males aged 18 to 64 in emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness.	Males aged 18 to 64 in any group quarters.
PCT19	Males aged 18 to 64 in group homes intended for adults.	Males aged 18 to 64 in any group quarters.

Suite and compare the results to the published 2010 tabulations. It should be noted, however, that because the published 2010 Census tables also include uncertainty resulting from the 2010 Census swapping mechanism, those comparisons, just like the analyses included in this factsheet, will likely overestimate the amount of uncertainty resulting from the 2020 Disclosure Avoidance System.

First, for each characteristic, we divided the 2010 Census published count by the 2010 published universe to get the share of the universe with that characteristic. Using table PCT15 and P16 as an example, a hypothetical tract with 3 female-female married-couple households and 100 total households yields a share of 3 percent. (This is also shown as the left three columns of the first row of Table 2.) We completed these percentages for every geography within each of three geographic types (block groups, tracts, and places).

Second, using the 2010 Census demonstration data, we repeated these calculations for each combination of characteristic and geography to calculate the same percentages. Using the same hypothetical tract as in the last paragraph, imagine that the 2010 Census demonstration data shows 2 female-female married-couple households and 99 total households for a share of 2.02 percent. (This is shown in the middle columns of the first row of Table 2.)

Third, we compared the 2010 Census published percentages to the 2010 Census demonstration data percentages and calculated the absolute difference between the two percentages as a measure of accuracy. For the example given in the last two paragraphs (and shown in the rightmost column of the first row in Table 2), the absolute difference in the shares is 0.98 percent.

Fourth, we grouped results for each geographic type into size categories based on the size of the universe and found the 90th percentile difference for each size category. Finally, we compared the 90th percentile difference for each size category to our 5 percentage-point and 3 percentage-point thresholds.

RESULTS

The results are summarized in Table 3 and Table 4 for the 5 and 3 percentage-point differences, respectively. Table 3 shows how large a universe should be for the calculated share to change by fewer than 5 percentage points for 90 percent of geographies. For example, for block groups, the share of males who are under 5 years old differs by less than 5 percentage points once the block group has 75 to 99 males. Block groups with fewer males experience more variability, but block groups with more males have increasing reliability. The share of households that are married-couple households requires 125 to 149 households at the block-group level to attain reliability, but require only 75 to 99 households at the tract level to attain the same degree of reliability. It should be noted that for this characteristic and others, places may require larger universe sizes to ensure reliability because of the way that noise was infused along a geographic spine. More information, is available on page 5 of the Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works.

Table 2.

Calculating Accuracy of 2010 Census Published Data Compared With Noise-Infused Demonstration Data (Hypothetical Example)

Tract	Published data			Demonstration data			Absolute percentage-point difference
	Female-female married-couple households	Households	Female-female married-couple households (share)	Female-female married-couple households	Households	Female-female married-couple households (share)	
1	3	100	3.00%	2	99	2.02%	0.98
2	0	10	0.00%	1	11	9.09%	9.09
3	2	250	0.80%	2	249	0.80%	0.00
4	1	1000	0.10%	2	1003	0.20%	0.10
5	7	200	3.50%	8	199	4.02%	0.52

Source: U.S. Census Bureau, Hypothetical Data.

Certain characteristics meet the reliability threshold even at small counts. For example, the share of households that are female-female married-couple households stays within 5 percentage points even for the least populous block groups, places, and tracts (between 0 and 24 households). Likewise, same-sex unmarried partners as a share of people in households in the demonstration data is always within 5 percentage points of its swapped share. These findings suggest that while counts may differ between the noise-infused data and the unpublished, enumerated counts, the distributions of characteristics are often accurate even for geographies with small populations.

Other characteristics require higher universe counts to bring variability below the 5-percentage point threshold. For example, the share of households that are headed by a male householder with no spouse present needs a universe of at least 175 to 199 households to reach the threshold for tracts. The share of AIAN households who own their house free and clear (without a mortgage) requires 225 to 249 AIAN households at the block group level, 575 to 599 at the place level, and 200 to 249 at the tract level. For these cases, users should consider aggregating geographies to reach this minimum size—for example, users should combine several tracts together until the number of AIAN households reaches the 200 to 249 size. If combining geographies is not feasible, users may instead consider adding demographic groups together to reach these population sizes.

Table 4 repeats the analysis but for a narrower threshold of 3 percentage points. In general, ensuring that a share changes by less than 3 percentage points in 90 percent of areas requires a larger universe than ensuring the share changes by less than 5 percentage points. Again, data users should consider aggregating geographies or demographic groups in order to reach these universe population counts for reliable numbers.

Table 3.
Minimum Universe Populations to Ensure 5 Percentage Point or Less Change in Share

Numerator	Denominator	Universe sizes		
		Block groups	Places	Tracts
Owned free and clear, AIAN householder.	AIAN households.	225–249	575–599	200–224
Housing units for seasonal, recreational, or occasional use.	Vacant housing units.	150–174	125–149	75–99
Owner-occupied with own children under age 6 only.	Households.	50–74	50–74	50–74
Males under age 5.	Males.	75–99	75–99	125–149
Male householder, no spouse present households.	Households.	150–174	125–149	175–199
Married-couple households.	Households.	125–149	225–249	75–99
Same-sex unmarried partners.	People in households.	0–24	0–24	0–24
Grandchildren under age 3.	Children under age 3.	175–199	175–199	175–199
Female-female married-couple households.	Households.	0–24	0–24	0–24
Own children of householder under age 3.	People in households.	75–99	50–74	100–124
Males aged 18–64 In emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness.	Males aged 18–64 in group quarters.	0–24	0–24	75–99
Males aged 18–64 in group homes intended for adults.	Males aged 18–64 in group quarters.	50–74	150–174	100–124

Note: Public release of this data product has been authorized by the Census Bureau's Disclosure Review Board (clearance number CBDRB-FY22-DSEP-004).
Source: U.S. Census Bureau, Calculations from 2010 Demonstration Data Products Suite (v2023-04-03) and 2010 Census Hundred Percent Detail File (HDF).

CONCLUSION

In general, the relative accuracy of a share or proportion depends on both the overall size of the base population or universe and also how broad the distribution is. Proportions based on broad distributions, such as household type among all households or relationship among everyone in households, will tend to be more accurate for a given base population than proportions based on more narrow distributions, such as homeownership among households in a particular race group or type of vacancy among vacant units. Proportions that tend to have a relatively small numerator and relatively large denominator, such as same-sex unmarried partners among everyone in households, may be very accurate even for less populous geographies. Finally, more populous geographies will tend to have more relative accuracy for all measures. Future research will investigate the effects on accuracy and bias of population shares by combining neighboring geographies into aggregate geographies.

Table 4.

Minimum Universe Populations to Ensure 3 Percentage Point or Less Change in Share

Numerator	Denominator	Universe sizes		
		Block groups	Places	Tracts
Owned free and clear, AIAN householder.	AIAN households.	400–424	975–999	575–599
Housing units for seasonal, recreational, or occasional use.	Vacant housing units.	450–474	225–249	125–149
Owner-occupied with own children under age 6 only.	Households.	125–149	125–149	75–99
Males under age 5.	Males.	200–224	200–224	150–174
Male householder, no spouse present households.	Households.	325–349	250–274	350–374
Married-couple households.	Households.	225–249	375–399	175–199
Same-sex unmarried partners.	People in households.	0–24	0–24	50–74
Grandchildren under age 3.	Children under age 3.	475–499	475–499	375–399
Female-female married-couple households.	Households.	0–24	0–24	50–74
Own children of householder under age 3.	People in households.	125–149	100–124	125–149
Males aged 18–64 in emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness.	Males aged 18–64 in group quarters.	100–124	250–274	250–274
Males aged 18–64 in group homes intended for adults.	Males aged 18–64 in group quarters.	75–99	275–299	150–174

Note: Public release of this data product has been authorized by the Census Bureau's Disclosure Review Board (clearance number CBDRB-FY22-DSEP-004).

Source: U.S. Census Bureau, Calculations from 2010 Demonstration Data Products Suite (v2023-04-03) and 2010 Census Hundred Percent Detail File (HDF).