Understanding Disclosure Avoidance-Related Variability in the 2020 Census Redistricting Data

While the U.S. Census Bureau used block group-level statistics for much of the tuning of the Disclosure Avoidance System (DAS) (and for which we have published analyses of the resulting accuracy), we have received requests from several data users to provide more information on the expected variation in block-level data for blocks of varying sizes.

To provide more clarity on the impact of the DAS, we conducted an analysis of expected data variation for different geographies and populations at different sizes. We did this by comparing the most recent demonstration data (vintage 2021-06-08), which uses DAS for confidentiality protection, to the published 2010 Census results, which uses swapping. While this method is not as accurate as a full simulation since it relies on geographic variation and not the intrinsic randomness of the DAS, we feel that it illustrates the general scope and magnitude of the population changes resulting from the DAS.

Block-Level Relative Impact of Confidentiality Protections is Minimal, Especially for More Populous Blocks

Table 1 shows mean absolute error statistics for all blocks with housing units or group quarters (GQ) population in the first row with subsequent blocks grouped by population size. The first column of data shows the count of blocks for this category.

Table 1.

Error Statistics for Total Population for Blocks (Blocks with Housing Units or GQ Population, Excluding Puerto Rico)

Blocks by size	Number of blocks	Mean absolute error (number of people)[1]	Error: middle 90 percent (counts of people)[1]	
			Minus	Plus
All blocks with housing units or GQs	**6,398,202**	**4.89**	**–11**	**+10**
Blocks with total population between 0–249	6,221,561	4.61	–10	+10
Blocks with total population between 250–749	156,251	13.50	–34	+7
Blocks with total population between 750–1,249	15,294	23.37	–53	+3
Blocks with total population between 1,250–1,749	3,515	28.16	–64	+3
Blocks with total population between 1,750–1,949	524	31.08	–69	+3
Blocks with total population between 1,950–2,049	197	30.00	–73	+2
Blocks with total population between 2,050–2,249	265	29.71	–78	+2
Blocks with total population between 2,250–2,749	323	30.34	–74	+4
Blocks with total population between 2,750–3,249	142	28.61	–81	+3
Blocks with total population at or above 3,250	130	22.32	–80	+3

[1] A block's error is calculated by taking the difference between its published, swapped total population from the 2010 Census and the same block's total population after the application of formal privacy protection. The mean absolute error shows the average amount of change (whether positive or negative), while the middle error shows the range of error experienced by the middle 90 percent of blocks.

Source: U.S. Census Bureau, Population Division, calculations from 2010 Demonstration Privacy-Protected Microdata File 2021-06-08.

United States® Census Bureau

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

The second column shows the Mean Absolute Error. For example, we expect the average block that has housing or GQs to gain or lose about five people. Blocks with a total population near 2,000 (between 1,950 and 2,049) had an average gain or loss of about 30 people.

The right-hand column shows a measure of error variability (published 2010 Census tabulations minus demonstration data tabulations). Sorting all the blocks for a particular size category in order, from the block having the largest negative error to the one having the largest positive error, then taking the middle 90 percent (leaving off the top 5 percent and bottom 5 percent), provides an approximate "90 percent confidence interval" for how much error to expect to see in blocks in that size category. For all block sizes, 90 percent of blocks with housing or occupied GQs have between an 11-person loss and a 10-person gain. For blocks close to 2,000 in total population, 90 percent of those have between a 73-person loss and a 2-person gain. Note that while these comparisons address ranges for the 2010 Census data, statisticians at the Census Bureau saw substantially similar results when comparing the impacts of disclosure avoidance on 2010 and 2020 Census data.

As the data in Table 1 show, the average variability in total population counts across all blocks resulting from confidentiality protections is 4.8 people, with an approximate 90 percent confidence interval of a loss of 11 people to a gain of 10. As block population increases, the relative error (mean absolute error as a percent of the block's total population) decreases. For blocks with total populations above 750, the mean absolute errors and the approximate 90 percent confidence interval are mostly consistent. Therefore, the error as a share of the overall population declines as the total population rises. For blocks near 2,000 total population, the mean absolute error is 30.0 people (which is about 1.5 percent of the overall population of 2,000), and 90 percent of the blocks are between a loss of 73 people (–3.7 percent of the overall population of 2,000) and a gain of 2 (~0.1 percent).

Table 2.

Error Statistics for White and Black or African American, Non-Hispanic Population for Blocks (Blocks with Housing Units or GQ population, Excluding Puerto Rico)

Blocks by size	Number of blocks	Mean absolute error (counts of people)[1]	Error: middle 90 percent (counts of people)[1]	
			Minus	Plus
All blocks with housing units or GQs	**6,398,202**	**0.31**	**–1**	**+1**
Blocks with total population between 0–249	6,221,561	0.28	–1	+1
Blocks with total population between 250–749	156,251	1.46	–5	+2
Blocks with total population between 750–1,249	15,294	2.52	–7	+2
Blocks with total population between 1,250–1,749	3,515	2.91	–8	+2
Blocks with total population between 1,750–1,949	524	2.83	–7	+3
Blocks with total population between 1,950–2,049	197	2.96	–8	+3
Blocks with total population between 2,050–2,249	265	3.01	–10	+2
Blocks with total population between 2,250–2,749	323	3.04	–9	+3
Blocks with total population between 2,750–3,249	142	2.66	–10	+2
Blocks with total population at or above 3,250	130	2.03	–6	+4

[1] A block's error is calculated by taking the difference between its published, swapped total population from the 2010 Census and the same block's total population after the application of formal privacy protection. The mean absolute error shows the average amount of change (whether positive or negative), while the middle error shows the range of error experienced by the middle 90 percent of blocks.

Source: U.S. Census Bureau, Population Division, calculations from 2010 Demonstration Privacy-Protected Microdata File 2021-06-08.

While the approximate confidence intervals for larger blocks do tend to skew negative (for algorithmic reasons previously discussed[1]), the relative impact of these errors on the larger underlying populations of these blocks is minimal.

We see similar results for data on small demographic subgroups. Table 2 presents the same statistics for block-level counts of the non-Hispanic White *and* Black population, which is a relatively small two-race population.

As before, for blocks above 750 in total population, the mean absolute error and 90 percent confidence interval stabilizes, and the error as a share of the total population continue to decline as population size increases.

Algorithmic Tuning Controlled Variability Because of Confidentiality Protections for Legal and Political Geographies

The production settings for the TopDown Algorithm were specifically tuned to control disclosure avoidance variability for legal and political geographies such as places, minor civil divisions (MCDs), and counties. Table 3 (states) and Table 4 (counties) show similar approximated 90 percent confidence intervals for total population counts for MCDs in strong MCD states and places in weak MCD states.

Tables 3 and 4 demonstrate that the variability because of confidentiality protections is controlled as individual blocks are aggregated into larger geographies of interest. On average, these larger geographies see much lower error relative to their population size (±4 people for counties, ±6 people for places/MCDs).

How Do I Use These Data in My Work?

As explained in more detail above, the numbers in the far-right column of the charts represent the range of differences between the 2010 Census published data and the 2010 Census data with disclosure avoidance applied. Again, while these comparisons address ranges for the 2010 Census data, the results are substantially similar when comparing the impacts of disclosure avoidance on 2010 and 2020 Census data. Also, note that 2010 Census published data had swapping applied as the disclosure avoidance method.

[1] The newsletter, "Post-processing, Consistency, and the Challenge of Negative Numbers" can be found at <https://content.govdelivery.com/accounts/USCENSUS/bulletins/2924168>.

Table 3.

Error Statistics for Total Population for Places and Minor Civil Divisions (MCDs) (Excluding Puerto Rico)

Places and MCDs by size	Number of places/MCDs	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All places/MCDs	**34,495**	**2.73**	**–6**	**+6**
Places/MCDs with total population between 0–249	6,637	1.57	–3	+3
Places/MCDs with total population between 250–749	7,509	1.91	–4	+4
Places/MCDs with total population between 750–1,249	4,017	2.40	–5	+5
Places/MCDs with total population between 1,250–1,749	2,507	2.61	–5	+5
Places/MCDs with total population between 1,750–1,949	763	2.57	–4	+6
Places/MCDs with total population between 1,950–2,049	360	2.75	–6	+5
Places/MCDs with total population between 2,050–2,249	646	2.88	–6	+6
Places/MCDs with total population between 2,250–2,749	1,332	2.77	–5	+6
Places/MCDs with total population between 2,750–3,249	996	2.69	–5	+6
Places/MCDs with total population at or above 3,250	9,728	4.33	–10	+9

Source: U.S. Census Bureau, Population Division, calculations from 2010 Demonstration Privacy-Protected Microdata File

Table 4.

Error Statistics for Total Population for Counties (Excluding Puerto Rico)

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**1.75**	**–4**	**+4**
Counties with total population between 0–249	2	2.00	–1	+3
Counties with total population between 250–749	19	1.32	–2	+2
Counties with total population between 750–1,249	26	1.38	–2	+4
Counties with total population between 1,250–1,749	24	1.00	–2	+3
Counties with total population between 1,750–1,949	14	1.14	–1	+2
Counties with total population between 1,950–2,049	10	1.50	–1	+5
Counties with total population between 2,050–2,249	16	0.88	–1	+1
Counties with total population between 2,250–2,749	35	1.31	–2	+3
Counties with total population between 2,750–3,249	38	1.29	–2	+3
Counties with total population at or above 3,250	2,959	1.79	–4	+4

Source: U.S. Census Bureau, Population Division, calculations from 2010 Demonstration Privacy-Protected Microdata File 2021-06-08.

Statisticians will use this information in many ways, but the measure of variability ranges can be useful for everyday users of the data as well. One common use might be to assess the likely source of an unexpected result in local census data from 2020. If you suspect a geographic area has a discrepancy significantly larger than those shown in the tables below, disclosure avoidance is not likely to be the source of the error. More information can be found in "What to Consider if You Find an Unexpected Census Result" to help you identify other potential explanations for an unexpected result.[2]

The Impact of Confidentiality Protections is Minimal Compared to Variability Because of Typical Operational and Coverage Error

It should be noted that variability in published census statistics resulting from confidentiality protections is just one component of the overall uncertainty in these data.[3]

Although we undertake extensive efforts to accurately count everyone in the decennial census, sometimes people are missed or duplicated. Census errors can result in a smaller or larger population count than the actual number of people. The post-enumeration survey[4] is one of the many ways we estimate the quality of the census. For example, we also compare census counts to other population benchmarks as described in our recent "Using Demographic Benchmarks to Help Evaluate 2020 Census Results" blog.[5] More information can be found on the 2020 Census Data Quality page.[6]

Two sources of error common to any census or survey are operational sources of error (e.g., reporting error by respondents or census takers, also called "nonsampling variability") and coverage error (e.g., omissions and erroneous enumerations), which should also

[2] More information on "What to Consider if You Find an Unexpected Result" can be found at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/release/data-expectations.html>.

[3] More information from "2020 Census Data Quality" can be found at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html>.

[4] More information can be found in the blog "2020 Census Post-Enumeration Survey" at <www.census.gov/newsroom/blogs/random-samplings/2021/12/post-enumeration-measuring-coverage-error.html>.

[5] More information can be found in the blog "Using Demographic Benchmarks to Help Evaluate 2020 Census Results" at <www.census.gov/newsroom/blogs/random-samplings/2021/11/demographic-benchmarks-2020-census.html>.

[6] More information can be found from the "2020 Census Data Quality" webpage at <www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/data-quality.html#evaluating>.

be considered when evaluating the resulting statistics.

While not measurable directly, we can estimate the impact of coverage error and operational error on 2010 Census counts using simulations based on available operational metrics and the post-enumeration survey results from the 2010 Census (called the Census Coverage Measurement Studies for the 2010 Census,[7] and the Post-Enumeration Survey [PES] for the 2020 Census[8]). As with all simulations, the resulting estimates are dependent on the underlying assumptions of the simulations. Tables 5 and 6 report the estimated 90 percent confidence interval for errors in total population counts for counties from two different types of simulations. Technical details of these simulations, including their underlying assumptions and methodologies, as well as additional results, are available on census.gov.[9]

The estimated errors in population counts presented in Table 5 are the results from a set of simulations seeking to estimate the inherent variability in overall census operations to determine how different the census counts might be if the same data collection and processing methods were repeatedly applied to the same fixed population. This simulation used conservative assumptions about known census errors from the 2010 Census to estimate the overall variability one would expect to conclude if the same census were conducted repeatedly using the exact same methodologies. Table 5 shows that county total population counts can be expected to vary by a loss of 248 people to a gain of 230 because of the inherent nonsampling variability associated with conducting a census (90 percent confidence interval).

[7] More information on the 2010 Census Coverage Measurement Studies can be found at <www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.2010.html>.

[8] More information can be found in the blog "2020 Census Post-Enumeration Survey" at <www.census.gov/newsroom/blogs/random-samplings/2021/12/post-enumeration-measuring-coverage-error.html>.

[9] The Technical Paper, "Simulation Studies to Investigate Variation in Census Counts and in Census Coverage Error Using 2010 SF-1 Data and 2010 CCM Results" can be found at <https://www2.census.gov/adrm/CED/Papers/CY22/2022-01-simulation-studies.pdf>.

Table 5.

Simulated Error in Population in Households for Counties (Excluding Puerto Rico) From Nonsampling Variability

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**117.27**	**–248**	**+230**
Counties with housing unit population between 0–999	37	10.03	–10	+27
Counties with housing unit population between 1,000–9,999	691	28.23	–38	+71
Counties with housing unit population between 10,000–99,999	1,849	74.45	–131	+177
Counties with housing unit population between 100,000–999,999	527	292.21	–784	+545
Counties with housing unit population at or above 1,000,000	39	1,463.12	–3,659	+1,351

Source: U.S. Census Bureau, Research and Methodology Directorate, simulations based on 2010 Census Coverage Measurement research and 2010 Census housing unit population.

Table 6.

Simulated Error in Population in Households for Counties (Excluding Puerto Rico) From Census Coverage Error

Counties by size	Number of counties	Mean absolute error (counts of people)	Error: middle 90 percent (counts of people)	
			Minus	Plus
All counties	**3,143**	**964.00**	**–1,841**	**+2,048**
Counties with housing unit population between 0–999	37	23.00	–22	+54
Counties with housing unit population between 1,000–9,999	691	121.00	–146	+284
Counties with housing unit population between 10,000–99,999	1,849	446.00	–832	+1,053
Counties with housing unit population between 100,000–999,999	527	2,930.00	–7,222	+6,278
Counties with housing unit population at or above 1,000,000	39	14,848.00	–44,833	+20,007

Source: U.S. Census Bureau, Research and Methodology Directorate, simulations based on 2010 Census Coverage Measurement research and 2010 Census housing unit population.

Of the known sources of error in census statistics, coverage error is often the most significant. Table 6 presents the results of a different set of simulations seeking to estimate the variability in county population counts because of coverage error alone.

Using an alternative model and assumptions about the potential variability of coverage error than those included in the simulations presented in Table 5, the results of the simulations reported in Table 6 show that county-level population counts reflect substantial variability because of coverage error. This simulation estimates that population counts for the average county can be expected to vary by a loss of 1,841 people to a gain of 2,048 as a result of census coverage error as measured in the 2010 Census Coverage Measurement Program and reflected in this simulation (90 percent confidence interval).

Conclusion

Overall, as can be seen in the tables above, confidentiality protections do introduce variability in the published census statistics, and that variability is, by design, most pronounced at the block level where the risk of reidentifying individual respondents is greatest. As individual blocks are aggregated into larger geographies of interest (MCDs, places, and counties), the relative variability due to confidentiality protections decreases substantially. For more populous geographic units (e.g., counties), the variability due to confidentiality protections is negligible compared with the larger variability in census counts because of operational and coverage error.

This research is part of comprehensive efforts after each census to better understand sources of variability—both their scope and potential impacts—so we can design the next census in a way that reduces that variability to the greatest extent possible.