Differential Privacy

Implications for the 2020 Decennial Census Data

Arizona State Demographer's Office

Overview

- Commitment to data stewardship
- Protections over time
- New threats and real world examples
- Differential privacy
- Impact of the new method
- Effect on Arizona
- What's next?

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

History of Census Bureau & Data Privacy

1790 - Officials posted results of first census so residents could correct errors.

1850 - The interior secretary decreed the results were “not to be used in any way to the gratification of curiosity and census officials,” or “the exposure of any man’s business or pursuits.”

1954 - The Census Bureau’s confidentiality mandate was codified in Title 13, Section 9 of the US Code.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Title 13

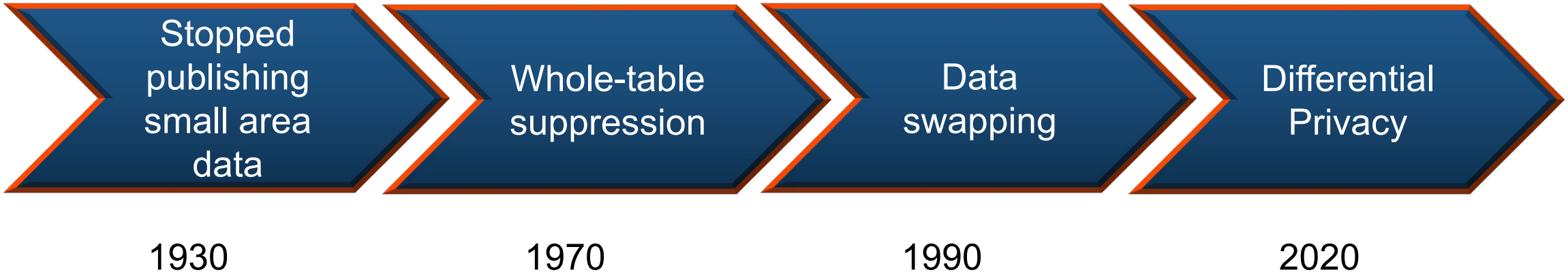
"To stimulate public cooperation necessary for an accurate census…Congress has provided assurances that information furnished by individuals is to be treated as confidential. Title 13 U.S.C. §§ 8(b) and 9(a) explicitly provide for nondisclosure of certain census data, and no discretion is provided to the Census Bureau on whether or not to disclose such data…" (U.S. Supreme Court, Baldrige v. Shapiro, 1982)

- Title 13, Section 9 of the United States Code prohibits the Census Bureau from releasing identifiable data "furnished by any particular establishment or individual."
- Census Bureau employees are sworn for life to safeguard respondents' information.
- Penalties for violating these protections can include fines of up to $250,000, and/or imprisonment for up to five years.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Privacy Protections Over Time

As the number and detail of Census Bureau data products has increased, the statistical techniques used to protect respondent data have improved.



ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Current Challenges in Keeping Public Trust

- Declining trust in government
- Increasingly common corporate data breaches
- Declining response rates to surveys
- Faster computers and availability of commercial data make safe-guarding private information more difficult

The Attack on Data

A *reconstruction attack* is an attempt to derive information about an individual using only published statistics. If a sizable number of aggregated statistics are published, then it is possible to deduce characteristics of the individuals that make up the aggregations.

A *re-identification attack* involves linking the information derived through reconstruction with other sources of data that contain identifiers like name, address, Social Security Number (SSN), and employer identification number, to reveal sensitive information about individuals.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Data Attacks in the Real World

Reconstruction and Re-identification are not just theoretical possibilities:

- Massachusetts Governor's Medical Records (Sweeney, 1997)
- AOL Search Queries (Barbaro and Zeller, 2006)
- Netflix Prize (Narayanan and Shmatikov, 2008)
- Washington State Medical Records (Sweeney, 2015)

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Massachusetts Group Insurance Commission

- The Commonwealth of Massachusetts Group Insurance Commission (GIC) released anonymous health records to encourage research to benefit society. The GIC took specific steps to protect privacy, such as suppressing street addresses and replacing people's names with random numbers.
- For $20, Latanya Sweeney (a PhD student at MIT) purchased a CD with the publicly available voter registration database for the city of Cambridge. By simply comparing the voter registration data with the GIC data, she was able to re-identify the health records in the GIC publication that belonged to the then governor of Massachusetts, William Weld.
- A few years later, Sweeney published a paper in which she concluded that up to 87% of individuals living in the United States can be uniquely identified by using the same 3 data features she used to identify the governor's records in the GIC data: birth date, ZIP code, and gender.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

AOL Data Release

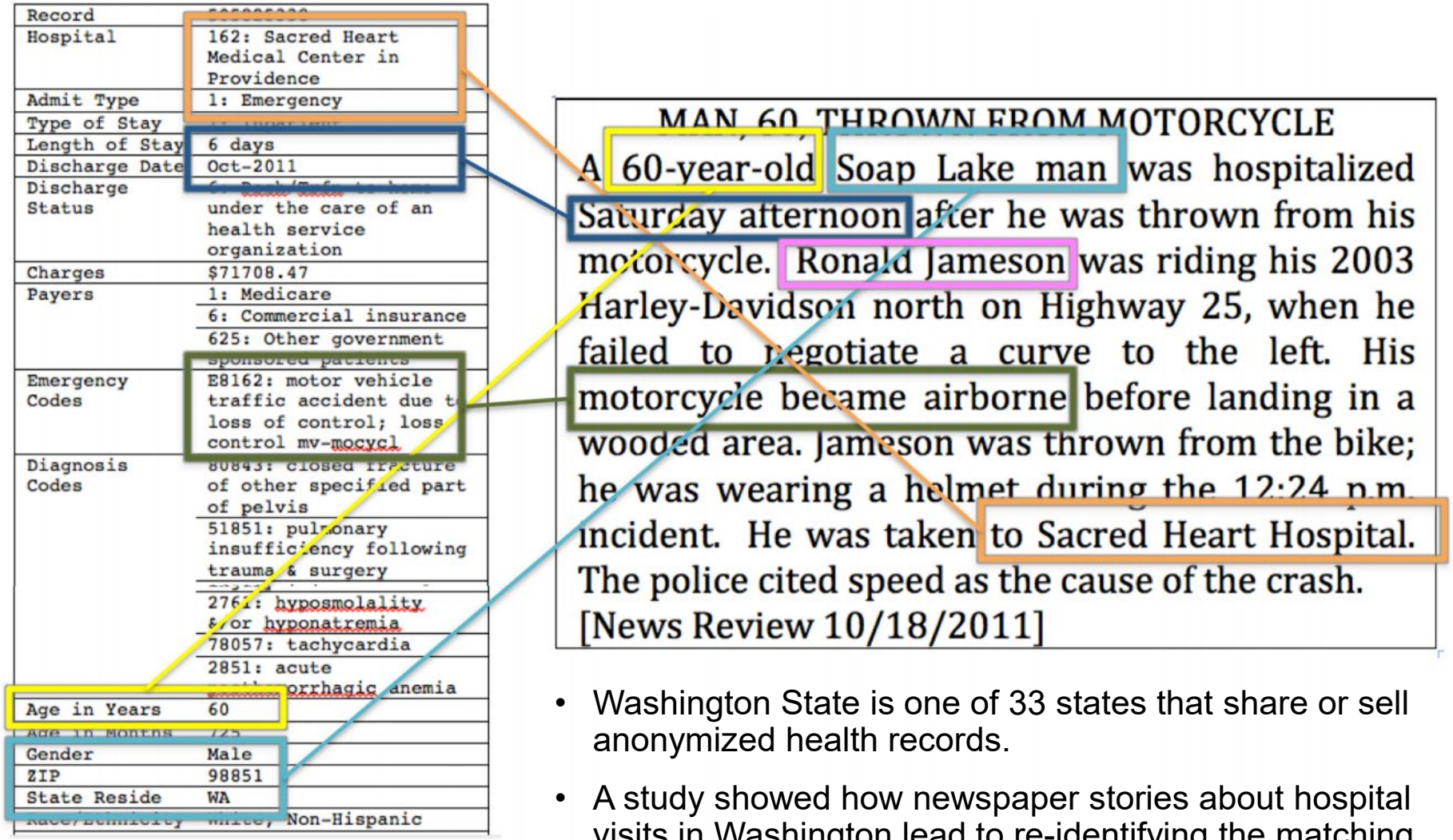
- To “embrac[e] the vision of an open research community,” AOL Research publicly posted to a website twenty million search queries for 650,000 users of AOL’s search engine, summarizing three months of activity.
- Bloggers pored through the data either attempting to identify users or “hunt[ing] for particularly entertaining or shocking search histories.”
  - User No. 3505202 asked about “depression and medical leave.”
  - User No. 7268042 typed “fear that spouse contemplating cheating.”
  - User No. 17556639 searched for “how to kill your wife” followed by a string of searches for things like “pictures of dead people” and “car crash photo.”
- Two *New York Times* reporters recognized clues to User 4417749’s identity in queries such as “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in Shadow Lake subdivision Gwinnett County, Georgia.” They quickly tracked down Thelma Arnold, a 62 year-old widow from Lilburn, Georgia who acknowledged that she had authored the searches.

Netflix Prize

- In 2006, Netflix ran a contest to improve its recommendation system. It released a sample of its subscribers' ratings histories. To protect their users' privacy, Netflix removed direct identifiers.
- Narayanan and Shmatikov re-identified a large share of the users in the Netflix Prize data by matching to data from IMDb.com, a comprehensive online database of films with information on casts, production, and crowd-sourced ratings.
- This attack harmed the re-identified users: *" …we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information."*
- In 2009, a few Netflix customers brought a class action lawsuit against the company for privacy violations stemming from the release of the Netflix Prize data.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Washington State Medical Records



Record	[redacted]
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	[redacted]
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	[redacted] under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due t[redacted] loss of control; loss control mv-mocycl
Diagnosis Codes	[redacted] of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	[redacted]
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute [redacted] orrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

- Washington State is one of 33 states that share or sell anonymized health records.
- A study showed how newspaper stories about hospital visits in Washington lead to re-identifying the matching health record 43% of the time.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Census Reconstruction & Re-identification

The Census Bureau performed database reconstruction for all 308,745,538 people enumerated in Census 2010 from public 2010 data products.

- Census block and voting age (18+) were correctly reconstructed in all 6,207,027 inhabited blocks.
- Block, sex, age, race (OMB 63 categories), and ethnicity were reconstructed:
  - Exactly for 46% of the population (142 million individuals)
  - Within +/-one year for 71% of the population (219 million individuals)
- Linking the reconstructed records to commercially available databases re-identified 17% of the population (52 million individuals)

Differential Privacy

- Also known as “Formal Privacy”
- Has roots in economic theory and incorporates cryptographic methods from computer science
- Is intended to quantify the precise amount of re-identification risk for all calculations/tables/data products produced no matter what external data is available now, or at any point in the future.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Who’s Using Differential Privacy?

- Google’s Chrome Browser
- Apple’s iOS 10 and macOS Sierra
- Microsoft’s Windows 10

Benefits of Differential Privacy

- Defines the maximum privacy “leakage” of each release of information independent of the attacker mode.
- Allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.

Data Quality |Bnae Kegouqe
Dada Qualitg |Vrkk Jzcfkdy
Data Qaality |Dncb PrhvBln
Dzte Qvality |Dncb Prtnavy
Dfha Quapyti |Tgta Ppijacy
Tgta Qucjity |Dfha Pnjvico
Dncb Qhulitn |Dzhe Njivaci
Ntue Quevdto |Dzte Privecy
Vrkk Zuhnvry |Dada Privacg
Bnaq Denorbe |Data Privacy

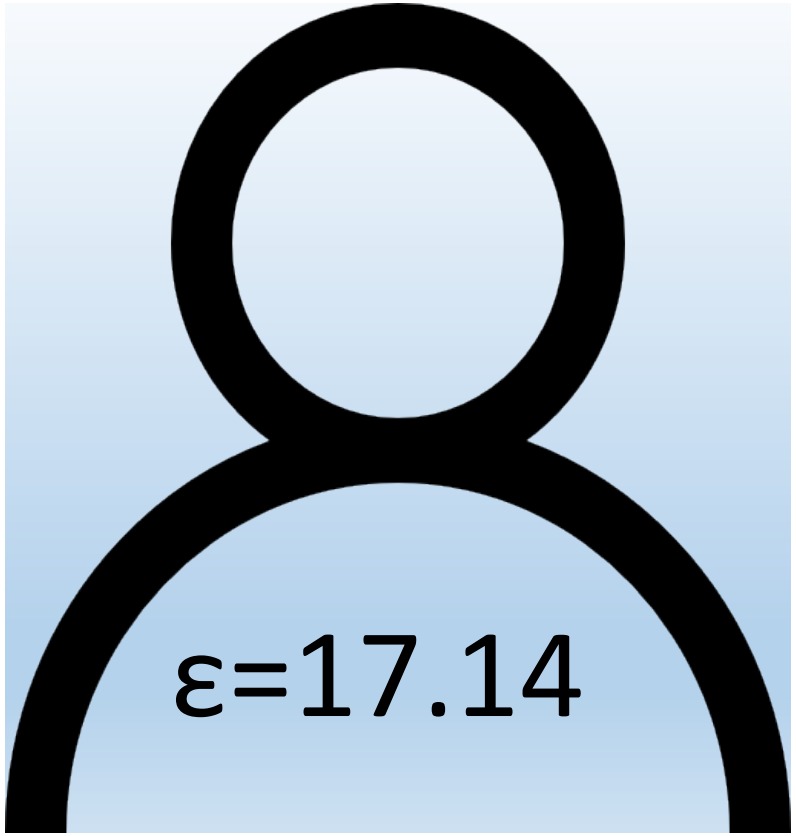
ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Choosing a Privacy Loss Budget

Differential privacy allows you to quantify a precise level of “acceptable risk” called the “Privacy Loss Budget” or “Epsilon.”

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Redistricting Privacy Loss Budget = 19.61



ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Allocation of Privacy Loss Budget

The total privacy loss budget is allocated to a specific combination of geographies and tabulations.

Production Settings:

Query	Per Query rho Allocation by Geographic Level					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		**3773/4097****	**3126/4097**	**1567/4102**	1705/4099	5/4097
CENRACE (63 cells)	52/4097	6/4097	10/4097	4/2051	3/4099	9/4097
HISPANIC (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHINSTLEVELS (3 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHGQ (8 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HISPANIC*CENRACE (126 cells)	**130/4097**	**12/4097**	**28/4097**	**1933/4102**	**1055/4099**	21/4097
VOTINGAGE*CENRACE (126 cells)	130/4097	12/4097	28/4097	10/2051	9/4099	21/4097
VOTINGAGE*HISPANIC (4 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE*HISPANIC*CENRACE (252 cells)	26/241	2/241	101/4097	67/4102	24/4099	71/4097
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	189/241	230/4097	754/4097	241/2051	1288/4099	3945/4097

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Invariants: Data without Noise

- Total population at the state level
- Total housing units at the census block level
- Number of group quarters facilities by type at the census block level

The Decennial Census

The U.S. Census Bureau is required to enumerate all the people living in the U.S. every 10 years (U.S. Constitution, Article 1, Section 2). The switch to Differential Privacy does not change the constitutional mandate.

By law, the Census Bureau will conduct the 2020 Census and deliver:

- Each state's population total, which determines each state's number of seats in the U.S. House of Representatives. (released April 26, 2021)
- The local counts each state needs to complete legislative redistricting. These PL94-171 redistricting statistics provide block-level population counts, including data on race and ethnicity, as mandated by the Office of Management and Budget (1997). (released August 16, 2021)

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Impact of DP-Induced Inaccuracy

Controls for federal surveys:

- American Community Survey
- Current Population Survey
- Survey of Income & Program Participation
- American Housing Survey

Denominators for vital rates and per capita statistics:

- Birth rates
- Death rates
- Incidence of disease

Allocation of federal funds:

- $675 Billion

Other:

- Academic research
- Business research
- Public information and education
- Program planning for public/private services

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Reaction: Against DP

“Differential privacy goes above and beyond what is necessary to keep data safe under census law and precedent … This is not the time to impose arbitrary and burdensome new rules that will sharply restrict or eliminate access to the nation’s core data sources.”

– Regents Professor of History & Population Studies,
University of Minnesota,
Steven Ruggles

“If the reliability of that data falls by the wayside or the data becomes so difficult to interpret that general users are unable to decipher it, we run the risk of basing decisions on no data at all or, perhaps worse, on inaccurate data.”

– Letter to Census Director, State of Maine

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Reaction: In Support of DP

"Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated that they can often "reidentify" or "deanonymize" individuals hidden in anonymized data with astonishing ease. By understanding this research, we realize we have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention. We must respond to the surprising failure of anonymization…"

– Paul Ohm, Professor, Georgetown University Law Center

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Impact on Arizona

Effect of Differential Privacy on Redistricting Data

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

2010 Demonstration Data Products

The Census Bureau released 6 iterations of “2010 demonstration data products” – Census 2010 data with the Differential Privacy algorithm applied to them. These are sometimes referred to as results of the Disclosure Avoidance System (DAS). Each iteration reflected improvements to the algorithm and/or changes in the privacy loss budget from ε=4.5 to ε= 19.61.

- The data were made available to the public.
- Demographers and other data users throughout the country analyzed the data, compared them with the original 2010 data tables, and shared their findings.
- Most people expressed the opinion that there was too much error in the data for them to be useful. Some users became more satisfied with the accuracy in later iterations.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Measures Used to Evaluate Accuracy

- Mean Absolute Error (MAE)
  - Provides an easy to interpret measure of the numeric error
- Mean Absolute Percent Error (MAPE)
  - An easy to interpret relative measure of error
- Mean Algebraic Percent Error (MALPE)
  - Identifies systematic bias

2020 Redistricting Data Error Profile

	Counties (N=15)			Incoporated Places (N=91)		
Population Size	MAE (#)	MAPE (%)	MALPE (%)	MAE (#)	MAPE (%)	MALPE (%)
All Sizes	2	0.01	0.00	6	0.08	-0.01
Total population <1,000	--	--	--	2.33	0.38	-0.17
Total population 1,000 to 4,999	--	--	--	3.31	0.12	0.01
Total population 5,000 to 9,999	4	0.05	0.05	4.8	0.07	0.01
Total population 10,000 to 49,999	2.67	0.01	0.00	6.39	0.03	0.00
Total population 50,000 to 99,999	2.5	0.00	0.00	11.83	0.02	-0.02
Total population >=100,000	1.67	0.00	0.00	15.1	0.01	0.00

Tracts (N=1,526)

- The mean absolute error over all tracts is 1.96.

- 3 tracts had a percent error exceeding 10%.

Incorporated Places with the Largest Absolute Percent Error

Place	DAS pop	Original pop	Difference	% Difference
Patagonia	913	917	4	0.44%
Fredonia	1,314	1,317	3	0.23%
Holbrook	5,053	5,064	11	0.22%
Mammoth	1,426	1,429	3	0.21%
Gila Bend	1,922	1,926	4	0.21%
Jerome	444	443	-1	-0.23%
Tombstone	1,380	1,376	-4	-0.29%
Parker	3,083	3,074	-9	-0.29%
Winkelman	353	351	-2	-0.57%
Duncan	696	690	-6	-0.86%

Potential Local Impact of Error: Phoenix

- The original demonstration data under reported the population for the City of Phoenix by 2,515, or 0.174 percent.
- Applying that percentage to state-shared revenue formula would mean that the city would lose nearly $1 million in shared revenue for this year alone.
- Similarly, the city would lose nearly $450,000 in federal funding for critical services for one year.

Over the course of a decade, the DP-induced under-reporting would cause the city to lose well over $10 million in state-shared revenue and federal funding.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Place	DAS pop	Original pop	Difference	% Difference
Navajo Nation	97,497	97,349	148	0.15%
Fort Apache	12,856	12,870	-14	-0.11%
Gila River	10,854	10,845	9	0.08%
San Carlos	9,846	9,835	11	0.11%
Tohono O'odham	9,163	9,139	24	0.26%
Hopi	6,865	6,857	8	0.12%
Salt River	4,540	4,496	44	0.98%
Pascua Pueblo Yaqui	3,140	3,154	-14	-0.44%
Colorado River	2,446	2,414	32	1.33%
Hualapai	1,258	1,264	-6	-0.47%
Fort McDowell Yavapai	853	852	1	0.12%
Maricopa (Ak Chin)	736	726	10	1.38%
Yavapai-Apache	558	557	1	0.18%
Cocopah	505	520	-15	-2.88%
Havasupai	435	436	-1	-0.23%
Fort Mojave	394	405	-11	-2.72%
Kaibab	197	203	-6	-2.96%
Yavapai-Prescott	115	115	0	0.00%
Tonto Apache	79	80	-1	-1.25%
Fort Yuma	0	2	-2	-100.00%
Zuni	0	0	0	--
All Reservations	**162,337**	**162,119**	**218**	**0.13%**

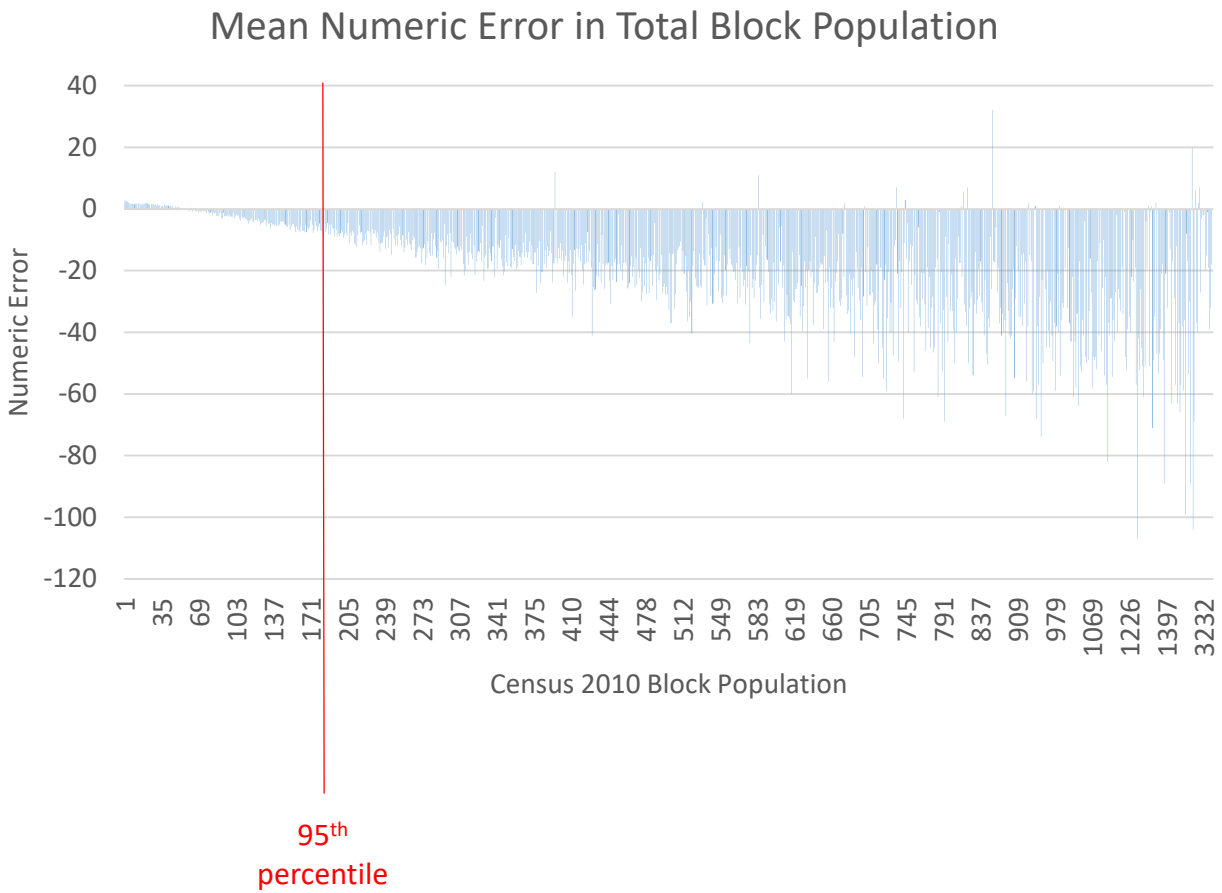
Tribal Population

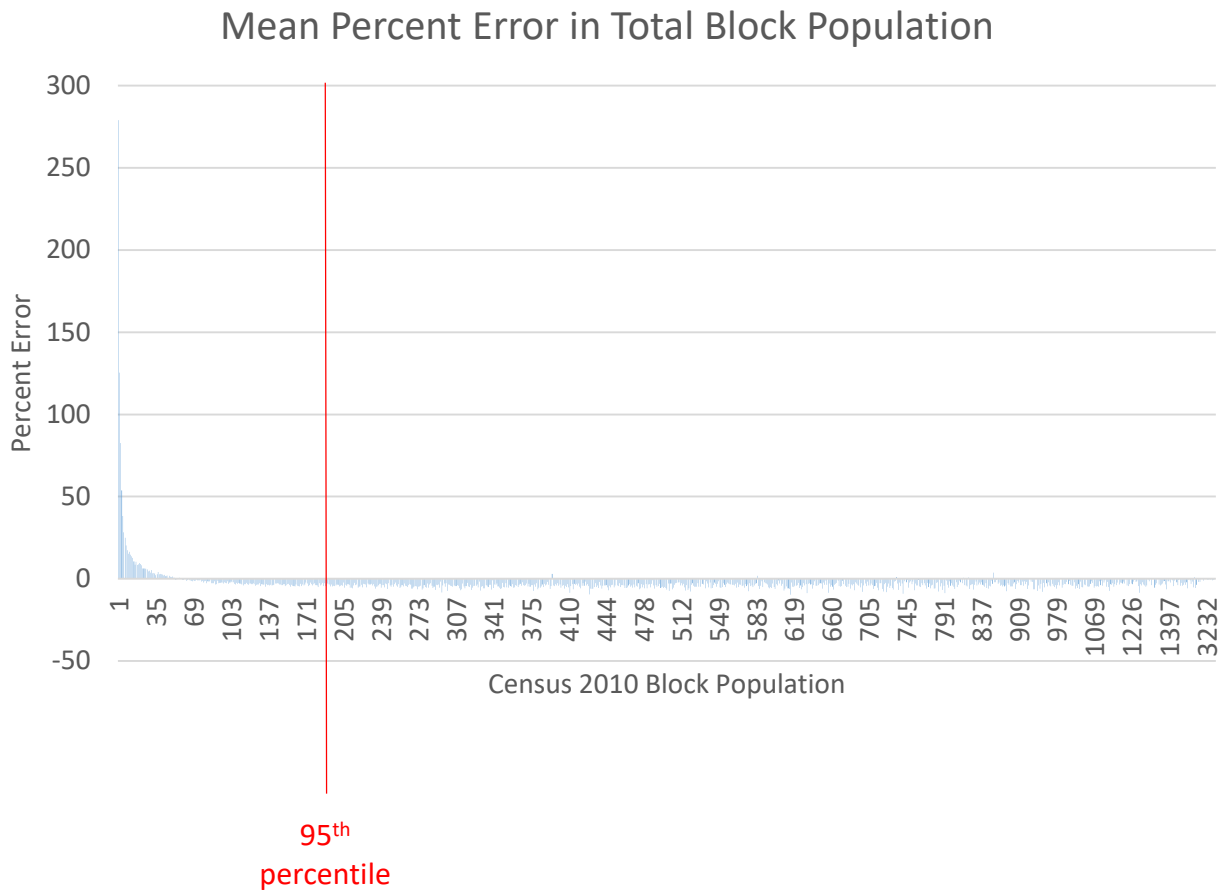
Arizona has the highest American Indian/Alaska Native population of any state at 332,273 persons (2019 ACS 1yr estimate).

This population in AZ tribal areas was undercounted in previous DAS runs but is now much more accurate.

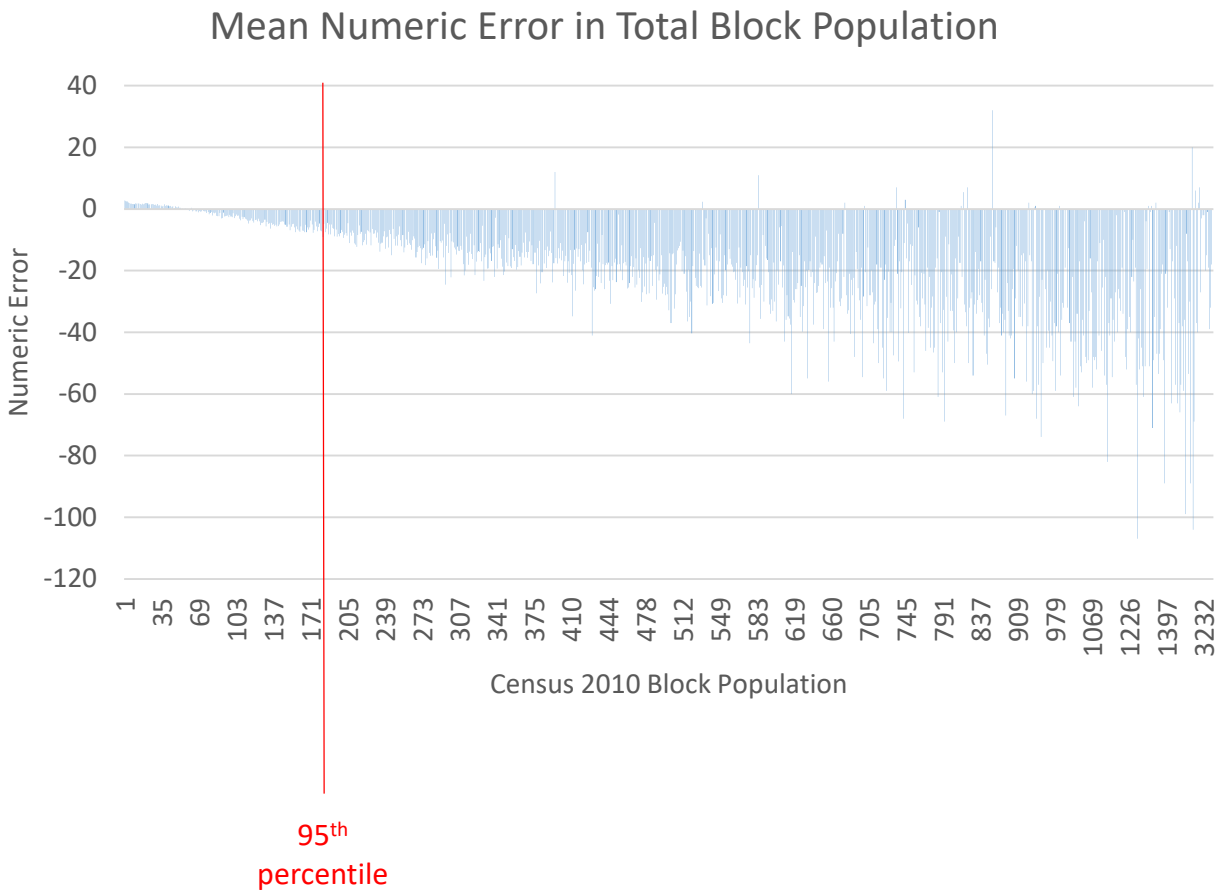
ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

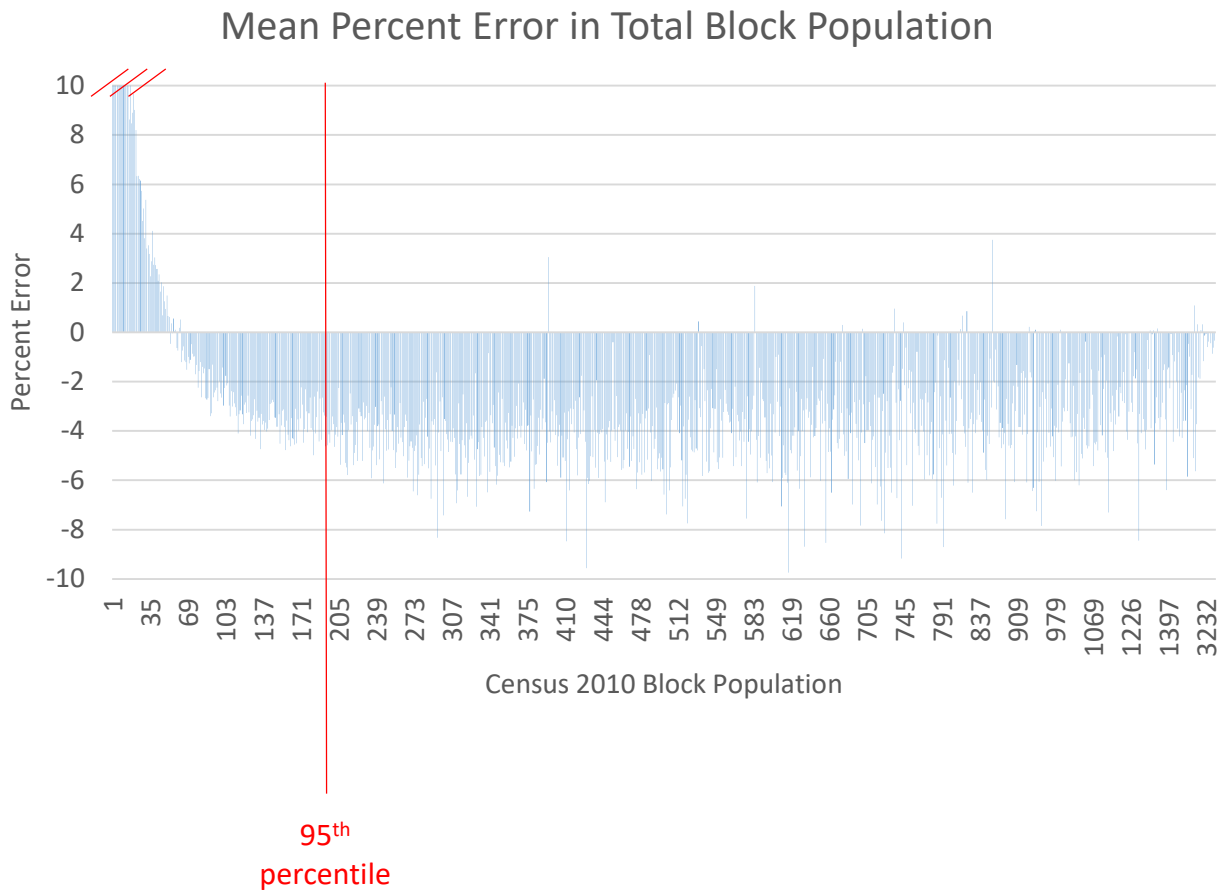
Error in Blocks with Nonzero Population

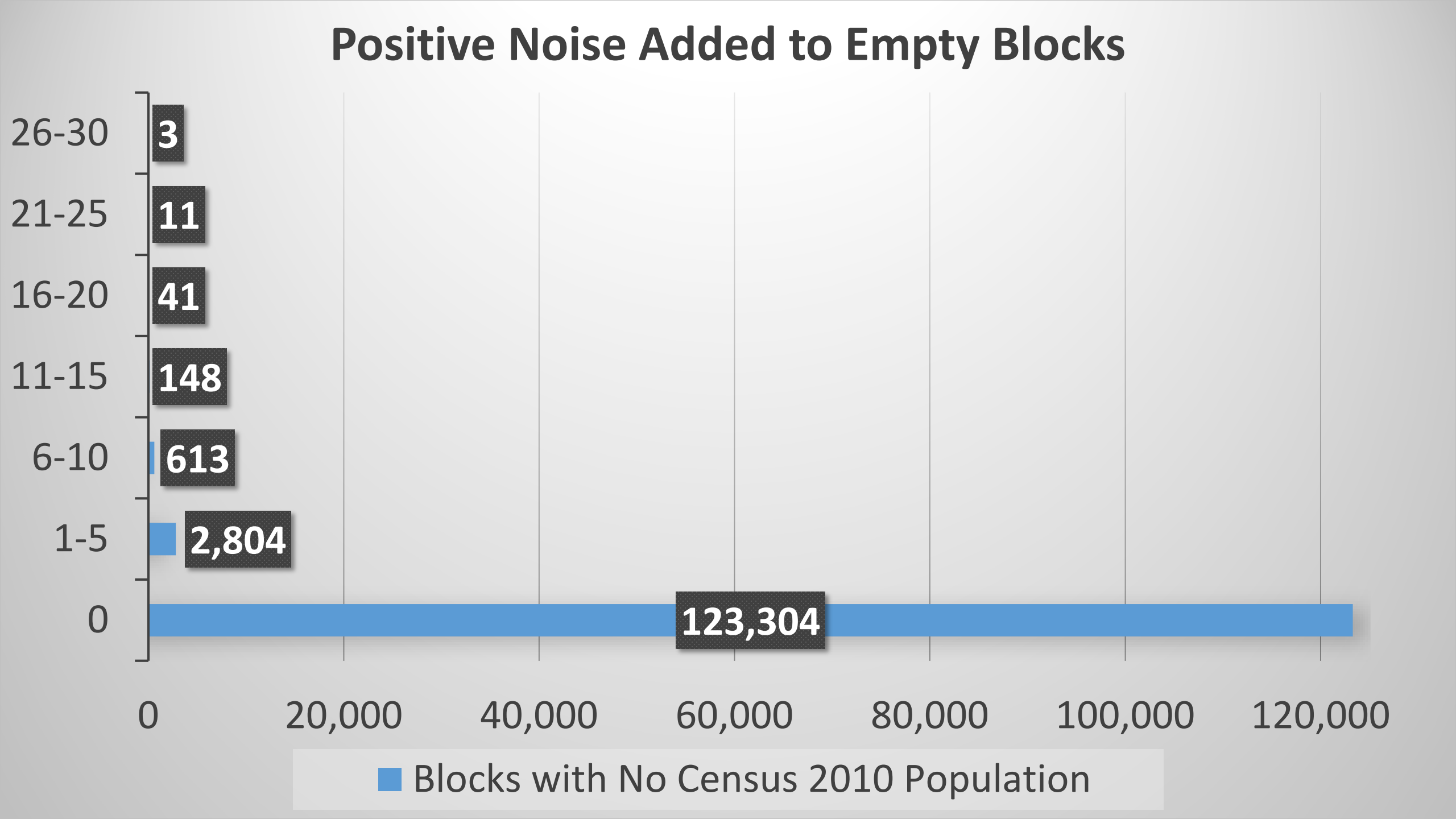




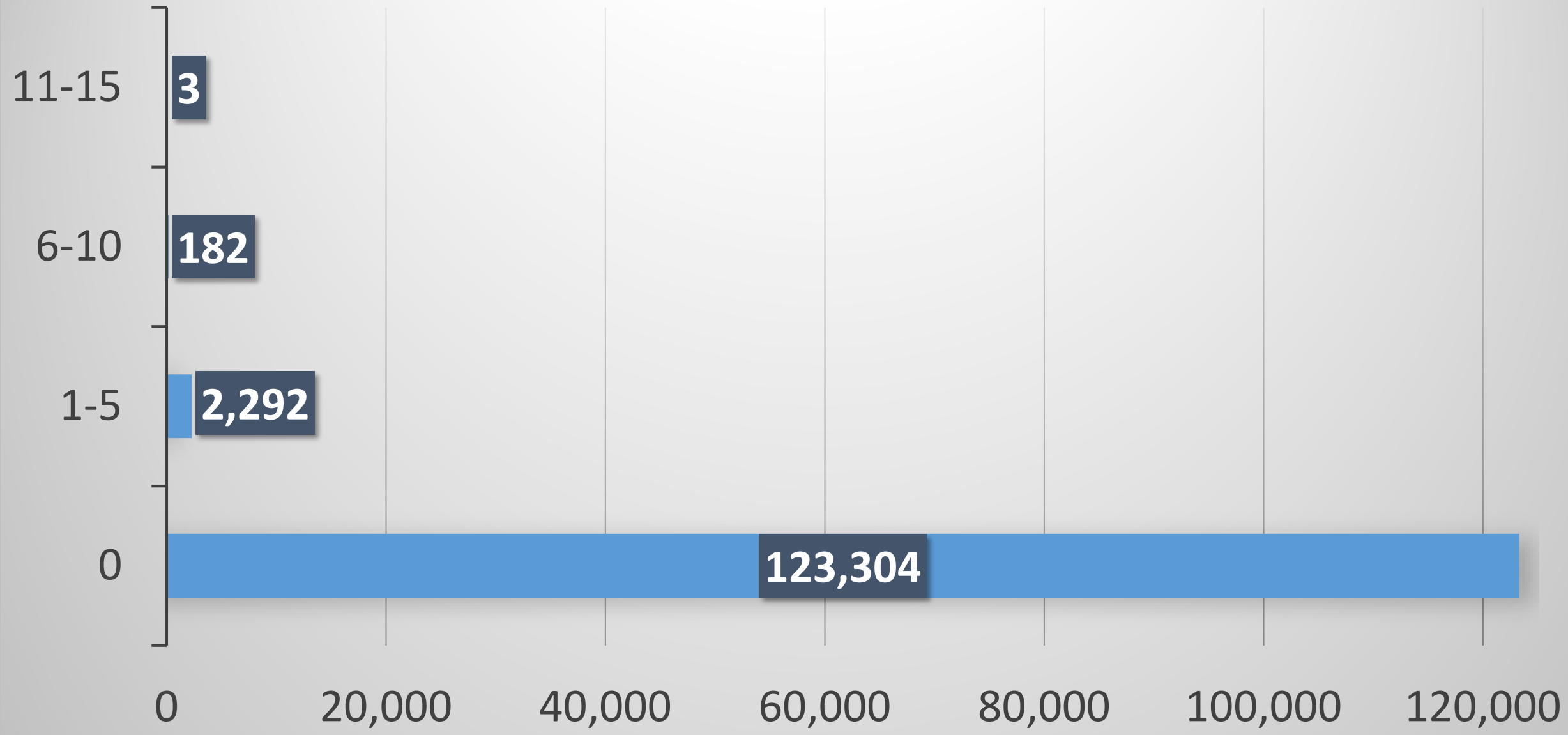
Error in Blocks with Nonzero Population






Positive Noise Added to Empty Blocks
26-30
3
21-25
11
16-20
41
11-15
148
6-10
613
1-5
2,804
0
123,304
0
20,000
40,000
60,000
80,000
100,000
120,000
Blocks with No Census 2010 Population

Negative Noise Causing Empty Blocks



Logical Inconsistencies: Blocks

Unlikely or Impossible Characteristics	Number of Blocks	Percent of Blocks
More Occupied Housing Units than Household Population	5,300	2%
No Occupied Housing Units, but Household Population Exists	11,190	5%
100% Occupancy Rate	65,101	27%
Persons Per Household is >=10 but Census Value is <10	3,151	1%

Block Example: 040134212013003 in Mesa

	Census 2010	DAS	Difference
Total Population	40	81	41
Non-Hispanic	23	47	24
White	18	22	4
Black	0	0	0
American Indian	0	4	4
Asian	3	10	7
Pacific Islander	0	0	0
Some Other Race	0	0	0
Two or More Races	2	11	9
Hispanic	17	34	17
Housing Units	16	16	0
Person Per Household	2.50	5.06	2.56
Occupancy Rate	100%	100%	--

Error in Group Quarters Population

		Metrics			
Geography	**N**	MAE (#)	MAPE (%)	MALPE (%)	Cases where Error >= 5%
Counties*	15	1	3.13	2.34	2
Incorporated Places	91	-2.13	17.99	12.76	35
Tracts	1,526	0.01	51.77	35.02	740
Block Groups	4,178	0.01	58.44	34.87	1,238
Blocks	241,666	0.01	65.72	36.45	2,009
Tribal Lands	21	2.64	15.04	10.40	7

*Greenlee County error = 28.6% and Santa Cruz County error = 5.7%

Reliability of Redistricting Data

Reliability is how well the differentially private data compares to the published 2010 Census data. The criteria for reliable data are described as follows:

“The difference between the Top Down Algorithm’s ratio of the largest demographic group and the corresponding swapping algorithm's ratio (used in the 2010 Census) for the largest demographic group is less than or equal to five percentage points at least 95% of the time.”

This applies to

- Block groups with 450-499 people
- Minor Civil Divisions and places with 200-249 people

https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates/2021-08-12.html

Reliability of Redistricting Data

		Total	Hispanic	NH W	NH B	NH AIAN	NH A	NH HPI	NH SOR	NH 2+
Census 2010	Phoenix	1,445,632	41%	47%	6%	2%	3%	0%	0%	2%
	Jerome	444	6%	90%	0%	0%	0%	0%	0%	3%
	Parker	3,083	42%	35%	1%	18%	1%	0%	0%	3%
DAS/TDA	Phoenix	1,445,639	41%	47%	6%	2%	3%	0%	0%	2%
	Jerome	443	7%	89%	1%	0%	0%	0%	0%	3%
	Parker	3,074	40%	35%	2%	19%	1%	0%	0%	3%

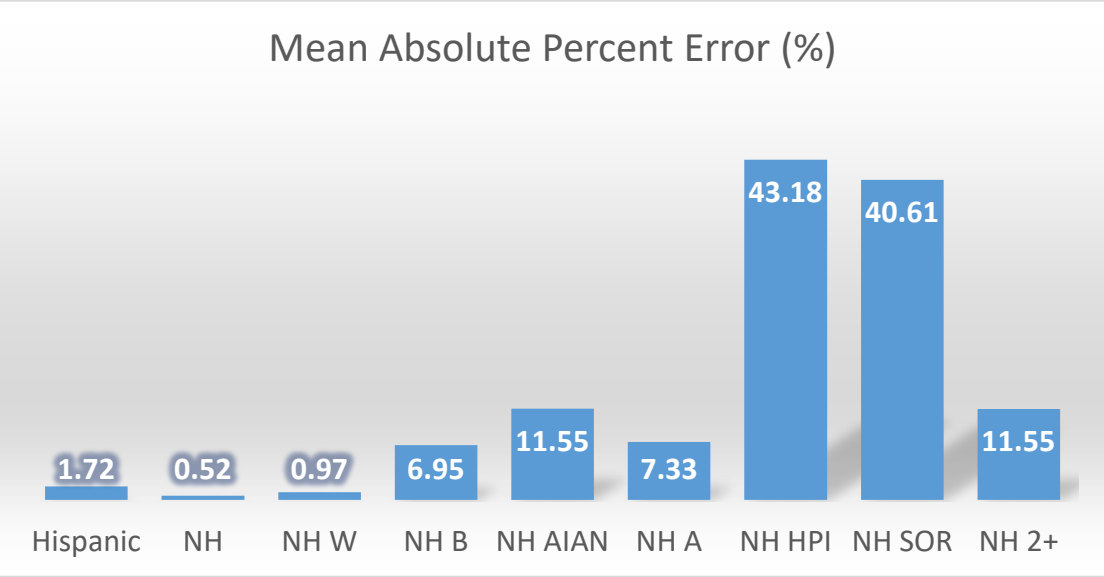
Absolute difference between ratios in Percentage Points

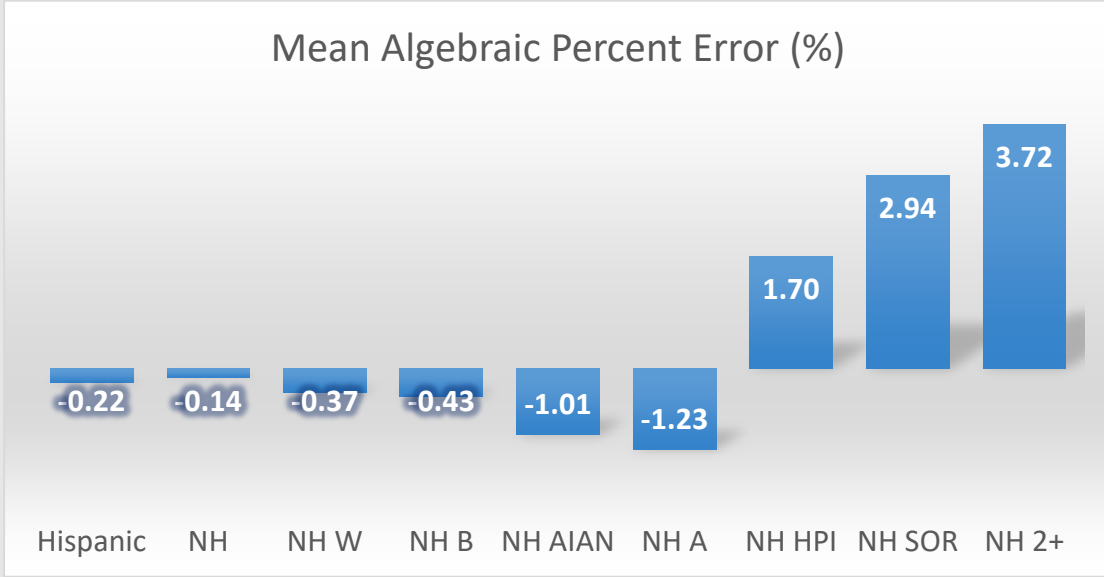
Phoenix 0

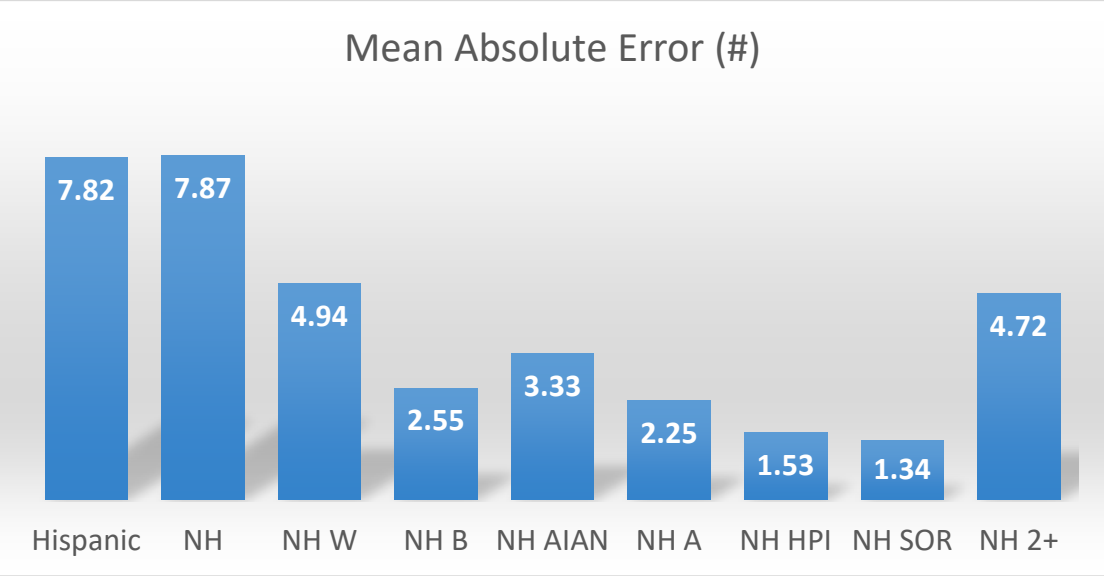
Jerome 1

Parker 2

Error in Race/Ethnicity: Tracts







- There are 1,526 tracts.
- 1,506 tracts have a population greater than 1,000 people.
- The numeric error across tracts is minimal.
- Large percentage errors mostly occur because the number of people in certain race groups is very small.

Next Steps in Differential Privacy

- Additional research continues on how to apply differential privacy to upcoming Census Data Products which include
  - Demographic Profile (DP)
  - Demographic and Housing Characteristics File (DHC)
  - Detailed Demographic and Housing Characteristics File (DDHC)
- New allocations of privacy loss budgets are being developed. Consistency with the P.L. 94-171 is planned, and improvements in the relationship between person data and housing data is expected.
- Feedback on the planned products may be sent to 2020DAS@census.gov through October 22, 2021.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Resources

A series of handbooks for each 2020 Census Data Product will be produced beginning with a guide for P.L. 94-171. Release dates have not been determined.

- 2020 Census Data Product Planning Crosswalk

https://www.census.gov/newsroom/press-releases/2021/2020-census-data-product-planning-crosswalk.html

- 2020 Census Data Products: Disclosure Avoidance Modernization

https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html

- Disclosure Avoidance Webinar Series

https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series.html

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Acknowledgements

In addition to our own original analysis, some of the information in this presentation was curated from various slide presentations, white papers, and news articles shared by the Census Bureau, private companies, and stakeholders from academic and government agencies.

The Arizona State Demographer's Office can provide the original sources upon request.

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY

Differential Privacy & Census 2020

OFFICE of ECONOMIC OPPORTUNITY

Thara Salamone, Demographer

602-771-1161, Thara.Salamone@oeo.az.gov

Jim Chang, State Demographer

602-771-1236, Jim.Chang@oeo.az.gov

ARIZONA
OFFICE OF ECONOMIC OPPORTUNITY