Differential Privacy and Accuracy

JAN VINK

NY SDC MEETING 2021

The opinions expressed in this presentation and on the following slides are solely those of the presenter

Cornell University

Program on Applied Demographics

CORNELL POPULATION CENTER

Background

Census Bureau is implementing new Disclosure Avoidance System

- More control over accuracy vs privacy
  - Differential Privacy adds noise
  - Post-processing makes all values non-negative and consistent
    - Can also affect accuracy
- Neither accuracy nor privacy is easily quantified
  - costs of less accuracy depend on use cases
- Stakeholder involvement to help find right balance

Definition of accuracy

From Statistics Canada:

Accuracy refers to the extent to which the data correctly describes the phenomenon they are supposed to measure.

- Accuracy is often decomposed into **precision**, which measures how similar are repeated measurements of the same thing, and **bias**, which measures any systematic departures from reality in the data.

Demonstration products

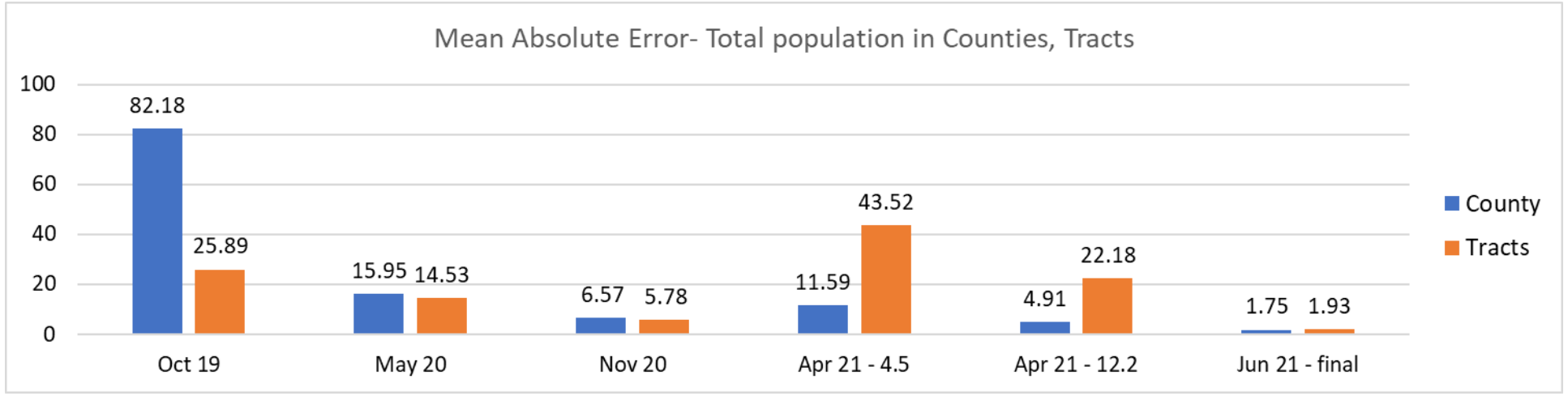
1. October 2019
   - Included most variables, ε = 6 (p:4 + hu:2)
2. May 2020
   - Included only person variables, ε = p:4
3. September/November 2020
   - Only PL variables, ε = 4.5 (p:4 + hu:0.5)
4. April 2021 (2 sets)
   - Only PL variables, ε = 4.5, ε = 12.2 (p:10.3 + hu:1.9)
5. June 2021 (production code)
   - Only PL variables, ε = 19.61 (p:17.14 + hu:2.47)

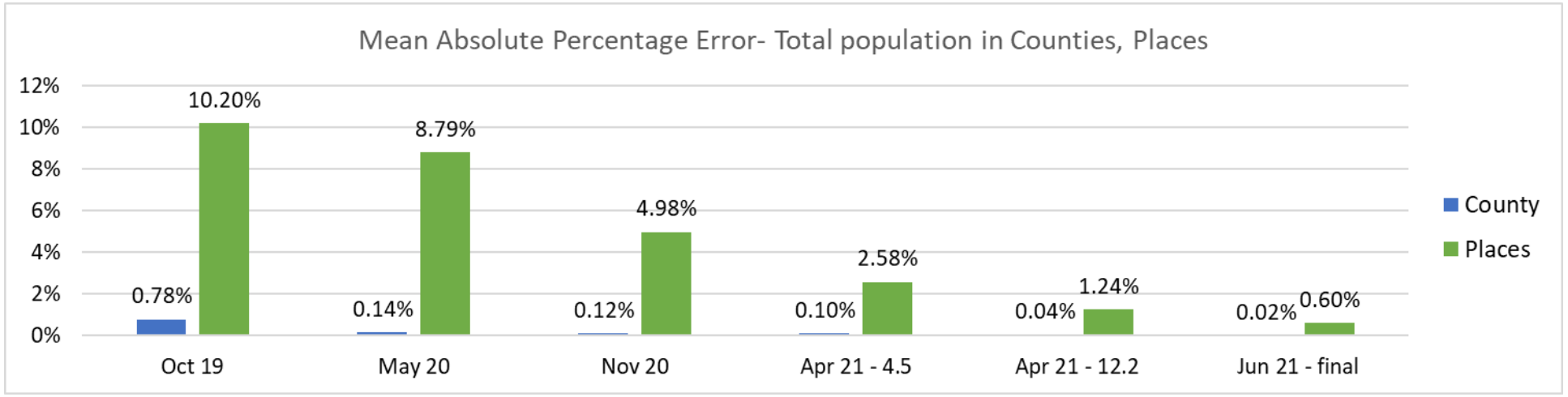
Metric tables

Produced for each demonstration product

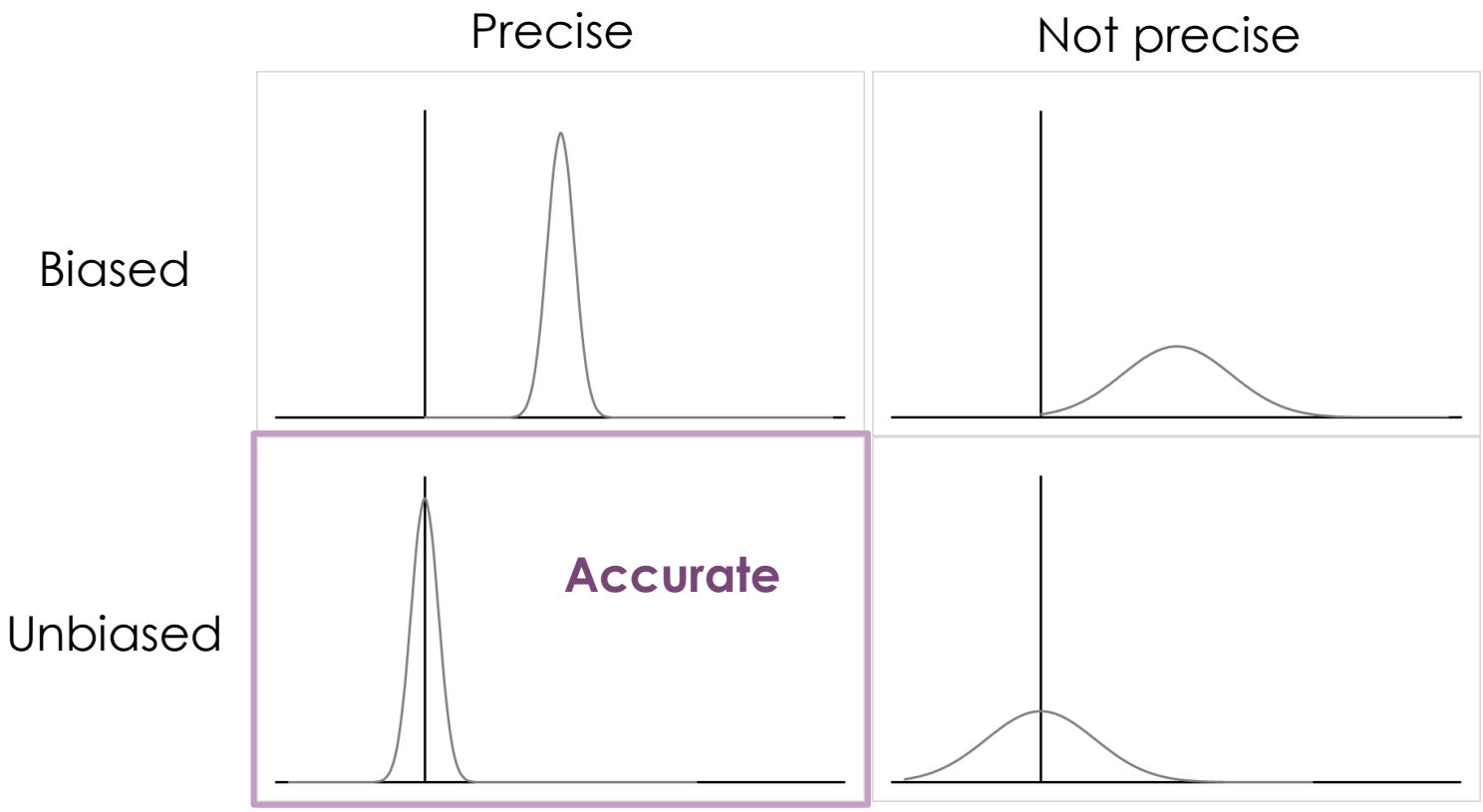
- Type of metrics
  - Mean errors, Mean Absolute Errors, Mean Percentage Errors, Mean Absolute Percentage Error, Frequency of outliers
- For different geographies
  - Sometimes also size categories
- For different race groups
- Goal: to be able to see the progress of DAS development

Metrics tables

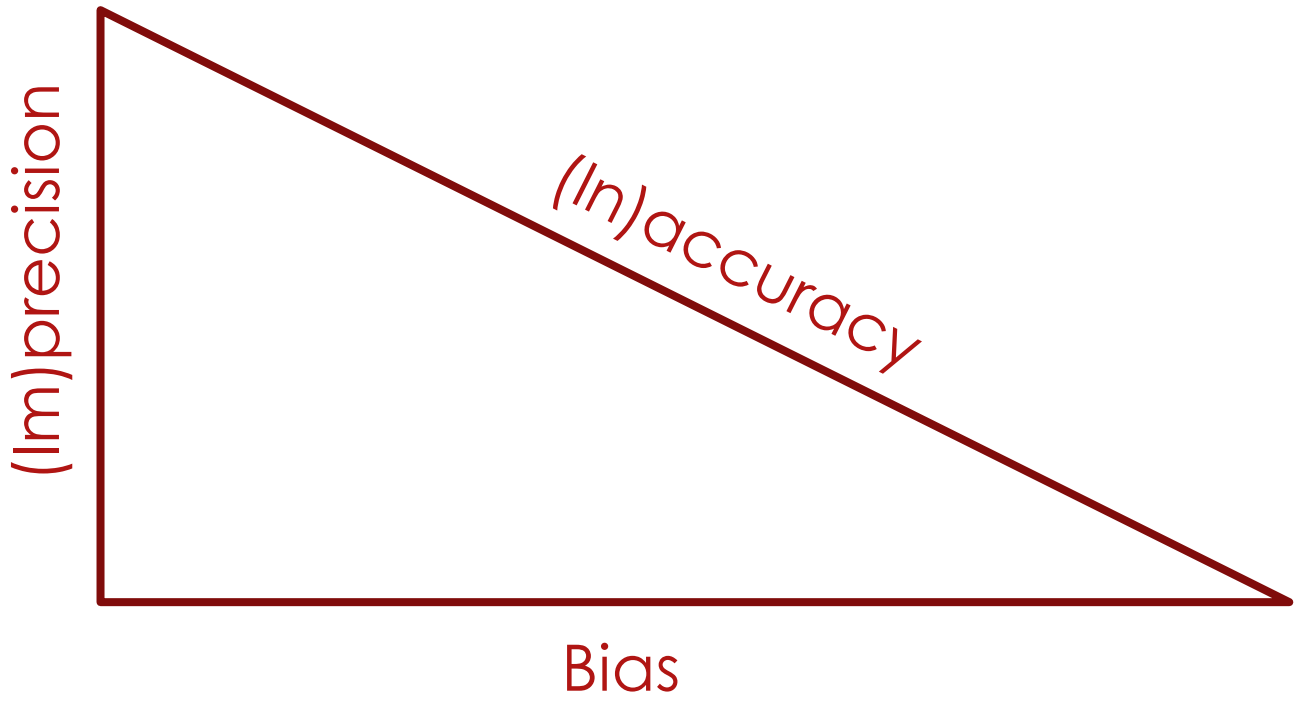




Bias, Precision and Accuracy



Bias, Precision and Accuracy



Final Demonstration Product: Total population in NY places

Total population

Group	N	SF1	DP	Difference in total	Count differences: Bias ME		Count differences: Precision StdDev	Count differences: Accuracy MAE	Percent differences: Bias MALPE		Percent differences: Precision StdDev	Percent differences: Accuracy MAPE	Extreme percent diff: APE >= 5%	Extreme percent diff: APE >= 10%
0 - 499	160	50,223	49,832	-391	-2.4	*	14.7	11.9	0.7%		14.5%	5.7%	49	16
500 - 4999	683	1,304,192	1,298,032	-6,160	-9.0	**	28.3	19.5	-0.6%	**	2.1%	1.4%	27	1
5000 - 49999	327	4,486,164	4,484,000	-2,164	-6.6	**	26.4	14.1	-0.1%	**	0.4%	0.2%	0	0
>=50000	19	9,867,359	9,867,405	46	2.4		15.2	12.4	0.0%		0.0%	0.0%	0	0
Cities	61	2,235,187	2,235,181	-6	-0.1		8.9	6.4	0.0%		0.0%	0.0%	0	0
Villages	556	10,080,714	10,074,725	-5,989	-10.8	**	27.4	17.9	-0.4%		7.2%	1.8%	30	8
CDPs	570	3,372,319	3,369,662	-2,657	-4.7	**	26.0	17.1	-0.3%		3.7%	1.6%	46	9
All places	1189	15,707,938	15,699,269	-8,669	-7.3	**	26.3	16.9	-0.3%		5.5%	1.6%	76	17
Remainder	1	3,670,164	3,678,833	8,669	8669.0	-	-	8669.0	0.2%	-	-	0.2%	0	0

Final Demonstration Product: Population by voting age in NY places

Voting age population

Group	N	SF1	DP	Difference in total	Count differences: Bias ME		Count differences: Precision StdDev	Count differences: Accuracy MAE	Percent differences: Bias MALPE		Percent differences: Precision StdDev	Percent differences: Accuracy MAPE	Extreme percent diff: APE >= 5%	Extreme percent diff: APE >= 10%
0 - 499	160	38,727	38,720	-7	0.0		9.7	7.7	0.9%		9.5%	4.4%	45	9
500 - 4999	683	1,012,832	1,010,241	-2,591	-3.8	**	18.3	13.0	-0.3%	**	1.8%	1.2%	17	1
5000 - 49999	327	3,438,660	3,437,563	-1,097	-3.4	*	24.2	17.2	-0.1%	**	0.3%	0.2%	0	0
>=50000	19	7,715,015	7,714,784	-231	-12.2		42.4	35.2	0.0%		0.1%	0.1%	0	0
All places	1189	12,205,234	12,201,308	-3,926	-3.3	**	19.9	13.8	-0.1%		3.8%	1.4%	62	10
Remainder	1	2,847,939	2,851,868	3,929	3929.0	-	-	3929.0	0.1%	-	-	0.1%	0	0

Non voting age population

Group	N	SF1	DP	Difference in total	Count differences: Bias ME		Count differences: Precision StdDev	Count differences: Accuracy MAE	Percent differences: Bias MALPE		Percent differences: Precision StdDev	Percent differences: Accuracy MAPE	Extreme percent diff: APE >= 5%	Extreme percent diff: APE >= 10%
0 - 499	160	11,496	11,112	-384	-2.4	**	9.3	7.5	19.2%		199.1%	35.1%	114	77
500 - 4999	683	291,360	287,791	-3,569	-5.2	**	17.0	12.7	-1.1%	*	13.0%	5.0%	203	77
5000 - 49999	327	1,047,504	1,046,437	-1,067	-3.3	**	21.1	15.3	0.4%		8.5%	1.3%	4	2
>=50000	19	-4,342,696	-4,345,122	-2,426	14.6		38.6	34.6	0.1%		0.2%	0.2%	0	0
All places	1189	3,502,704	3,497,961	-4,743	-4.0	**	18.1	13.1	2.1%		73.9%	7.9%	321	156
Remainder	1	822,225	826,965	4,740	4740.0	-	-	4740.0	0.6%	-	-	0.6%	0	0

Final Demonstration Product: Total population in NY Cities/Towns

Total population

Group	N	SF1	DP	Difference in total	Count differences: Bias ME		Count differences: Precision StdDev	Count differences: Accuracy MAE	Percent differences: Bias MALPE		Percent differences: Precision StdDev	Percent differences: Accuracy MAPE	Extreme percent diff: APE >= 5%	Extreme percent diff: APE >= 10%
City	61	2,235,187	2,235,181	-6	-0.1		8.9	6.4	0.0%		0.0%	0.0%	0	0
Town	932	8,958,225	8,958,233	8	0.0		4.3	3.1	0.0%		0.5%	0.1%	1	1
Village (part)	632	1,905,581	1,899,598	-5,983	-9.5	**	25.8	16.4	0.6%		15.3%	3.5%	60	25
CDP (part)	632	3,372,319	3,369,662	-2,657	-4.2	**	24.4	15.9	0.6%		11.3%	2.6%	70	19
Remainder of town	911	3,660,607	3,669,272	8,665	9.5	**	26.4	15.6	0.6%	**	4.4%	0.9%	15	2

Average errors in block groups by diversity index quintiles

	April, 12.2 Mean error	Final Mean error
20% with lowest diversity	5.05	1.43
Group 2	4.24	1.67
Group 3	0.99	0.67
Group 4	-2.22	-0.60
20% with highest diversity	-8.07	-3.11

Census blocks

Limited Privacy Loss Budget assigned to blocks

- Much noise added
- Big impact of post-processing
  - Many instances where count + noise < 0
  - Number have to be made consistent
    - Within block, e.g. Hispanic + Non Hispanic = Total
    - With higher levels of geography:
      sum of blocks in block group = block group

If noise is random, noise get cancelled out in aggregation

Number of living quarters was held invariant (no noise added)

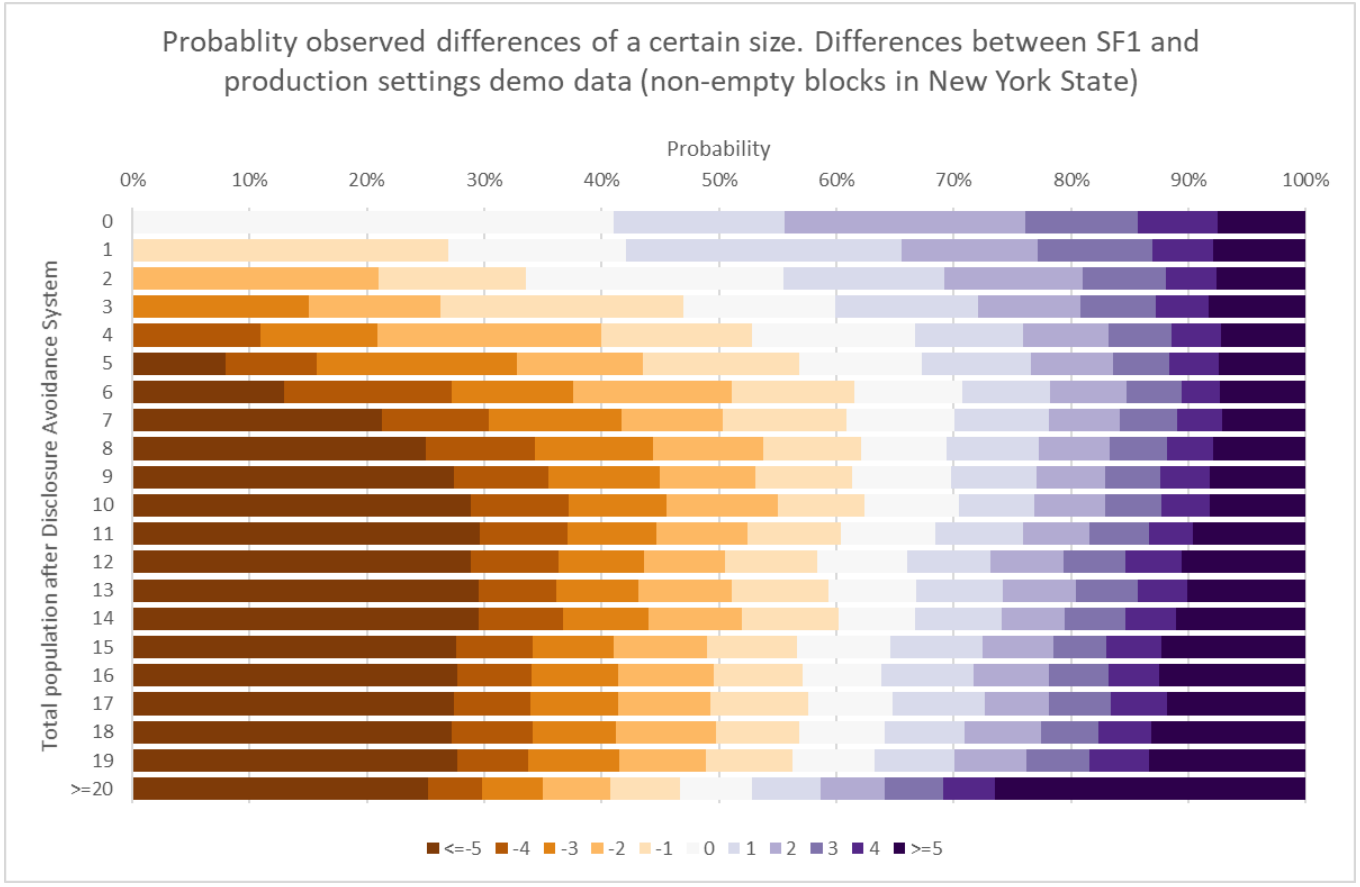
My block - 2010

1. SF1
   - **8 NH white adult + 1 NH White youth**
2. October 2019
   - 10 NH White adult
3. May 2020
   - 5 NH White adult
4. November 2020
   - 9 NH White adult + 8 NH Black adult
5. April 2021, ε = 4.5
   - 18 NH White adult + 4 NH White+Asian adult + 2 NH Black youth
6. April 2021, ε = 12.2
   - 8 NH white adult + 1 Hisp Other youth
7. June 2021 (production code)
   - 8 NH white adult + 1 Hisp White adult + 1 NH Asian adult + 1 NH Asian youth

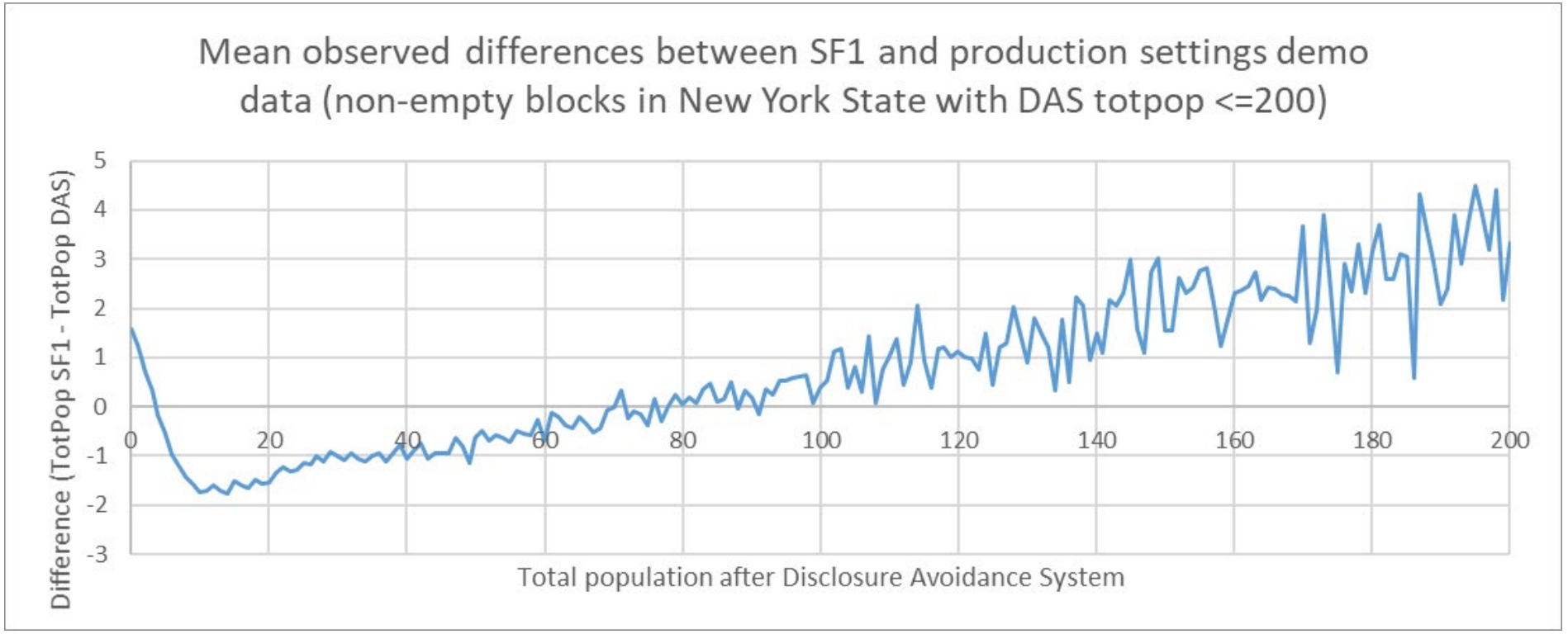
My block - 2020

1. My own count
   - **7 NH white adult + 4 NH White youth**
2. Published PL94-171
   - 4 NH White adults + 6 NH White youth + 3 NH Asian youth

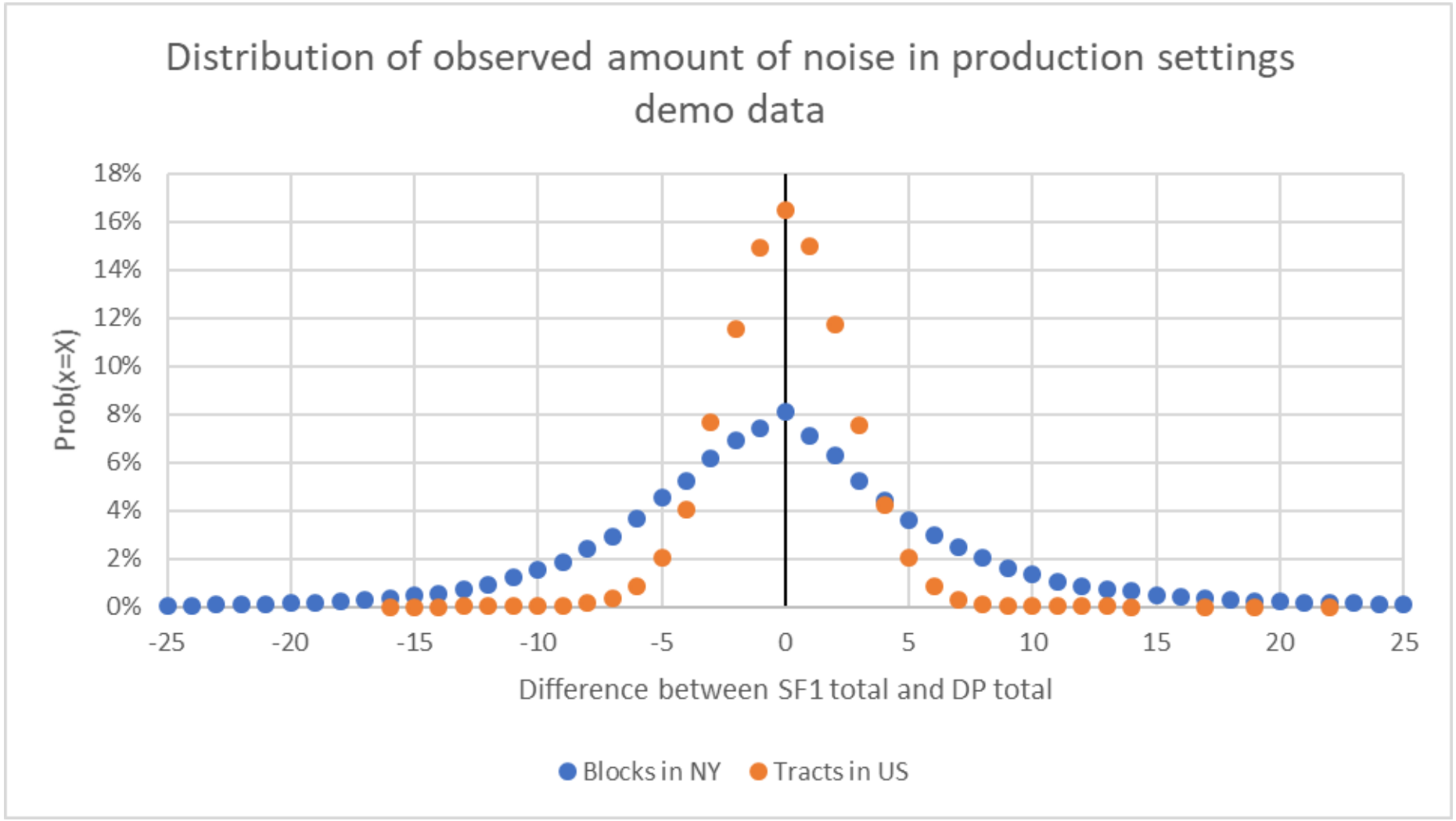
Block count differences



Differences in block counts



Error distribution (tracts and blocks)



Impossible and improbable blocks

	2010		2020	
	Count	% of all	Count	% of all
Non empty blocks	250,070		233,182	
Households (occupied houses) and household population				
Household population > 0, but occupied houses = 0	Impossible in 2010		14,276	6.1%
Household population < occupied houses (Persons per household < 1)			5,764	2.5%
Household population = 0, but occupied houses > 0			1,834	0.8%
PPH > 10	53	0.0%	4,510	1.9%
Youth only				
Only 0-17	21	0.0%	2,808	1.2%
Without GQ and only 0-17	1	0.0%	2,795	1.2%

Accuracy in future products

DAS for Demographic and Housing Characteristics (DHC) file is in development

- 2 Demonstration products
  - National workshop (CNSTAT)
- Consistency not decided yet
- Tables and geographic details not decided yet
  - GIVE FEEDBACK!
- Current time line indicates publication in summer 2022

Accuracy in future products

DAS for Detailed Demographic and Housing Characteristics (Detailed DHC) file is in development

- Not Top-Down
- Probably not consistent with other products
- Tables and geographic details not decided yet

GIVE FEEDBACK!

Handbooks and Guidance

The Census Bureau asked Population Reference Bureau (PRB) to produce handbooks that explain what Differential Privacy is

- Expected soon!

Census Bureau is looking into producing some guidance as far as uncertainty of a certain count