Elementary School Enrollment

Jan Vink

Cornell Program on Applied Demographics

jkv3@cornell.edu

Cornell University

Program on Applied Demographics

CORNELL POPULATION CENTER

Overview

- Big picture: Data quality and accuracy
- Results use cases
  - Projecting change in enrollment
  - Catchment ratio's
- Statistics on 6-11 year old for School Districts
- Statistics on 6-11 year old for Tracts
- Thoughts

// Census.gov > About the Bureau > **What We Do**

What We Do

Our Mission

The Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy.

Statistics Canada / Statistique Canada

Quality attributes related to the data and metadata

Relevance and usefulness: The extent to which the data pertain to the desired phenomenon. Data would be considered less relevant if they are too old, or do not include information about topics of interest. Usefulness of metadata refers to the extent to which it describes the data in terms of methods, concepts, limitations, assumptions made, and quality assurance practices followed.

Coverage: The extent to which the data represent the entire desired phenomenon. This could be assessed in terms of temporal or geographic coverage, or coverage of population units (i.e., people, households, businesses). Coverage is sometimes referred to as **completeness** (particularly when referring to metadata).

Granularity: Granularity refers to the unit or level of a single record in the dataset. For example a highly granular dataset could contain records of people, medical procedures or lakes, while a less granular dataset could contain records aggregated to the level of a province, or a year. The more granular or local a dataset, the greater the perceived value, balanced by greater need to protect data from unauthorized disclosure. It is usually straight-forward to aggregate or roll-up from granular data to a less granular level, but rolling down from an aggregate level is not usually possible.

Accuracy and reliability: Accuracy refers to the extent to which the data correctly describes the phenomenon they are supposed to measure. Reliability is the extent to which the data are accurate consistently over time. 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. Other factors contributing to accuracy and reliability are **validity**, the extent to which variables in the dataset have values that correspond to expected outcomes, and **consistency**, the extent to which the data are free of contradiction.

Standardization or conformance: The extent to which the data and metadata follow recognized standards in terms of formats and naming conventions, and conform to recognized dissemination standards such as SDMX for statistical products. Other aspects of standardization and conformance are the use of industry-standard software and file formats, and controlled vocabulary for data values where appropriate.

Protection of sensitive information: Unless consent has been explicitly given, it is not acceptable to disclose sensitive information in datasets made available to users beyond those granted specific access. Sensitive information includes, but is not limited to, identifiers that would associate granular data to a person, household or business, or sufficient detail in aggregate data such that one could deduce attributes of a person, household or business. There are various methods for protecting data against disclosure of sensitive information, depending on the nature and granularity of the data. Examples include suppression of sensitive information and introduction of random disturbance to data values. Many disclosure control algorithms provide diagnostics of the level of protection achieved.

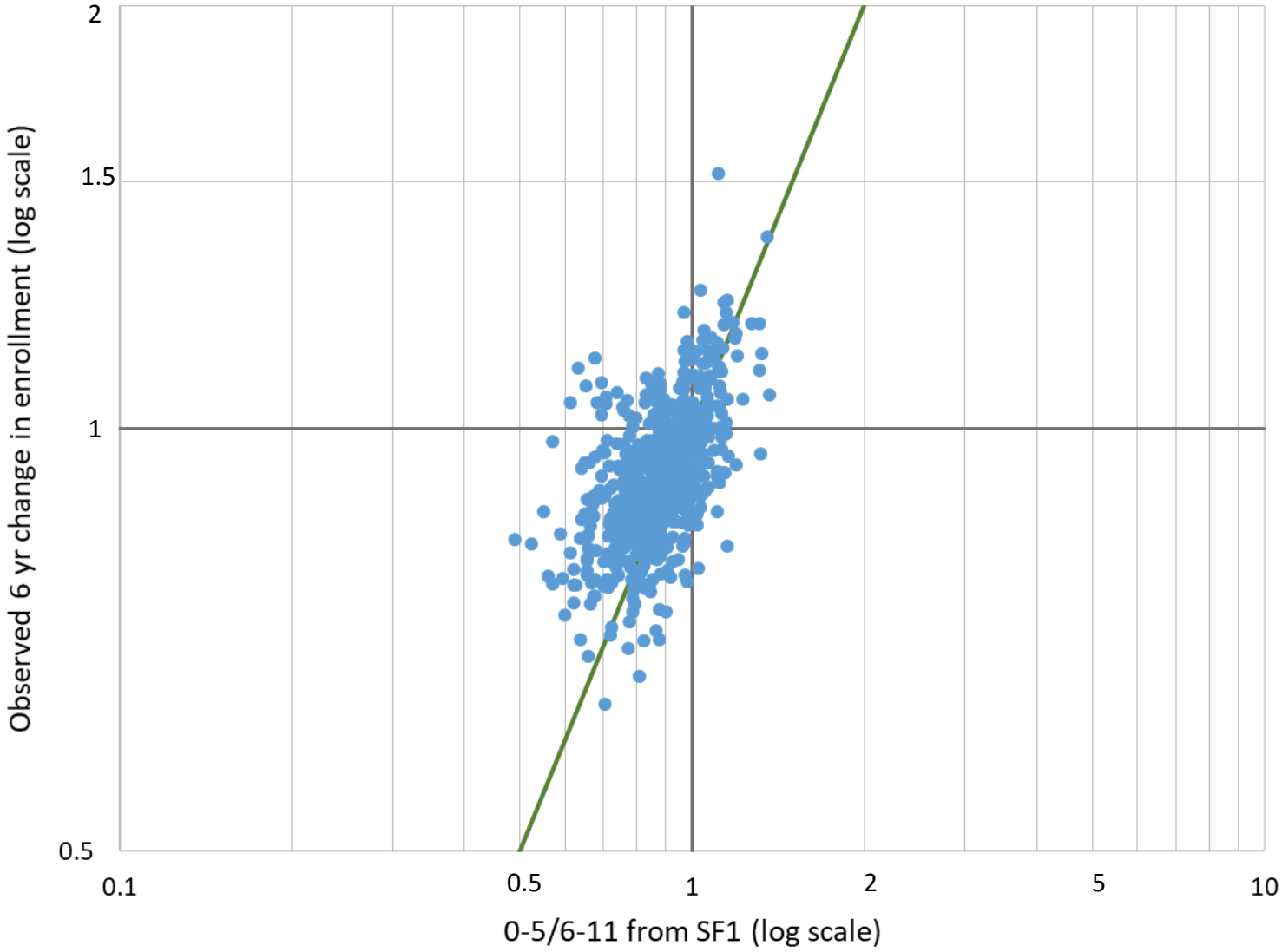
Accuracy and reliability

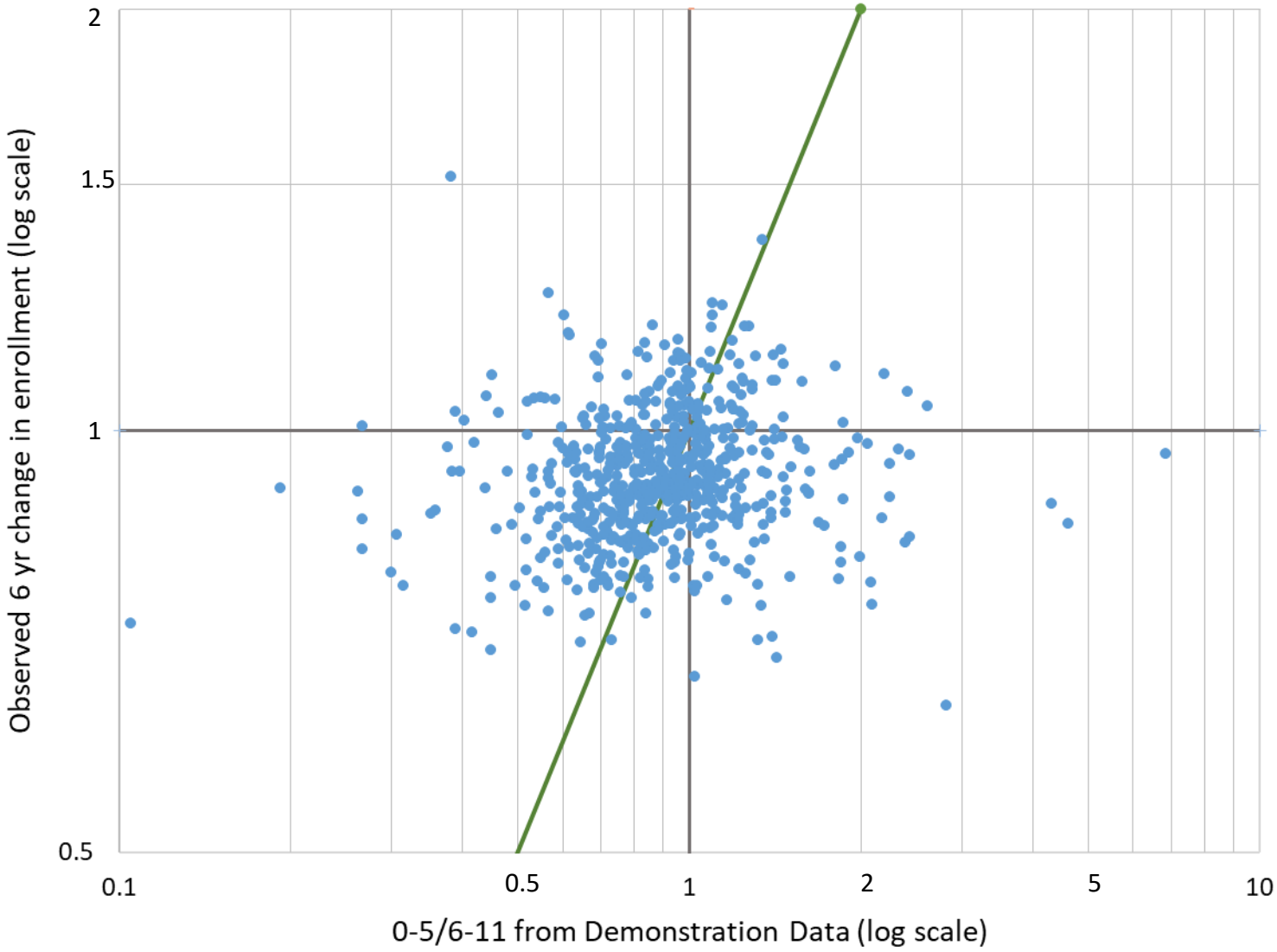
- **Accuracy** refers to the extent to which the data correctly describes the phenomenon they are supposed to measure.
  - Accuracy is often decomposed into **precision**, which measures how similar are repeated measurements of the same thing, and **bias**, which measures any systematic departures from reality in the data.
- **Reliability** is the extent to which the data are accurate consistently over time.

Other factors contributing to accuracy and reliability are **validity**, the extent to which variables in the dataset have values that correspond to expected outcomes, and **consistency**, the extent to which the data are free of contradiction.

Projecting change in enrollment

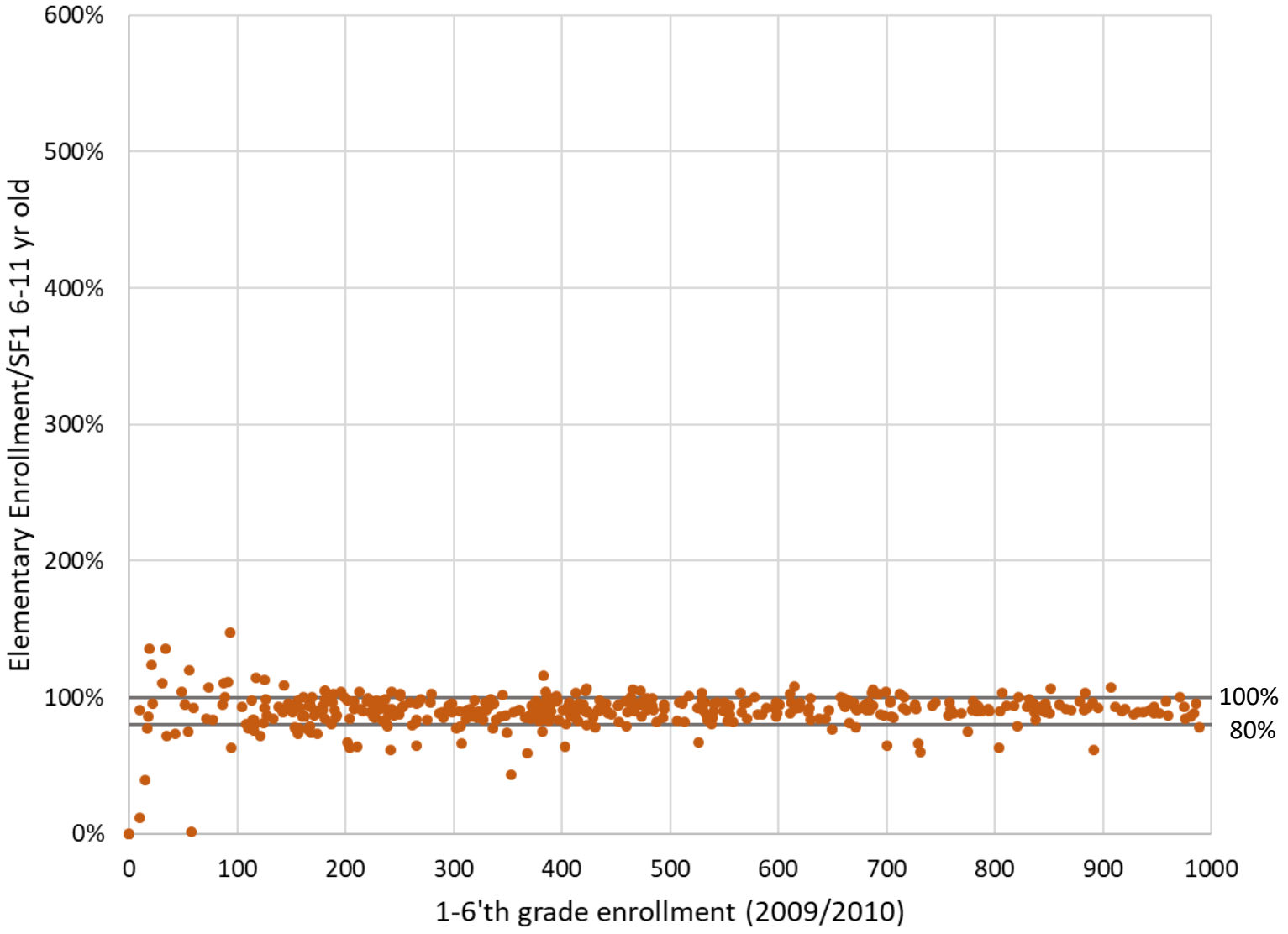
- Current elementary schools (1'st through 6'th grade) are about 6-11 year old
- In 6 year time the current 0-5 yr old will be 6-11 yr old
- $\frac{0-5\ yr\ old}{6-11\ yr\ old}$ is a good indicator of expected change in enrollment
- Compare with observed change in enrollment between 2009/2010 and 2015/2016
- Universe: 425 NY school districts with at least 50 enrolled in 2009/2010

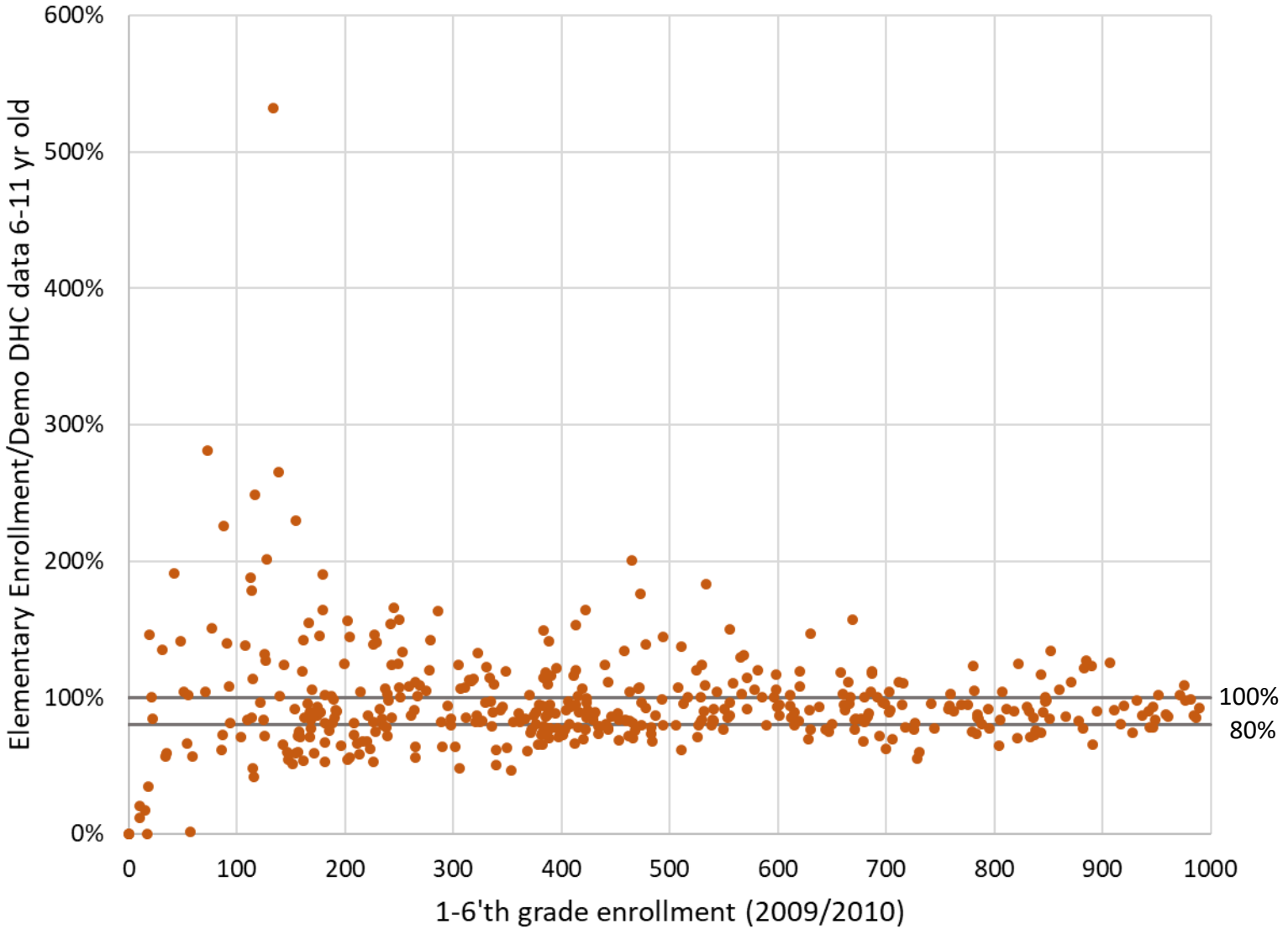

Observed 6 yr change in enrollment (log scale)
2
1.5
1
0.5
0.1
0.5
1
2
5
10
0-5/6-11 from SF1 (log scale)


Observed 6 yr change in enrollment (log scale)
2
1.5
1
0.5
0.1
0.5
1
2
5
10
0-5/6-11 from Demonstration Data (log scale)

Catchment ratios

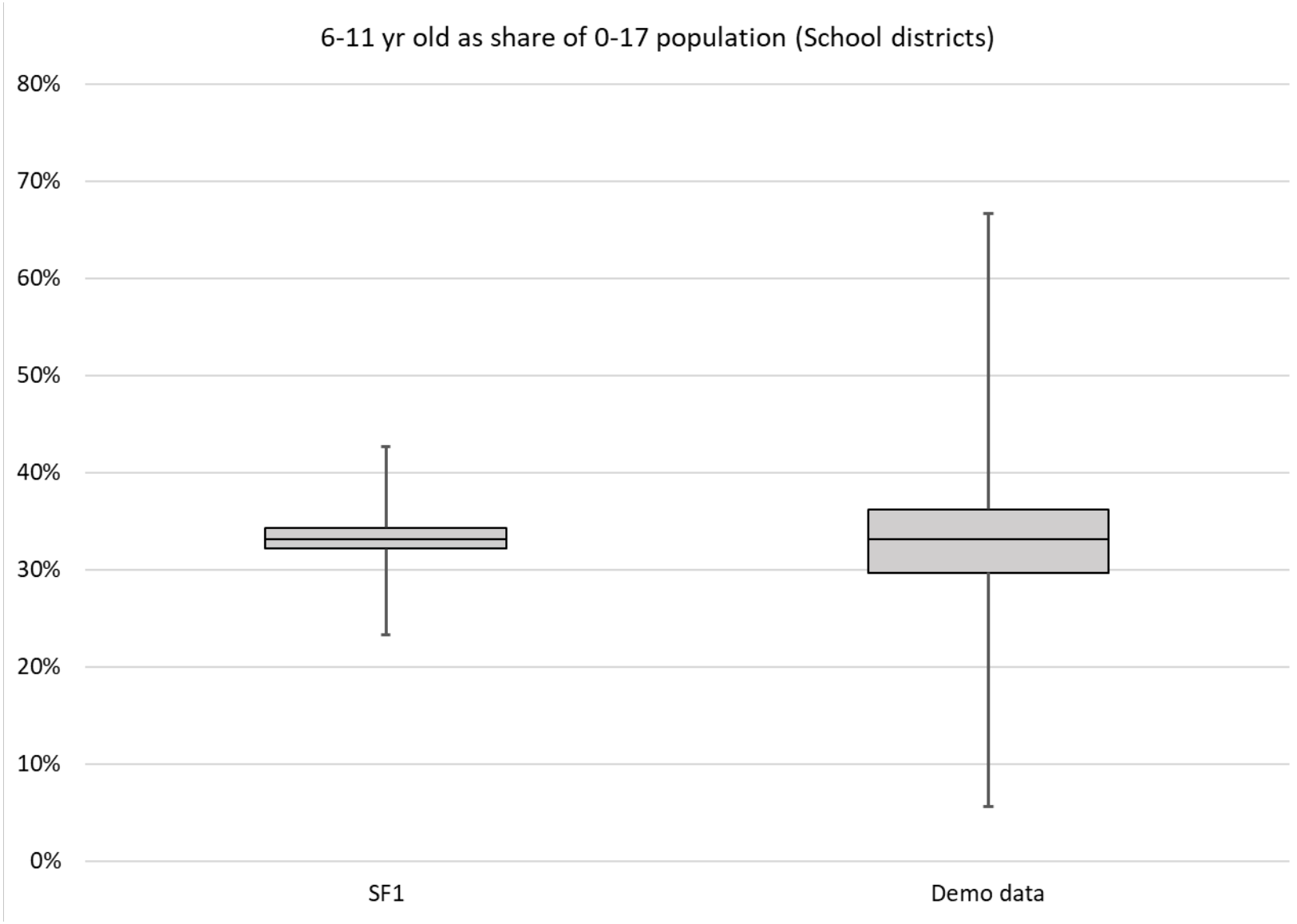
- How many of the kids 6-11 yr old in the school district are enrolled in elementary school (grade 1-6)?
- Universe: NY school districts with 1,000 or less students in elementary school (444)

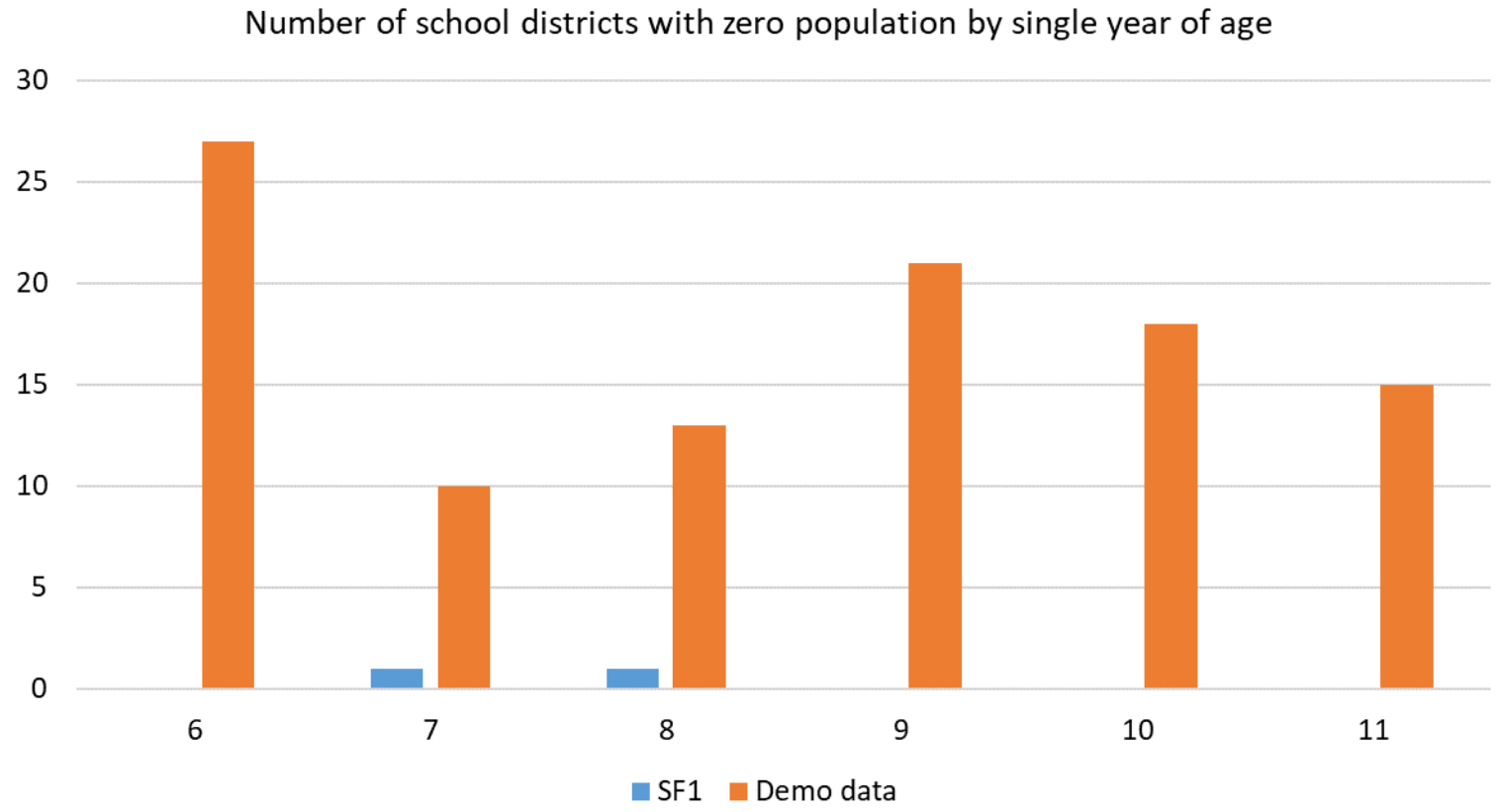

600%
500%
400%
300%
200%
100%
0%
Elementary Enrollment/SF1 6-11 yr old
0
100
200
300
400
500
600
700
800
900
1000
1-6'th grade enrollment (2009/2010)
100%
80%


600%
500%
400%
300%
200%
100%
0%
Elementary Enrollment/Demo DHC data 6-11 yr old
0
100
200
300
400
500
600
700
800
900
1000
1-6'th grade enrollment (2009/2010)
100%
80%

Stats on 6-11 year old

- Comparing share of the 0-17 yr old population in each school district that is 6-11 yr old
- Count number of zeroes for each single year of age
- Universe: NY school districts with at least 50 children age 0-17 (N=684)


6-11 yr old as share of 0-17 population (School districts)
80%
70%
60%
50%
40%
30%
20%
10%
0%
SF1
Demo data

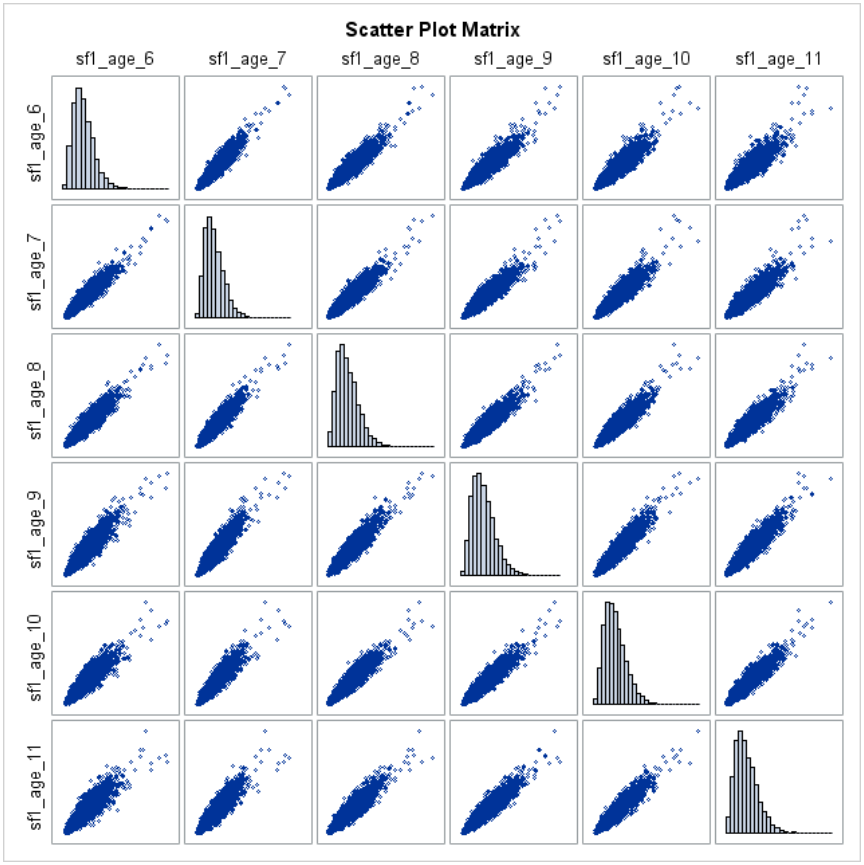


	# SD with at least 1 zero	Max 6-11 pop for SD with at least 1 zero
SF1	2	25
Demo data	62	393

Stats on 6-11 year old for Tracts

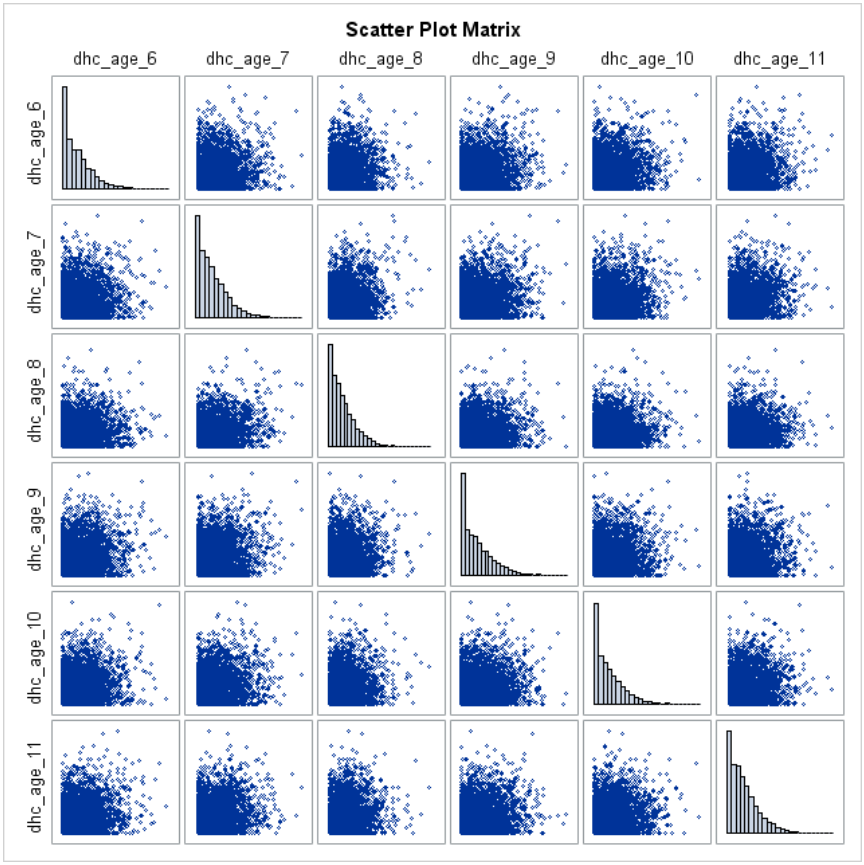
- Comparing correlation between single years of age
- Universe: NY tracts with at least 50 children age 0-17 (N=4775 (SF1), N=4790 (Demo data))

SF1



Covariance Matrix, DF = 4774						
	sf1_age_6	sf1_age_7	sf1_age_8	sf1_age_9	sf1_age_10	sf1_age_11
sf1_age_6	753.5955451	696.2766229	695.7526907	701.9613036	705.8387955	693.3538153
sf1_age_7	696.2766229	749.2440664	694.7432954	705.0641527	711.5661265	701.9891252
sf1_age_8	695.7526907	694.7432954	749.5012140	708.0295953	720.2631937	712.5130243
sf1_age_9	701.9613036	705.0641527	708.0295953	775.4407861	733.3400115	731.2236409
sf1_age_10	705.8387955	711.5661265	720.2631937	733.3400115	801.6438849	744.4950579
sf1_age_11	693.3538153	701.9891252	712.5130243	731.2236409	744.4950579	798.7011803

Demo data



Covariance Matrix, DF = 4789						
	dhc_age_6	dhc_age_7	dhc_age_8	dhc_age_9	dhc_age_10	dhc_age_11
dhc_age_6	2311.851960	529.425070	552.567698	629.862151	639.274834	633.621290
dhc_age_7	529.425070	2313.548270	660.986924	618.210839	680.651565	622.045279
dhc_age_8	552.567698	660.986924	2354.772433	522.264827	545.775478	544.683236
dhc_age_9	629.862151	618.210839	522.264827	2260.175485	563.856135	539.333804
dhc_age_10	639.274834	680.651565	545.775478	563.856135	2459.309288	603.507021
dhc_age_11	633.621290	622.045279	544.683236	539.333804	603.507021	2305.866850

Simple phenomenon

- Random group of N people, each 50% chance of being male and female
- What is the probability both sexes are represented in this group
- Before DP: Prob = $1 - (2/2^N)$
- DP process: Add noise to the possible histograms (#males, #females) by applying two random draws from a Laplace function $f(x|0,b)$
- $\epsilon = 2/b$
- $\epsilon = 0.1$ ($b=20$), $N = 50$ => Prob = 0.73
- $b = 0.3$ ($\epsilon=6.7$) => probabilities similar as before DP for different values of N between 10 and 100

Optimization process

- Does it lead to bias?
- One of the optimization steps (quota rule) looks a lot like Hamilton's method for reapportionment, which was biased towards larger states

Plan for accuracy thresholds

From Count Review Operational Plan, process: Conduct Microdata Detail File (MDF) Review [23-3.5]

"Purpose: Prior to releasing census data for external use, conduct a review of final counts after tabulation recodes of the demographic data are applied and the application of disclosure avoidance procedures are performed to ensure that changes in the counts at multiple levels of geography ***are reasonable***."

High level agreement necessary for metrics on:

- Precision: is MDF still measuring the same thing?
- Bias: is there a bias introduced?
- Validity: are demographic metrics within valid range?
- Consistency: did DAS introduce inconsistencies?

"Concern for man and his fate must always form the chief interest of all technical endeavors. Never forget this in the midst of your diagrams and equations."

Einstein, 1931