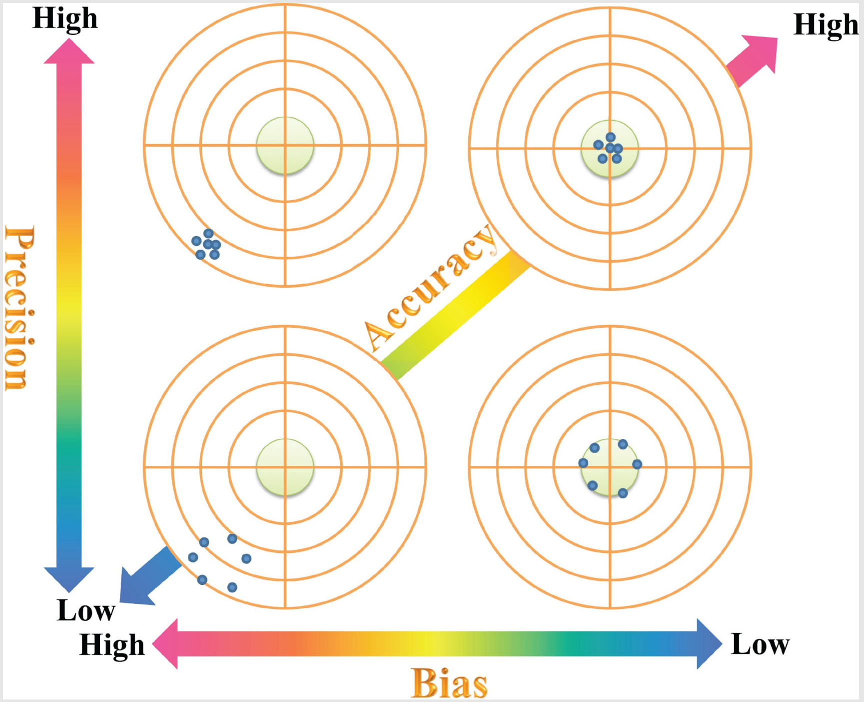
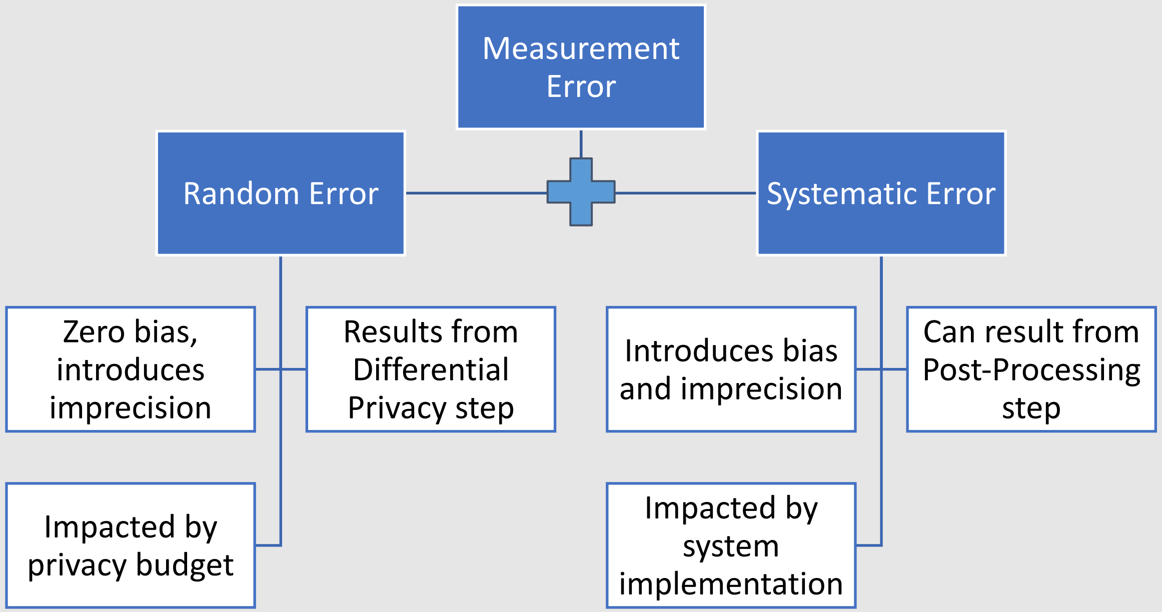
Measuring data quality

With examples from the Census Bureau Disclosure Avoidance System tabulations





Cornell University

Jan Vink

Email: jkv3@cornell.edu

Twitter: @JanVink18

Program on Applied Demographics

CORNELL POPULATION CENTER

Presented at the PAA Applied Demography Conference 2021

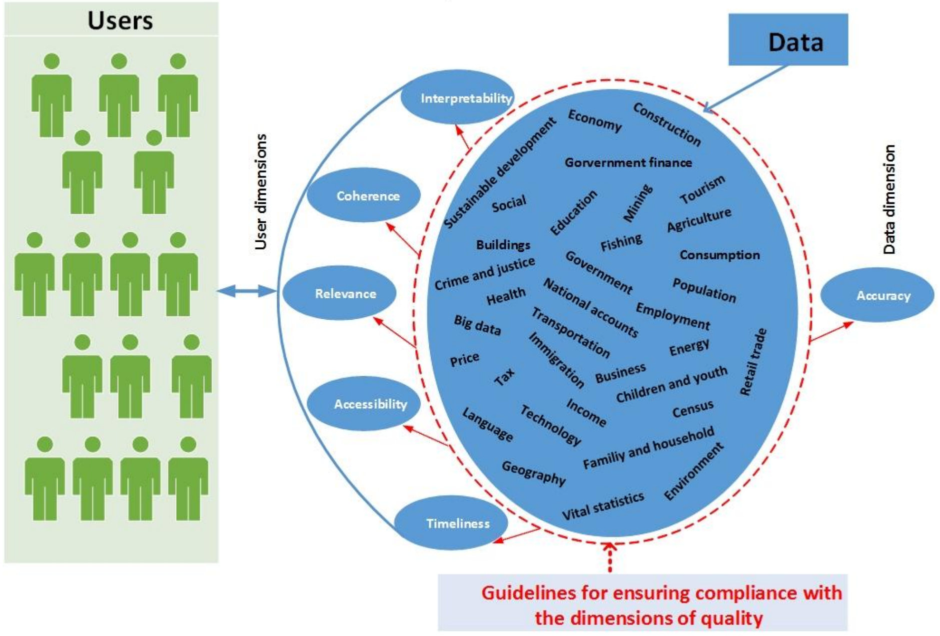
Content

What is data quality?

From Statistics Canada publications on data quality:

- There are six dimensions of quality; namely relevance, accuracy, coherence, interpretability, timeliness and accessibility

Figure1



NOTE:
Accuracy is seen as a data dimension and NOT a user dimension, which makes it possible to measure accuracy without defining use cases

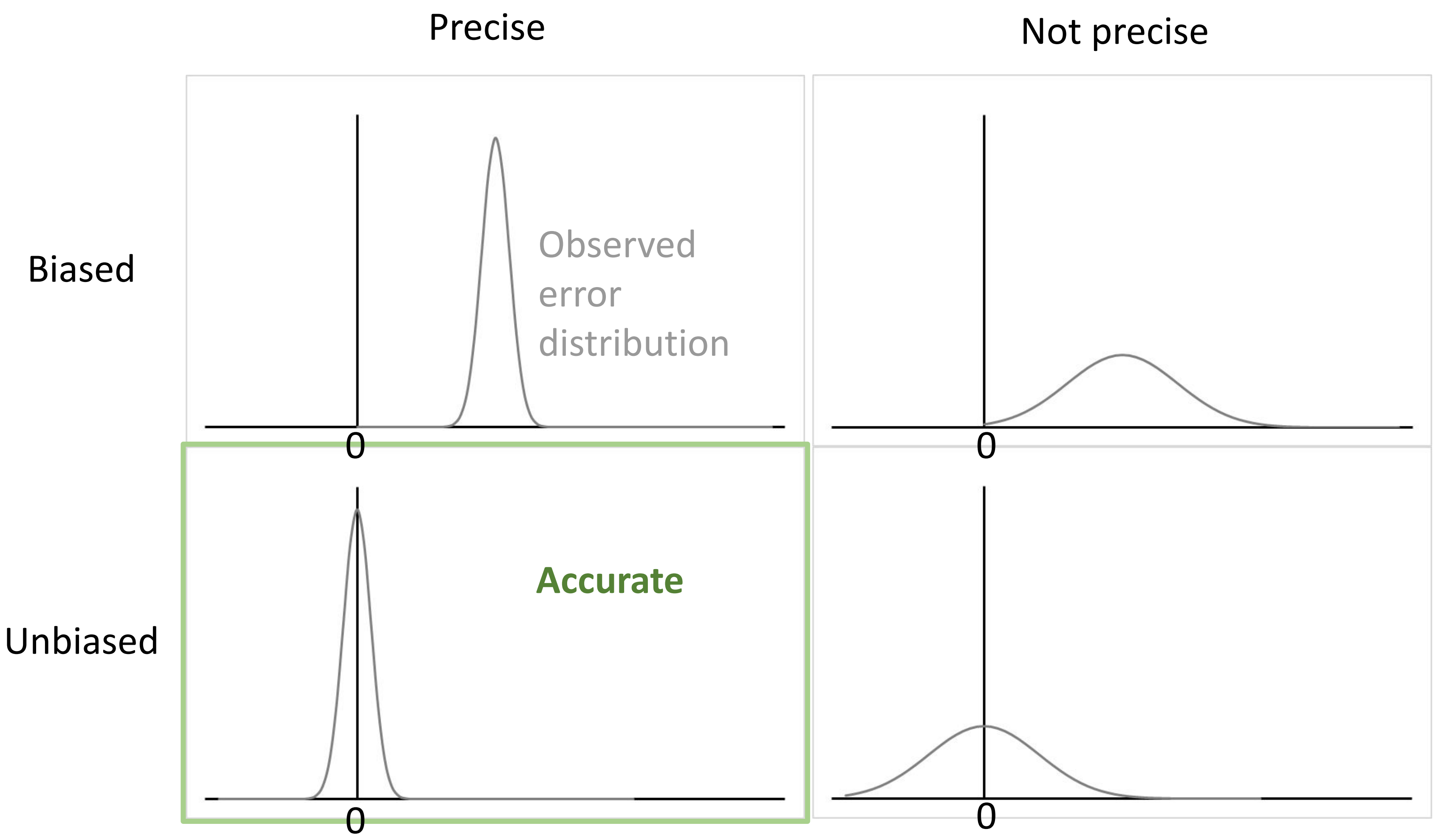
What is data accuracy

Again 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.

Dimensions of accuracy: bias and precision



Error distribution, bias and precision

$$\text{Observed value}_i = \text{True value}_i + \text{Error}_i,$$

where Error_i are observations from an unknown error distribution

- **Bias** is related to the location of this distribution, the expected value
- **Precision** is related to the spread of this distribution, the variability
- **Accuracy** is a function of BOTH bias and precision

Count errors: measuring bias

Common metrics to estimate bias (location) of the error distribution

- Mean of observations
- Median of observations
- We can scale the mean with the mean of the true count to get a measure of the relative bias

Example: Measures of bias in the published counts of persons of American Indian or Alaska Native race on American Indian Home Land (SUMLEV = 250, n=692)

	Demo (release Oct 2019)	PPMF5 (release May 2020)	PPMF11 (release Nov 2020)
Mean Error	-48	-55	-13
Median Error	-22	-22	-1
Scaled Mean Error	-3.4%	-4.0%	-0.9%

Count errors: measuring precision

Common metrics to estimate precision (spread) of the error distribution

- Standard deviation of observations
- Range of outcomes (maximum – minimum)
- Distance between 2 percentiles, e.g. p95 - p5
- Presence of outliers

Count errors: measuring precision

Example 1: Measures of precision in the published counts of persons of American Indian or Alaska Native race on American Indian Home Land (SUMLEV = 250, n=692)

	Demo	PPMF5	PPMF11
Standard deviation	123	138	61
Range	2,234	2,137	910
# outliers (abs(error) >= 25)	365	376	229

Example 2: Measures of precision in the published counts of persons of Non Hispanic Asian race Alone, age 0-17 for Census tracts in New York State (n=4919)

	Demo	PPMF5	PPMF11
Standard deviation	N/A	14.3	34.5
Range	N/A	169	331
# outliers (abs(error) >= 25)	N/A	449	1730

Count errors: measuring accuracy

- Common metrics to estimate accuracy of the error distribution
  - $Mean\ Absolute\ Error = \frac{\sum|Count\ Error_i|}{n}$
  - $Root\ Mean\ Square\ Error\ (RMSE) = \sqrt{\frac{(Count\ Error_i)^2}{n}}$
  - $CV = \frac{RMSE}{\sum True\ Count_i/_n}$
- If $Error_i \sim N(\mu, \sigma)$ then $|Error_i|$ is a folded normal distribution with

$$\mu_Y = \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\mu^2}{2\sigma^2}} + \mu\left[1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right]$$

where Φ is the normal cumulative distribution function:

- One can prove that $RMSE^2 = \mu^2 + \sigma^2$
- Both Mean Absolute Error and RMSE are functions of bias AND precision

Count errors: measuring accuracy

Example: Measures of precision in the published counts of persons of American Indian or Alaska Native race on American Indian Home Land (SUMLEV = 250, n=692)

	Demo	PPMF5	PPMF11
Bias: Mean Error	-48	-55	-13
Spread: σ	123	138	61
Accuracy: MAE	58	64	30
Accuracy: RMSE	132	148	62
Accuracy: CV	9.4%	10.6%	4.5%

Count errors: measuring accuracy

Some thoughts

- **Accuracy metrics are more sensitive to improvements in precision than in bias**, whereas bias might cause more problems
- Outliers can influence metrics for location and for precision.
  - **One can consider using robust metrics for bias and let outliers only influence precision metrics**
- Since accuracy is a function of both bias and precision, publishing metrics on just bias and accuracy masks the precision dimension
  - **Consider making precision metrics more prominent and explicit**

Percent errors: definition

$$Observed\ count_i\ =\ True\ count_i\ + Count\ Error_i$$

$$\frac{Observed\ count_i}{True\ count_i} = 1 + \frac{Count\ Error_i}{True\ count_i},$$

$$Percent\ Error_i = 100\%\ * \frac{Count\ Error_i}{True\ count_i}$$

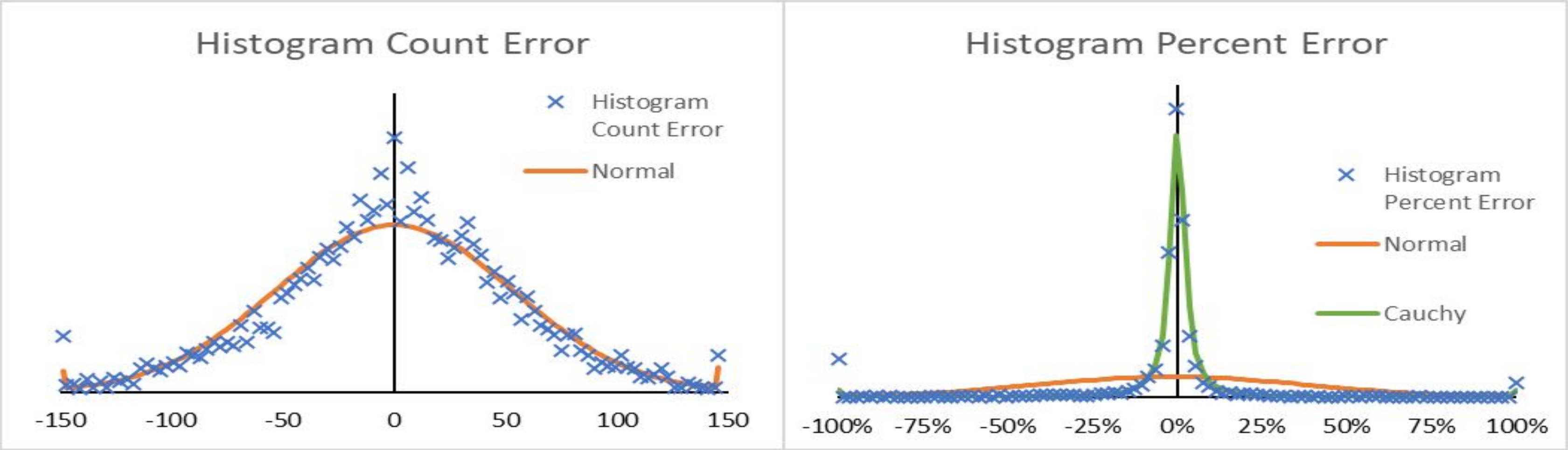
$$If\ True\ count_i = 0\ than\ Percent\ Error_i = 100\%\ * \frac{Count\ Error_i}{t}$$

$$t\ is\ small\ constant, Census\ Bureau\ uses\ t = 0.5$$

Percent error: distribution

- Distributions of count errors and percent errors have very different shapes

Example: count and percent error distributions for Voting age Non Hispanic White alone population in tracts in New York



- Percent error is result of division of two stochastic distributions
  - Quotient of two normal distributions is a Cauchy distribution

Percent error: measuring bias and precision

- The heavy tails of the percent distribution can make average and standard deviation inconsistent estimators of location (bias) and spread (precision) of the distribution
- Alternative measures for bias:
  - Median percentage error
  - Average of the middle quarter of the observations (consistent estimator for location parameter in Cauchy distributions)
- Alternative measures for precision
  - 75’th percentile - 25’th percentile
    - 50% of observations fall within x percentage points of each other
  - 95’th percentile – 5’th percentile
    - 90% of observations fall within y percentage points of each other

Percent error: measuring bias and precision

Example: percent error distribution for voting age Non Hispanic White alone population in tracts in New York

	Demo	PPMF5	PPMF11
Bias: Mean Error	7.2%	2.7%	-1.7%
Bias: Median	0.00%	0.00%	0.03%
Bias: average middle quartile	0.01%	-0.02%	0.02%
Spread: σ	127.3pp	40.9pp	40.1pp
Spread: (p75-p25)	1.36pp	1.79pp	4.6pp
Spread: (p95-p5)	26.6pp	32.2pp	88.5pp

Percent error: measuring accuracy

- Calculating Mean Absolute Percentage Error (MAPE) and RMSE might also suffer problems that arise from the distribution shape
- Alternative measures of accuracy include
  - Median Absolute Percentage Error
  - MAPE-R (using transformations to better deal with the non-symmetric shape)
  - Percent of observations where the percent error exceeds a certain threshold
  - 90’th percentile of the absolute percent error distribution

Percent error: measuring accuracy

Example: percent error distribution for voting age Non Hispanic White alone population in tracts in New York

	Demo	PPMF5	PPMF11
Accuracy: MAPE	10.7%	8.4%	13.9%
Accuracy: Median APE	0.68%	0.90%	2.24%
Accuracy: MAPE-R	0.86%	1.16%	2.9%
Accuracy: PE >=10%	10.7% of observations	13.1% of observations	20.2% of observations
Accuracy: p90	11.1%	16.7%	40.3%

Other aspects of accuracy (describing reality)

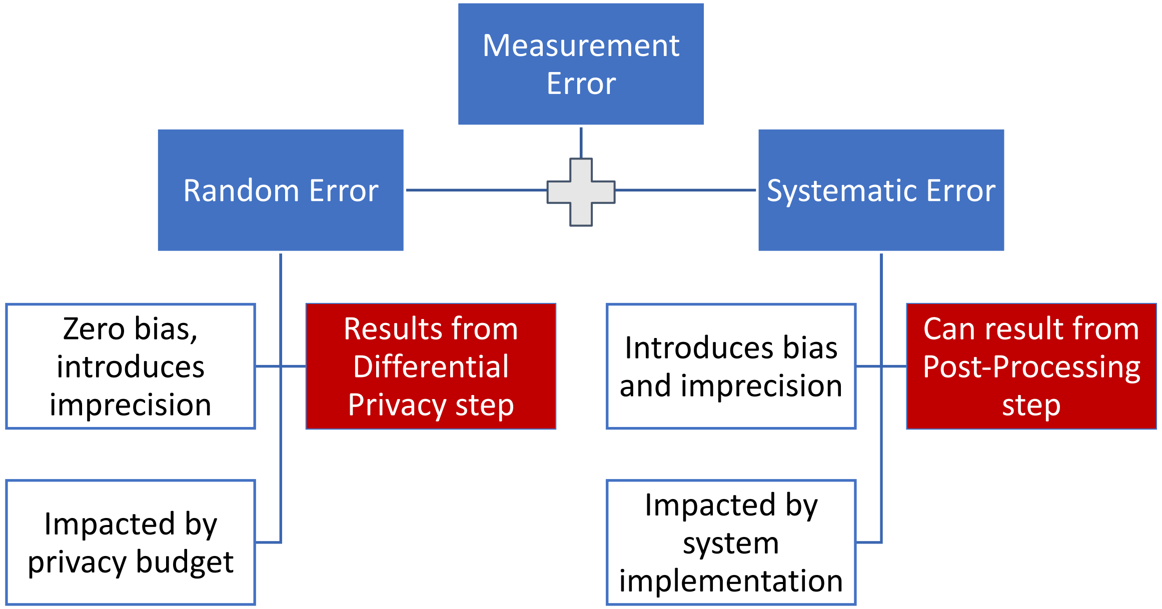
- Accurate **composition** of the population
  - Metric: Similarity index
- Accurate **correlation** between subgroups counts
  - E.g. Count of youth compared to count of adults or count of 4 yr old compared with count of 5-year old
  - Metric: Compare Pearson's correlation coefficient
- Demographically **impossible** or **improbable** observations
  - E.g. toddlers without mothers, sex-ratios equal to 0 or 1, occupied houses without population, population without occupied houses, children in military barracks, seniors in juvenile institutions, etc.
  - Metric: frequency of observation

Sources of error

Measurement theory recognizes two sources of error:

- Random errors: all errors are drawn from the same distribution with zero bias
- Systematic errors: the measurement instrument has a constant bias or the parameters of the error distribution depend on circumstances of the measurement

Sources of error in the Disclosure Avoidance System



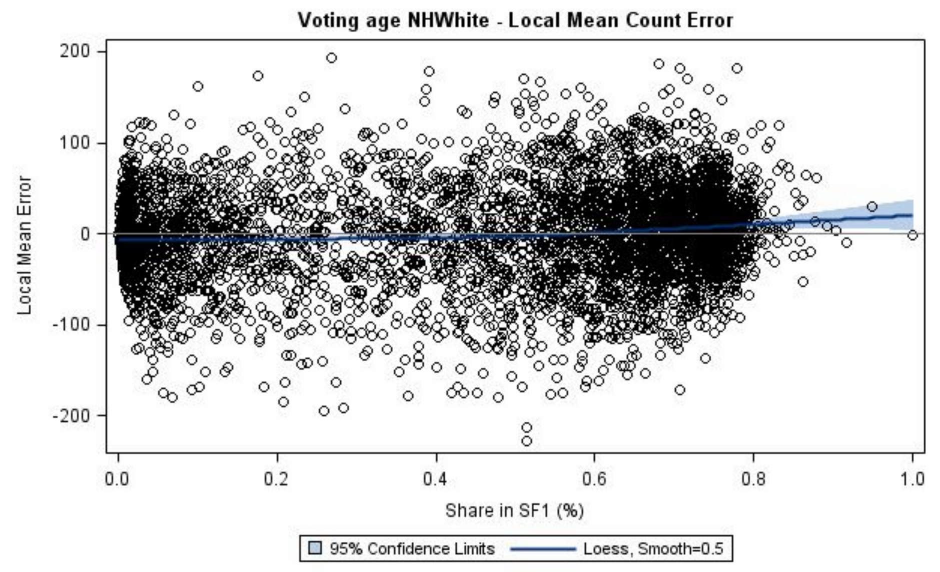
Finding systematic bias

- It helps to have knowledge about the system and potential circumstances that influence the error distribution
- Methods:
  - Split the observations by value of some variable that might cause systematic errors and examine bias for each sub-group.
    - For example, split by population size or % change in population in the case of estimates evaluation
  - Split the observations by geography and think through why some geographies have higher/lower bias than others
  - Locally Estimated Scatterplot Smoothing (**LOESS**)
  - Order the observations by some variable and plot **cumulative errors**

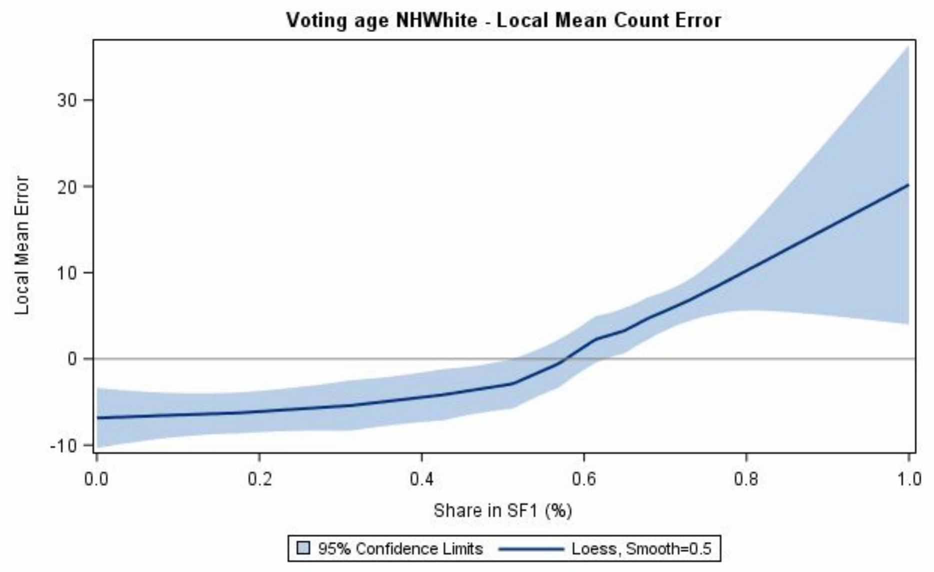
Finding systematic bias: LOESS

Locally Estimated Scatterplot Smoothing (LOESS)

Example: count errors for voting age Non Hispanic White alone population in tracts in New York, share of total population as independent variable



LOESS with scatter plot of individual observations



LOESS without scatter plot of individual observations

Finding Systematic bias: cumulative errors

Step 1: Rank all observations, e.g. share of total population that have a certain characteristic

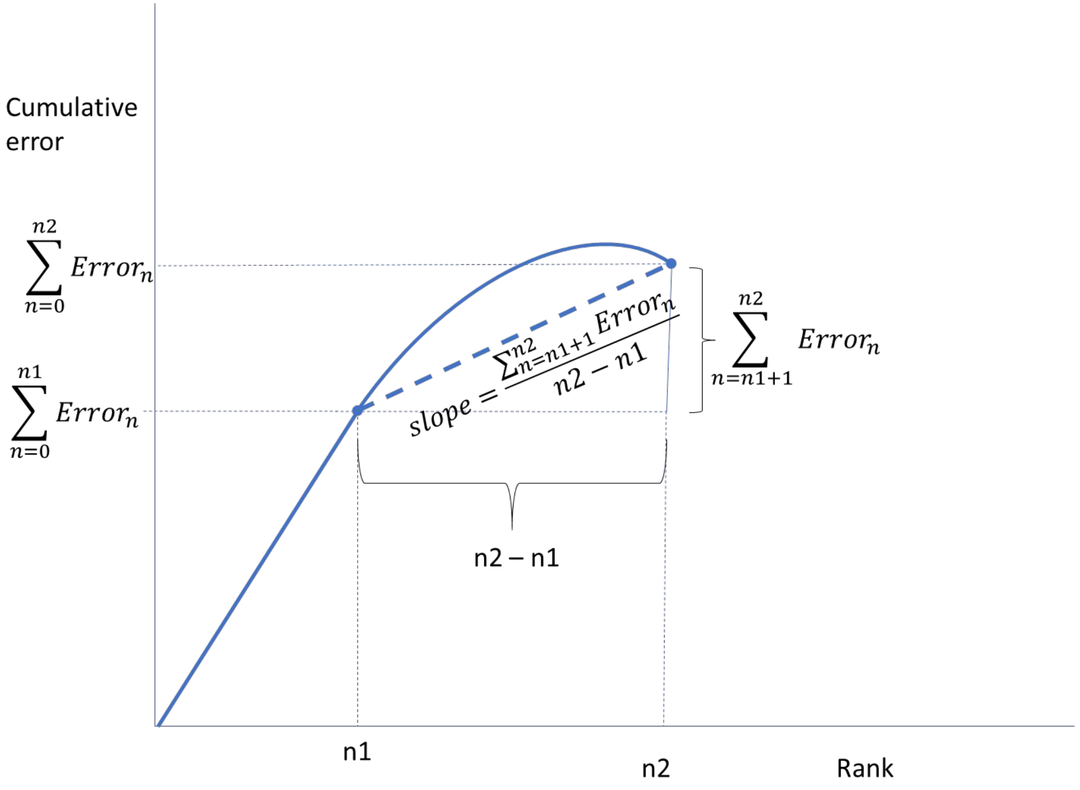
Step 2: For each rank r, calculate

$$Cumulative\ error_r = \sum_{i=1}^{r} count\ error_i$$

Step 3: Add (0,0) and plot (r, cumulative error_r)

If the error is solely random, the cumulative error would be a random walk

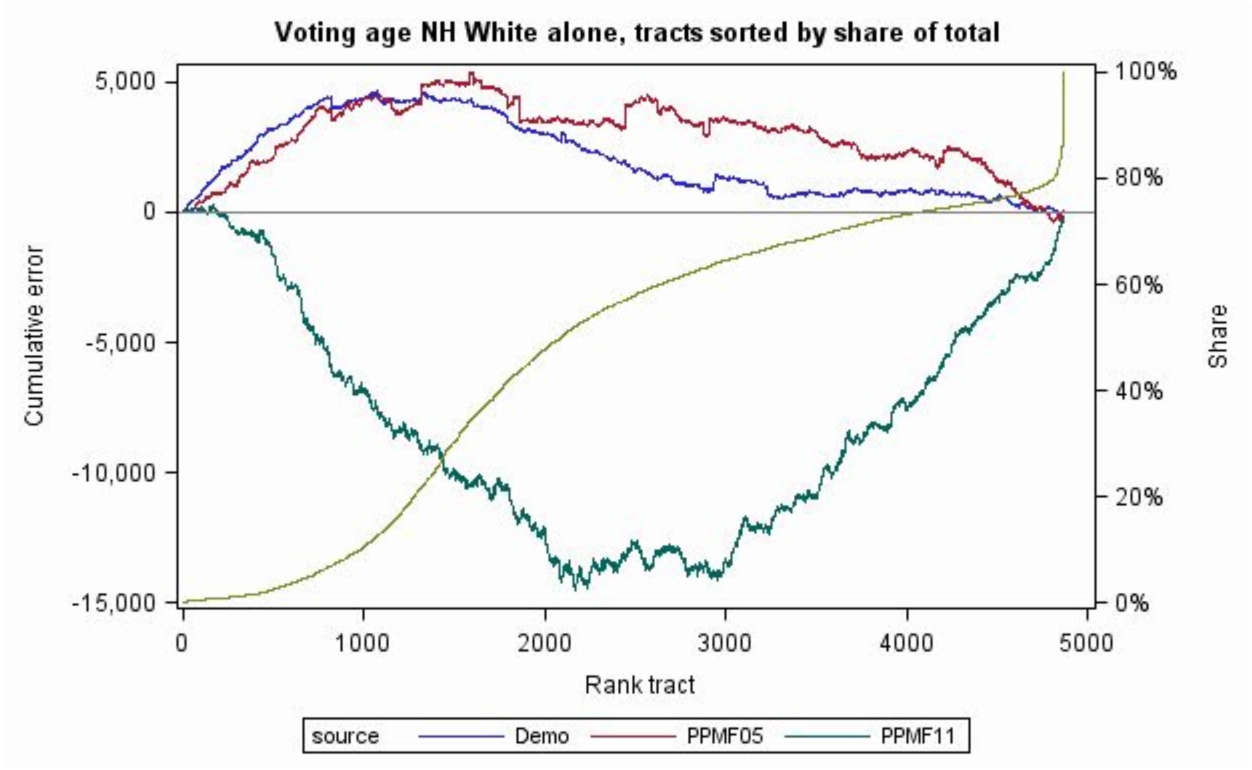
Finding Systematic bias: cumulative errors



- The **slope** of the line between points is the **average error** of the observations between those two points
- The slope of the line connecting (0, 0) with the last point is the overall mean error
- Maximum cumulative errors are related to CUSUM tests

Finding Systematic bias: cumulative errors

Example: Cumulative count errors for voting age Non Hispanic White alone population in tracts in New York, share of total population as independent variable



The PPMF11 line corresponds with the LOESS example and shows again the negative bias (negative slope) for observations with relatively few persons of the subgroup and a positive bias for tracts with relative many persons of this subgroup.

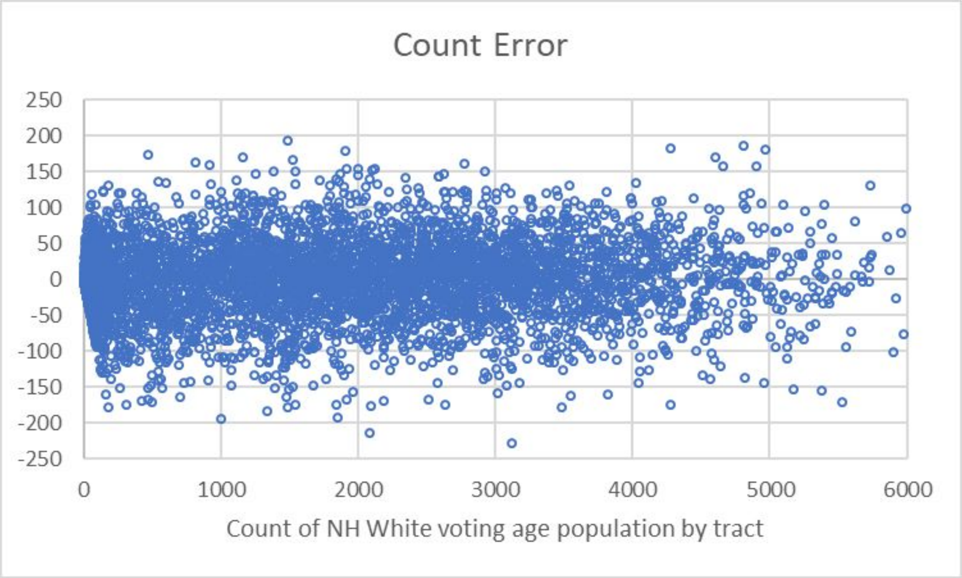
The direction of the systematic and the magnitude was different in the PPMF11 file than in the previous releases

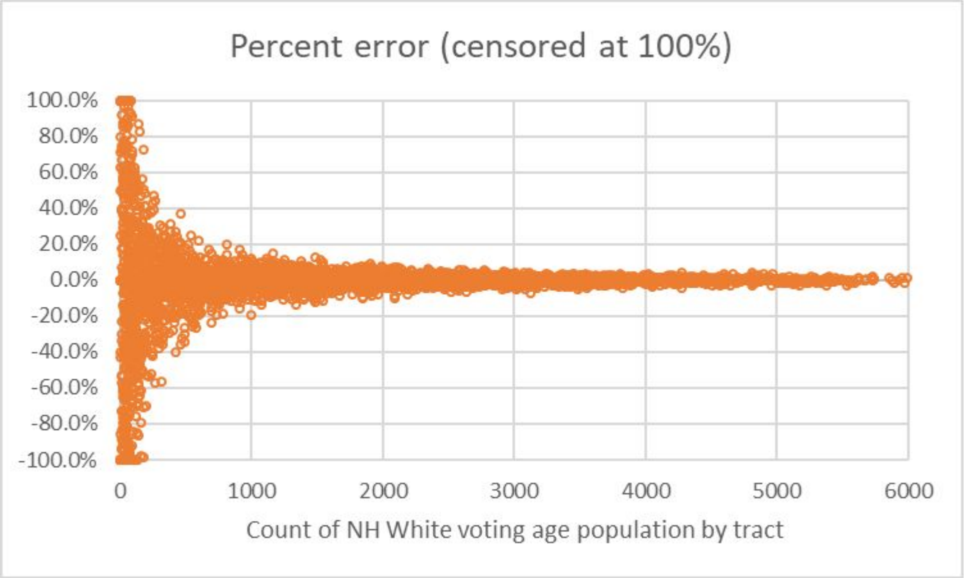
Systematic imprecision

- Count error distribution and percent error distribution can NOT both have constant precision for all True Value$_i$
  - This implies that one can expect more variation in percent errors at smaller X values and thus more imprecision and less accuracy
- Testing for heteroscedasticity in count errors can bring systematic imprecision to light as can finding patterns in squared or absolute errors

Systematic imprecision

Example: count and percent errors for voting age Non Hispanic White alone population in tracts in New York (tracts with count less than 6,000)





Conclusions

- There is added value in examining precision as a dimension of accuracy
- Outliers can cause average errors to mask true bias (location parameter of the error distribution)
- Count errors and percent errors have very different shapes and cannot both have constant precision (and accuracy) for different values of the true count
- It is possible and important to detect systematic errors and compare system variants based on the size of systematic errors