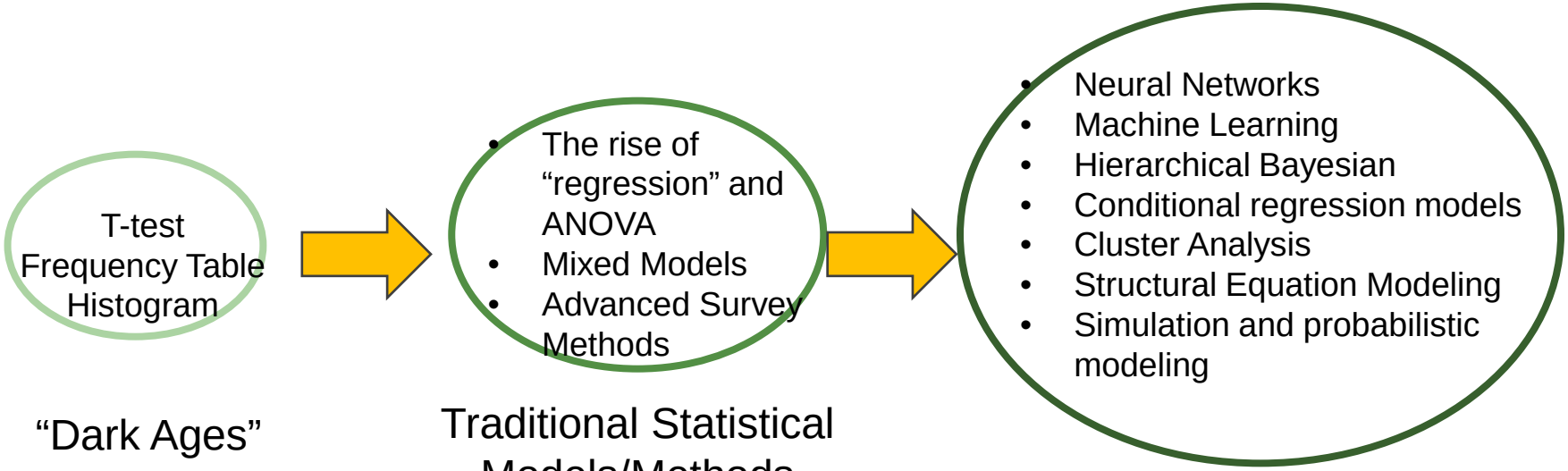
Ben Pierce
Battelle
pierceb@battelle.org
(614) 315-8911

Adopting and Fusing Administrative Records with Traditional Survey Based Data to Increase Accuracy and Timeliness

Datapalooza: Policy Development in the Era of Big Data

June 17, 2015

Its Not only Data that is Getting Bigger!



With advanced computing power it is possible to do many things that were only theoretical statistical exercises before.

100 Years of Survey Statistics

1. Avoid bias in your estimate
2. Seek to minimize the variance in your sample estimates
   - Increase your sample size
   - Improve your sample design (e.g., stratification)
3. It is generally impossible to sample the universe
4. Do the best you can to minimize under-coverage, frame miss-coverage, etc.

   Use post-stratification weighting techniques

The Golden Rule of Statistical Estimation:

Minimize Mean Squared Error

Mean Square Error (or uncertainty) of the survey estimate, which is the Expected value of the squared difference between the estimate and its true value:

$= E[(\theta - \hat{\theta})^2]$ *or*

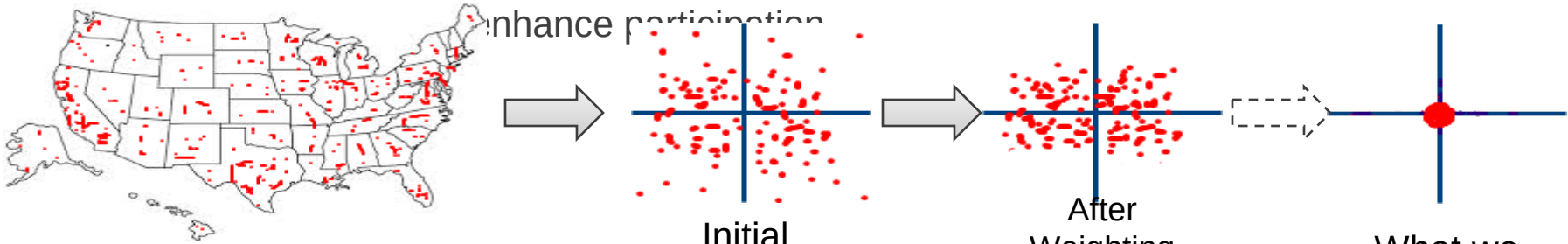
$= E[(\theta - E(\theta))^2] + (E(\theta - \hat{\theta}))^2$ *or*

$=$ *Variance + Bias*^{2}

So, what has this led to?

- Large, national surveys with sample sizes in the tens of thousands
  - Complex sample designs using multi-stage sampling
  - Extensive post-stratification and weighting methods

$$Variance(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N})$$



Now, a quick story.

Who’s a better shot?

Me

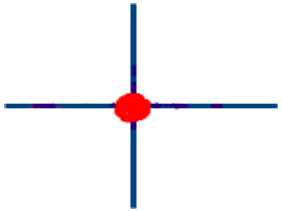
Not that Accurate, but pretty Precise

My Son

Not that Precise, but relatively Accurate

Where is this going?

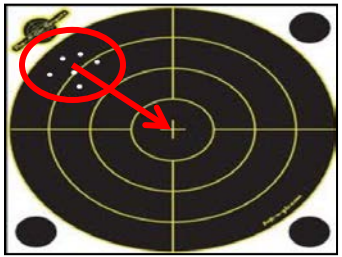
- Remember what we want in our estimates:



- How can we get there in the two previous outcomes?

Try to squeeze out more of the variation

- Larger sample sizes
- Stratification



Recalibrate to reduce or eliminate bias

- Find a reference point
- Use better math

So how does this relate to fusing administrative data with survey?

Survey Data

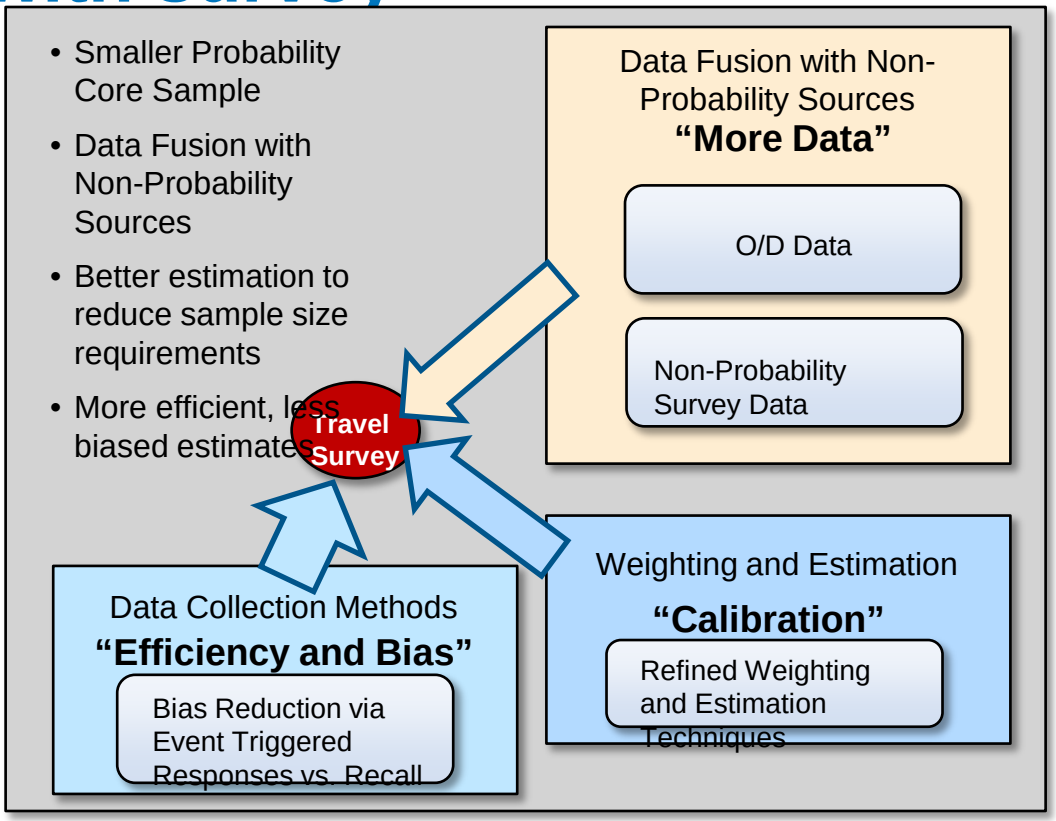
- Unbiased
- Representative
- Not necessarily precise
- Take time/resources to obtain

Administrative Data

- Biased
- Not necessarily representative
- Usually much more precise
- Usually obtainable in much less time/resources

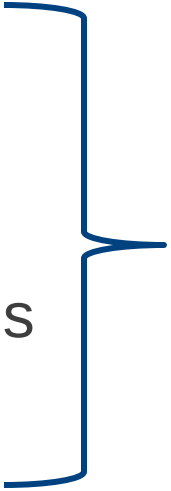
So how does this relate to fusing administrative data with survey

- Collect a large amount of highly-precise travel occurrence data from cell-phone usage or other administrative, "non-probability" records
- Collect a smaller amount of unbiased, accurate, but not as precise travel data via probabilistic survey



Other non-probabilistic data ripe for the plucking

- “Opt-in” panels
- Online surveys open to anyone
- Convenience or judgmental samples
- Network sampling
- Cell phone O/D data



“Cheap”
“Quick”
Fairly Precise

But

Biased

Will incorporating non-probabilistic data work?

- A test of the methodology using the 1995 American Travel Survey data*
  - Long distance trips (>75 miles one-way) over a 12-month period
- Used computer simulation to sample person-records repeatedly from this dataset
  - Obtain a **probabilistic core sample**
  - Obtain a **opt-in sample under various degrees of sampling bias**
    - Simple random sampling
    - Higher proportion of young (<35 years) frequent travelers
    - Higher proportion of young earners and young educated persons
  - **Different sample sizes** (1,000; 7,000; 15,000 persons)

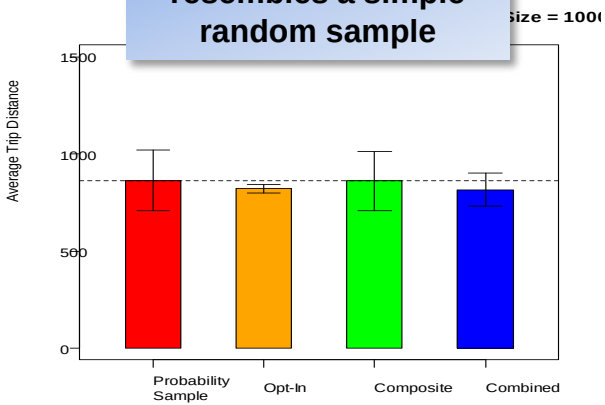
* Research sponsored by FHWA Office of Policy

Methods

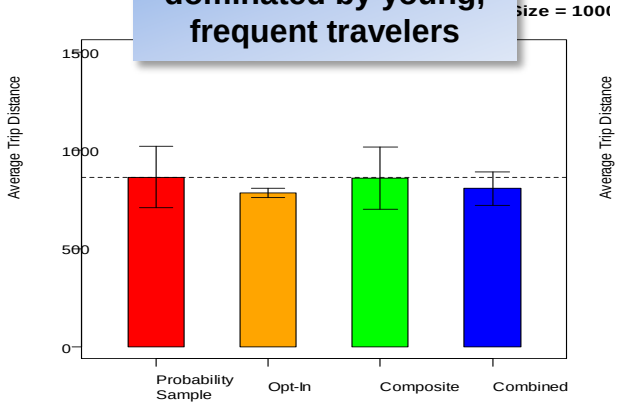
- Examined two simple methods for statistical estimation
  - ***Combined approach*** (Disogra et al., 2009): Iterative post-stratification and reweighting of combined probabilistic/non-probabilistic data
  - ***Composite approach*** (Ghosh-Dastidar et al., 2009): Combine mean estimates from both data sources
- Trip parameters of interest:
  - Average # trips, average trip distance, proportion of LD travelers
- Calculate bias and mean squared error (MSE) of the mean estimates
  - ***Probabilistic core sample estimate used as the reference in defining bias*** (thus, it has bias 0)

Summary of Research or Data Experience

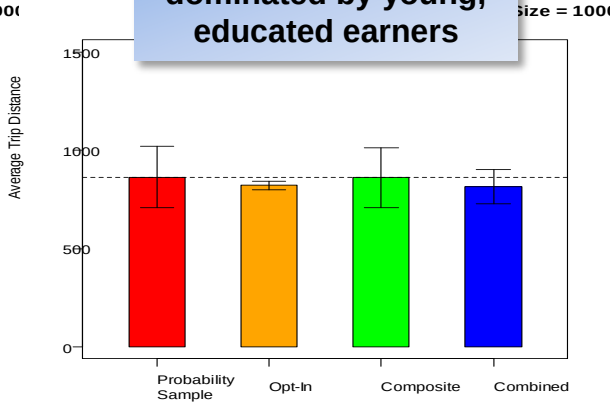
Opt-in sample resembles a simple random sample



Opt-in sample dominated by young, frequent travelers



Opt-in sample dominated by young, educated earners



Average Trip Distance

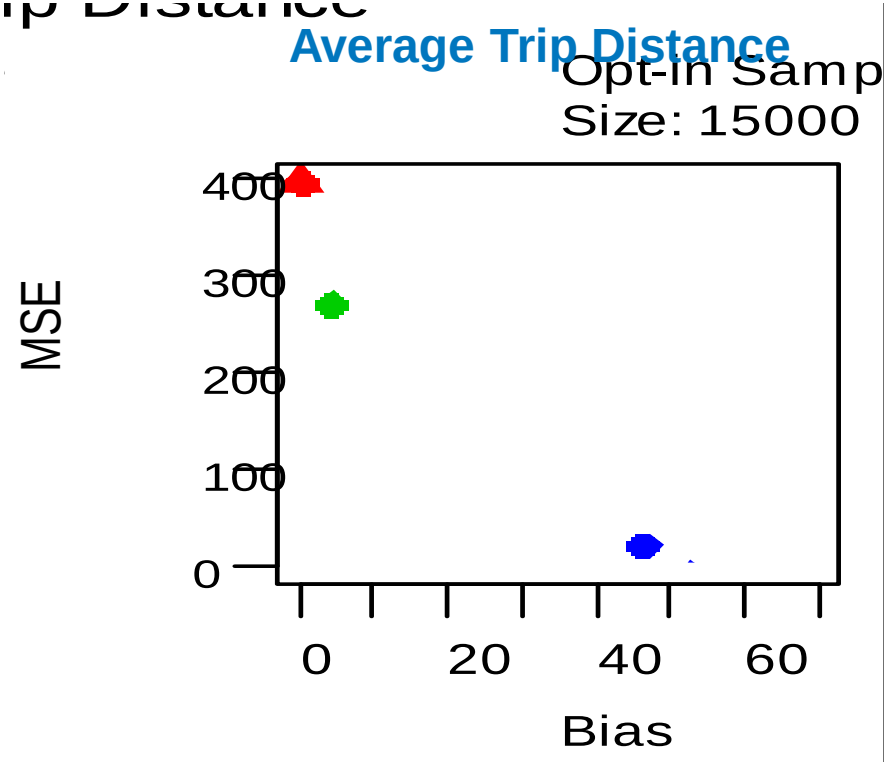
Probabilistic-based sample size = 1,000
Opt-in sample size = 15,000

Equal or slightly less variability, with an acceptable level of bias

Summary of Research or Data Experience

Composite estimation approach

- **Reduces MSE (always)**
- **Introduces vary little bias even when 15 to 1 ratio**



Legend:

= Probabilistic core sample

= Composite estimation approach

= Combined estimation approach

The Bottom Line

The Good

- Can be a very powerful way to combine data from different sources to improve survey estimates
  - Leverage data to improve estimates
  - Good for non-sensitive estimates
- Potential for a much more robust database, without an exponential increase in resources needed
- Will continue growth of efficiency and applicability as technology growth continues

The Bad

- The statistical methods can get complicated quickly
  - Processor intense = more resources and time for analysis
  - More challenging to interpret to a layperson
- Auxiliary data usually has a lower level of detail
  - Lacking trip purpose, driver, etc.
- Bias can still be an issue for sensitive estimates

What Does this Mean for Policy Decisions?

The use of non-probabilistic records/convenience samples can be used to enhance traditional survey results

- Extend the “life” of the data
- Enable “rapid” investigation and support of emerging policy questions
- May not provide a “perfect” answer – but it may well be “good enough”

Resource Intense Survey Data

and

“Big Data” and Administrative Records

plus

Some Fancy Statistical Methods

=

Support for Policy Decisions

Contact Information/Publications

- **Ben Pierce** (614-424-3905; PierceB@battelle.org)
- **Bob Lordo** (614-424-4516; LordoR@battelle.org)

DiSogra, C, Cobb, C, Chan, E, and Dennis, JM. (2011) Calibrating non-probability internet samples with probability samples using early adopter characteristics. In: *Section on Survey Research Methods, JSM Proceedings.* Alexandria, VA: American Statistical Association.

Ghosh-Dastidar, B, Elliott, MN, Haviland, AM, and Karoly, LA. (2009) Composite estimates from incomplete and complete frames for minimum-MSE estimation in a rare population: an application to families with young children. *Public Opinion Quarterly.* 73(4):761-784.

Project Report: "Design of a Completely New Approach for a Household-Based Long Distance Travel Survey Instrument" http://www.fhwa.dot.gov/advancedresearch/pubs/13081/