Internet-use Propensity for Matching Probability and Non-Probability Samples: the “Fac-sample”

Charles DiSogra
Andrew Burkey

AAPOR – Austin TX
May 14, 2016

A Very Specific Situation

- You already have non-probability Web panel cases
- You used a non-probability source because
  - “rare” target population
  - efficient reaching a large number of the target population
  - rapid data collection needed
  - cost-effective
- You had no better alternative to study the target population

Dilemma

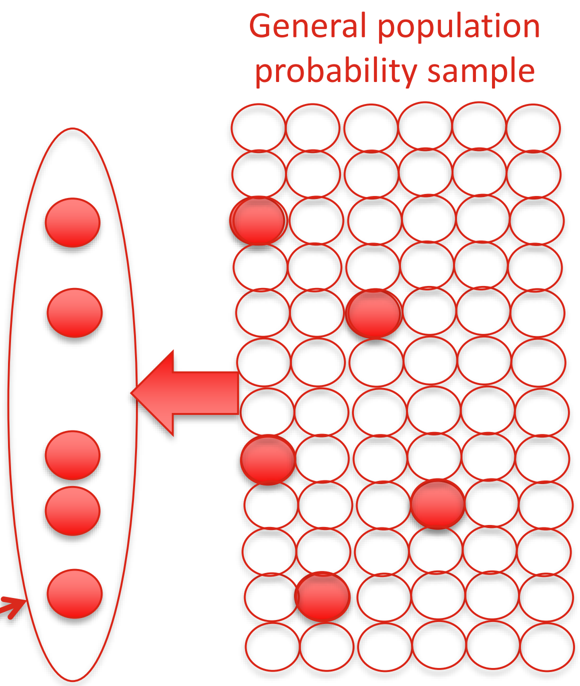
- You need to compute a confidence interval for your data
- But these are non-probability cases!

Possible solution

- Make a facsimile of a probability sample as follows:
  - ✓ Treat your large number of non-probability cases as a source pool
  - ✓ Match cases from the pool to existing probability sample cases
  - ✓ Use a propensity score as the matching metric
  - ✓ Propensity to be a non-daily Internet user

Find a probability sample! This is key

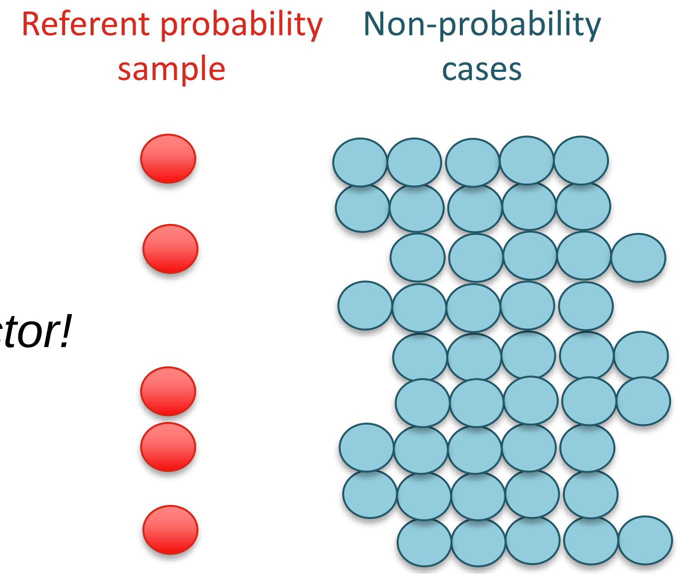
- Identify a probability sample from a population which includes your target group
  - a domain within the larger sample
  - Examples: pregnant women, teachers, a specific health condition, healthcare personnel, LGBT
- Identify eligible cases in the probability sample that meet your survey criteria
  - These cases become your referent sample



Identify common variables in your "sample" and the probability sample

- Examples of common variables are
  - age
  - gender
  - education
  - home ownership
  - children in household
  - income
  - …. etc.

Note: common variables are a constraining factor!



Designate a propensity variable to model – Non-daily Inernet User

- The non-probability source is an opt-in Web panel
  - ALL cases have Internet access
  - Assumed to be Daily Internet users
  - Coverage error = Non-daily users and users not on panels
- The probability source may be a general population sample
  - Cases consist of “Daily” and “Non-daily” Internet users
- Step-wise regression tells us which of our common variables are significant for predicting Non-daily Internet users

Compute a propensity score

Using the combined referent and non-probability cases

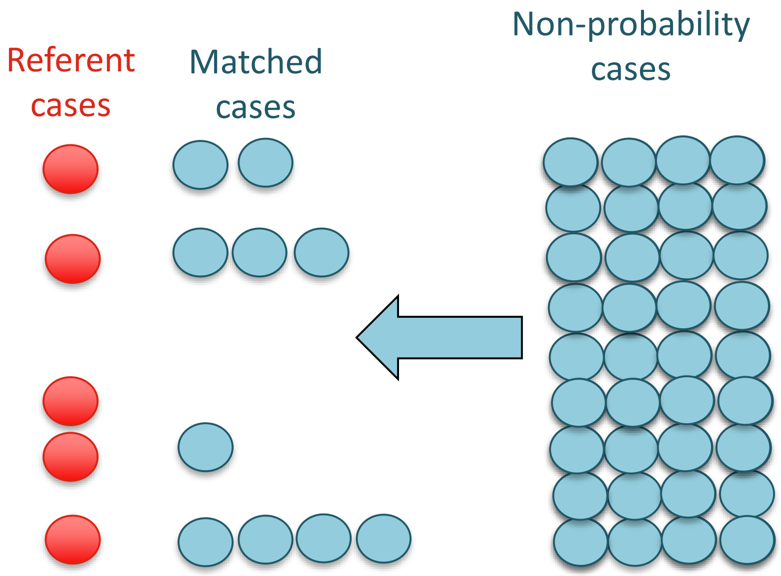
- Compute the probability of a Non-daily Internet user (p)

$$\text{Log}\ (p/(1\text{-}p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots. + \beta_k X_k$$

- Round the propensity scores to achieve a robust match rate *(>80%)* of the referent cases to best approximate the probability cases

Find your matched cases

- Make exact matches based on the rounded propensity score
- Use only the non-probability cases that match
- Note that you may find multiple matches



Resolve multiple matches

- A weight share* adjustment (ω_i) for each matched non-prob. case

$$\omega_i = \frac{(\text{number of } y_i \text{ referent cases in a match})}{(\text{number of } x_i \text{ non-prob. cases in same match})}$$

Example: 1 referent matches to 2 non-prob. = 1/2 = 0.50 $= \omega_i$

$$\sum \omega_i x_i = \sum y_i$$

Sum weighted matched cases = number of referent cases

* Deville JC, Lavallée, P. Indirect Sampling: The Foundations of the Generalized Weight Share Method. *Survey Methodology*, 32:2 pp165-176, 2006.

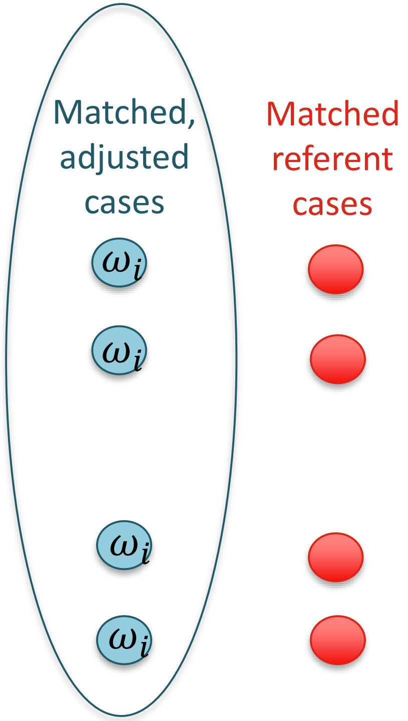
Our "Fac-sample" is made!

- The Fac-sample is theoretically

  "one of any number of possible samples that can be drawn from the population of interest"

- Fac-sample is next weighted to target population benchmarks

- *An approximated SE of the estimate and confidence interval for the population value can now be calculated!*



Some limitations

- Must have identified a suitable probability referent sample
- Results hinge on Internet usage propensity
- Propensity score matching is restricted to available variables
- Not all referent cases are matched
- Possible mode effects between probability referent sample and the non-probability sample

Conclusions

- Non-probability cases can be made a facsimile of a probability sample using a propensity matching procedure
- A confidence interval around the "Fac-sample" can be calculated
- Inherent bias likely still exists in the non-probability sample
- More work needs to be done with this *ex post facto* approximation

Internet-use Propensity for Matching Probability and Non-Probability Samples: the “Fac-sample”

Charles DiSogra
c.disogra@srbi.com

Apples to Oranges or Gala vs. Golden Delicious?

Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples

David Dutwin, Ph.D., SSRS
Trent D. Buskirk, Ph.D., MSG

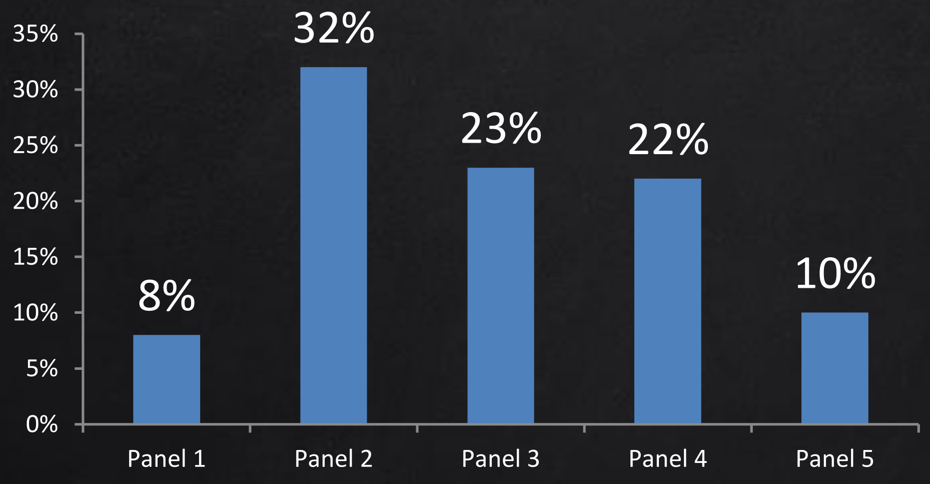
AAPOR, 2016

What’s the Problem, Exactly?

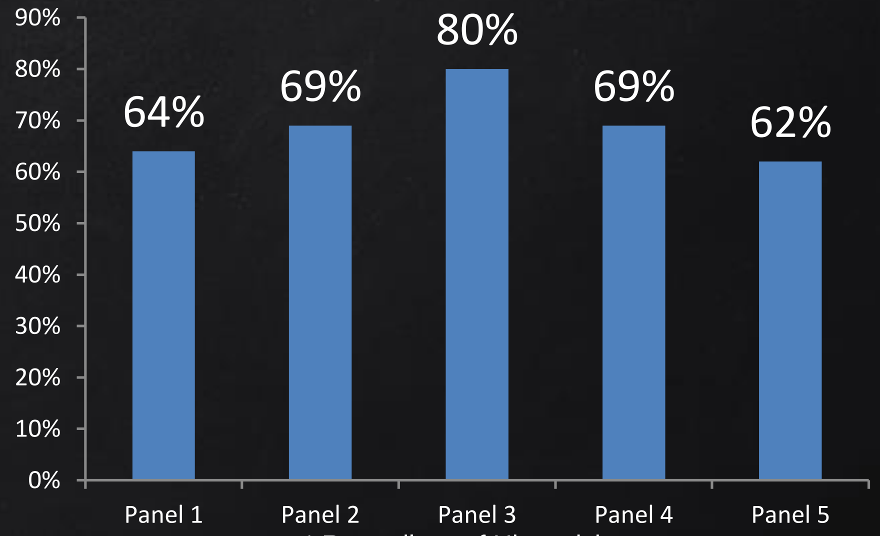
- Non-probabilistic data sources can have self-selection bias over and above what probability panels might have:

- As such, considerable variance in the universe of web panels:

Percent of Web Panelists 18-24 from 5 Leading Convenience Panels



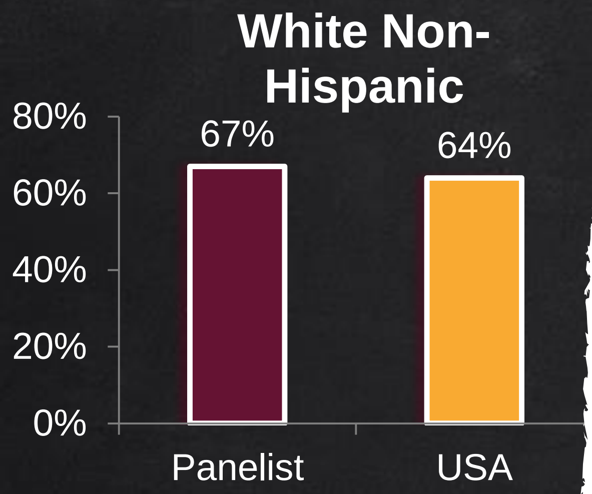
Percent of Web Panelists White* from 5 Leading Convenience Panels

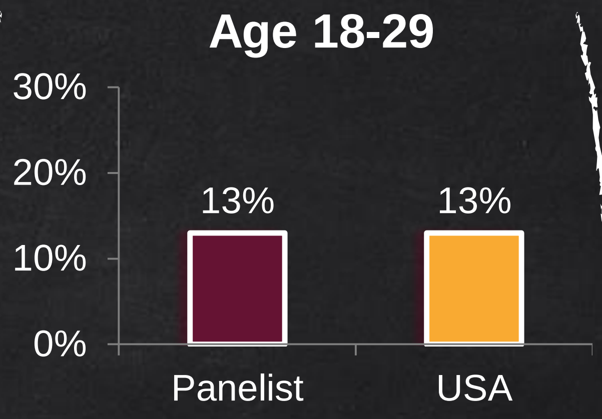


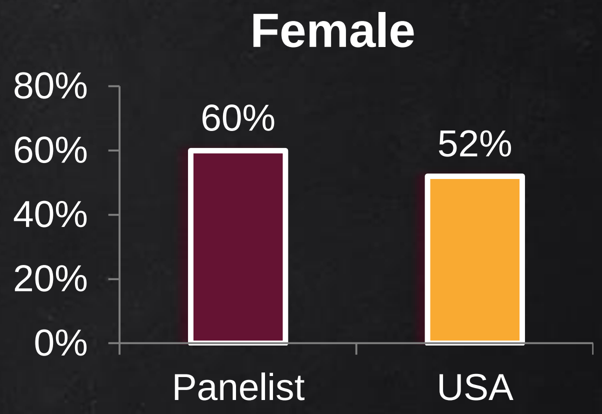
* Regardless of Hispanicity

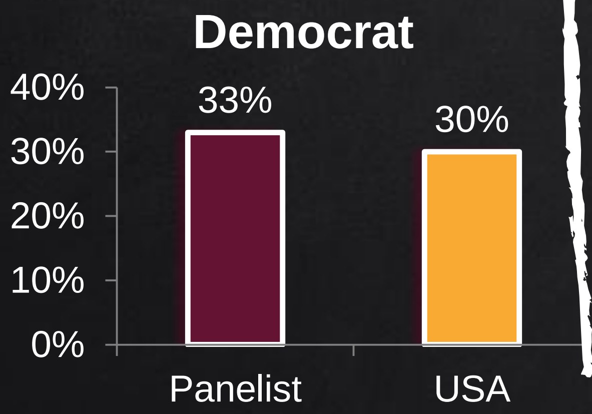
Who are the Web Panelists Anyway?

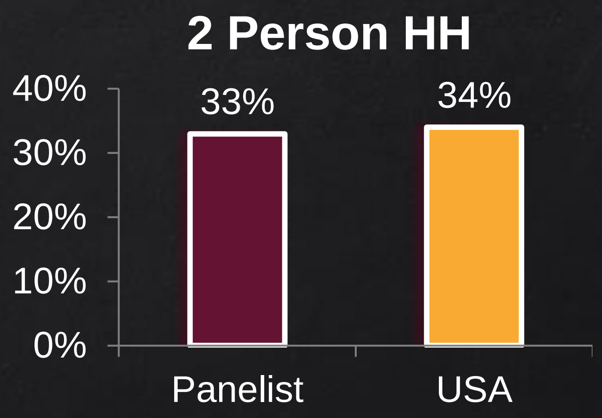
- Well, they look a lot like Americans…

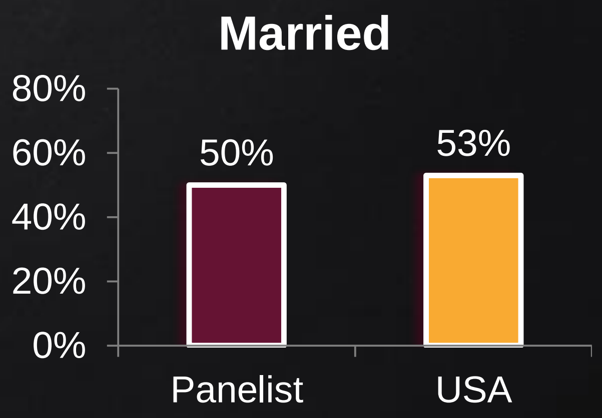






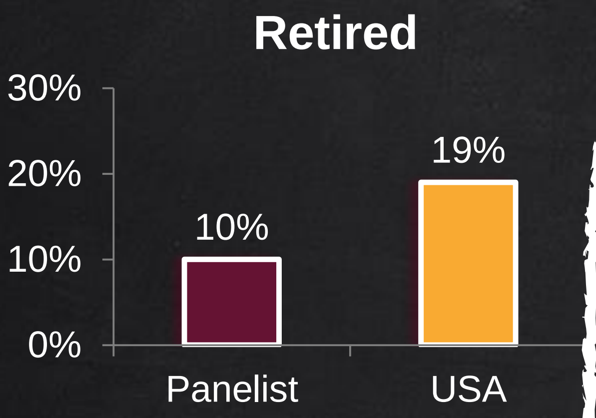


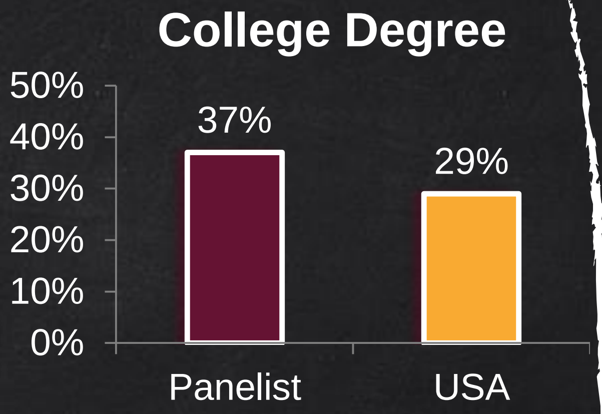


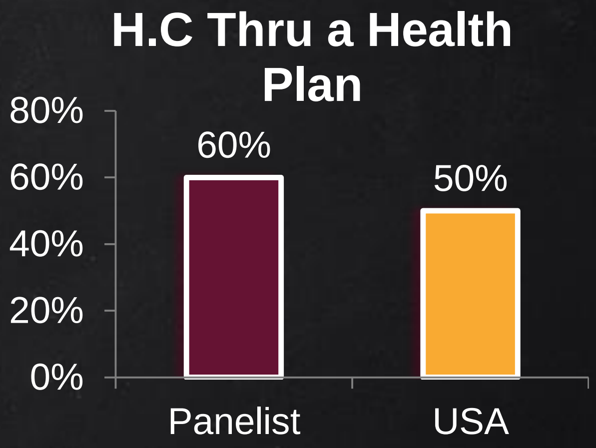


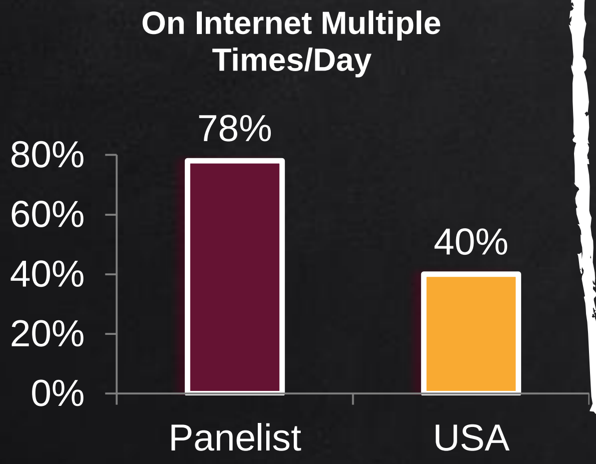
Who are the Web Panelists Anyway?

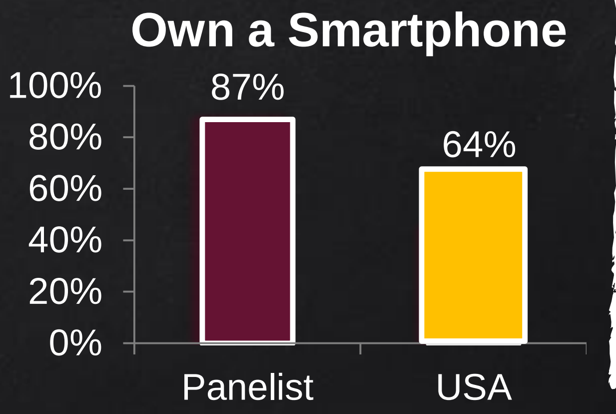
- Except that they don't…

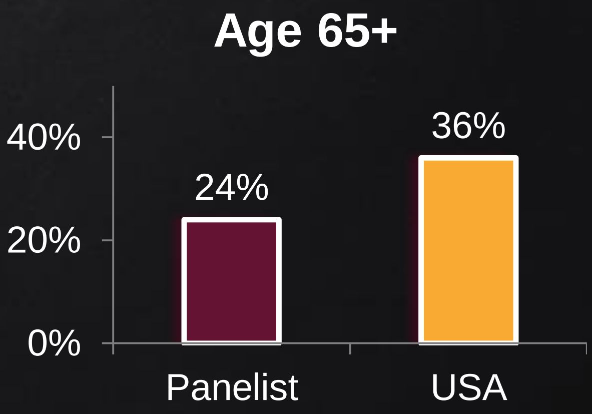












Cost vs. Quality?

- Probability Samples cost more
- When compared head to head, estimates from probability samples tend to be more accurate than those from non-prob samples
  - ☒ Callegaro et al., 2014;
  - ☒ Yeager et al., 2011
  - ☒ Krosnick and Chang, 2009
  - ☒ Walker et al., 2009
- Response rates continue to decline
- Non-Probability seductively cheaper
- Non-probability vary in execution, recruitment and quality
  - ☒ Pew Research Report, 2016
- Methods based on modeling, weighting and matching continue to emerge to improve quality of estimates from non-prob samples.
  - ☒ Terhanian et al., 2016
  - ☒ Dever et al., 2015
  - ☒ DiSogra et al., 2015
  - ☒ Rivers and Bailey, 2009
  - ☒ Dutwin and Buskirk, 2016

Our main Research Questions

- How does the quality of non-probability samples compare to that of low-response rate probability samples?

- Can we improve the quality of estimates from non-probability samples using alternative adjustment methods like calibration, propensity weighting or sample matching?

Data Sources

- Dual-Frame RDD Telephone Sample 1
  - ☒ n = 107,347; ~9% Response Rate.
- Non-probability Web Panel 1
  - ☒ n = 82,478; Response Rate unknown.

1st Sample Set

- Dual-Frame RDD Telephone Sample 2
  - ☒ n = 29,153; ~12% Response Rate.
- Non-probability Web Panel 2
  - ☒ n = 61,782; Response Rate unknown.

2nd Sample Set

- All samples selected and fielded between October 2012 thru 2014

Methods for Adjusting Non-Probability samples

Raking/ Calibration

- Base Weight = 1 other than Dual-Frame which gets single frame estimator (Best and Buskirk, 2012).
- Raking to Education, Region, Gender, Age, Race/Ethnicity
- No Trimming

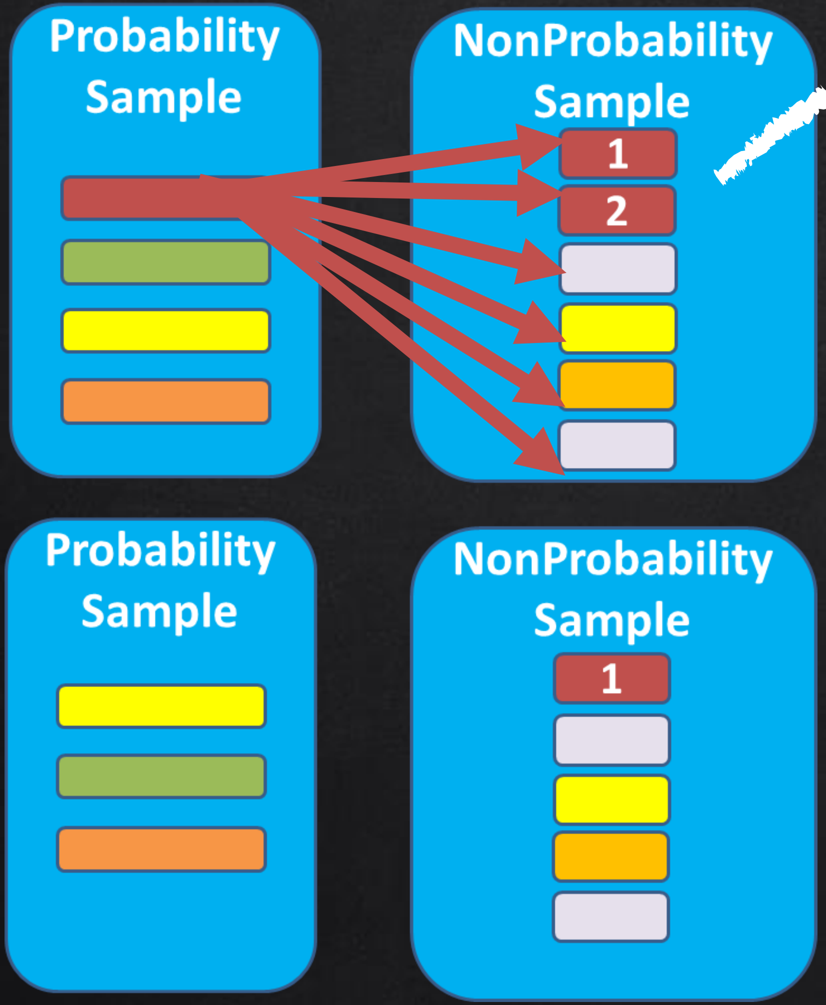
Propensity Adjustments

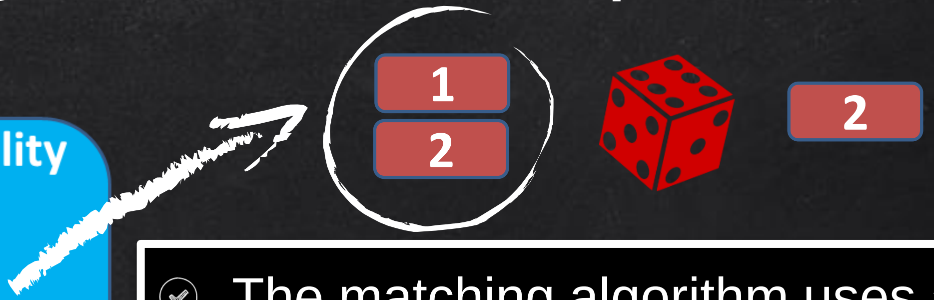
- Conducted only in NP-Panel #2 which included webographics
- Logistic Regression Model: Education, Region, Gender, Age, Race/Ethnicity, Metro Status, # of Adults, and Webographics
- Tested both raw scores and a weighting class (5) variant

Sample Matching

- Based on pairing non-probability cases with members of a probability sample
- Sample matches based on similarity across core set of common variables (Rivers and Bailey, 2009)
- Applied to both NP panel samples

Creating Matched Samples





- The matching algorithm uses a collection of categorical variables from each case in the Prob. sample and computes a similarity index defined as the simple matching coefficient (SMC) to each case in the NP sample.
- The matched case is randomly selected from among those identified as the most similar.
- See: http://bit.ly/1Fb2Jhs for specific details of computing the SMC for categorical variables.

Generating the Matched Sample

An 3.5% SRS of the Adults contained in the 2013/2014 CPS Public Release Data file were matched to sampled units in the combined NP Panel 1/Panel 2, respectively sample based on 8 common demographic variables including:

- Region (North, South, East, West)
- Male
- Age Group (18-29, 30-49, 50-64 and 65+)
- Race Group (White (NH), Black (NH), Hispanic, Other (NH))
- Education Level (<HS, HS, Some College, BS, BS+)
- Own/Rent (Own, Rent/Board)
- Marital Status (Married, Single, Partnered, Divorced/ Widow/ Separated)
- Employment Status (Currently Employed or not)

Primary Metrics

Absolute Bias

- The mean absolute bias (MAB) – computed as the arithmetic mean of absolute value of the difference between the table estimate and the corresponding benchmark estimate
  - ☒ the mean is taken over the total number of estimates within the variable set

St. Deviation of Biases

- The standard deviation of the absolute biases computed from each variable set was also computed.
  - ☒ Provides a sense of the variability in the level of biases.

Overall Average MAB

- The overall average MAB – computed as the mean of the 12 MAB statistics computed across the 12 variable sets for each sample

Key Outcomes We Consider

- Common set of survey variables across the sources of data include household and person-level demographics
- External benchmarks for media related information contained in the main survey source are not commonly available
- Given this scenario, we will focus our evaluation and computation of bias metrics on distributions of one demographic variable ***within*** levels of a second
  - ☒ What is the distribution of Education ***within*** each level of Race
  - ☒ What is the distribution of Race ***within*** each level of Education
- The reference/benchmark values are computed using the 1 Year PUMS Data from the 2012 American Community Survey.

Evaluated Outcome Variable Sets

Specific Demographic Variable Sets of interest include:

- Education (5 levels) within Race (4 levels) and
  - ☒ and Race within Education
- Education (5 levels) within Age-group (4 levels)
  - ☒ and Age-group within Education
- Education (5 levels) within Region (4 levels)
  - ☒ and Region within Education
- Age-group (4 levels) within Race (4 levels)
  - ☒ and Race within Age-group
- Age-group (4 levels) within Region (4 levels)
  - ☒ and Region within Age-group
- Race (4 levels) within Region (4 levels)
  - ☒ and Region within Race

Computing the Primary Metrics

Consider the demographic cross tabulation of Race and Region producing a 4-by-4 table. Taking the absolute value of the difference between the row percentages and the corresponding benchmarks from CPS produces a total of 16 absolute bias measures. (Distribution of Region within Race)

Race	Midwest	South	West	Northeast
White				
Black				
Other				
Hispanic				

Row Percentages

4 absolute bias measures

4 absolute bias measures

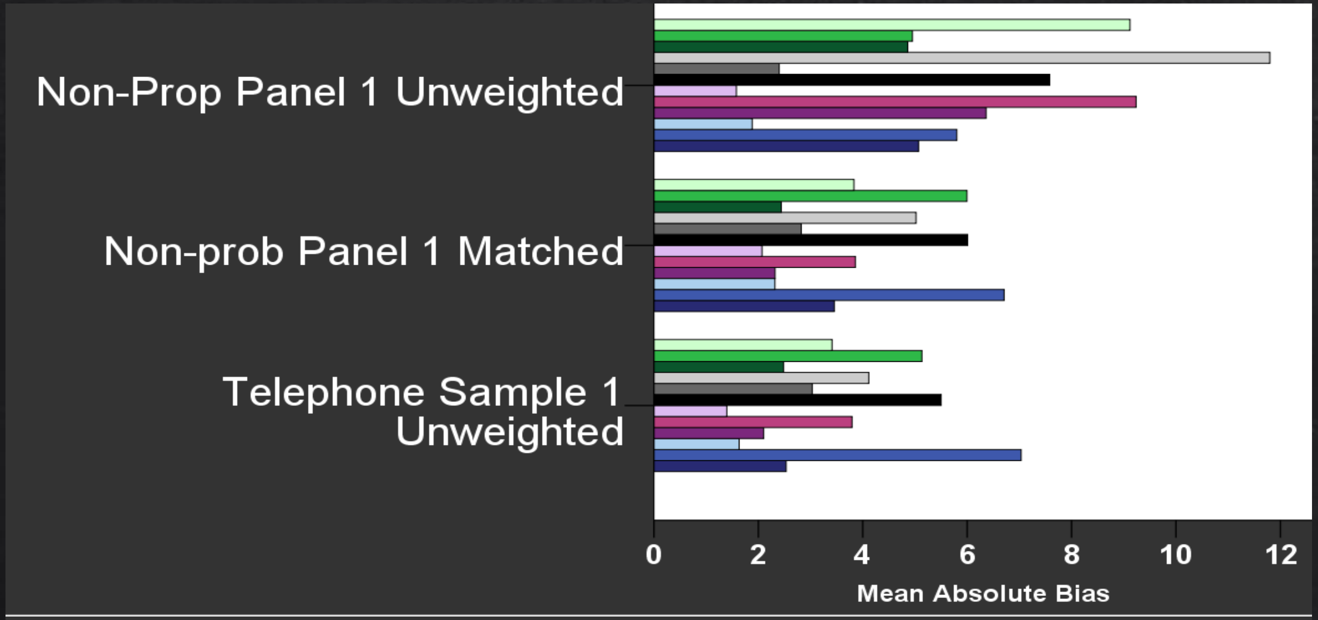
4 absolute bias measures

4 absolute bias measures

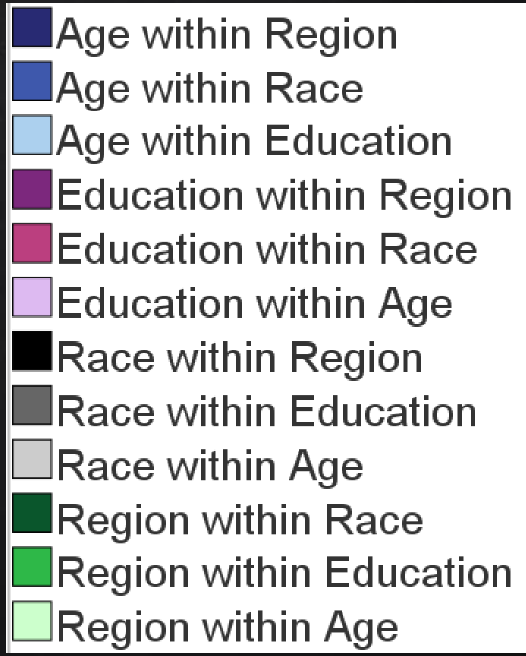
The average of these 16 bias measures represents the Mean Absolute Bias (MAB) of **Region within Race.**

Repeating the calculations for each of the column percentages (Distribution of Race within Region) yields the MAB for Race within Region.

MAB Statistics for Demographic Cross-Tabulations

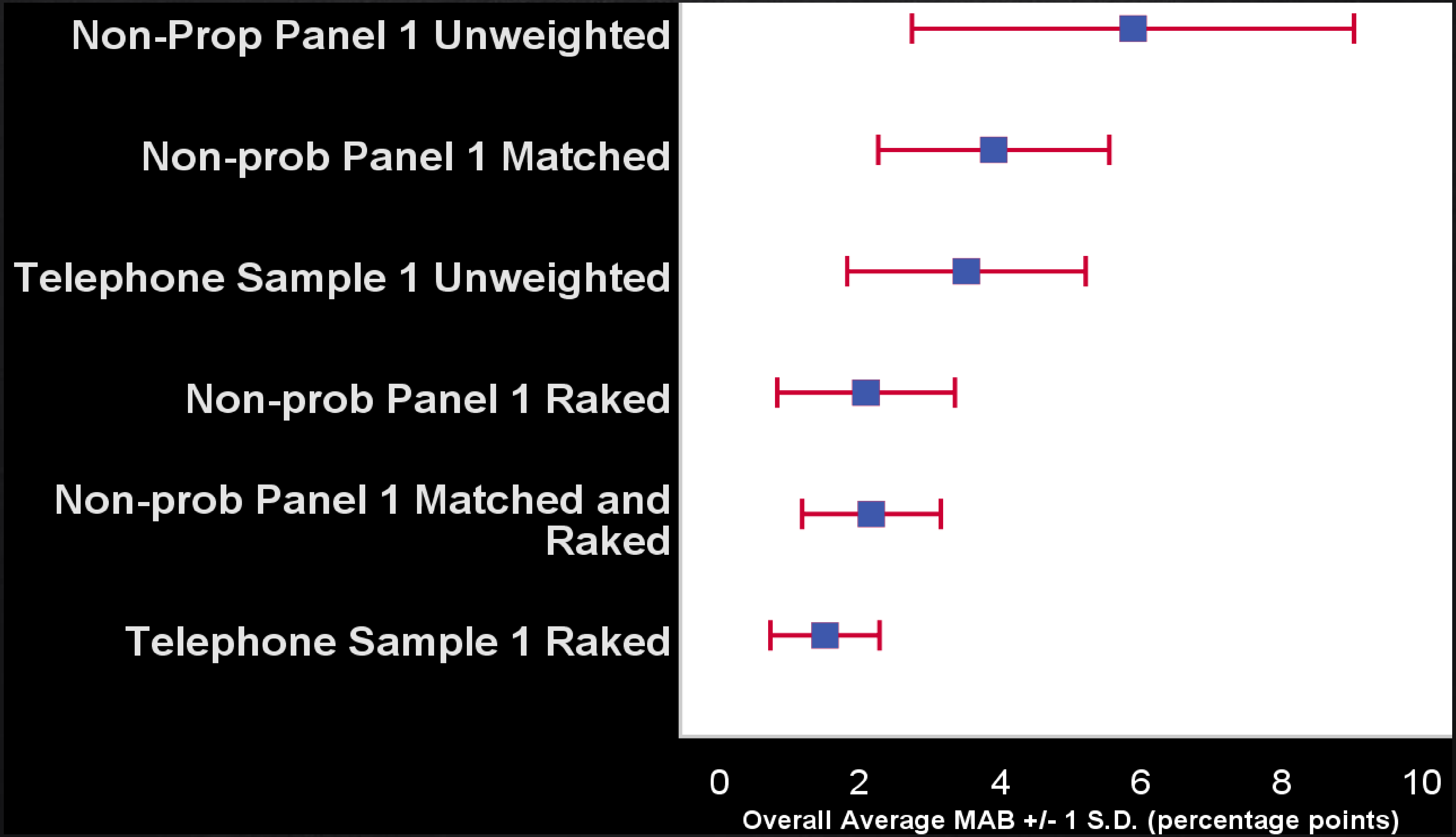


1st Sample Set

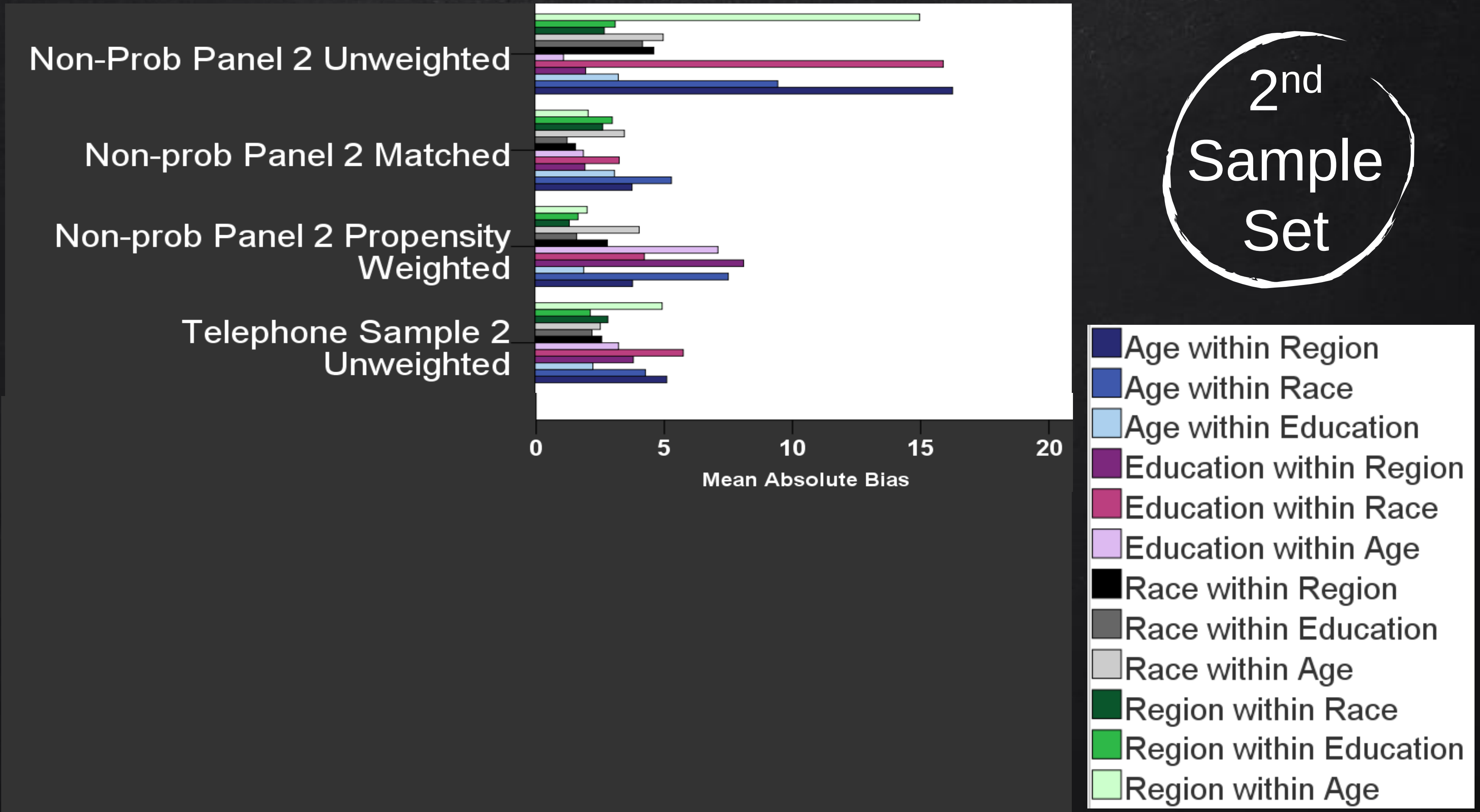


Unweighted Overall Average Biases

1st Sample Set



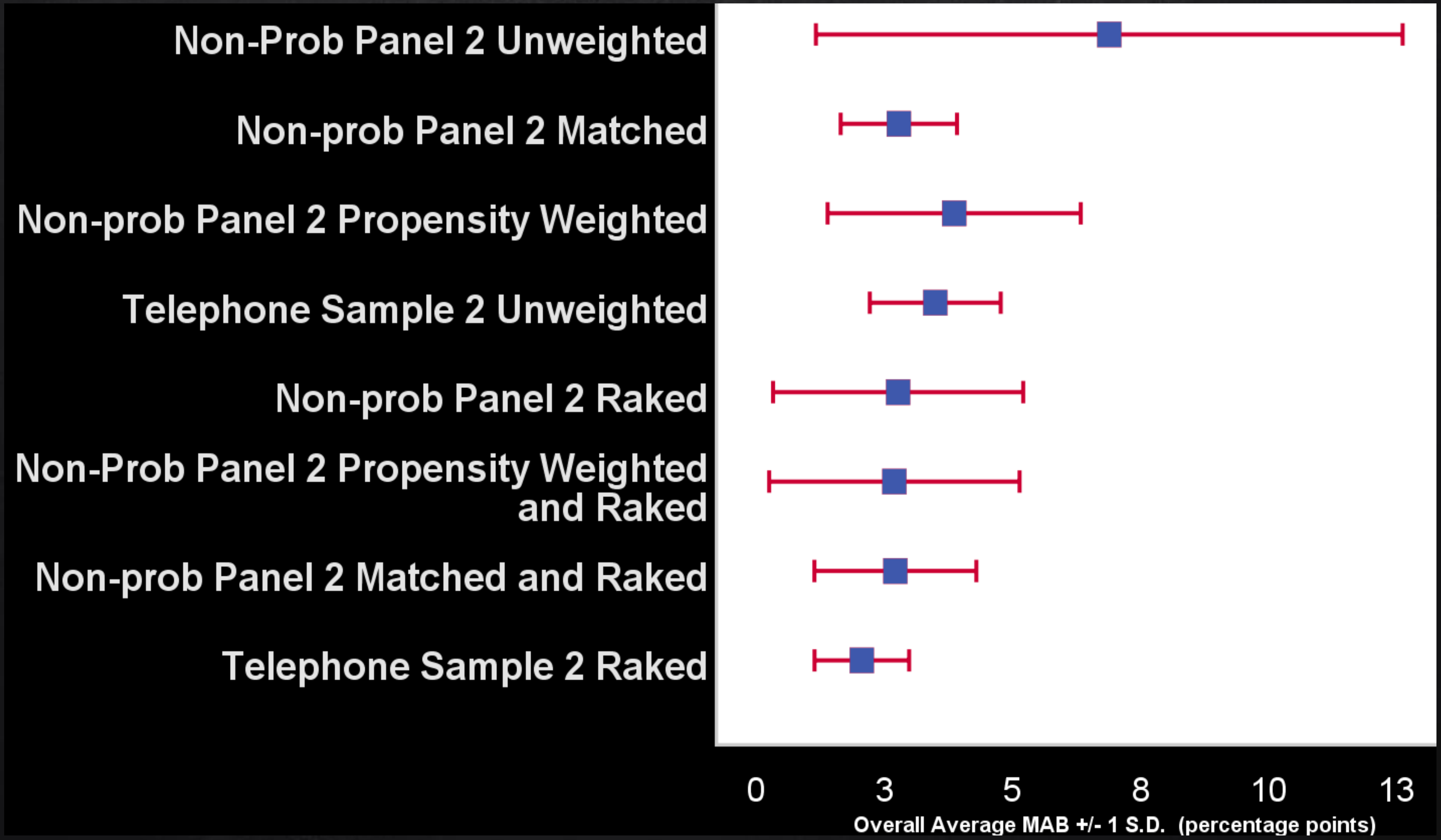
MAB Statistics for Demographic Cross-Tabulations



2nd Sample Set

2nd Sample Set

Unweighted Overall Average Biases



Unequal Weighting Effects

1st Sample Set

Unequal Weighting Effects

- Telephone 1: 1.36
- ABS: 1.74
- NP Panel 1: 2.71
- NP Panel Matched and Raked: 1.18

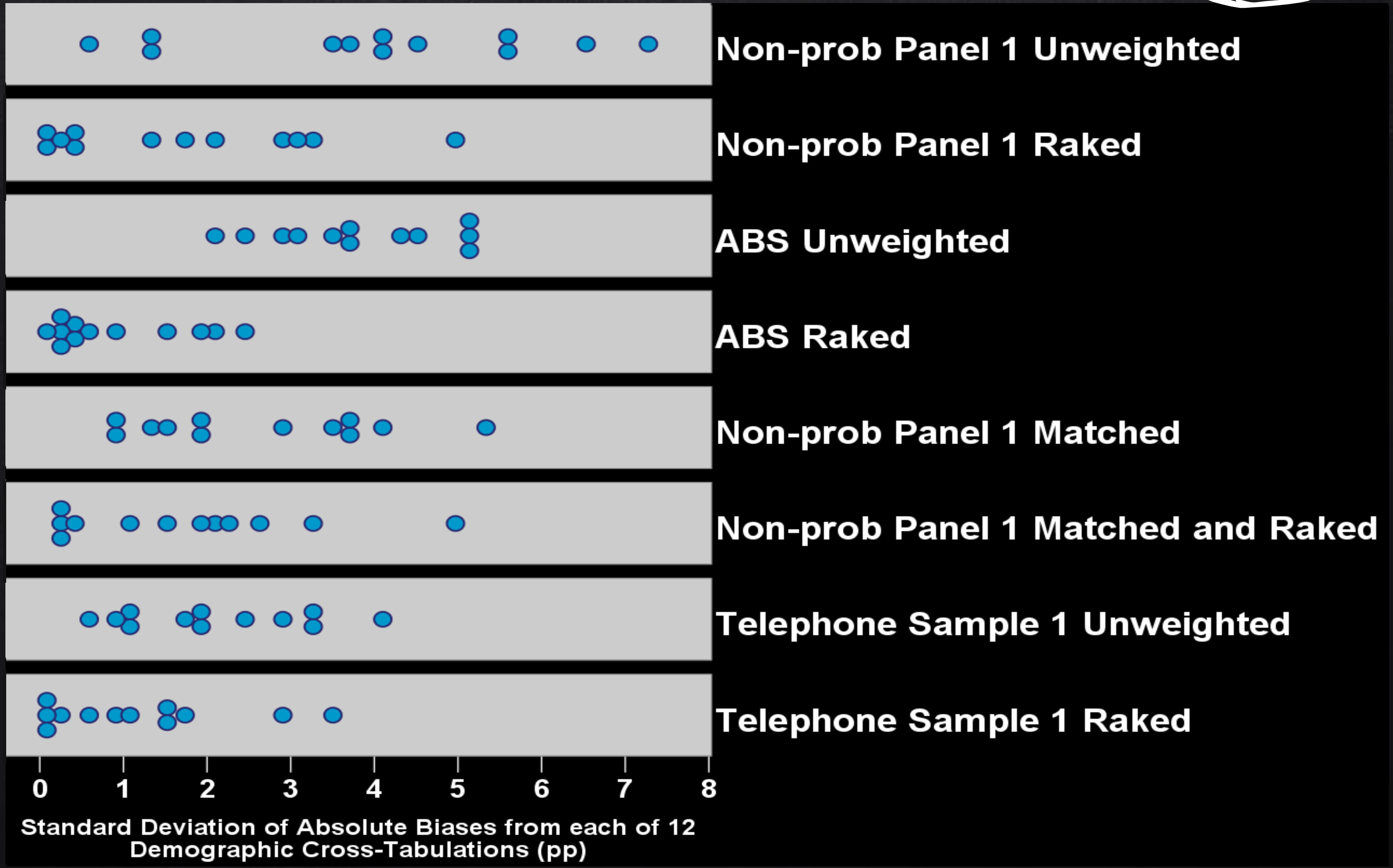
2nd Sample Set

Unequal Weighting Effects

- Telephone 2: 1.21
- NP Panel 2 Rake: 2.83
- NP Panel 2 Propensity: 6.35
- NP Panel 2 Propensity and Rake: 5.43

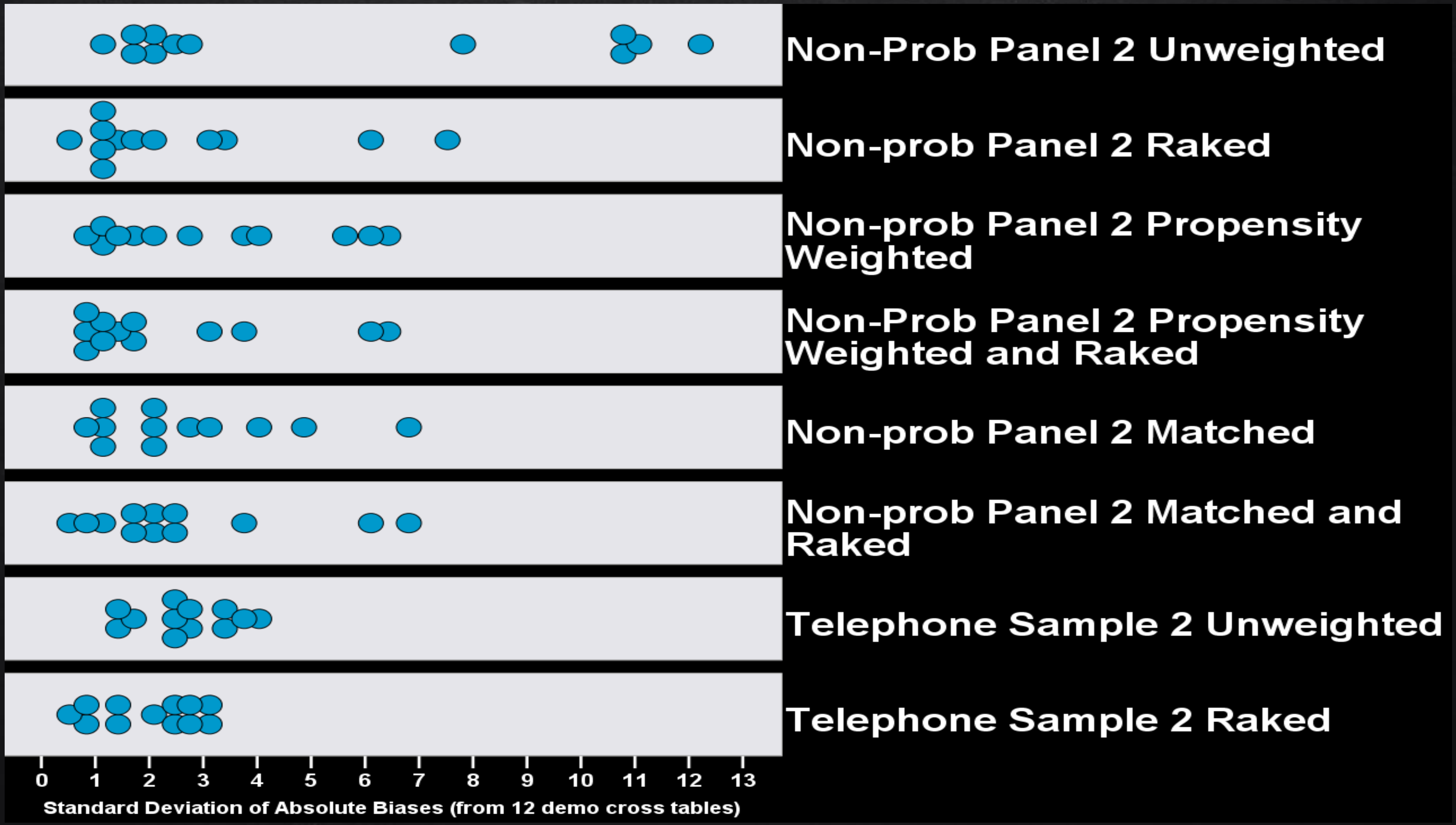
Variability in Absolute Bias Measures

1st Sample Set



Variability in Absolute Bias Measures

2nd Sample Set



Discussion

- Methods for improving the quality of nonprobability panels continue to be made including extensive use of modelling/optimization methods for selecting candidate variables for weighting (Terhanian et al., 2016)
- While we saw that matched samples (based on a simple matching coefficient) tended to move the absolute bias measures downward, compared to unweighted and unmatched nonprobability samples, they still produced estimates with inherently more bias with more variability than probability samples.
- More work is needed to better understand how to optimize the matching process including:
  - ☒ How to incorporate a mixture of categorical and continuous variables
  - ☒ Optimal combination of matching and raking variables
  - ☒ Incorporation of sampling weights into the matched process
  - ☒ Optimal relative size of non-probability and probability samples used for matching.

What does a 9% response rate get you that a web panel cannot?

- Unweighted, substantially less bias
- Weighted, significantly, but not substantially, less bias
- A much lower design effect from weighting
- Generally less variability in the absolute biases
- That said, matched samples attain the lowest design effect of any weighting
- And that said, matched samples “close” to weighted telephone in terms of lower bias and lower variability in the absolute biases….but still substantially inferior
- And…there is still the issue of cost.

- References Available Upon Request

- Please contact us with questions

TBuskirk@m-s-g.com | 314-695-1378 | @trentbuskirk
DDutwin@ssrs.com | 484-840-4406 | @ddutwin

Thank you!

Calculating Standard Errors for Non-probability Samples when Matching to Probability Samples

Adam Lee
Randal ZuWallack

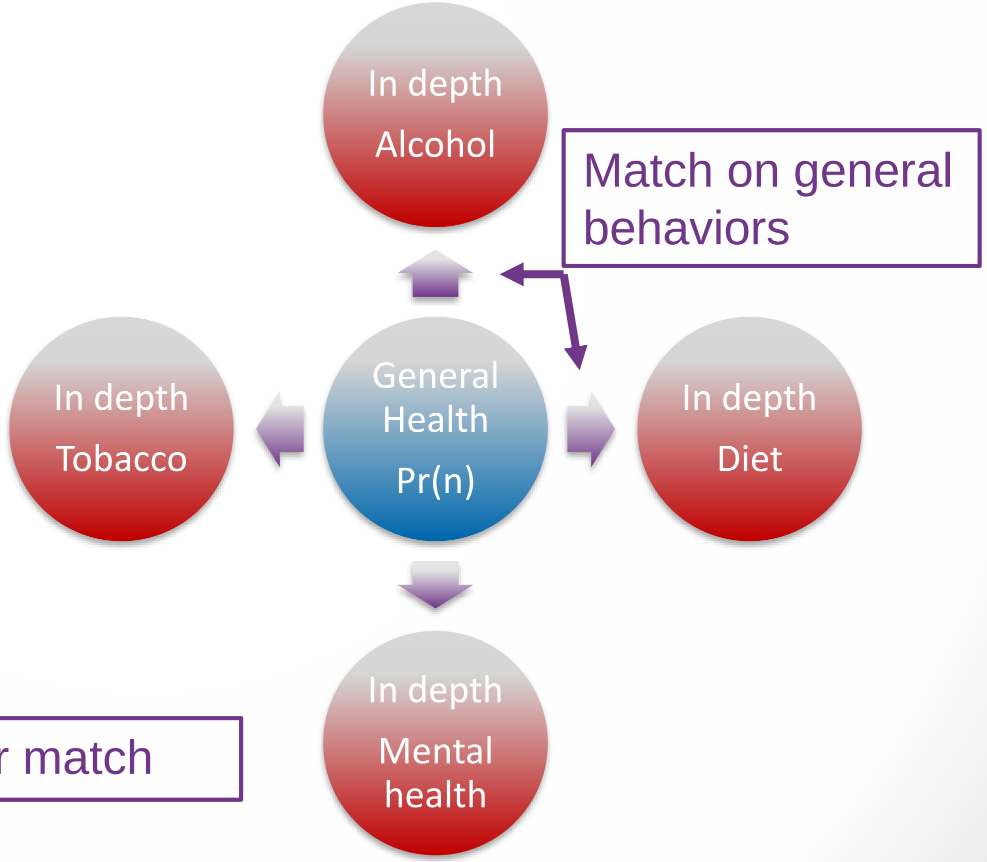
ICF International

AAPOR
Austin, TX
May 15, 2016

Motivation

- Matching probability samples to nonprobability samples is emerging as a popular methodology
  - Reduce costs
  - Rare populations
  - Split surveys
  - Quick turnaround
  - Web data collection

- Expand depth

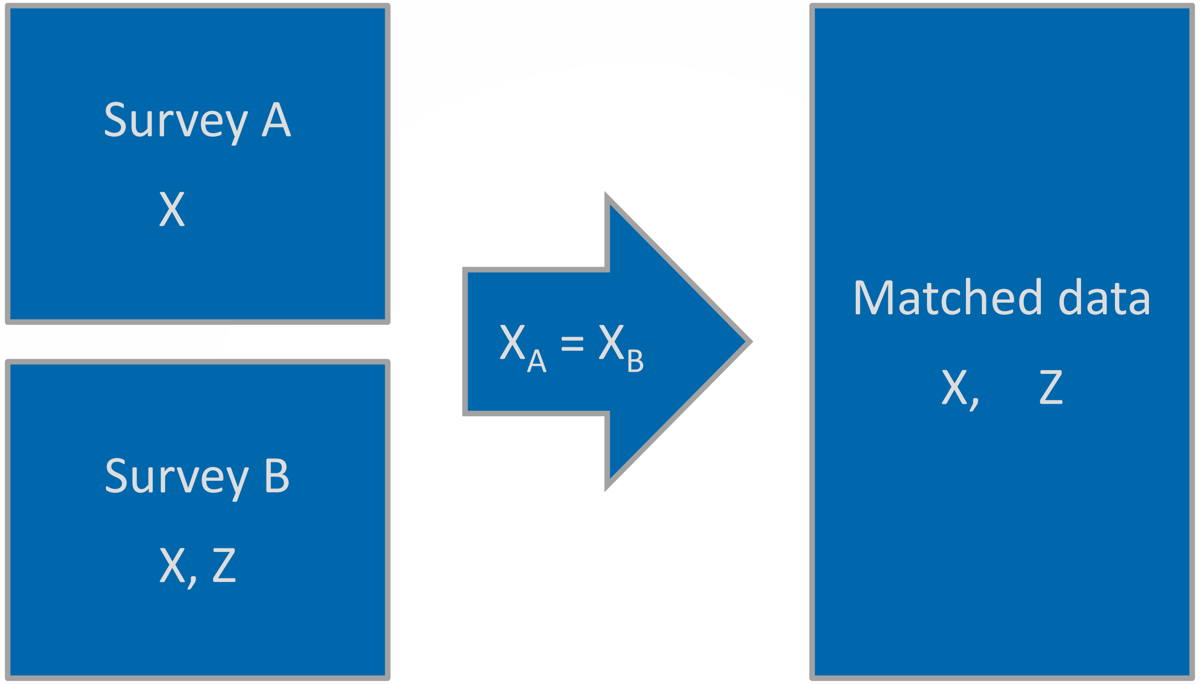


Pr(n) inherits NPS data from their match

AAPOR 2015: NPS matching

1. Selected or Self-Selected? Part 1: A Comparison of Methods for Reducing the Impact of Self-Selection Biases from Non-Probability Surveys
   – Dutwin and Buskirk
2. Selected or Self-Selected? Part 2: Ex ploring Non-Probability and Probability Samples from Response Propensities to Participant Profiles to Outcome Distributions
   – Buskirk and Dutwin
3. Matching an Internet Panel Sample of Health Care Personnel to a Probability Sample
   – DiSogra, Greby, Srinath, Andrew Burkey, Black, Sokolowski, Yue, Ball, Donahue
4. Matching an Internet Panel Sample of Pregnant Women to a Probability Sample
   – Burkey, DiSogra, Greby, Srinath, Black, Sokolowski, Ding, Ball, Donahue
5. Weighting and Sample Matching Effects for an Online Sample
   – Brick, Cohen, Cho, Scott Keeter, McGeeney, Mathiowetz,
6. Can Surveys Posted on Government Websites Give Fair Representations of Public Opinion?
   – Kobayashi
7. Combining a Probability Based Telephone Sample with an Opt-in Web Panel
   – ZuWallack, Dayton, Freedner-Maguire, Karriker-Jaffe, Greenfield

Statistical matching



- In a nonprobability to probability matching application, the focus may be less on joint distributions and more on weighting.
- The probability sample, which represents the population, provides the distribution to calibrate the nonprobability sample. Each person selected in the probability sample is assigned a statistical match from the nonprobability sample and inherits the nonprobability data from that match.

Present Research

- Matching Pr(n) with NPS
  - Is this a probability sample?
  - Can we calculate sampling variance?
    - How? $$\frac{\sigma_{NPS}^{2}}{n_{\mathrm{Pr}}}$$
  - Are there other forms of variance that must be included?

SIMULATION

PROBABILITY TO PROBABILITY MATCHING

Data

- Using National Health Interview Survey (NHIS) data, we drew two SRS
  - “Receiver” Sample: Which would serve as our “Probability” sample, would have retain the demographic variables, but did not have any analysis variables
  - “Donor” Sample: Which would serve as our “Non-Probability Panel” sample. Would have both demographics variables and analysis variables.
- We would use the common demographics variables (Age & Race) and input the analysis variable on the receiver sample using the donor sample. This would create a “Matched” dataset.
- We varied the size of the donor sample to change the variance but held the receiver sample size constant.

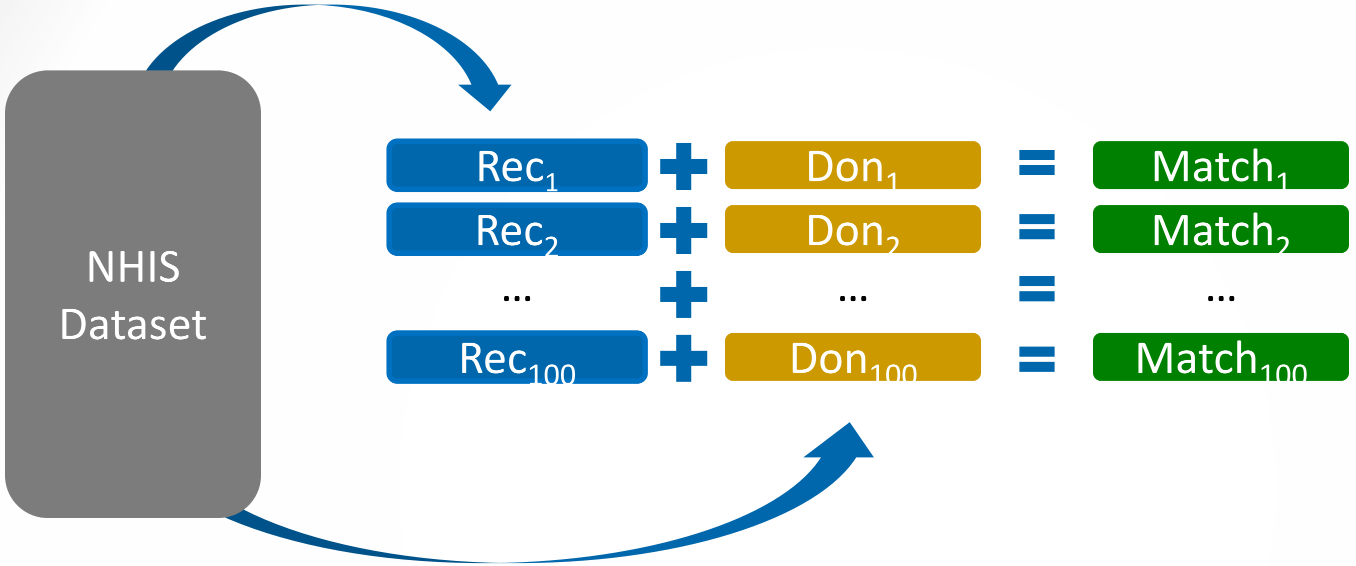
Methodology

- Iteratively draw to 100 samples at each donor sample size level (from 800 to 5,800)
  - Each Receiver Sample
    - Sample Size Fixed (n = 1,000)
    - Includes Demographic Variables (Sex & Race) for Matching
    - “Missing” Key Variable of Interest (i.e., Alcohol Drinking Status)
  - Each Donor Sample
    - Sample Size Iteratively Increased (m = 800 - 5,800)
    - Includes Demographic Variables (Sex & Race) for Matching
    - Includes Key Variable of Interest (i.e., Alcohol Drinking Status)
- Match the Donor sample to the Receiver sample based on demographic variables (Sex by Race) using a Random Hot Deck Matching procedure
  - Create a “Matched” dataset that has Receiver demographics and Donor variable of interest.

StatMatch Package in R

- *"Integration of two data sources referred to the same target population which share a number of common variables (aka data fusion). Some functions can also be used to impute missing values in data sets through hot deck imputation methods. Methods to perform statistical matching when dealing with data from complex sample surveys are available too."*
- **Random Hot Deck** - Finds a donor record for each record in the recipient data set. The donor is chosen at random in the subset of available donors.
  - Identify the Receiving and Donor Datasets
  - Requires the identification of "donation classes" (e.g., Sex by Race).
  - Variables must be shared by both datasets

Statistical Matching Approach



$Receiver_1$	
Sex_R	$Race_R$
M	White
F	Asian

+

$Donor_1$		
Sex_D	$Race_D$	$ALCSTAT_D$
M	White	0
M	White	1
F	Asian	0
F	Asian	0

=

$Matched_1$		
Sex_R	$Race_R$	$ALCSTAT_D$
M	White	1
F	Asian	0

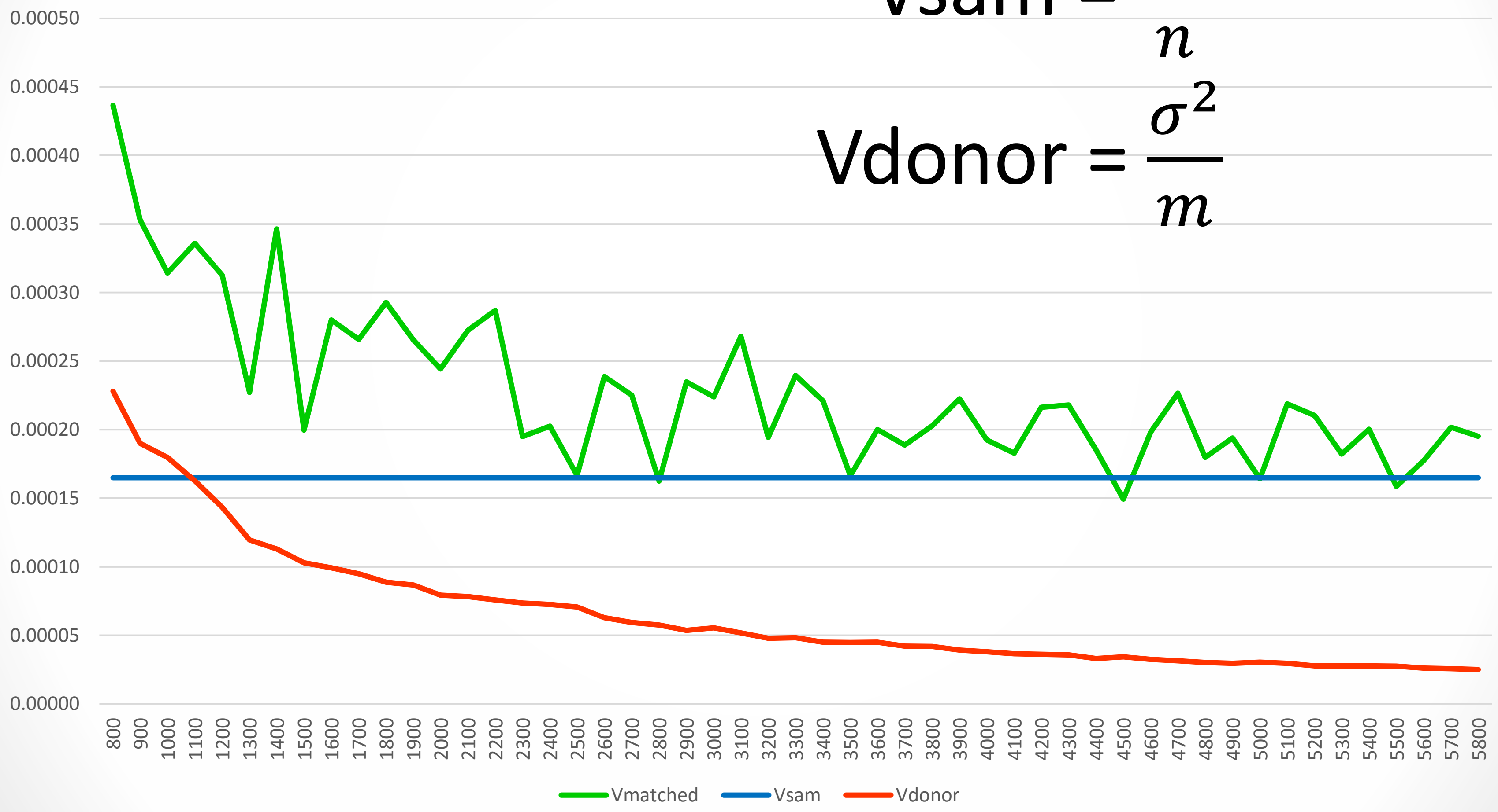
Results

- Vsam = Variance of the 100 original sample estimates (n = 1000)
- Vdonor = Variance of the 100 donor samples (m = 800 to 5800)
- Vmatched = Variance of 100 matched samples

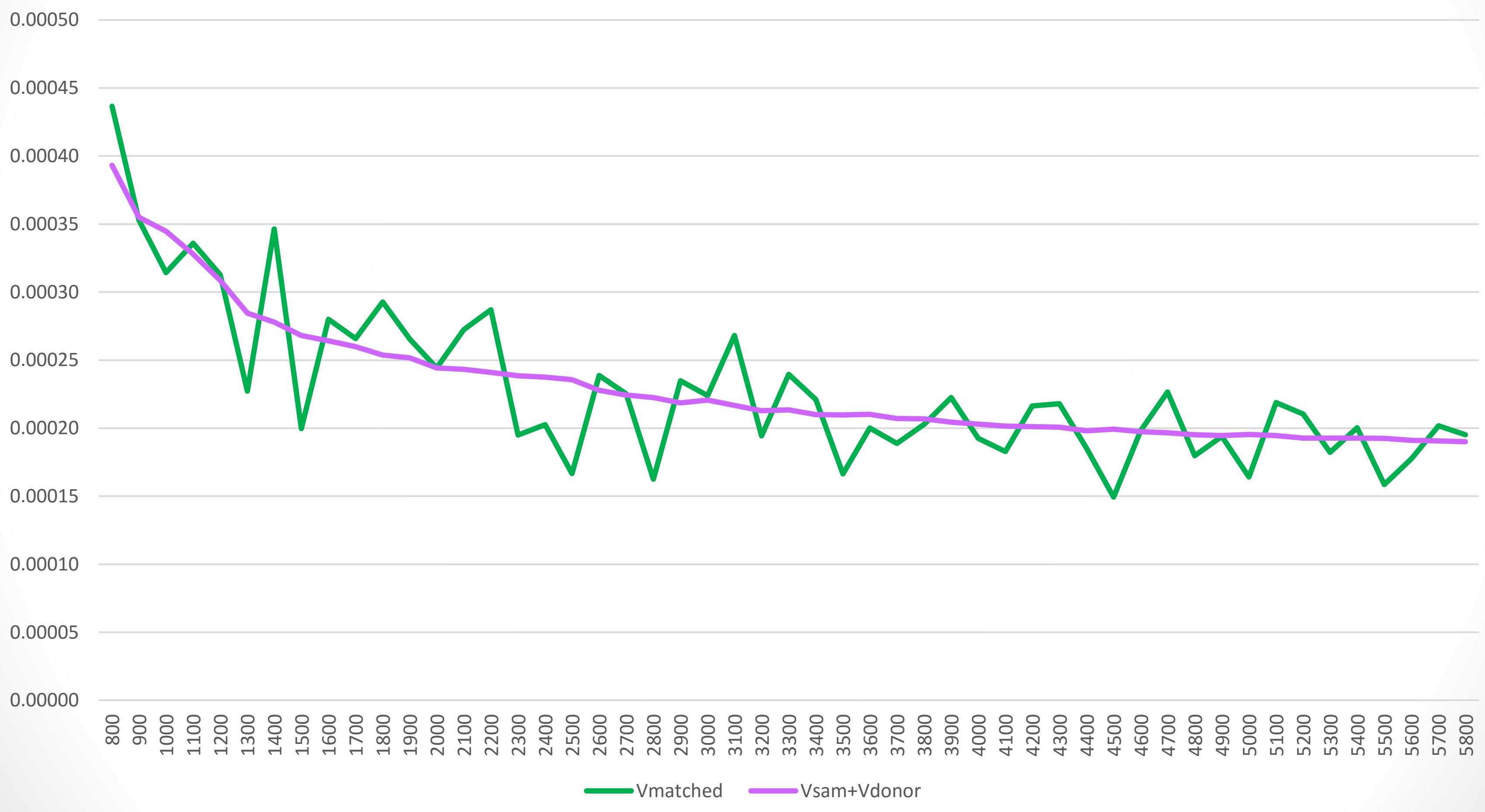
Variances by Donor Size

$$Vsam = \frac{\sigma^2}{n}$$

$$Vdonor = \frac{\sigma^2}{m}$$



Relationship between Vmatch and Vsam+Vdon



SIMULATION

PROBABILITY TO NON-PROBABILITY MATCHING

Experiment

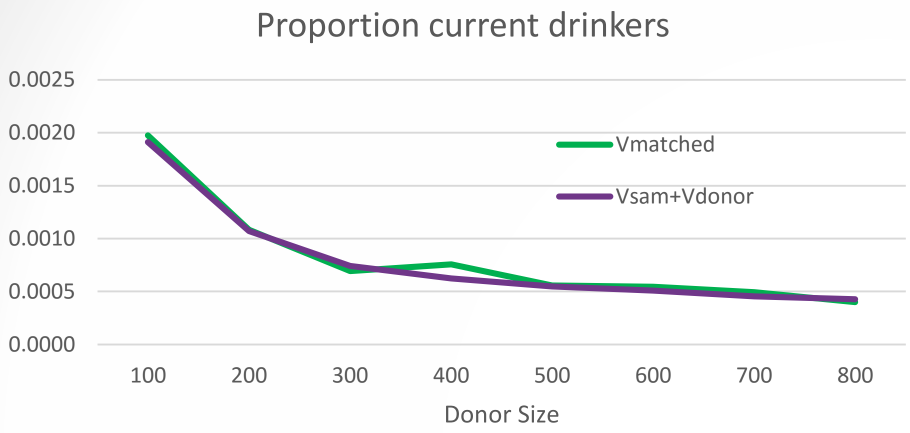
- National Alcohol Survey
  - Dual-frame RDD, CATI (3874 landline; 2749 cell)
  - NAS Web experiment (n=841)
- 100 RDD samples matched to 100 Web samples
  - Sample size: 400
  - Donor size: 100-800
  - All SRS
- StatMatch based on age and gender
- Vmatched = Variance of 100 matched samples

Panel/RDD means and variances

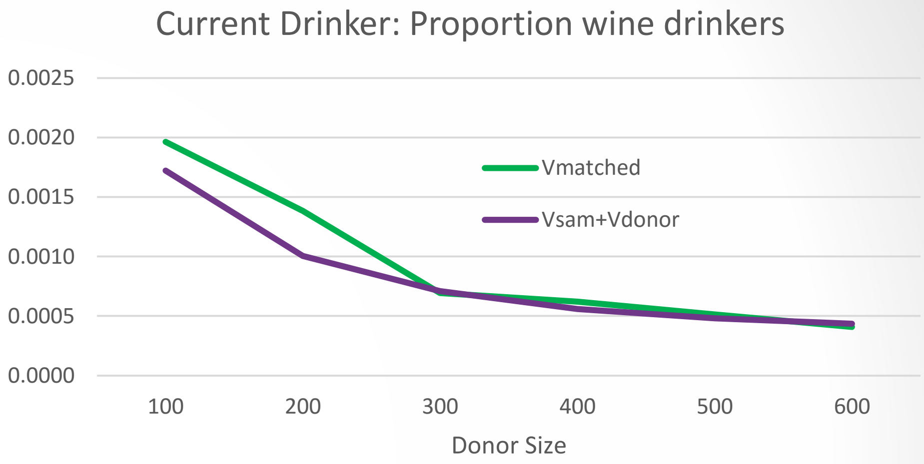
	RDD			Web panel		
	n	Mean	SD	n	Mean	SD
Percentage of current drinkers	6623	0.60	0.49	841	0.79	0.41
Current drinkers: proportion who drink wine	3973	0.74	0.44	663	0.85	0.36
Current drinkers: proportion who drink beer	3973	0.61	0.49	663	0.70	0.46
Current drinkers: typical number of drinks when drinking on a quiet evening at home (0-8)	3973	1.27	1.29	663	1.71	1.57

$$\text{Vsam+Vdonor} = \left(\frac{1}{n_{\text{Pr}}} + \frac{1}{m_{\text{NPS}}}\right) \sigma_{NPS}^2$$

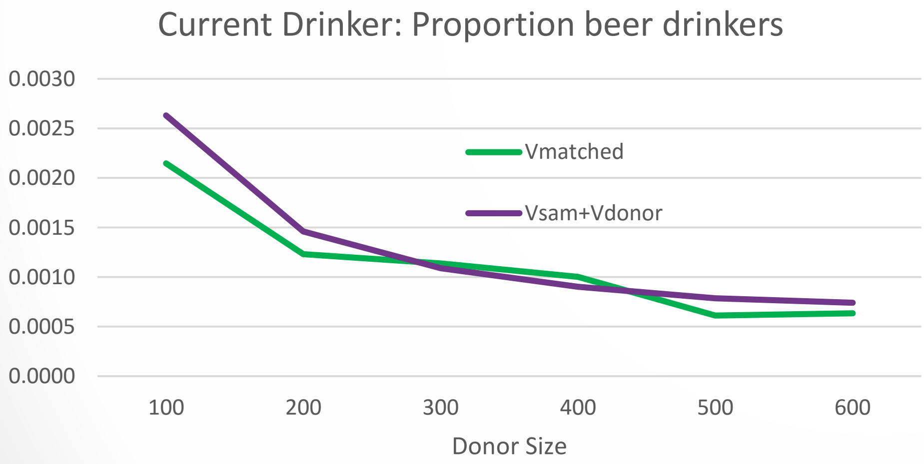
Results



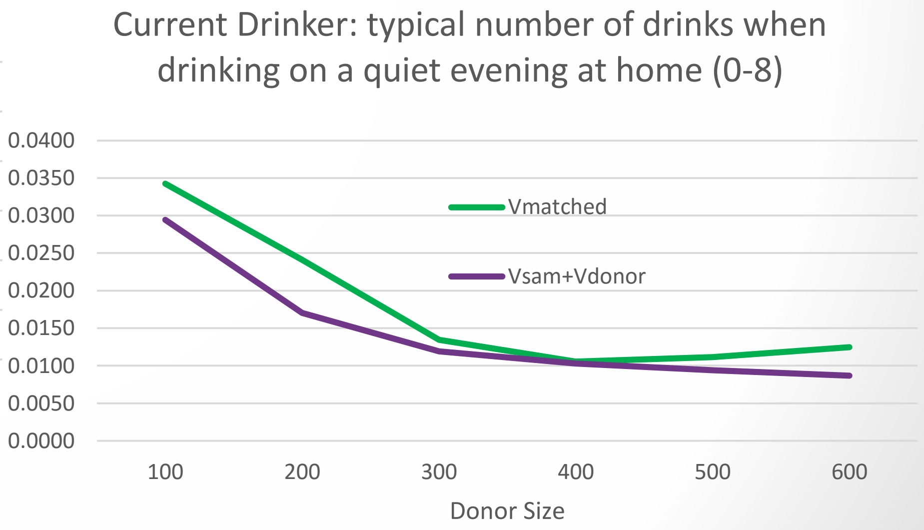
Proportion current drinkers
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
200
300
400
500
600
700
800
Donor Size
Vmatched
Vsam+Vdonor



Current Drinker: Proportion wine drinkers
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
200
300
400
500
600
Donor Size
Vmatched
Vsam+Vdonor



Current Drinker: Proportion beer drinkers
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
200
300
400
500
600
Donor Size
Vmatched
Vsam+Vdonor



Current Drinker: typical number of drinks when drinking on a quiet evening at home (0-8)
0.0400
0.0350
0.0300
0.0250
0.0200
0.0150
0.0100
0.0050
0.0000
100
200
300
400
500
600
Donor Size
Vmatched
Vsam+Vdonor

SUMMARY

Summary

$$\text{Var} = \cancel{\frac{\sigma_{NPS}^2}{n_{\text{Pr}}}} \; ? \qquad \text{Var} \approx \left(\frac{1}{n_{\text{Pr}}} + \frac{1}{m_{\text{NPS}}}\right) \sigma_{NPS}^2$$

- Variance must account for matching and the donor pool
  - Large donor pool: $\text{Var} \approx \frac{\sigma_{NPS}^2}{n_{\text{Pr}}}$

- Other variance increases/decreases
  - Matches used more than once
  - Good matching model

Summary

- Still NPS sample
  - Probability matches provide weights

- Variability is based on the NPS
  - Need a random sample from panel to estimate σ_{NPS}^2

Thank you

- For more information, please contact:
  - Adam.Lee@icfi.com
  - Randy.Zuwallack@icfi.com

Non-Probability Samples at AAPOR 2016
- Selected papers

Charles DiSogra
c.disogra@srbi.com