A Tree-Based Dual-Frame Estimation Approach for Combining Probability and Non-probability Samples

Chien-Min Huang and Jay Breidt

NORC at the University of Chicago

Joint Statistical Meetings

August 8, 2024

Motivation

- Streamlined combination of probability and non-probability samples
  - maximal use of non-probability to reduce variance and cost
  - minimal introduction of bias due to non-probability
- Seamless handling of multiple scenarios
  - probability samples with sparsity or undercoverage
  - non-probability samples with undercoverage or overcoverage
  - highly targeted non-probability samples
- Current methods are built with the presumption of parametric models, conditional on the covariates selected
  - sensitive to assumed statistical model
  - model selection could be difficult in production environment

Notation

- Population, $U = \{1, 2, \ldots, k, \ldots, N\}$
- Probability sample, $A \subset U$
  - inclusion probabilities

$$\pi_k^A = \mathrm{P}\left[k \in A\right] = \mathrm{P}\left[A_k = 1\right] = \mathrm{E}\left[A_k\right]$$

    where $A_k = 1$ if $k \in A$, $A_k = 0$ if $k \notin A$
  - π_k^A **positive** for all $k \in U$
  - π_k^A **known** for all $k \in A$
- Non-probability sample, $B \subset U$
  - often, $B \subset U \setminus A$ so that

$$\pi_k^B = \mathrm{P}\left[k \in B\right] = \mathrm{P}\left[k \in B | \, k \notin A\right] \mathrm{P}\left[k \notin A\right] = \rho_k (1 - \pi_k^A)$$

  - ρ_k is **unknown** and **might be zero**
  - write $B_k = 1$ if $k \in B$, $B_k = 0$ if $k \notin B$

Dual-frame approach

- Elements can enter the combined sample, $A \cup B$, via two paths:

$$\begin{aligned} \mathrm{P}\left[k \in A \cup B\right] &= \mathrm{P}\left[k \in A\right] + \mathrm{P}\left[k \in B\right] - \mathrm{P}\left[k \in A \cap B\right] \\ &= \pi_k^A + \rho_k(1 - \pi_k^A) - 0 \end{aligned}$$

- If we knew the combined probability above for all $k \in A \cup B$, we could construct the unbiased **dual-frame estimator**

$$\widetilde{T}_y = \sum_{k \in U} \frac{A_k + B_k(1 - A_k)}{\pi_k^A + \rho_k(1 - \pi_k^A)} y_k$$

Dual-frame approach, II

- **Advantages:**
  - no need to choose a "tradeoff parameter" between probability and non-probability samples
  - even if ρ_k are small or zero, dual-frame weights are stable:

$$1 \leq \frac{1}{\pi_k^A + \rho_k(1 - \pi_k^A)} \leq \frac{1}{\pi_k^A}$$

- **Challenges:** To use the dual-frame approach, we need
  - to impute the unknown π_k^A for $k \in B$
  - to estimate the unknown ρ_k for $k \in A \cup B$
- We use a tree-based approach for both problems

Tree-based approach

- Fit a weighted tree using (any) observed data $\boldsymbol{x}_k$
  - leaves are homogeneous with respect to ρ_k
- Non-probability model: B_k independent Bernoulli random variables with success probabilities

$$\mathrm{P}\left[B_k = 1 \mid \boldsymbol{x}_k\right] = \rho(\boldsymbol{x}_k).$$

- Feasible pseudo-log-likelihood

$$\begin{aligned} L(\rho_k) &= \sum_{k \in U \cap L_h} (1 - A_k) B_k \ln\{\rho_k\} \\ &+ \sum_{k \in U \cap L_h} \left\{ \frac{A_k}{\pi_k^A} (1 - \pi_k^A) \right\} (1 - B_k) \ln\{1 - \rho_k\}. \end{aligned}$$

- The score of likelihood function is unbiased for 0

Tree-based approach, continued

- Estimate ρ_k as piecewise constant over leaf h:

$$\begin{aligned}
\widehat{\rho}_k &= \frac{\text{“successes”}}{\text{“successes”} + \text{“weighted failures”}} \\
&= \frac{\sum_{k \in L_h} B_k}{\sum_{k \in L_h} B_k + \sum_{k \in L_h} (1/\pi_k^A - 1) A_k} \\
&= \frac{\sum_{k \in L_h} B_k}{\sum_{k \in L_h} B_k + \widehat{\beta}_h}
\end{aligned}$$

  - weights of one for non-probability sample “successes” (in the absence of other information, these cases only represent themselves)
  - design weights minus one for probability sample “failures”

Tree-based approach, continued

- Because leaves are expected to be homogeneous, impute

$$\tilde{\pi}_k^A = \frac{\sum_{k \in L_h} \pi_k^A A_k}{\sum_{k \in L_h} A_k},$$

  the leaf average design probability, for non-probability cases

- We then have the initial combined weights

$$w_k = \frac{A_k}{\pi_k^A + \widehat{\rho}_k(1 - \pi_k^A)} + \frac{B_k}{\tilde{\pi}_k^A + \widehat{\rho}_k(1 - \tilde{\pi}_k^A)}$$

Ratio adjustment of initial combined weights

- Any leaf with probability cases represents

$$\widehat{\alpha}_h = \sum_{k \in L_h} \frac{1}{\pi_k^A} A_k$$

  such cases in the population
- We just added $\sum_{k \in L_h} B_k$ non-probability cases to that leaf
  - these do not change the representation of the leaf: they just add further detail
  - accordingly, we ratio-adjust the initial combined weights

$$w_k = \frac{A_k}{\pi_k^A + \widehat{\rho}_k(1 - \pi_k^A)} + \frac{B_k}{\tilde{\pi}_k^A + \widehat{\rho}_k(1 - \tilde{\pi}_k^A)}$$

    for leaf L_h, redistributing the prob weights across prob and non-prob cases:

$$\omega_k = \left(\sum_{k \in L_h} \frac{1}{\pi_k^A} A_k \right) \frac{w_k}{\sum_{k \in L_h} w_k}$$

Special case: A-only leaves

- **Interpretation:** leaf with only elements from A corresponds to part of the population not covered by the non-probability sample
- In this case, $\widehat{\rho}_k = 0$ and probability elements in this leaf retain their (unmodified) probability weight:

$$\frac{1}{\pi_k^A + \widehat{\rho}_k(1 - \pi_k^A)} = \frac{1}{\pi_k^A + 0(1 - \pi_k^A)} = \frac{1}{\pi_k^A}$$

- Ratio adjustment does not change the weight, so the ratio-adjusted combined weight for A-only leaf is the original design weight

Special case: B-only leaves

- **Interpretation:** leaf with only elements from B corresponds to either ...
  - elements that are not part of the population of interest?
  - part of the population of interest not covered by the probability sample?
- In either case, we do not know what these elements represent and cannot "borrow representation" from the probability sample
  - imputed $\tilde{\pi}_k^A = 0$ and $\widehat{\rho}_k = 1$, so non-probability elements in this leaf have their (unmodified) non-probability weight:

$$\frac{1}{\tilde{\pi}_k^A + \widehat{\rho}_k(1 - \tilde{\pi}_k^A)} = \frac{1}{0 + 1(1 - 0)} = 1$$

  - No A-information to use in ratio adjustment, so the ratio-adjusted weight for B-only leaf is still equal to one

Dual-frame estimator

- Dual-frame estimator of the total:

$$\begin{aligned}\widehat{T}_y &= \sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} \widehat{\alpha}_h \frac{w_k}{\sum_{k\in L_h} w_k} y_k \\ &= \sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} \omega_k y_k\end{aligned}$$

- Dual-frame estimator of the ratio:

$$\widehat{R} = \frac{\widehat{T}_y}{\widehat{T}_z} = \frac{\sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} \omega_k y_k}{\sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} \omega_k z_k}$$

Variance estimation for dual-frame estimator

- Act as if tree behaves like parametric model of size H
  - (though the number of leaves H is not fixed)
- Joint estimating equations

$$\Phi(\eta) = \begin{bmatrix} J(\theta_1, \rho_1) = \sum_{k \in (A \cup B) \cap L_1} w_k(\widehat{\alpha}_1 y_k - \theta_1) \\ J(\theta_2, \rho_2) = \sum_{k \in (A \cup B) \cap L_2} w_k(\widehat{\alpha}_2 y_k - \theta_2) \\ \vdots \\ J(\theta_H, \rho_H) = \sum_{k \in (A \cup B) \cap L_H} w_k(\widehat{\alpha}_H y_k - \theta_H) \\ G(\rho_1) = \sum_{k \in U \cap L_1} B_k/\rho_1 - \widehat{\beta}_1/(1-\rho_1) \\ G(\rho_2) = \sum_{k \in U \cap L_2} B_k/\rho_2 - \widehat{\beta}_2/(1-\rho_2) \\ \vdots \\ G(\rho_H) = \sum_{k \in U \cap L_H} B_k/\rho_H - \widehat{\beta}_H/(1-\rho_H) \end{bmatrix}$$

where θ_h is the population total in each leaf

Variance estimation for dual-frame estimator, continued

- Asymptotic variance has the sandwich form

$$\{\boldsymbol{\phi}(\boldsymbol{\eta})\}^{-1} \operatorname{Var}[\Phi(\boldsymbol{\eta})] \left\{\boldsymbol{\phi}(\boldsymbol{\eta})^{-1}\right\}^{\top}$$

  where $\boldsymbol{\phi}(\boldsymbol{\eta}) = \partial\Phi(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$

- The derivative $\boldsymbol{\phi}(\boldsymbol{\eta})$

$$\begin{bmatrix} \frac{\partial J(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial \boldsymbol{\theta}^{\top}} & \frac{\partial J(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}^{\top}} \\ \mathbf{0} & \frac{\partial G(\boldsymbol{\rho})}{\partial \boldsymbol{\rho}^{\top}} \end{bmatrix}$$

- The variance $\operatorname{Var}(\Phi(\boldsymbol{\eta}))$ can be derived from

$$\mathrm{E}[\operatorname{Var}(\Phi(\boldsymbol{\eta})|A)] + \operatorname{Var}(\mathrm{E}[\Phi(\boldsymbol{\eta})|A])$$

Variance estimation for dual-frame estimator, continued

- Estimate the variance by plugging in estimators

$$\{\phi(\hat{\eta})\}^{-1} \widehat{V}\left(\Phi(\hat{\eta})\right) \left\{\phi(\hat{\eta})^{-1}\right\}^{\top}$$

- Variance of the ratio:

$$\widehat{R} \approx \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} + \sum_{h\in H} \sum_{k\in(A\cup B)\cap L_h} \frac{v_k w_k \widehat{\alpha}_h}{\sum_{k\in(A\cup B)\cap L_h} w_k}$$

where $v_k = \left\{ \frac{1}{\sum_{k\in U} z_k} \left(y_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} z_k \right) \right\}$

  - replace y_k by $\hat{v}_k$ to get the variance estimates of the ratio

Implementation details: choice of leaf size

- Compute a **bias-variance score** for each potential leaf size:
  - compute a **squared bias score** by comparing ratio-adjusted combined estimates to (unbiased) prob-only estimates
  - compute a **variance score** using the usual WEFF (DEFF due to weighting)
- Normalize each set of scores to $[0, 1]$ and add them together

$$\text{score}_\ell = \frac{\text{bias}_\ell - \text{min bias}}{\text{max bias} - \text{min bias}} + \frac{\text{var}_\ell - \text{min var}}{\text{max var} - \text{min var}}$$

- Choose the leaf size that minimizes score_ℓ

Implementation details: bias score

- Choose a set of J **key variables**
- At each leaf size ℓ, compare unbiased prob-only estimates $\widehat{\theta}_{A,j}$ to ratio-adjusted combined estimates $\widehat{\theta}_{A\cup B,\ell,j}$:

$$\text{bias}_\ell = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{\widehat{\theta}_{A\cup B,\ell,j} - \widehat{\theta}_{A,j}}{\widehat{\theta}_{A,j}}\right)^2$$

Empirical assessments: Monte Carlo experiments

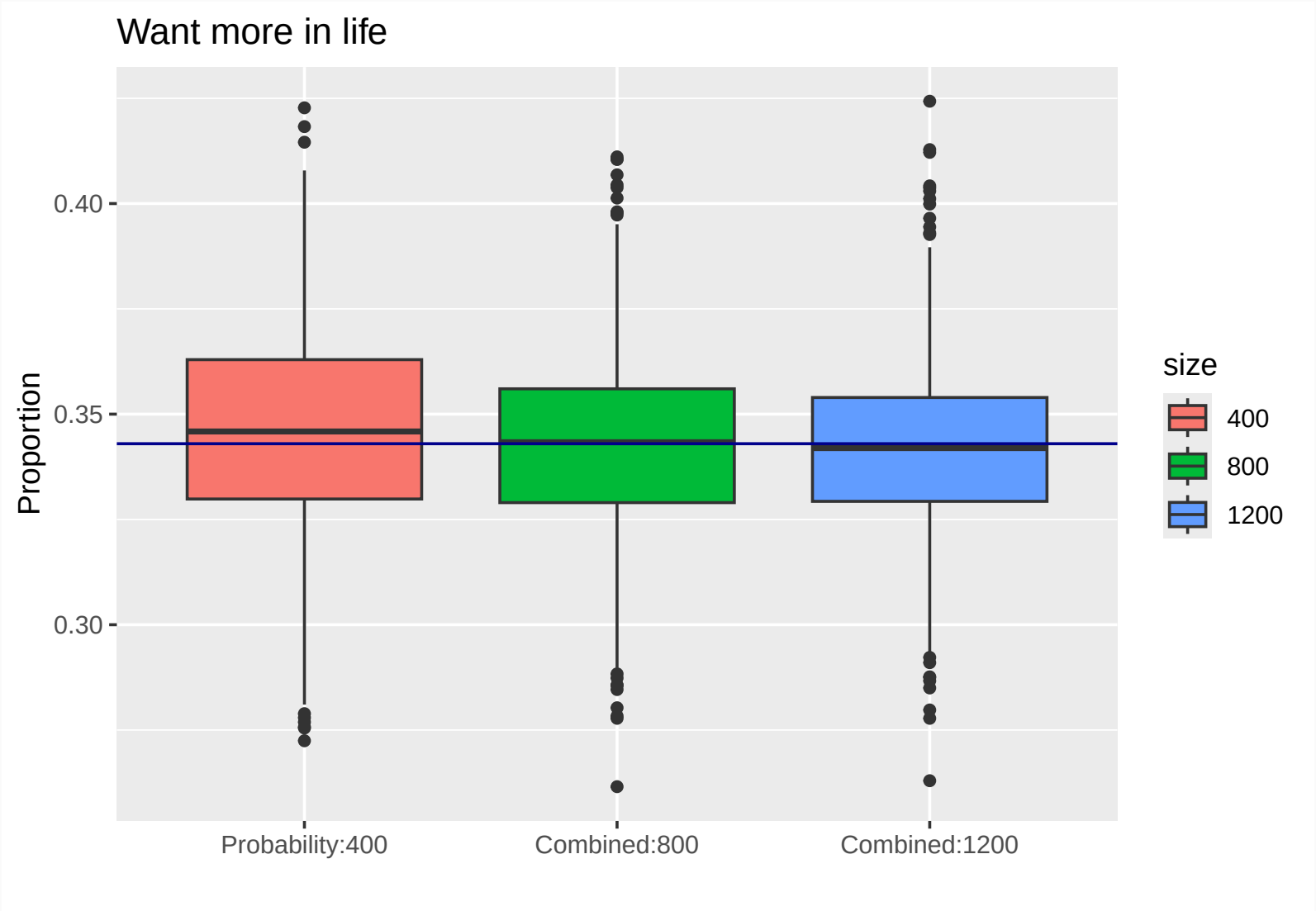
- Artificial population created from Culture and Community in Time of Crisis (CCTC) survey
  - $N = |U| = 123,757$ records
  - probability sample frame (F_0): $117,276$ records from $|U|$
  - non-probability sample frame: $74,202$ records from all of F_0^C and a subset of F_0
  - 20 binary variables of interest: want more in life, music festival, online exhibitions, ...
- 1000 simulated replicates:
  - probability sample A with size n_A =400 via stratified sampling, and known inclusion probabilities π_k^A
  - non-probability sample B with size n_B =400 or 800 and unknown inclusion probability
  - no overlaps in probability and non-probability samples
- Many possible covariates for tree-based modeling

Monte Carlo experiments, continued

- Use `R` function `rpart` to fit the tree
  - allows specification of leaf size
  - include all possible variables
- For each replicate at each non-probability sample size, construct tree-based dual-frame estimator with leaf size given by minimum number of observations
  - across a range of leaf sizes: (total sample size)$^{0.4}$ to (total sample size)/10
  - choose leaf size to minimize score$_\ell$
- Estimate the proportion for 20 variables, using A only, and dual-frame estimator
- Variance estimation: estimating equations

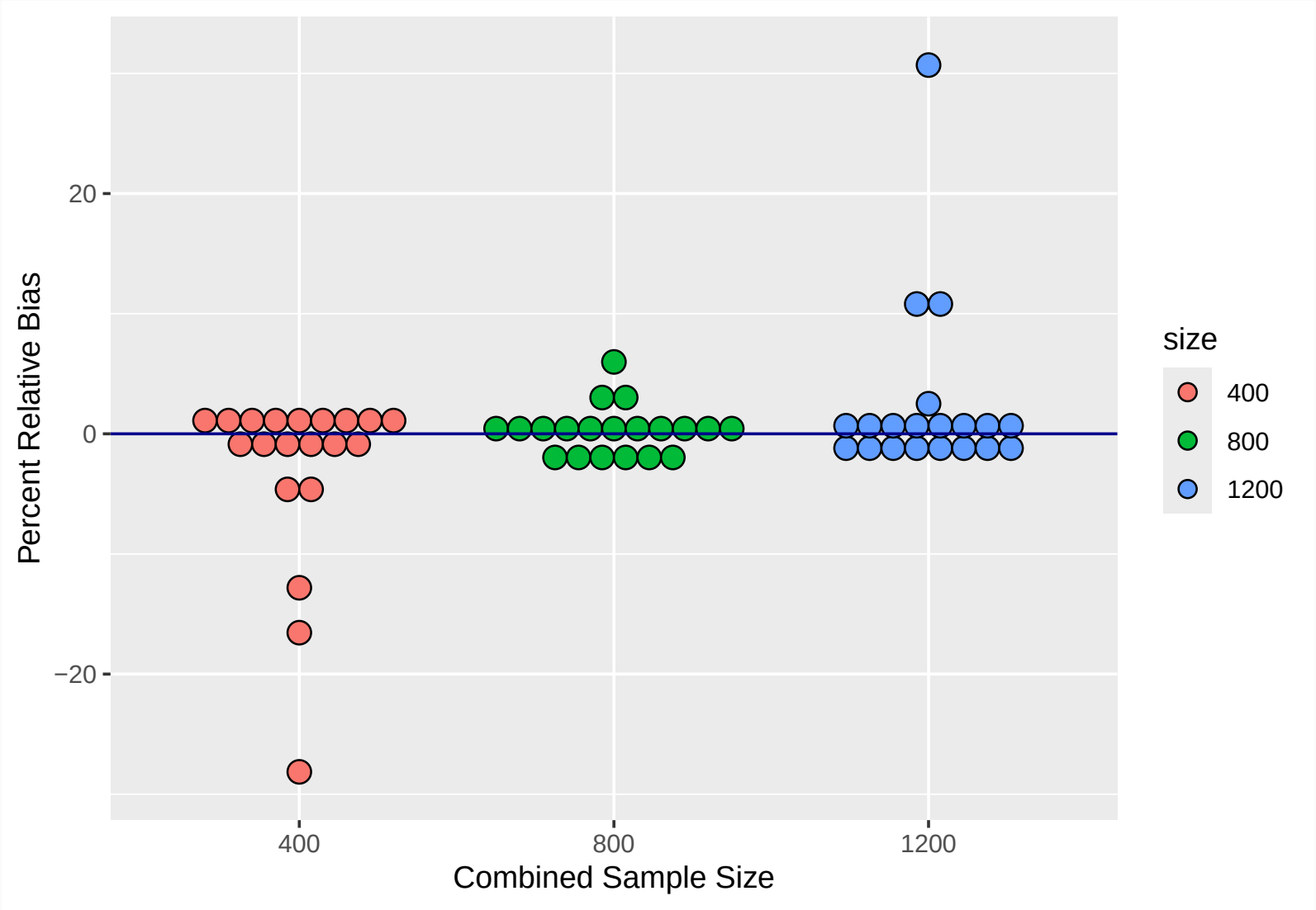
Simulation results: point estimates

- Point estimates across 1000 replicates: want more in life



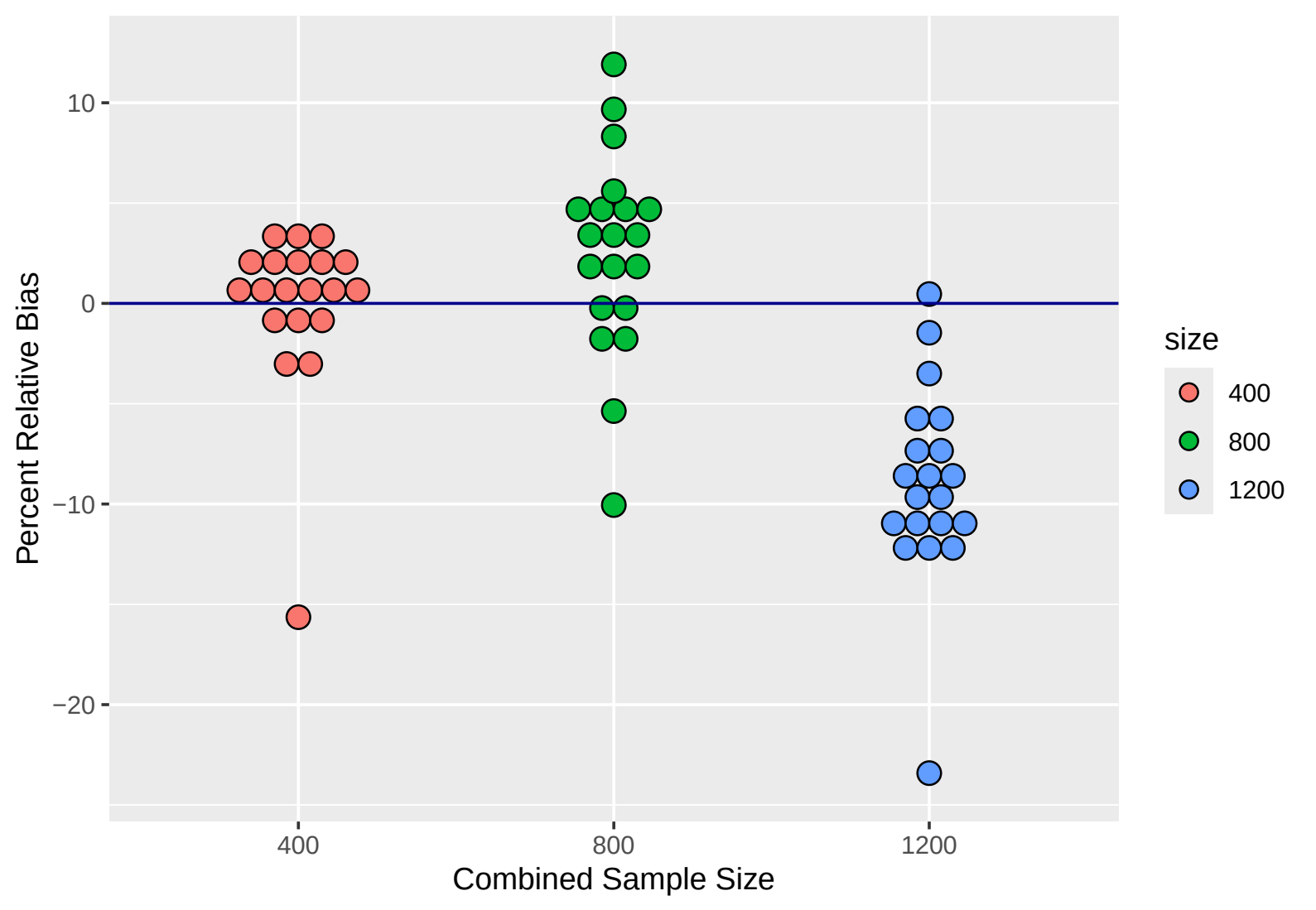
Simulation results: bias of point estimates

- Percent relative bias of point estimates across 20 characteristics



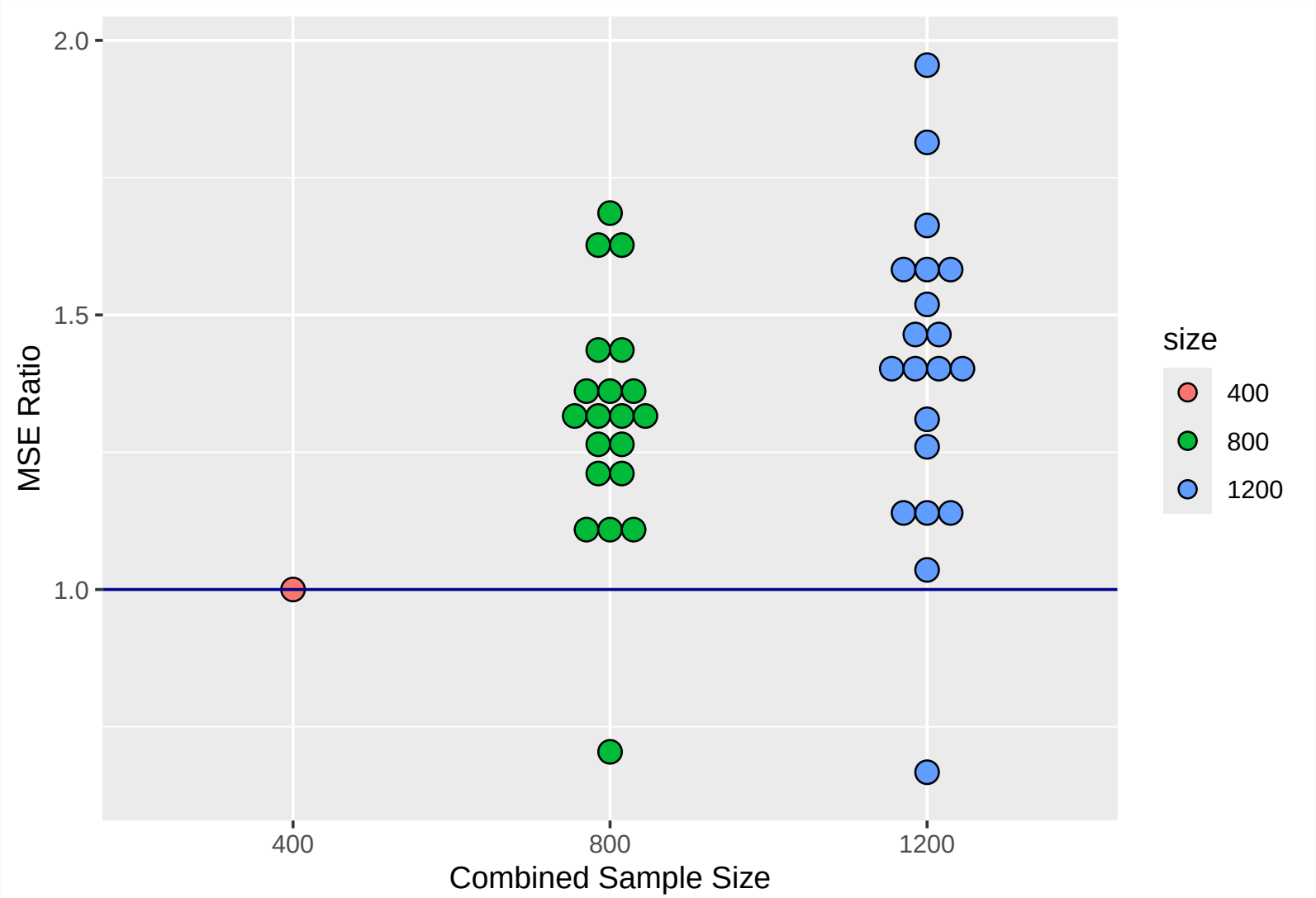
Simulation results: bias of standard deviation

- Percent relative bias of standard deviation across 20 characteristics



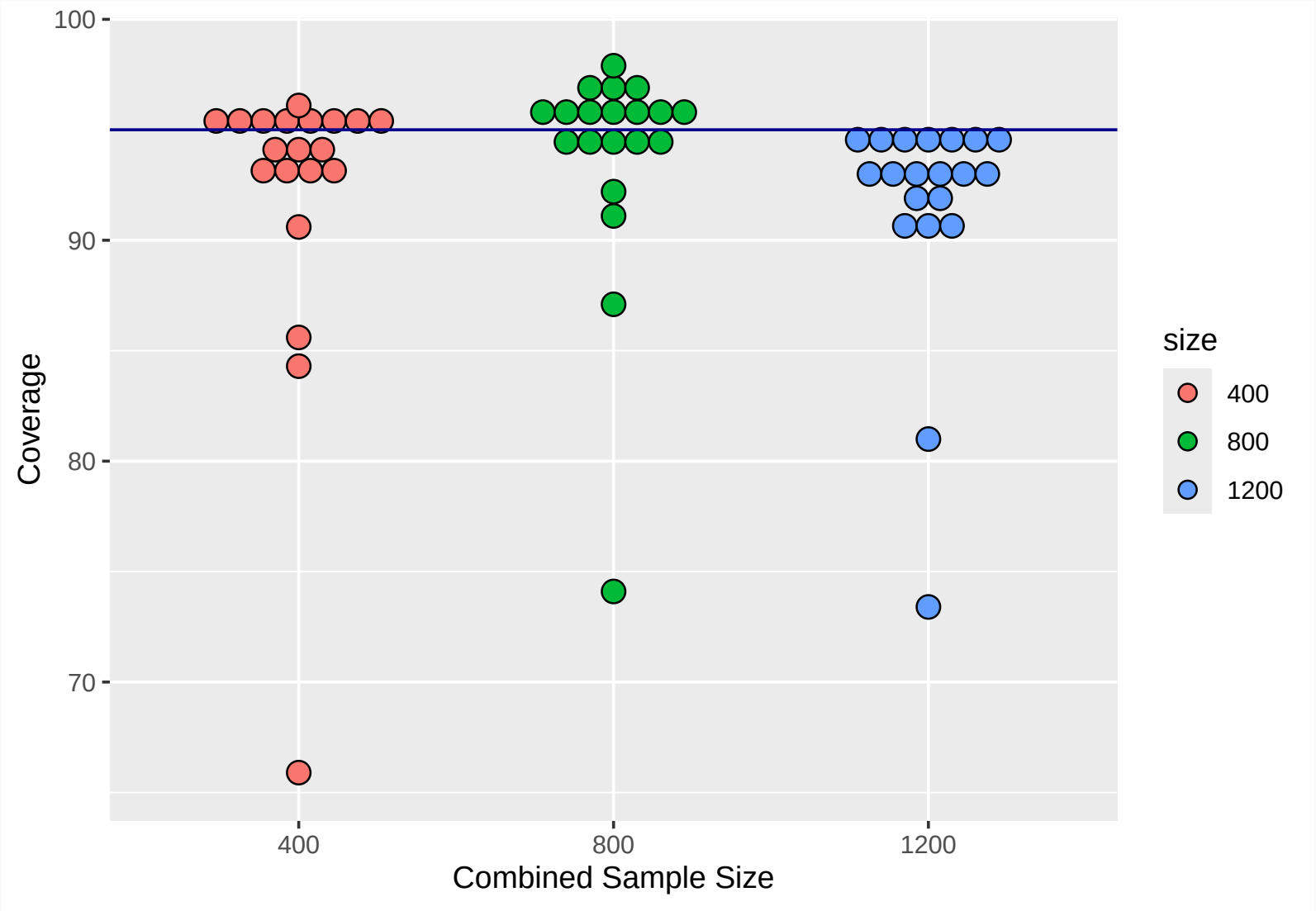
Simulation results: MSE ratios

- Effective sample size ratio across 20 characteristics: (MSE for prob only)/(MSE for combined)



Simulation results: confidence interval coverage

- 95% confidence interval coverage across 20 characteristics

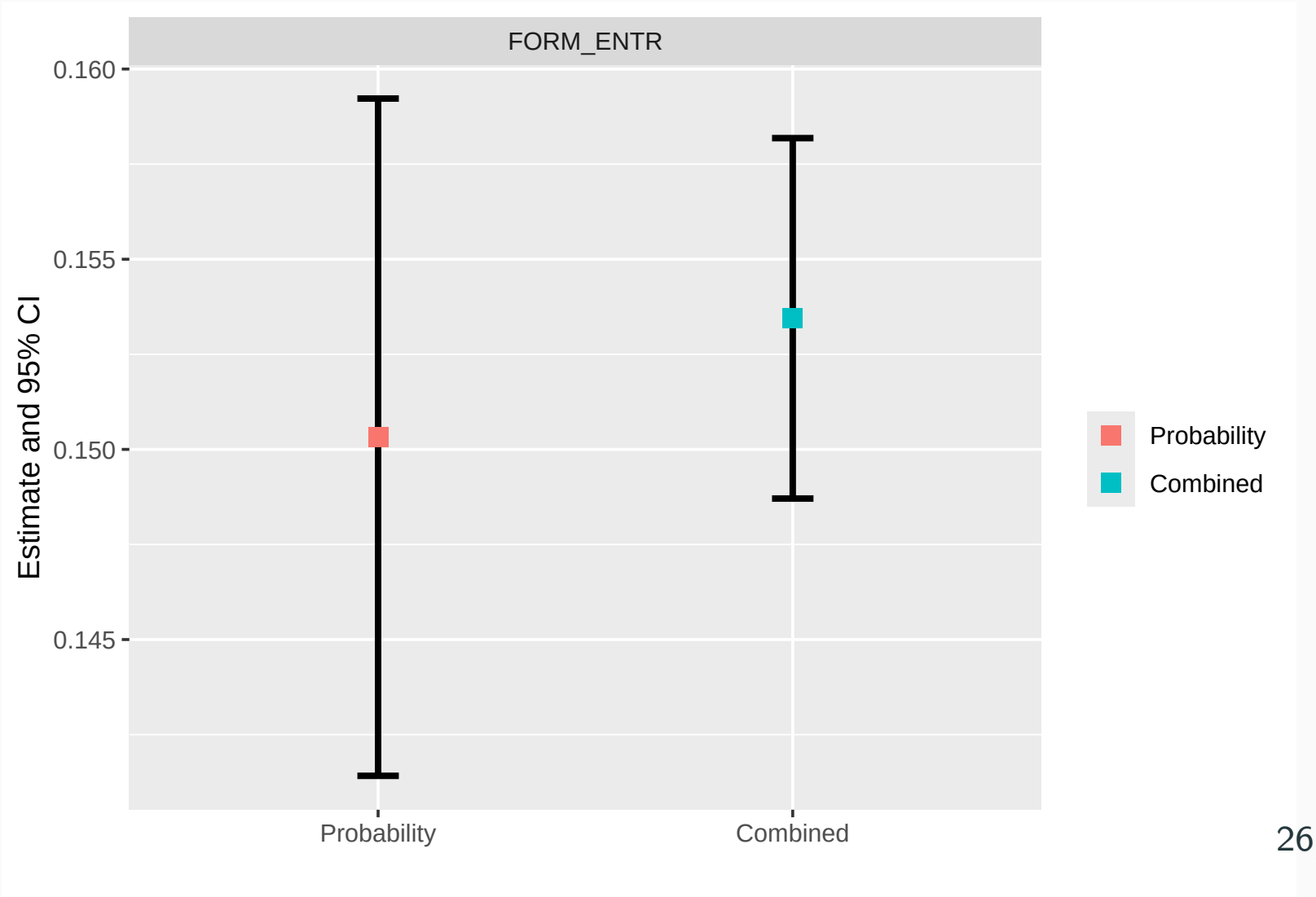


Empirical assessments: real data application

- Entrepreneurship in the Population (EPOP) Survey
  - $|A| = 12,691$, $|B| = 17,941$
  - probability sample came from AmeriSpeak probability panel and address-based sampling, non-probability sample came from web data
- Treat survey weights (with nonresponse adjustments and calibration) as inverse design probabilities
- Choose appropriate leaf size with score$_\ell$
- Get ratio-adjusted combined weights ω_k and dual-frame estimates

EPOP results

- Point estimate and confidence interval
  - non-probability estimate = 0.175



Summary on tree-based dual-frame estimator

- The estimator is robust across a range of Monte Carlo experiments and real data application:
  - yields stable weights by construction
  - fixes bias from non-probability sample
  - improves upon using only the probability sample
  - handles multiple scenarios seamlessly
- Variance estimation is challenging
  - proposed estimating equations approach reduces bias compared to other alternatives
  - confidence interval coverage close to the nominal 95%
- **Thank you!**