Dual-frame estimation approaches for combining probability and nonprobability samples

Jay Breidt

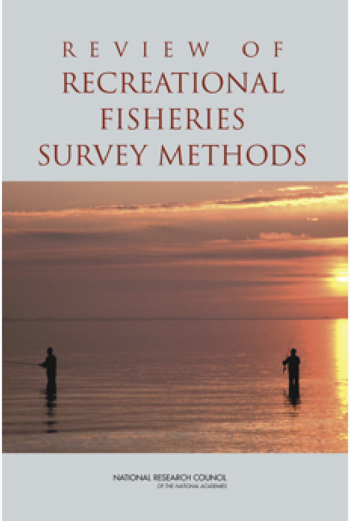
NORC at the University of Chicago

BIRS Workshop: Kelowna, BC

May 24, 2022

Joint work with Chien-Min Huang, Colorado State

Background: Recreational fisheries surveys



- (National Academies Press, 2006)
- Coordinated by NOAA's National Marine Fisheries Service
- Typically, catch estimate is

$$\widehat{(\text{effort})}\,\widehat{\left(\frac{\text{catch}}{\text{effort}}\right)}$$

from (off-site survey) (on-site survey)

This talk: motivated by Large Pelagics Intercept Survey

- Interested in fishing trips that target pelagic species (tuna, sharks, billfish, etc.)
- How many Wahoo were caught by recreational anglers along the US Atlantic coast in 2021?

Survey statistics: sampling from a finite population

- Make inference about a numerical characteristic of a real and well-defined **finite population**

$$
\begin{aligned}
U_{trips} &= \{1, 2, \ldots, N_{trips}\} \\
&= \{\text{all Atlantic large pelagics trips in 2021}\}
\end{aligned}
$$

- y_k = number of Wahoo caught on kth trip
  - *unknown real numbers, not random variables*
- Total Wahoo caught = $T_y = \sum_{k \in U} y_k$
- Infeasible to obtain data on all N large pelagics trips: instead, use a **probability sample** $s \subset U$

Sampling the large pelagics fishery

- No frame U_{trips} of all large pelagics boat-trips
- Instead, sample from frame of site-days: $s \subset U$ $= \{\text{access sites}\} \times \{\text{days in season}\}$
- Count the number of pelagics trips, $\{z_k\}_{k \in s}$
- Collect catch by species for pelagics trips, generically denoted $\{y_k\}_{k \in s}$

Probability sampling: design-based inference

- Universe of elements $U = \{1, 2, \ldots, N\}$
- Variables of interest: y_k, z_k (unknown real numbers)
- Population parameters: $T_y = \sum_{k \in U} y_k$; $T_z = \sum_{k \in U} z_k$; $T_y / T_z = \sum_{k \in U} y_k / \sum_{k \in U} z_k$
- Draw probability sample $s \subset U$ via design with known, positive inclusion probabilities $\pi_k = \Pr[k \in s] > 0$
- Sample membership indicators $I_k = 1$ if $k \in s$, $I_k = 0$ otherwise

$$\mathrm{E}[I_k] = \pi_k$$

Probability sampling: estimation for the population

- Under repeated sampling, the Horvitz-Thompson (1952) estimator
$$\widehat{T}_y = \sum_{k \in U} y_k \frac{I_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{\pi_k}$$
is unbiased for T_y;
$$\widehat{T}_z = \sum_{k \in U} z_k \frac{I_k}{\pi_k} = \sum_{k \in s} \frac{z_k}{\pi_k}$$
is unbiased for T_z
- $\widehat{T}_y / \widehat{T}_z$ is asymptotically unbiased for T_y / T_z

Motivation for nonprobability sampling: LPIS

- Large Pelagics Intercept Survey (LPIS) data are used to estimate **catch rate**: average recreational catch per large pelagic trip, by species: T_y/T_z
- **Problem:** Many site-days have no pelagics trips: $z_k = 0$
    - Field crews want to choose their own site-days!
- **Designed compromise**: select an initial probability sample of site-days $s_o \subset U$ and randomly divide it into s_A and s_B
    - s_A is maintained as a strict probability sample, with **known** inclusion probabilities $\pi_k^A > 0$
    - field crew can leave s_B as-is or move anywhere in $U \setminus s_A$
    - s_B has **unknown** inclusion probabilities π_k^B

Other applications?

- Many surveys involve screening for domain of interest
  - U = households, z_k = age-eligible children, y_k = immunization status
  - U = hospitals, z_k = radiation oncologists, y_k = number of cancer patients
  - U = land segments, z_k = farms served by well water, y_k = pesticide contamination
- Nonprobability sampling methods might be used to build out the initial probability sample

Expert judgment probabilities

- Expert judgment "selection mechanism" is unknown; s_B is no longer a probability sample
- Field crew choose s_B after seeing s_A, so $s_A \cap s_B = \emptyset$

$$
\begin{aligned}
\pi_k^B &= \Pr[k \in s_B \mid k \in s_A] \Pr[k \in s_A] \\
&\quad + \Pr[k \in s_B \mid k \notin s_A] \Pr[k \notin s_A] \\
&= 0 + \rho_k (1 - \pi_k^A)
\end{aligned}
$$

- Need to estimate ρ_k, which may depend on site-day characteristics $\boldsymbol{x}_k$, including trips z_k or catch y_k
- Specify a parametric model for ρ_k and fit using s_A, s_B

Logistic regression model for selection

- Judgment model is Poisson sampling: I_k^B independent Bernoulli(ρ_k) for $k \notin s_A$, with

$$\mathrm{logit}(\rho_k) = \text{linear function of covariates}$$

- Feasible pseudo-log-likelihood is unbiased for log-likelihood:

$$\sum_{k \in U \setminus s_A} I_k^B \ln\left(\frac{\rho_k}{1-\rho_k}\right) + \sum_{k \in U} \ln(1-\rho_k)(1-\pi_k^A)\frac{I_k^A}{\pi_k^A}$$

  - Similar approach if we replace Poisson by with-replacement
- Maximize with respect to parameters in ρ_k and obtain $\widetilde{\rho}_k$
  - Obtain $\widehat{\rho}_k$, normalized version of $\widetilde{\rho}_k$, to match expected sample size n_B
- Estimated inclusion probabilities for s_B are then

$$\widehat{\pi}_k^B = \widehat{\rho}_k\left(1-\pi_k^A\right)$$

Dual-frame judgment sample 1: separate estimator

- Similar to probability sampling from two frames: multiple valid estimators
- Compute HT estimator from sample s_A:

$$\widehat{T}_A = \sum_{k \in s_A} \frac{y_k}{\pi_k^A}$$

- Compute approximate HT estimator from sample s_B:

$$\widehat{T}_B = \sum_{k \in s_B} \frac{y_k}{\left(1 - \pi_k^A\right) \widehat{\rho}_k}$$

- Convex combination, with $\psi \in (0, 1)$:

$$\widehat{T}_{\text{sep}} = \psi \, \widehat{T}_A + (1 - \psi) \, \widehat{T}_B$$

Dual-frame judgment sample 2: combined estimator

- Combine the sample as $s = s_A \cup s_B$
- Compute a combined inclusion probability,

$$\begin{aligned}\Pr[k \in s] &= \Pr[k \in s_A] + \Pr[k \in s_B] - \Pr[k \in s_A \cap s_B] \\ &= \pi_k^A + \left(1 - \pi_k^A\right) \rho_k - 0\end{aligned}$$

- Plugging in $\widehat{\rho}_k$, the resulting HT-like estimator is

$$\widehat{T}_{\mathrm{com}} = \sum_{k \in s_A \cup s_B} \frac{y_k}{\pi_k^A + \left(1 - \pi_k^A\right) \widehat{\rho}_k}$$

- Ensures some weight stability because denominator $\geq \pi_k^A$

Asymptotic properties: combined estimator

- Under some mild assumptions, the combined estimator is design mean square consistent
- $\widehat{V}\left[N^{-1}\widehat{T}_{y,\mathrm{com}}\right]$ is design consistent for $\mathrm{Var}\left(N^{-1}\widehat{T}_{y,\mathrm{com}}\right)$
- The combined estimator is asymptotically normal almost surely (a.s.)

$$\left.\frac{\widehat{T}_{y,\mathrm{com}} - T_{y,N}}{\sqrt{V_A + V_B}}\right| F_N \xrightarrow{\mathcal{L}} N(0,1) \text{ a.s.}$$

- Theoretical support relies on an assumed but possibly wrong model
- Robustness?

Dual-frame doubly-robust estimation

- Possible misspecification of ρ_k
- Consider constructing **doubly-robust** catch estimator by specifying two models:
  - model for the selection probability ρ_k
  - model for the outcome $\mathrm{E}_\xi\left[y_k \mid \boldsymbol{x}_k\right] = m(\boldsymbol{x}_k)$
- Requires auxiliary variable available at population level

$$\widehat{T}_{\mathrm{DR}} \quad = \quad \sum_{k \in s_A \cup s_B} \frac{y_k - \widehat{m}(\boldsymbol{x}_k)}{\pi_k^A + \left(1 - \pi_k^A\right) \widehat{\rho}_k} + \sum_{k \in U} \widehat{m}(\boldsymbol{x}_k)$$

- Consistent (and approximately unbiased) if at least one of the two models is correctly specified

Estimation of catch rate

- But (1) we do not have great covariates available for the whole frame and (2) we are interested in **catch rate**, not catch
- For either separate or combined, estimated catch rate is

$$\widehat{R} = \frac{\widehat{T}_y}{\widehat{T}_z} = \frac{\sum_{k \in s} w_{ks} y_k}{\sum_{k \in s} w_{ks} z_k},$$

  where the weights w_{ks} do not depend on y_k (but may depend on z_k)

Doubly-robust property for rate

- Rate is doubly-robust by construction under a plausible outcome model:

$$\mathrm{E}_\xi[y_k \mid z_k] = \phi z_k$$

- If weights depend on z_k but not y_k and the outcome model is correct, then

$$\mathrm{E}_\xi\left[\frac{\widehat{T}_y}{\widehat{T}_z} - \frac{T_y}{T_z}\right] = \frac{\phi\widehat{T}_z}{\widehat{T}_z} - \frac{\phi T_z}{T_z} = 0,$$

whatever the quality of the probability model

- If the probability model is correct, then

$$\mathrm{E}_p\left[\frac{\widehat{T}_y}{\widehat{T}_z} - \frac{T_y}{T_z}\right] \simeq \frac{T_y}{T_z} - \frac{T_y}{T_z} = 0,$$

whatever the quality of the outcome model

Simulation experiments

- Use historical LPIS data to create population with 30 strata and 57,388 site-days, each with known pressure
- Simulate trips for each site-day using zero-inflated Poisson (matching trip features from LPIS data)
- Given trips, simulate catch for **11 different species** with (possibly truncated)(possibly zero-inflated) Poisson with various relationships with trips:

$$\mathrm{E}[\mathrm{catch} \mid \mathrm{trips}] \propto (\mathrm{trips})^{p}, \quad p \in \{0.5, 1, 2\}$$

Simulation experiments, continued

- Use traditional LPIS sampling design to select original stratified unequal-probability sample, $s_o = \cup_{h=1}^{H} s_{oh}$, of size 865 site-days
- Within each stratum h, divide s_{oh} at random:
  - 75% strict probability sample s_{Ah}
  - 25% movable sample s_{Bh} (can use judgment or leave as-is)
- **Two methods** = sets of constraints on movement of s_{Bh}
  - stratum method: moves remain strictly within stratum
  - bucket method: moves maintain the same allocation by state, month, and kind-of-day (weekday or weekend), but modes can change

Simulated judgment behaviors of field staff

- **No Move (with judgment):** choose to keep sample as originally selected
- **Unskilled:** random moves (simple random sampling)
- Change distribution of zeros only
  - **Expert Jump:** successfully avoids all zero-trip site-days
  - **Skilled Jump:** reduces number of zero-trip site-days
- Change distribution of non-zeros only
  - **Pure Tilt:** increase probability of more trips when there are non-zero trips
- Change distribution of both zeros and non-zeros
  - **Jump and Tilt:** shift the entire distribution toward fewer zeros and higher-value non-zeros
  - **Skilled Shift:** leave half unmoved and move the other half to highest-trip site-days

Nine simulated judgment behaviors, continued

- Generate logistic inclusion probabilities as function of trips

$$\mathrm{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}(z_k = 0) + \beta_2 z_k \mathbf{1}(z_k > 0),$$

  and then
  - **Logistic:** ...draw without-replacement sample using (approximately) these unequal probabilities
  - **With replacement:** ...draw with-replacement sample using (exactly) these unequal probabilities
- **No Move (without judgment)** yields the original probability sample with original (known) weights
  - can we beat this classic design/estimator strategy?
- For all nine judgment behaviors, estimate the unknown conditional inclusion probabilities, $\rho_k = \Pr\left[k \in s_B \mid k \notin s_A\right]$

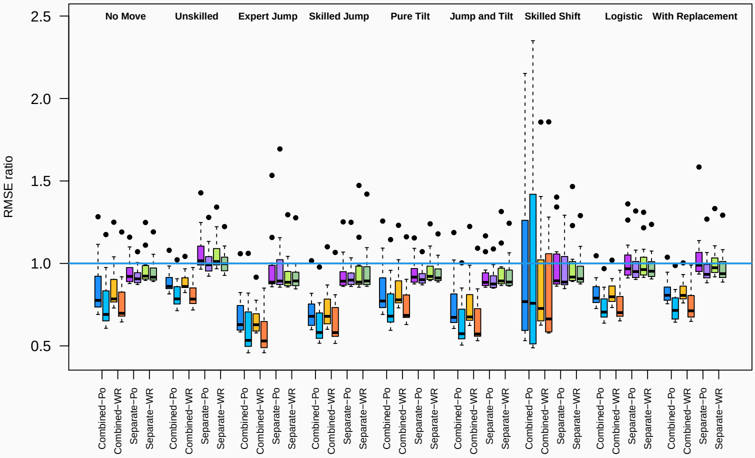
Estimation for each judgment behavior

- For each of 1000 replicated original samples s_o, generate all nine judgment samples under two movement methods
- Model ρ_k as function of trips, z_k:

$$\mathrm{logit}(\rho_k) = \beta_0 + \beta_1 \mathbf{1}(z_k = 0) + \beta_2 z_k \mathbf{1}(z_k > 0)$$

- For all (9 judgment) $\times$ (2 movement method) samples, estimate catch rates for 11 species, using four estimators:
  - **Combined-Po:** Poisson estimates of ρ_k
  - **Combined-WR:** with-replacement estimates of ρ_k
  - **Separate-Po:** Poisson estimates of ρ_k
  - **Separate-WR:** with-replacement estimates of ρ_k
- For **No Move (without judgment)**, compute the weighted estimator using the original design weights

RMSE ratios: less than one favors expert judgment



Variance estimation for the combined estimator

- $\widehat{V}_1$: treat final (combined) weights as if they are traditional survey weights and use Taylor linearization in standard software (easy!)
- $\widehat{V}_2$, $\widehat{V}_3$, $\widehat{V}_4$: derived using Poisson sampling and with replacement sampling approximation
- Replication methods: considered jackknife and grouped balanced repeated replication (BRR)
- Among these variance estimators, $\widehat{V}_1$ has best mean square error property and best confidence interval coverage

Summary for expert judgment sampling

- Estimator using simple (and wrong) model for judgment probabilities works in almost all cases, fixing most of the bias due to judgment sampling
- Combined estimator beats classic strategy (probability sample/weighted estimator) in almost all cases
    - across range of catch characteristics (11 different types)
    - across range of judgment behaviors (9 different types)
    - across two different sets of movement constraints
    - for both Poisson and with-replacement likelihoods
- Combined beats separate estimator in almost all cases
- Simple variance estimator gives good confidence interval coverage

Application in other nonprobability sampling contexts?

- Dual-frame estimation approach works well in our specific context of expert judgment sampling
- Try out this system on two other problems:
    - Respondent-driven sampling with initial probability sample of "seeds" and nonprobability sample of "sprouts"
    - Probability sample with supplemental convenience sample

Dual-frame for respondent-driven samples

- Link-tracing design in research of hidden populations
- Start with a set of initial respondent "seeds" (probability sample), who recruit their peers (nonprobability sample), these in return refer their peers (nonprobability sample), and so on
- Need to estimate the unknown probability of the recruitment process
  - Existing methods make strong modeling assumptions on how recruitment works
- We propose to apply the dual-frame estimator directly to RDS
  - Assess robustness to misspecified recruitment model via simulation

Simulation experiment of RDS

- Artificial population: Use **Project 90** network sample data, from a study of heterosexuals' transmission of HIV
  - 4430 individuals and 18407 edges
  - 13 binary attributes (including retired, female, pimp, $\cdots$)
- Simulated respondent-driven sampling design: mimics a real LGBT study in Michaels et al. (2019, *J. Official Stat*)
  - Start with 100 random seeds, seeds selected randomly or proportional to degree
  - Target sample size is 130 or 150
  - Each respondent recruits up to 3 peers

Estimators for comparison

- SH (Salganik and Heckathorn 2004) estimator: restricted to categorical outcomes

$$\widehat{\mu}_A^{\mathrm{SH}} = \frac{\widehat{d}_B \widehat{C}_{BA}}{\widehat{d}_A \widehat{C}_{AB} + \widehat{d}_B \widehat{C}_{BA}}$$

- VH (Volz and Heckathorn 2008) estimator:

$$\widehat{\mu}_y^{\mathrm{VH}} = \frac{\sum_{k\in s} d_k^{-1} y_k}{\sum_{k\in s} d_k^{-1}}$$

- SS (Gile 2011) estimator:

$$\widehat{\mu}_y^{\mathrm{SS}} = \frac{\sum_{k\in s} \widehat{\pi}^{-1}(d_k) y_k}{\sum_{k\in s} \widehat{\pi}^{-1}(d_k)}$$

Estimators for comparison, continued

- Combined estimator: dual-frame approach

$$\widehat{\mu}_y^{\mathrm{com}} = \frac{\sum_{k\in s} \frac{y_k}{\pi_k^A + \left(1-\pi_k^A\right)\widehat{\rho}_k}}{\sum_{k\in s} \frac{1}{\pi_k^A + \left(1-\pi_k^A\right)\widehat{\rho}_k}}$$

- Convex estimator: convex combination of VH and combined estimator

$$\widehat{\mu}_y^{\mathrm{cnvx}} = \frac{\sum_{k\in s} \left[\frac{n_A}{n_A+n_B} \frac{1}{\pi_k^A+(1-\pi_k^A)\rho_k} + \frac{n_B}{n_A+n_B} \frac{N d_k^{-1}}{\sum_{k\in s} d_k^{-1}}\right] y_k}{\sum_{k\in s} \left[\frac{n_A}{n_A+n_B} \frac{1}{\pi_k^A+(1-\pi_k^A)\rho_k} + \frac{n_B}{n_A+n_B} \frac{N d_k^{-1}}{\sum_{k\in s} d_k^{-1}}\right] 1}$$

Simulated recruitment behavior of respondent

- **Random:** acquaintances are recruited at random (standard assumption)
- **Recruit fraction:** 0, 1, 2, or 3 acquaintances are recruited at random, with probabilities $(1/6, 1/6, 1/6, 1/2)$
- **Degree:** recruitment probabilities are proportional to the degrees of acquaintances
- **Inverse degree:** recruitment probabilities are proportional to the inverse degrees of acquaintances
- **Prefer female:** females must recruit female, males recruit males
- **Prefer pimp:** pimps must recruit pimp, non-pimps recruit non-pimps
- **Expert female:** everyone must recruit female
- **Expert pimp:** everyone must recruit pimp

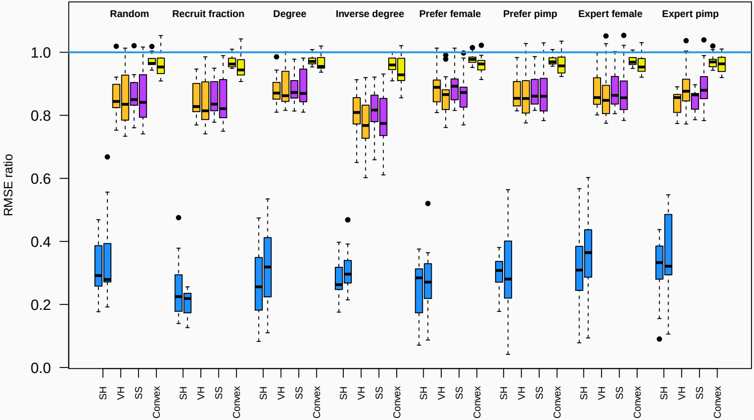
Estimation of recruitment behavior

- Requires a model of recruitment behavior for the combined estimator, simple model of degree:

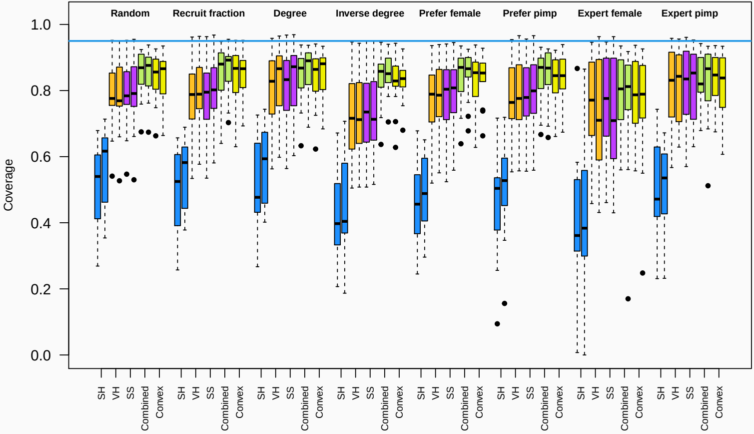
$$\text{logit}(\rho_k) = \beta_0 + \beta_1 \text{degree},$$

  fitted by maximizing pseudo-log-likelihood
- For each of 1000 replicated probability samples,
  - generate all eight versions of the recruited sample
  - estimate rates and variances for 13 attributes using SH, VH, SS
  - estimate ρ_k and rates for 13 attributes using Combined and Convex, with variances computed by treating final combined weights as if they are traditional survey weights

RMSE ratios: less than one favors combined estimator



95% confidence interval coverage across all attributes



Summary for respondent-driven samples

- In our limited simulation setting, the combined estimator dominates the existing estimators
    - robust across a range of attributes and across a range of recruitment behaviors
    - no strong assumptions required
    - simple variance estimator of the combined estimator gives good confidence interval coverage
- In other settings, like fewer random seeds or longer waves of recruitment, the existing estimators are more competitive

Dual-frame for convenience samples

- Increasingly common as response to surveys decreases, the cost of obtaining probability sample is high
  - Small, representative probability sample drawn from the whole population U; large, biased convenience sample drawn from the sub-population U_B
- Example: **Culture and Community in a Time of Crisis** (CCTC)
  - probability sample s_A from U.S. general population, with known inclusion probabilities $\pi_k^A > 0$
  - nonprobability sample s_B from art organization mailing list, with unknown inclusion probabilities $\pi_k^B = (1 - \pi_k^A)\rho_k$
- **Goal:** Combine these two resources using dual-frame method
- **Challenge:** For $k \in U_B$, π_k^A and ρ_k are unknown

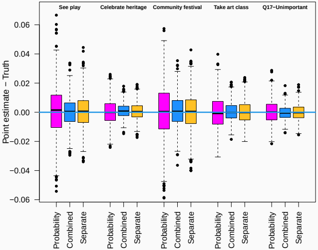
Modifications to dual-frame methodology

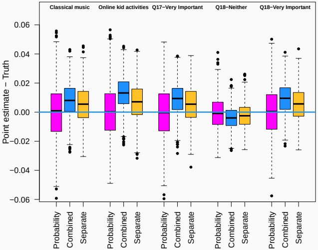
- Since π_k^A is unknown for $k \in s_B$, use covariates available in both samples to find a matching record $\ell \in s_A$ and assign its inclusion probability
- U_B is a strict subset of U, hence part of s_A will not match s_B
- Use the matched part of s_A plus s_B in dual-frame estimation for the matched part of the frame, U_B
    - includes likelihood-based estimation of ρ_k, $k \in s_B$
- Use the unmatched part of s_A in single-frame estimation for the unmatched part of the frame, $U \setminus U_B$

Simulation experiment from CCTC

- For a JSM 2021 competition, NORC used CCTC to create a simulation platform to study prob/nonprob combination
  - population U consists of 113,459 records
  - subpopulation U_B consists of 74,202 records
  - 22 binary variables of interest: see a play, celebrate heritage, take art class, ...
  - 1000 simulated probability samples s_A of size $n_A = 1000$
  - 1000 simulated nonprobability samples s_B of size $n_B = 4000$
- Known inclusion probabilities for s_A
- Unknown inclusion probabilities for s_B
- Many possible covariates for matching and propensity estimation

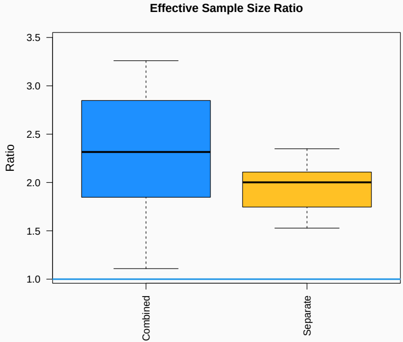
Estimation summary across all variables





- Best five and worst five responses (among 22)
- Combined has lower MSE than separate in most cases
- Using nonprobability data dominates probability only
- Simple variance estimator yields confidence intervals with coverage close to nominal

Effective sample size ratios across all variables



- Ratio of MSE for combined sample to MSE of probability sample only
- Ratio $\simeq 5$ if nonprobability contains as much information as probability and we are fully efficient in extracting the information
- Combined mostly dominates separate
- Either dominates probability only

Conclusions and thanks

- Dual-frame is simple and effective method for combining probability and nonprobability samples
  - single set of weights with some weight stability by construction
  - some double robustness if estimating rates
  - simple variance estimation and confidence intervals
- Considerable robustness across a range of situations
  - nonprobability types include expert judgment, respondent-driven samples, or convenience samples
  - wide variety of simulated settings within nonprobability samples
- Ongoing work: further development of matching and estimation methods for convenience samples
- Thank you!