Synergies between Survey Statistics and Causal Inference: Moving from the Pipette to the Population

Michael Elliott[1,2]

[1] Department of Biostatistics, University of Michigan
[2] Survey Methodology Program, Institute for Social Research

Overview

- Review of finite population inference
- Review of causal inference
- Commonalities in problems faced and solutions provided
- Extending these synergies: from Pipette to Patient to Patient to Population
    - Generalizing (“transporting”) causal inference from randomized trials to a target population

Finite Population Inference

- Making inference about a fixed and well-defined population of size N.

$$(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_N, \mathbf{X}_N)$$

  - US population resident April 1, 2020.
  - Michigan residents who received a COVID-19 diagnosis between March 1, 2020, and February 28, 2021.
  - North American auto parts manufactures with a gross income in excess of $1M between 2010 and 2020

Finite Population Inference

- Making inference about a fixed and well-defined population of size N.

$$(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_N, \mathbf{X}_N)$$

  - US population resident April 1, 2020.
  - Michigan residents who received a COVID-19 diagnosis between March 1, 2020, and February 28, 2021.
  - North American auto parts manufactures with a gross income in excess of \$1M between 2010 and 2020

- Focus is on inference about a population quantity: a descriptive statistic such as a population mean $\overline{Y} = N^{-1} \sum_{i=1}^{N} Y_i$, or a model parameter such as linear regression coefficients $\mathbf{B} = \left(\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i Y_i$.

Finite Population Inference

Overview of the population

i	I	Y	X
1	1	Y_1	X_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	Y_n	X_n
n+1	0	?	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	0	?	?

Two major approaches to finite population inference:

- Randomization or “design-based” inference
- Finite population Bayesian inference

Randomization or “Design-based” Inference

- Population data treated as fixed and sampling indicators I as random.
- (Asymptotically) unbiased estimators of the population quantity of interest.
- (Asymptotically) unbiased estimators of variance of these population quantity estimators with respect to repeated sampling of I.

Finite Population Bayesian Inference

- Finite population Bayesian inference imputes the unobserved portions of the population using posterior predictive distributions (Ericson 1969):

$$P(Y_{nobs}, \mathbf{X}_{nobs} \mid y_{obs}, \mathbf{x}_{obs}) =$$

$$\int P(Y_{nobs}, \mathbf{X}_{nobs} \mid \theta, y_{obs}, \mathbf{x}_{obs}) P(\theta \mid y_{obs}, \mathbf{x}_{obs}) d\theta \propto$$

$$\int P(Y_{nobs}, \mathbf{X}_{nobs} \mid \theta, y_{obs}, \mathbf{x}_{obs}) P(y_{obs}, \mathbf{x}_{obs} \mid \theta) P(\theta) d\theta$$

- The model for $(y_{obs}, \mathbf{x}_{obs} \mid \theta)$ should incorporate sensible design features
    - In an unequal probability of selection design, the means and possibly the variances should be a function of sampling probabilities.
- Can use hierarchical modeling to smooth effects of design and do bias-variance tradeoffs to minimize mean square error (Elliott and Little 2000).

Finite Population Inference

- Design based inference is “model” and “distribution free” (although some estimators can be derived from models).
- Bayesian inference uses models to reduce variance but can be susceptible to model misspecification.
- Use models that incorporate design features and robust models (splines, Dirichlet processes, etc.) (Elliott and Little 2000; Elliott 2007; Elliott and Xia 2021).

Finite Population Inference

- “Model assisted’ estimators can bring together elements of both (usually in a design-based framework) (Särndal et al. 2003).
  - Suppose $\mathbf{X}$ is known in the population, and a model for $Y_i \mid \mathbf{X}_i$ is developed with $E(Y_i \mid \mathbf{X}_i) = m_i$. The “doubly robust” estimator:

    $$\overline{y}_{DR} = n^{-1} \sum_{i=1}^{n} (y_i - \hat{m}_i) + N^{-1} \sum_{i=1}^{N} \hat{m}_i$$

    is unbiased for $\overline{Y}$ with respect to repeated sampling even if the mean is misspecified but becomes more efficient than $\overline{y}$ as $E(Y_i \mid \mathbf{X}_i) \to m_i$

Causal Inference: Potential Outcomes

- “We may define a cause to be an object, followed by another. . . where, if the first object had not been, the second had never existed.” (Hume 1748).
- Rubin Causal Model (Holland 1986): consider “potential outcomes” for the same subject Y_i under different treatment levels $Z = 1, ...T$: $Y_i^1, ..., Y_i^T$.
- Average casual treatment effect comparing treatment level $Z = z$ to $Z = z'$: $ACE = N^{-1} \sum_{i=1}^{N} (Y_i^z - Y_i^{z'})$

Causal Inference: Potential Outcomes

- “We may define a cause to be an object, followed by another. . . where, if the first object had not been, the second had never existed.” (Hume 1748).
- Rubin Causal Model (Holland 1986): consider “potential outcomes” for the same subject Y_i under different treatment levels $Z = 1, ... T$: $Y_i^1, ..., Y_i^T$.
- Average casual treatment effect comparing treatment level $Z = z$ to $Z = z'$: $ACE = N^{-1} \sum_{i=1}^{N} (Y_i^z - Y_i^{z'})$
- Fundamental Problem of Causal Inference: We only observe the outcome for the actual treatment given: $Y_i^{Z_i = z}$. All others are counterfactual. If Z is binary:

i	I	Z	Y^0	Y^1	X
1	1	1	?	Y_1^1	X_1
1	1	0	Y_2^0	?	X_2
⋮	⋮	⋮	⋮	⋮	⋮
n	1	1	?	Y_n^1	X_n
n+1	0	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮
N	0	?	?	?	?

Causal Inference: Assignment Mechanism

- Focus is typically on assignment mechanism: $P(Z \mid Y^1, Y^0, \mathbf{X})$
- Control over treatment assignment: randomized assignment breaks all associations and confounding, so that

  $$P(Z \mid Y^1, Y^0, \mathbf{X}) = P(Z) \text{ (typically } 1/T \text{ to maximize power).}$$

- If treatment assignment is uncontrolled but is a function of covariates unaffected by treatment, then

  $$P(Z \mid Y^1, Y^0, \mathbf{X}) = P(Z \mid \mathbf{X}) = p(\mathbf{X}).$$

  (sometimes termed the “propensity score”) (Rosenbaum and Rubin 1983).
- If treatment assignment is a function of the potential outcomes conditional in X (or equivalently unobserved confounders), then $P(Z \mid Y^1, Y^0, \mathbf{X})$ cannot be reduced and inference will typically depend on unidentified parameters.

Causal Inference: Randomization vs. Bayesian Inference

- Randomization inference: considers the possible assignments of Z_i across the population given the assignment mechanism of Z_i
  - If treatment assignment is randomized, the observed mean difference $\overline{y}_1 - \overline{y}_0 = n_1^{-1} \sum_{i=1}^n Z_i y_i - n_0^{-1} \sum_{i=1}^n (1 - Z_i) y_i$ is unbiased for sample ACE.
  - If treatment assignment is not randomized but depends on **X**, the propensity weighted mean difference $\overline{y}_{w1} - \overline{y}_{w0} = \frac{\sum_{i=1}^n Z_i/p(\mathbf{X}_i) y_i}{\sum_{i=1}^n Z_i/p(\mathbf{X}_i)} - \frac{\sum_{i=1}^n (1-Z_i)/(1-p(\mathbf{X}_i)) y_i}{\sum_{i=1}^n (1-Z_i)/(1-p(\mathbf{X}_i))}$ is unbiased for sample ACE.
  - Proof of unbiasedness is similar to that of SRS/weighted means in probability sampling setting.

Causal Inference: Randomization vs. Bayesian Inference

- Bayesian inference proceeds similar to the survey setting:

$$P(Y_{nobs} \mid y_{obs}, \mathbf{X}) =$$

$$\int P(Y_{nobs} \mid \theta, y_{obs}, \mathbf{X}) P(\theta \mid y_{obs}, \mathbf{X} \mid \theta) d\theta \propto$$

$$\int P(Y_{nobs} \mid \theta, y_{obs}, \mathbf{X}) P(y_{obs}, \mathbf{X} \mid \theta) P(\theta) d\theta$$

Obtain draws of $\overline{Y}^{1^{(b)}}$ as $n^{-1} \sum_{i=1}^{n} z_i y_i + (1 - z_i) Y_i^{1^{(b)}}$ where $Y_i^{1^{(b)}}$ is a posterior predictive draw of the control outcome in subjects assigned to the treatment arm, and similarly $\overline{Y}^{0^{(b}}$ as $n^{-1} \sum_{i=1}^{n} z_i Y_i^{0^{(b)}} + (1 - z_i) y_i$. A draw of the sample ACE is then given by $\Delta^{(b)} = \overline{Y}^{1^{(b)}} - \overline{Y}^{0^{(b)}}$.

  - Differences: only outcome is missing; often model will involve unidentified parameters (e.g, $C(Y_i^1, Y_i^0)$).
  - Modeling still needs to be sensitive to the design.

Causal Inference: Doubly Robust Estimators

- Doubly robust model assisted estimators are available in the casual inference setting as well.
- Since the propensity score is a balancing score – it summarizes all the information about the association between treatment Z and covariates $\mathbf{X}$ – we need only model the potential outcome as a function of $p(\mathbf{X})$ to obtain a consistent estimator.
- But we can also add an standard mean model to predict the potential outcome using covariates directly (Bang and Robins 2005):

$$E(Y_i \mid Z_i, \mathbf{X}_i) = \mathbf{X}_i^T\beta + \phi_1 Z_i p_i(\mathbf{X})^{-1} + \phi_2(1 - Z_i)(1 - p_i(\mathbf{X}))^{-1}$$

  - If either the mean $\mathbf{X}_i^T\beta$ or the propensity score $p_i(\mathbf{X})$ is correctly specified, then a consistent estimator of the ACE is given by

$$\Delta = n^{-1}\sum_{i=1}^{n}(\hat{E}(Y_i \mid 1, \mathbf{X}_i) - \hat{E}(Y_i \mid 0, \mathbf{X}_i))$$

Causal Inference: Doubly Robust Estimators

- Doubly robust model assisted estimators are available in the casual inference setting as well.
- Since the propensity score is a balancing score – it summarizes all the information about the association between treatment Z and covariates $\mathbf{X}$ – we need only model the potential outcome as a function of $p(\mathbf{X})$ to obtain a consistent estimator.
- But we can also add an standard mean model to predict the potential outcome using covariates directly (Bang and Robins 2005):

$$E(Y_i \mid Z_i, \mathbf{X}_i) = \mathbf{X}_i^T \beta + \phi_1 Z_i p_i(\mathbf{X})^{-1} + \phi_2 (1 - Z_i)(1 - p_i(\mathbf{X}))^{-1}$$

  - If either the mean $\mathbf{X}_i^T \beta$ or the propensity score $p_i(\mathbf{X})$ is correctly specified, then a consistent estimator of the ACE is given by

$$\Delta = n^{-1} \sum_{i=1}^{n} (\hat{E}(Y_i \mid 1, \mathbf{X}_i) - \hat{E}(Y_i \mid 0, \mathbf{X}_i))$$

- An alternative approach replaces $\phi_1 Z_i p_i(\mathbf{X})^{-1} + \phi_2 (1 - Z_i)(1 - p_i(\mathbf{X}))^{-1}$ with a treatment-specific spline on $p_i(\mathbf{X})$ and uses a Bayesian multiple imputation approach for improved efficiency (Zhou et al. 2019).

Survey and Causal Inference: Overlap

- Both involve selection bias: one into a population and the other into a treatment assignment.
- Both involve missing data: one for the unsampled component of the population and the other for the unassigned outcome in the sample.
- In both cases the randomization approach uses weighting and generalized estimating equations to compute point estimates and confidence intervals.
- In both cases the Bayesian approach uses multiple imputation to impute missing data while account for potential selection bias in the modeling.
  - Randomization offers robustness at sometimes extreme efficiency costs; Bayesian approach offers efficiency but always requires careful model considerations.

Pipette to Patient to Population

- In the clinical trial world we discuss “bench to bedside” (or “pipette to patient”), bringing the results of biological research to improve patient health.
- But a missing piece is in step from the patient to the population
- Note that I previously referred to the *sample* ACE, not the *population* ACE, when discussing the ACE estimators.
- “Transporting” the sample ACE estimators to the population ACE requires understanding the relationship between the treatment effect in the sample and the treatment effect in the population.

Pipette to Patient to Population

Survey Inference

i	I	Z	Y^0	Y^1	X
1	1	1	?	Y_1^1	X_1
1	1	0	Y_2^0	?	X_2
⋮	⋮	⋮	⋮	⋮	⋮
n	1	1	?	Y_n^1	X_n
n+1	0	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮
N	0	?	?	?	?

Pipette to Patient to Population

Causal Inference: Trial

i	I	Z	Y^0	Y^1	X
1	1	1	?	Y_1^1	X_1
1	1	0	Y_2^0	?	X_2
⋮	⋮	⋮	⋮	⋮	⋮
n	1	1	?	Y_n^1	X_n
n+1	0	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮
N	0	?	?	?	?

Pipette to Patient to Population

Causal Inference: Population

i	I	Z	Y^0	Y^1	X
1	1	1	?	Y_1^1	X_1
1	1	0	Y_2^0	?	X_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	1	?	Y_n^1	X_n
n+1	0	?	?	?	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	0	?	?	?	?

Effect Modification

- A key issue is that the clinical trial sample is rarely a probability sample from the population, so standard survey methods cannot typically be applied to clinical trials samples.
- This leads to the issue of *transportability*: making inference to a relevant reference population rather than a generalization of the trial population (Westreich et al. 2017).
- Why do we care? In clinical trials (putting non-compliance aside) we are in the happy situation were treatment assignment is randomized, so $p(\mathbf{X}) \equiv p$ is constant, and confounding is not an issue.
- Randomization of treatment eliminates the effect of unobserved *confounders*, but it does not eliminate the effect of unobserved *effect modifiers*.

Effect Modification

- Suppose our true model relating an outcome to a treatment involves an unobserved variable U that is both a confounder and effect modifier (Elliott 2016):

$$E(Y|U,Z) = \beta_0 + \beta_1 Z + \beta_2 U + \beta_3 ZU$$

- $ACE = E_u(E(Y^1 - Y^0 \mid U = u)) = \beta_1 + \beta_3 \mu_U$.
- $E(\overline{y}_1 - \overline{y}_0)$ without randomization (but assuming a linear association between U and Z) is $\beta_1 + (\alpha_0 + 2\alpha_1)\beta_3$ where $\alpha_0 = E(U) - \sigma_{UZ}/\sigma_Z^2 E(Z)$ and $\alpha_1 = \sigma_{UZ}/\sigma_Z^2 E(Z)$.
- Randomization guarantees $U \perp Z$ and thus $E(\overline{y}_1 - \overline{y}_0) = \beta_1 + \beta_3 E(U)$.
- However to guarantee $E(U) = \mu_U$ requires either a probability sample or some type of adjustment to make the clinical trial result representative (realistically, more representative) of the population.

Generalizability Review

- Seminal works include Cole and Stuart (2010) and Stuart et al. (2011).
  - Cole and Stuart combined data from a RCT of HIV testing the effect of a protease inhibitor with data from US-wide surveillance of new HIV cases to develop inverse probability of selection weights.
    - Weighted Cox PH models found a marginally significant RR of 0.57 (95% CI 0.33-1.00) versus the highly significant RR of 0.51 (95% CI 0.33-0.77) in the RCT.
  - Stuart et al. developed a propensity matching method to complement the IPWT method, based on the propensity to be in population sample.
- Hartman et al. (2016) adapted the IPWT method by first pairing cases with controls within the RCT, and then weighting these pairs to better match the distribution of the population of interest.
  - Estimates treatment effect among the treated by weighting the pairs to the treated population.

Generalizability Review

- Kern et al. (2016) model outcomes as a function of covariates and treatment status, allowing for interactions between the two. These models predict outcomes under treatment and control within data from the population or a representative sample thereof.
- “Doubly-robust” methods that combine propensity score weights and outcome models have been the focus of recent developments.
  - Dahabreh et al. (2020) consider three versions of these estimators that combine predictions of the outcome under treatment or control in the representative sample with IPTW-weighted residuals of the outcome model in the RCT.
  - Schmid et al. (2022) consider a targeted maximum likelihood estimator (TMLE) that uses a “clever covariate” (the IPTW weight itself) together with the outcome model to predict the outcome under treatment and control in the representative sample.
- Degtiar and Rose (2023) provide a overview of the currents methods used for RCT generalizability.

Non-probability Inference Review

- Traditionally population inference has focused on probability inference (Neyman 1934).

  Cost, response rates, and new types of available data have led to a rethinking of rethinking of the role of non-probability samples (Baker et al. 2013).

- Valliant and Dever 2011 develop IPWTs to estimate a "true" probability of selection for the non-probability sample elements in a manner similar to Cole and Stuart. Elliott et al. (2010) developes IPWTs in a somewhat different manner.
- Similar to Stuart et al. 2011, Rivers (2006) matched subjects in the non-probability sample to subjects in the probability sample via a propensity score to be in the probability sample, with the matched nonprobability sample used for inference.

Non-probability Inference Review

- Direct outcome regression models that predict outcomes based on covariates are less common in the non-probability literature, perhaps because of survey statisticians' traditional aversion to fully model-based approaches.
- But "doubly robust" estimators have been developed: Chen et al. (2020) use estimators that combine model-based estimates from the probability sample with propensity-weighted residuals from non-probability sample.
- A review of estimation from non-probability samples is available at Wu (2022).

Distinctions between the Generalizability and Probability/Non-probability Sampling Literature

There are many similarities between the RCT generalizability literature and the combining of probability and non-probability samples literature, but there are also key distinctions.

- With the exception of Ackerman (2021), the generalizability literature has generally ignored complex sample design features such as weighting, clustering, or stratification in the benchmark probability sample, although these features are commonly present in both general population surveys.
- While the probability survey literature has a large section devoted to missing data, it usually does not face a setting where all observations have missing elements in a joint distribution of interest.
- The relevant patient population may be more difficult to define, let alone obtain a high-quality sample from.

Our Proposed Work: Notation and Assumptions

- Notation:
  - Defined population of size N.
  - Binary treatment $Z_i \in \{0, 1\}$, with potential outcomes $Y(0)_i$ and $Y(1)_i$.
  - Sampling indicators S_i^R (R=randomized trial) and S_i^B (B=probability/benchmark dataset)..
  - Probabilty of being sampled in B is known: $P(S_i^B = 1) = \pi_i^B$.
  - Common covariates X_i in B and R.
- Assumptions:
  - Randomization: $(Y(1)_i, Y(0)_i) \perp Z_i \mid S_i^R = 1$;
  - Stable Unit Value Treatment Assignment (SUTVA): the observed outcome $Y_i = z_i Y(1)_i + (1 - z_i) Y(0)_i$ for treatment assignment $Z_i = z_i$;
  - Positivity: $P(S_i^R = 1) > 0$ and $P(S_i^B = 1) > 0$ for all i;
  - Estimability: $P(S_i^R = 1) = \pi_i^R = g(X_i; \theta)$ for known g and unknown θ;
  - Ignorability: $(Y(1)_i, Y(0)_i) \perp S_i^R, S_i^B \mid X_i$.

Pseudo-weights

- Standard inverse probability weighting (Valliant and Dever (2011)):

$$\pi_i^{RB} = P\left(S_i^R = 1 | X_i = x_i, S_i^B = 1 \text{ or } S_i^R = 1\right).$$

  where π_i^{RB} is estimated by (weighted) logistic regression,

- Elliott et al. (2011) show via Bayes' rule that

$$\pi_i^R = P(S_i^R = 1 | X_i = x_i) \propto$$

$$P(S_i^B = 1 | X_i = x_i) \frac{P\left(S_i^R = 1 | X_i = x_i, S_i^B = 1 \text{ or } S_i^R = 1\right)}{1 - P\left(S_i^R = 1 | X_i = x_i, S_i^B = 1 \text{ or } S_i^R = 1\right)} = \pi_i^B \times \frac{\pi_i^{RB}}{1 - \pi_i^{RB}}$$

  The components of π_i^R can be estimated using generalized linear regression or Bayesian Additive Regression Trees (Chipman et al. 2010).

- Chen et al. (2020) argue that $\hat{\pi}_i^{RB}$ is not a consistent estimator of π_i^R unless π_i^R is a constant. Chen et al. suggest an maximum likelihood estimator of π_i^R that does provide a consistent estimator; however, it does not easily admit non-linear estimators such as BART.

Prediction

Under randomization, we have

$$E(Y(1)_i \mid X_i) = E(Y_i \mid X_i, Z_i = 1) \mid S_i^R = 1$$

$$E(Y(0)_i \mid X_i) = E(Y_i \mid X_i, Z_i = 0) \mid S_i^R = 1.$$

Thus a correct model of $E(Y_i \mid X_i, Z_i)$ allows prediction of $Y_i(1 - Z_i)$, and the following estimators of the PATE are

$$\hat{\Delta}_{WVD} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{RB}[Z_i(y_i - \hat{Y}(0)_i) + (1 - Z_i)(\hat{Y}(1)_i - y_i)]}{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{RB}}$$

$$\hat{\Delta}_{WE} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{R}[Z_i(y_i - \hat{Y}(0)_i) + (1 - Z_i)(\hat{Y}(1)_i - y_i)]}{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{R}}$$

Prediction

If outcome information is available in the probability sample, an alternative that only uses prediction is

$$\Delta_{PRED} = \hat{\hat{Y}}(1) - \hat{\hat{Y}}(0),$$

$$\hat{\hat{Y}}(Z) = \frac{\sum_{i=1}^{N} [I(S_i^R = 1) + (w_i^B - n_R/n_B) I(S_i^B = 1)][I(Z = z_i) y_i + I(Z = 1 - z_i) \hat{Y}_i(Z)]}{n_R + \sum_{i=1}^{N} I(S_i^B = 1)(w_i^B - n_R/n_B)}$$

Inference

- Since all of the methods we are consider involve estimating $Y(1-z_i)$ using BART, we will use a Bayesian approach for inference.
- Each draw of $Y(1-z_i)$ generates a draw of the relevant PATE estimator.
  - Point estimates are obtained as the posterior mean of these draws, with 1-α credible intervals obtained from the $\alpha/2$ and $1-\alpha/2$ empirical CDFs.
  - For Δ_{WE} we also consider an estimator of the variance (Δ_{WE2}) that incorporates uncertainty in the estimation of π_i^R.

Inference

- Because the prediction model uses a complex sample design for the probability sample, we use Rubin's Rules for combining multiple imputations:

$$\hat{E}(\Delta_{PRED} \mid \text{data}) = \frac{1}{B}\sum_{b=1}^{B} \Delta_{PRED}^{(b)}$$

$$v(\Delta_{PRED} \mid \text{data}) = \frac{1}{B}\sum_{b=1}^{B} v(\Delta_{PRED}^{(b)}) +$$

$$\frac{B+1}{B}\frac{1}{B-1}\sum_{b=1}^{B}\left(\Delta_{PRED}^{(b)} - \hat{E}(\Delta_{PRED} \mid \text{data})\right)^2$$

where $v(\Delta_{PATE}^{(b)})$ is estimated using a design-based estimator of variance that treats the imputed values of $Y(1-z_i)$ as observed.

Treatment effect among the treated

Simulations and example focus on population treatment effect among the treated (PATT):

$$\hat{\Delta}_{WVD,PATT} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{RB} Z_i (y_i - \hat{Y}(0)_i)}{\sum_{i=1}^{N} I(S_i^R = 1) Z_i / \hat{\pi}_i^{RB}}$$

$$\hat{\Delta}_{WE,PATT} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{R} Z_i (y_i - \hat{Y}(0)_i)}{\sum_{i=1}^{N} I(S_i^R = 1) Z_i / \hat{\pi}_i^{R}}$$

$$\hat{\Delta}_{PRED,PATT} = \frac{\sum_{i=1}^{N} [I(S_i^R = 1) + (w_i^B - n_R/n_B) I(S_i^B = 1)] Z_i [y_i - \hat{Y}_i(0)]}{n_{R_1} + \sum_{i=1}^{N} I(S_i^B = 1) Z_i (w_i^B - n_R/n_B)}$$

where n_{R_1} is the number of observations assigned to treatment in the RCT.

- Inference using BART for prediction proceeds as the in estimation of the PATE.

Simulation Study

- A linear model is used to generate each potential outcome $Y(Z)$ for a binary treatment Z. The linear predictor for $Y(1)$ has two normally distributed covariates, X_1 and X_2: $N = 20,000$, $n = 1000$.

$$Y(Z) \sim \mathcal{N}(\mu_Z, 1)$$

$$\mu_1 = \beta_0 + \delta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2, \mu_0 = \beta_0$$

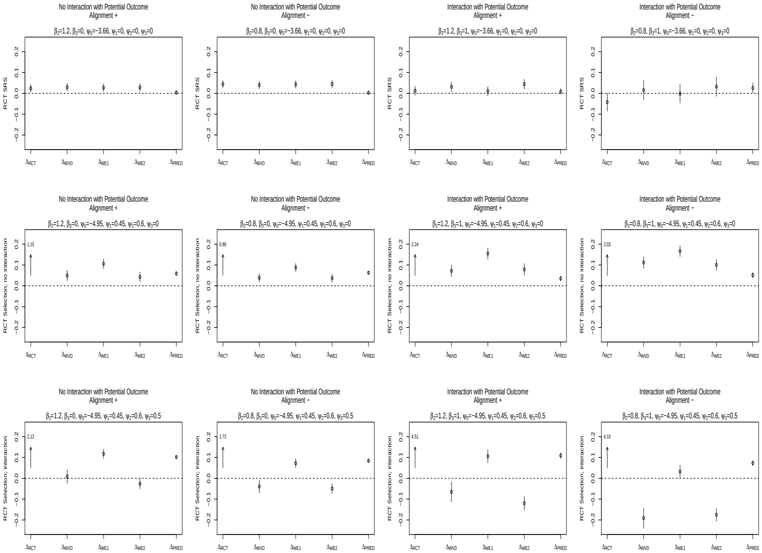
- Poisson sampling is used to allocate observation i into the RCT (R; non-probability) data or benchmark (B; probability) sample:

$$\Pr(S_i^B = 1) = \text{expit}(\psi_0^B)$$

$$\Pr(S_i^R = 1) = \begin{cases} 0, & S_i^B = 1 \\ \text{expit}(\psi_0^R + \psi_1^R X_{1,i} + \psi_2^R X_{2,i} + \psi_3^R X_{1,i} X_{2,i}), & S_i^B = 0 \end{cases}$$

- Consider a $2 \times 2 \times 3$ design:
  - Outcome with and without quadratic term
  - RCT SRS, with and without interaction
  - Alignment + (Effect of X in same direction for outcome and selection) and - (different directions) (Kern et al. 2016).

Simulation Study: Bias



No Interaction with Potential Outcome
Alignment +

β_2=1.2, β_3=0, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

No Interaction with Potential Outcome
Alignment −

β_2=0.8, β_3=0, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

Interaction with Potential Outcome
Alignment +

β_2=1.2, β_3=1, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

Interaction with Potential Outcome
Alignment −

β_2=0.8, β_3=1, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

No Interaction with Potential Outcome
Alignment +

β_2=1.2, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

No Interaction with Potential Outcome
Alignment −

β_2=0.8, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

Interaction with Potential Outcome
Alignment +

β_2=1.2, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

Interaction with Potential Outcome
Alignment −

β_2=0.8, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

No Interaction with Potential Outcome
Alignment +

β_2=1.2, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

No Interaction with Potential Outcome
Alignment −

β_2=0.8, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

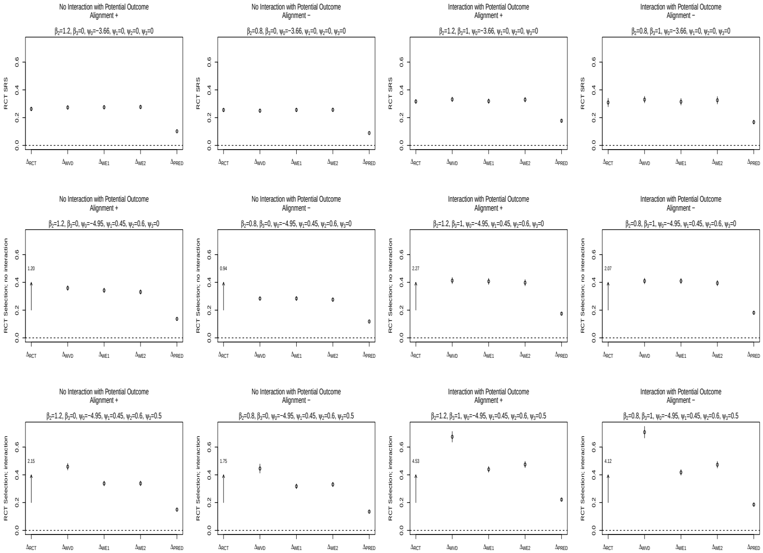
Interaction with Potential Outcome
Alignment +

β_2=1.2, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

Interaction with Potential Outcome
Alignment −

β_2=0.8, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

Simulation Study: RMSE



No Interaction with Potential Outcome
Alignment +
β_2=1.2, β_3=0, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

No Interaction with Potential Outcome
Alignment −
β_2=0.8, β_3=0, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

Interaction with Potential Outcome
Alignment +
β_2=1.2, β_3=1, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

Interaction with Potential Outcome
Alignment −
β_2=0.8, β_3=1, ψ_0=−3.66, ψ_1=0, ψ_2=0, ψ_3=0

No Interaction with Potential Outcome
Alignment +
β_2=1.2, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

No Interaction with Potential Outcome
Alignment −
β_2=0.8, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

Interaction with Potential Outcome
Alignment +
β_2=1.2, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

Interaction with Potential Outcome
Alignment −
β_2=0.8, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0

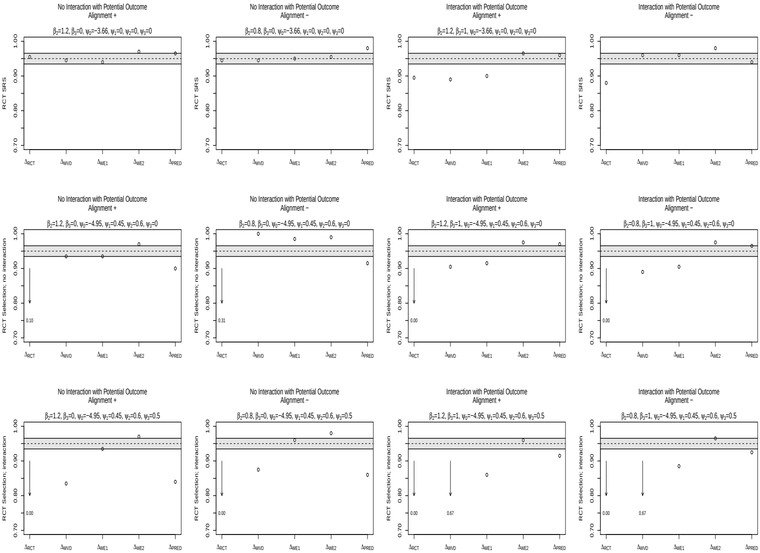
No Interaction with Potential Outcome
Alignment +
β_2=1.2, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

No Interaction with Potential Outcome
Alignment −
β_2=0.8, β_3=0, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

Interaction with Potential Outcome
Alignment +
β_2=1.2, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

Interaction with Potential Outcome
Alignment −
β_2=0.8, β_3=1, ψ_0=−4.95, ψ_1=0.45, ψ_2=0.6, ψ_3=0.5

Simulation Study: Coverage



Simulation Study: Summary

- SATT good if RCT is simple random sample; poor otherwise.
- WVD (estimated with logistic regression and no interaction) not too bad for bias until prediction model is complex; coverage is poor is prediction model is misspecified.
- WE1 (treating pseudo-weight as fixed) generally works reasonable well with respect to bias but has modest undercoverage with more variable selection probabilities; WE2 (incorporating variance of pseudo-weight) somewhat overcorrects for more conservative coverage except when prediction is complex, in which case bias effects coverage.
- PRED has best bias properties and, because it utilizes predictions from benchmark data, much smaller RMSE. Generally good coverage though some undercoverage occurs when prediction model is simple and positively aligned.

Study of pulmonary artery catheterization (PAC) in critical care

- PAC is an invasive and controversial cardiac monitoring device that is used in critical care. “PAC-Man” randomized trial (Harvey et al. 2005):
  - 1,013 subjects at 65 United Kingdom intensive care units.
  - Outcome=in-hospital mortality.
- Concerns about differences between the study sites and the general population in which PAC is used (Sakr et al. 2005).
- Obtain data from the Intensive Care National Audit Research Centre (ICNAR) database (Harrison et al. 2004)
  - 1.5 million admissions to 250 critical care units in the UK.
- Restricting to same inclusion and exclusion criteria as PAC-Man yields 1052 PAC population cases
- Population control group not exchangeable with RCT controls, even conditional on available covariates.
  - Restricted their analysis to the treated only: PATT
  - Approximate as being a SRS from a superpopulation by assigning a small sampling fraction value: 0.01 so $\pi_i^B \equiv 0.01$ and thus $w_i^B \equiv 100$.

Covariates

Variable	RCT	INCAR	*p*-value
Age	64.5	61.9	$<$0.001
% Female	41.8	39.0	0.22
% Elective	6.3	9.3	0.007
% Emergency	27.4	23.1	
% Medical	66.2	67.6	
% Ventilator	90.3	86.2	0.006
% Teaching Hosp.	21.5	41.2	$<$0.001
Survival Prob.	54.1	52.5	0.15
AP2 score	17.8	17.5	0.32
% Cardio event	3.8	3.2	0.60
% Renal failure	1.2	1.2	1.00
% Resp problems	3.6	2.5	0.19
% Liver failure	2.5	2.2	0.78
% Immunte disorder	7.8	6.8	0.46
Glasgow coma score	3.95	3.77	0.042

Results

- Adjusted SATT obtained from a BART model trained on the observed data in the RCT assigned to treatment assigned to control:-4.3% (95% CI -9.5%,1.0%)
- The PATT estimated under the pseudo-weighting method of WE1 is 0.2% with a 95% CI of (-4.2%,4.4%).
- The PATT estimated under WE2 is 0.2% with a 95% CI of (-9.5%,10.1%).
- The PATT estimated under PRED was 6.8% with a 95% CI of (-1.2%,14.8%).
    - While none of the effects significant, the PATT expected direction of the effect, in contrast to the SATT.

Model Checking: Testing for ignorability

- Transportability relies on the ignorabilty assumption: potential outcomes are independent of sampling indicator given covariates.
  - Impute $Y(Z)_i$ when $Z_i = 1 - z$ in the probability sample (or $Y(Z)_i$ for $Z = 0, 1$ if Y is not observed in the probability sample).
- Assumption testable when Y is observed in the probability sample
  - Test the reduced version $Y(1)_i \perp S_i^R, S_i^B \mid X_i$ in PAC-Man since only treatment outcomes are available.
- Posterior predictive distribution p-value: $T^{rep} = \sum_{i=1}^{N} I(S_i = 1) Y(1)_i^{rep}$ versus $T^{obs} = \sum_{i=1}^{N} I(S_i = 1) I(Z_i = 1) y_i$.
  - $P(T^{rep} < T^{obs} \mid \text{ data }) = 0.159$,
  - Overestimate the success of PAC in the population, or, equivalently, subjects in the RCT were more likely to have a good outcome even after controlling for available covariates.

Ignorability-corrected PATT

- The impact on the failure of ignorability in this setting depends on how the joint distribution of $(Y(1)_i, Y(0)_i) \mid X_i, S_i = 2$ differs from $(Y(1)_i, Y(0)_i) \mid X_i, S_i = 1$.
- If $\delta(1, X_i)^S - \delta(0, X_i)^S = 0$ for all X_i, $\delta(z, X_i)^S = P(Y(z)_i \mid X_i, S_i = 2) - P(Y(z)_i \mid X_i, S_i = 1)$ then the PATT estimate remains unbiased
- Other extreme:ignorability holds on the control arm, so that $\delta(0, X_i)^S = 0$ or, more generally $E(\delta(0, X_i)^S) = 0$.
  - $E(\delta(1, X_i)^S) = E(T^{rep} - T^{obs} \mid \text{ data }) = 8.4\%$
  - Estimate corrected PATT of 6.8%+8.4%=15.1%.

Discussion

- Econometricians, epidemiologists, and biostatisticians have independently invented and reinvented the wheel of causal inference for the past several decades, in the process following or borrowing the tools of population inference from survey statistics.
- Survey statistics can return the favor by adapting recently developed methods for non-probability samples for the important task of transporting randomized trials to better understand how novel treatments can work in a larger population.
- “Tip of the iceberg” of research opportunities:
  - Accommodating non-compliance.
  - Mediation; confounding by indication in longitudinal studies.
  - Adaptive trial design to ferret out key interactions.

Contact info

THANK YOU!

I want to thank Richard Grieve, Orlagh Carroll, and James Carpenter at the London School of Hygiene and Tropical Medicine for their assistance and introduction to the Pac-Man trial data.

Please feel free to contact me at mrelliot@umich.edu if you would like to discuss any of the material in this presentation further.

References

Ackerman, B., Lesko, C.R., Siddique, J., et al. Generalizing randomized trial findings to a target population using complex survey population data. *Statistics in Medicine*, 40, 1101-1120.

Bang, H., Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-973.

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Cole, S.R., Stuart, E.A. (2010) Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Americal Journal of Epidemiology*, 172, 107-115.

References

Dahabreh, I.J., Robertson, S.E., Steingrimsson, J.A., et al. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39, 1999-2014.

Degtiar, I., Rose, S. (2023) A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application*, 10, 7.1-7.24.

Elliott, M.R. (2007). Bayesian Weight Trimming for Generalized Linear Regression Models. *Survey Methodology*, 33, 23-34.

Elliott, M.R. (2016). Discussion of ‘Perils and Potentials of Self-Selected Entry to Epidemiological Studies and Surveys’. *Journal of the Royal Statistical Society A: Statistics in Society*, 179, 357.

References

Elliott, M.R., Little, R.J.A. (2000). Model-based Alternatives to Trimming Survey Weights. *Journal of Official Statistics*, 16, 191-209.

Elliott, M.R., Resler, A., Flannagan, C.A., Rupp, J.D. (2010). Appropriate analysis of CIREN data: using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Elliott, M.R., Xia, X. (2021). Weighted Dirichlet Process Mixture Models to Accommodate Complex Sample Designs for Linear and Quantile Regression. *Journal of Official Statistics*, 37, 71-95.

References

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, B31, 195–233.

Harrison D.A., Brady A.R., Rowan K. (2004) Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit Research Centre Case Mix Programme Database. *Critical Care*, 8, 1-3.

Hartman, E., Grieve, R., Ramsahai, R., et al. (2016). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society A*, 178, 757-778.

References

Harvey S., Harrison D.A., Singer M., et al. (2005) Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomised controlled trial. *The Lancet*, 366, 472-477.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81, 945-960.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*, London.

Kern, H., Stuart, E.A., Hill, J., Green, D.P. (2016) Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9, 103-127.

References

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *J R Stat Soc*, 97, 558–625.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings (Vol. 4)*, Alexandria, VA: American Statistical Association.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Sakr Y., Vincent J.L., Reinhart K., et al. (2005). Sepsis Occurrence in Acutely Ill Patients Investigators. Use of the pulmonary artery catheter is not associated with worse outcome in the ICU. *Chest*, 128, 2722-2731.

References

Särndal, C. E., Swensson, B., Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science Business Media.

Schmid, I., Rudolph, K.E., Nguyen, T.Q., et al. (2022). Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations. *Communication in Statistics: Simululation and Computation*, 51, 4326-4348.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 369-386.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 369-386.

References

Stuart, E.A., Ackerman, B., Westreich, D. (2018). Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on Social Work Practice*, 28, 532-537.

Valliant, R., Dever, J.A. (2011) Estimating propensity adjustments for volunteer web surveys. *Sociological Methods Research*, 40, 105-137.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186, 1010-1014.

Wu, C. Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 283-311.