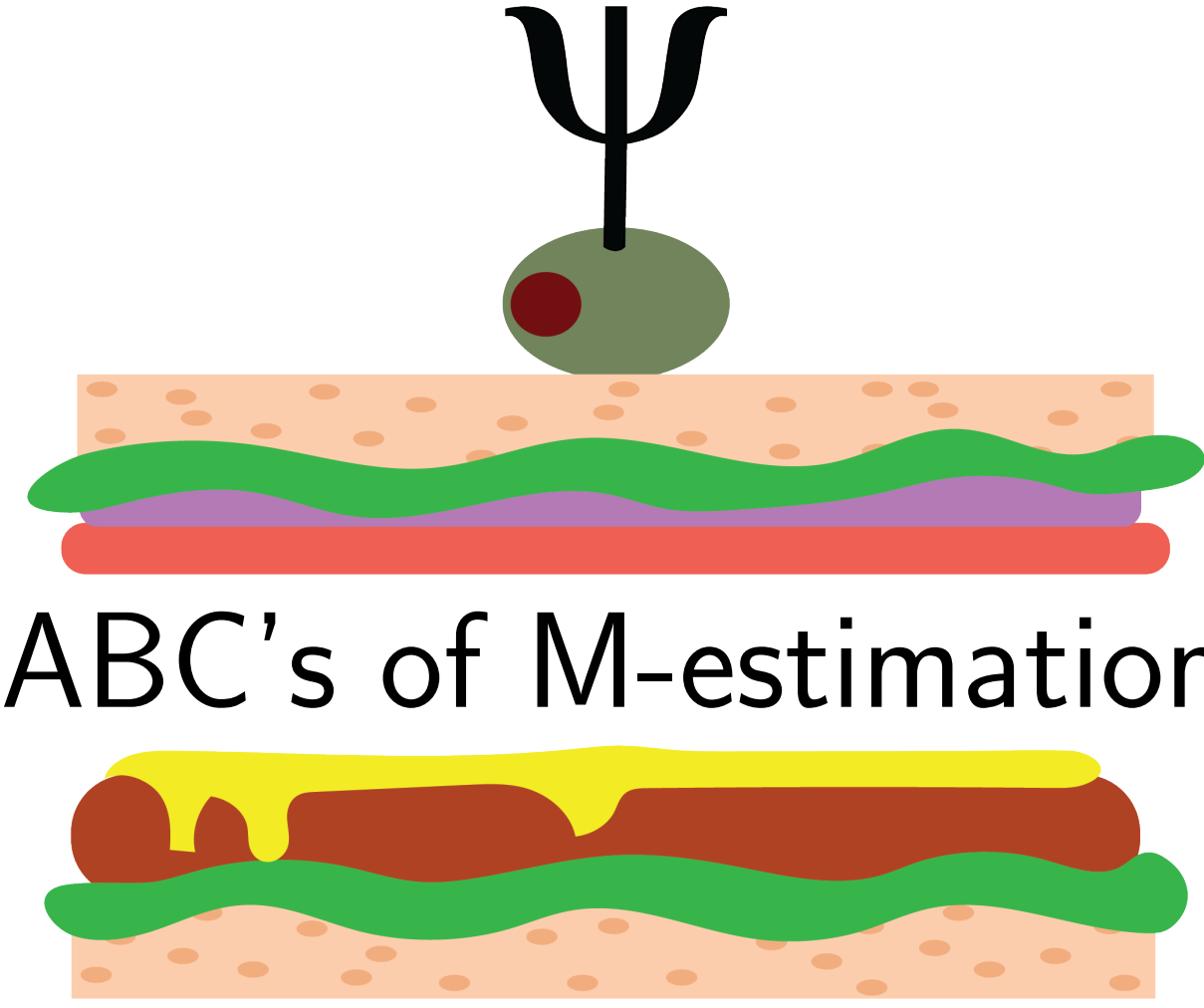


ABC's of M-estimation

Paul Zivich, Rachael Ross, Bonnie Shook-Sa

University of North Carolina at Chapel Hill and Columbia University

Acknowledgements

Supported by NIH K01-AI177102 (PNZ), R01-DA056407 (RKR), R01-AI157758 (BES).

bshooksa@email.unc.edu

Overview

Section 1: introduction

Break (15min)

Section 2: applied examples

Break (15min)

Section 3: in context

Advantages of M-estimation

Why M-estimation?

- Provides a method for point and variance estimation when fitting **multiple models simultaneously**
  - Appropriately propagates error from estimating parameters in earlier models
  - Special case is fitting "nuissance" models (e.g., propensity or outcome models for standardization approaches presented in Part II)
- Automation of Delta method

Why M-estimation?

Broad range of use cases, e.g.,

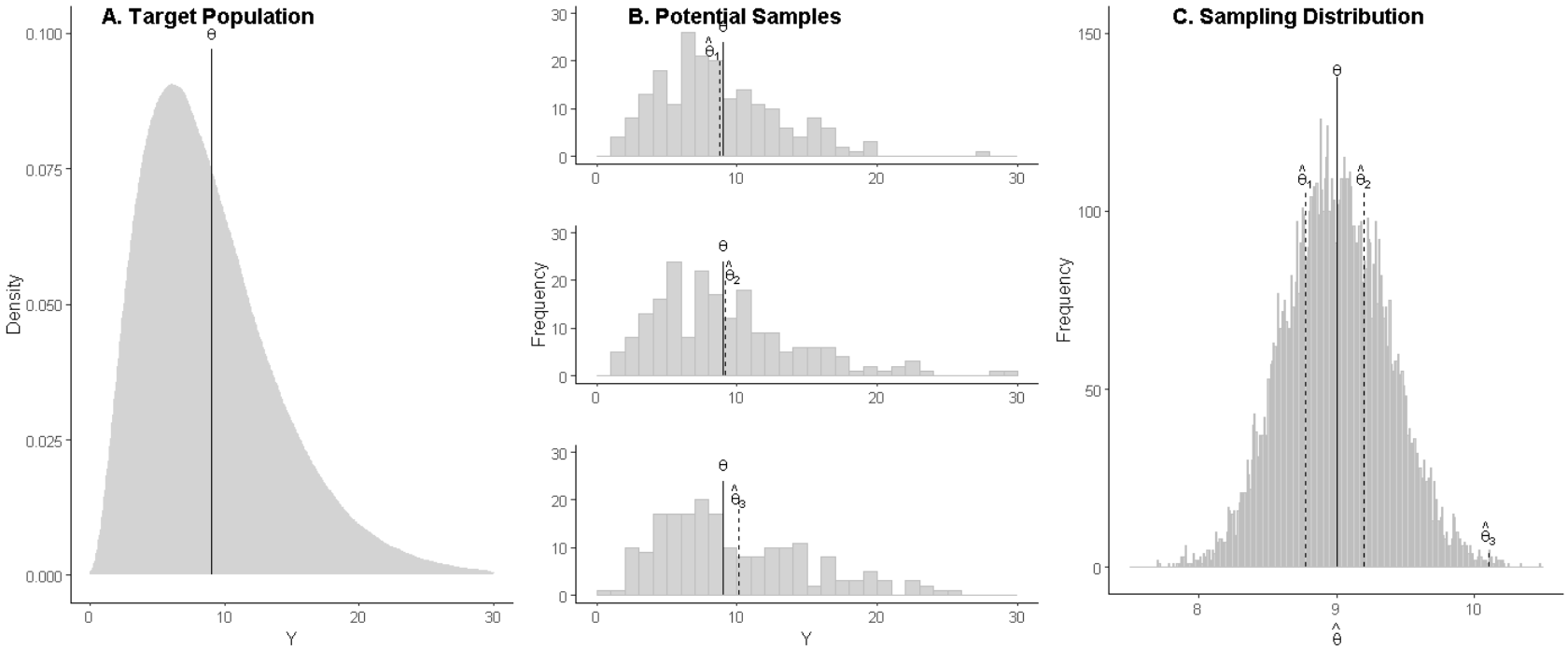
- IPTW and g-computation estimators for point exposures, e.g., Lunceford and Davidian (2004)
- Measurement error corrections, e.g., Cole et al. (2024)
- Sensitivity analyses, e.g., Dahabreh et al. (2023)
- Difference in difference approaches, e.g., Tchetgen Tchetgen et al. (2024)
- Target trial emulation, e.g., DeMonte et al. (2024)
- Data fusion, e.g., Cole et al. (2023)
- Bridged treatment comparisons, e.g., Shook-Sa et al. (2024b)
- Iterative conditional expectation (ICE) g-computation for time-varying exposures, e.g., Zivich et al. (2023)

Statistical Properties of M-estimators

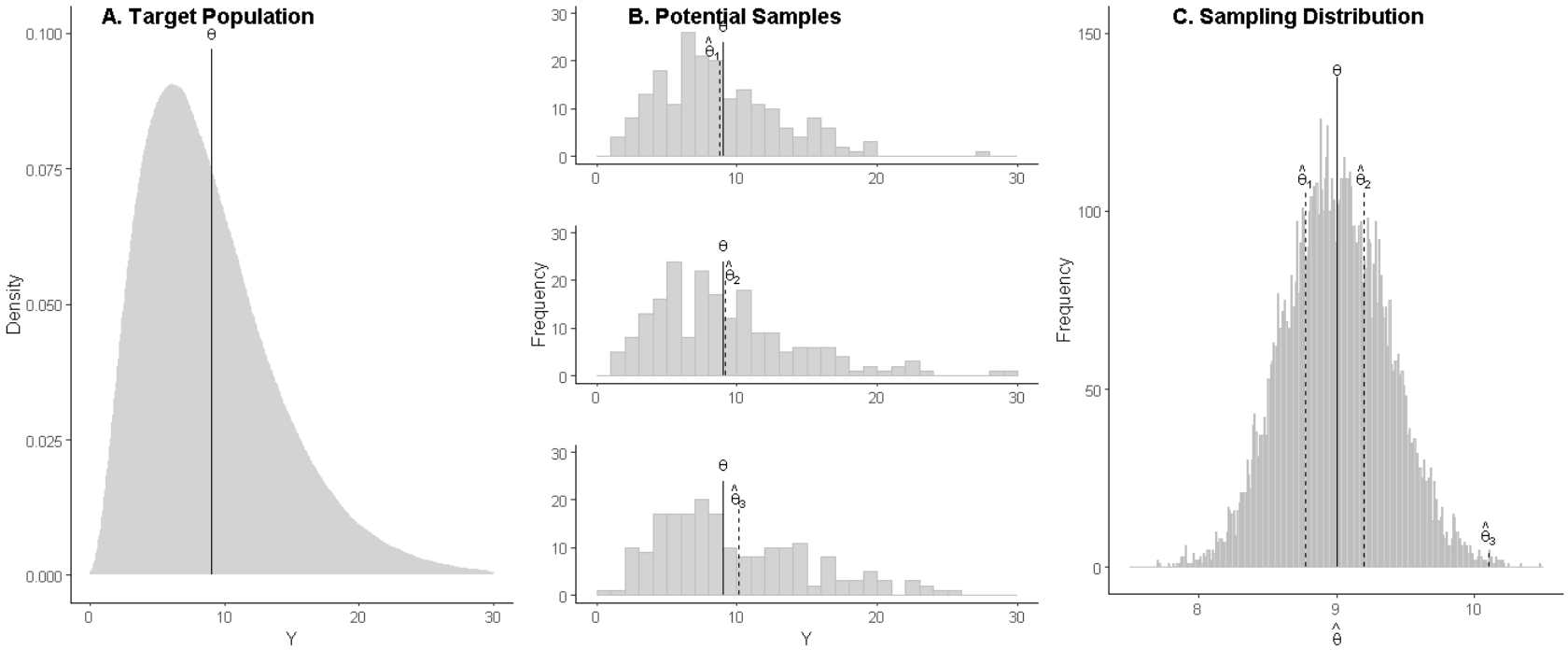
M-estimation Theory

- Statistical theory provides justification for the use of estimators and approaches, like M-estimation
- M-estimation is rooted in the large sample frequentist inferential paradigm
- Theory tells us about the behavior of estimators in **large samples**, i.e., in the limit as the sample size n tends towards infinity
- Many of these concepts are about characterizing an estimator's **sampling distribution**
- For more complete coverage of large sample statistical theory, see e.g., Lehmann (1999); Casella and Berger (2024)

Sampling Distribution

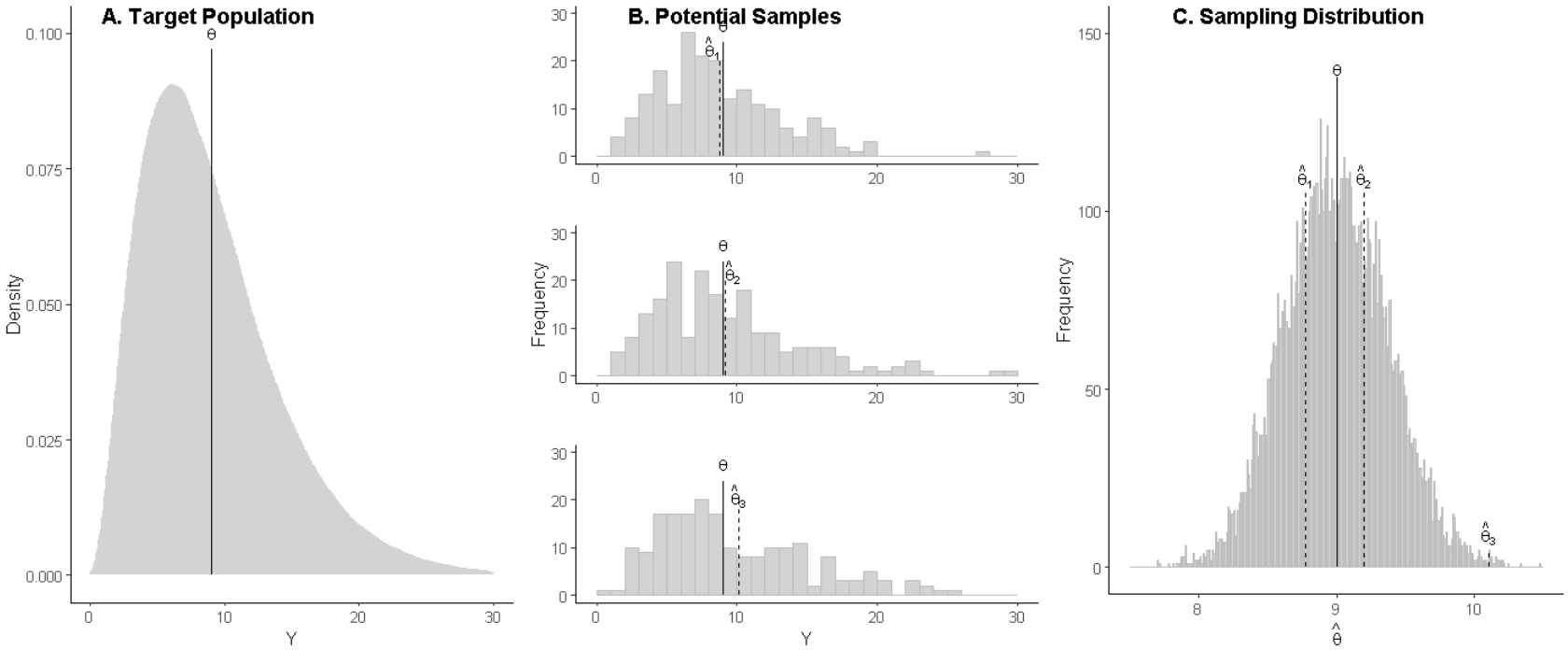


Properties of Estimators



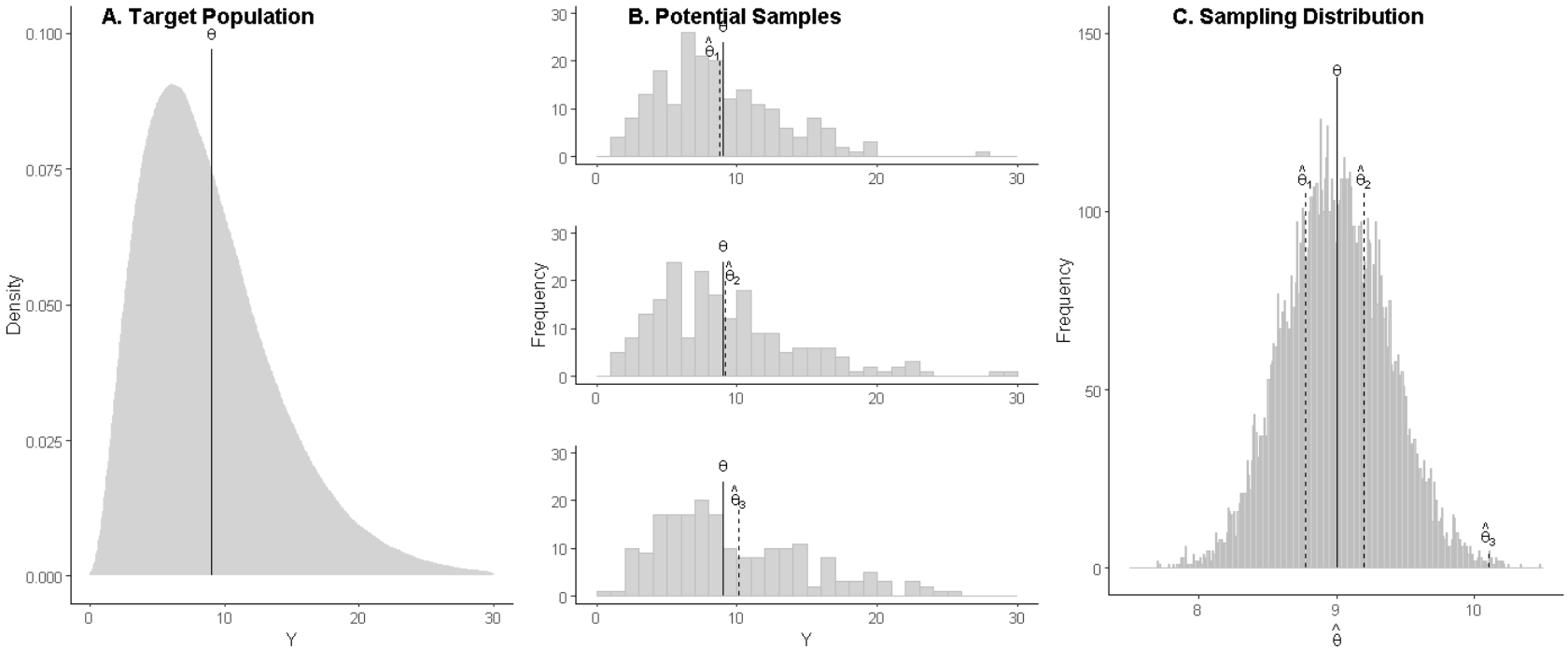
1. Bias
2. Variance

Properties of Estimators



1. **Bias: where is the sampling distribution centered?**
2. Variance

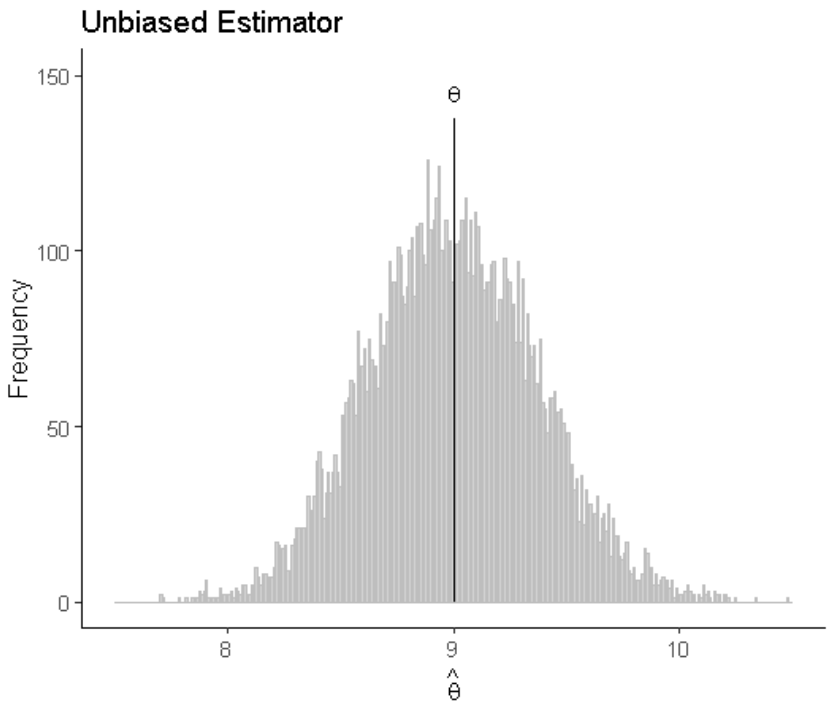
Properties of Estimators

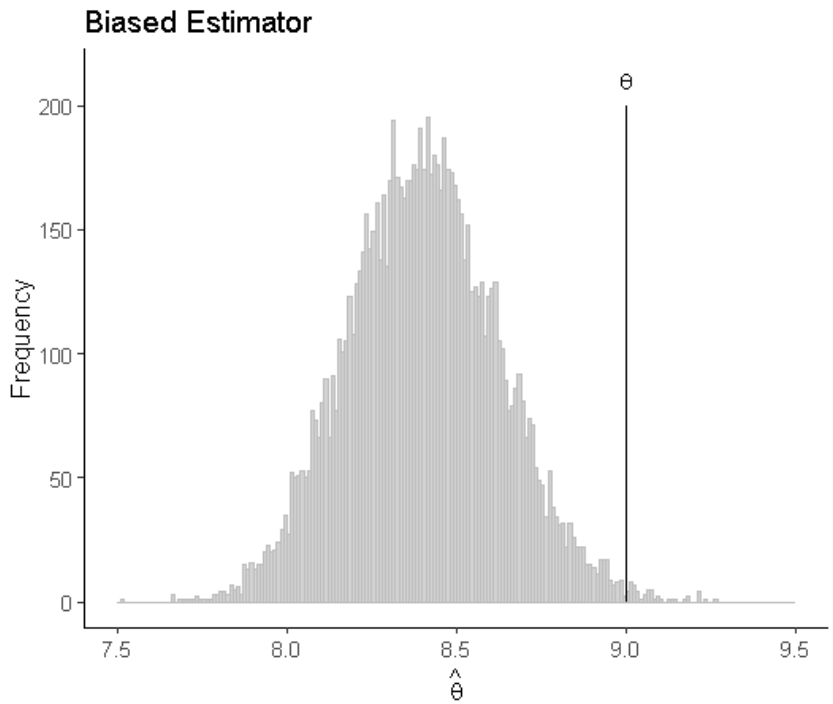


1. Bias: where is the sampling distribution centered?
2. **Variance: how much spread is there in the sampling distribution?**

Properties of Estimators: Bias

- $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$
- Persists regardless of the sample size
- Unbiased estimators give us the right answer on average (across possible samples from the target population)
- Biased estimators give us the wrong answer on average (across possible samples from the target population)





Properties of Estimators: Variance

- We typically only observe one sample resulting in a single estimate
- This single value is rarely exactly equal to θ, even with an unbiased estimator
- The variance of the estimator is: $Var(\hat{\theta}) = E\{\hat{\theta} - E(\hat{\theta})\}^2$
- $Var(\hat{\theta})$ and $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$ quantify how much $\hat{\theta}$ differs from its expected value, on average

Properties of Estimators: Asymptotic Behavior

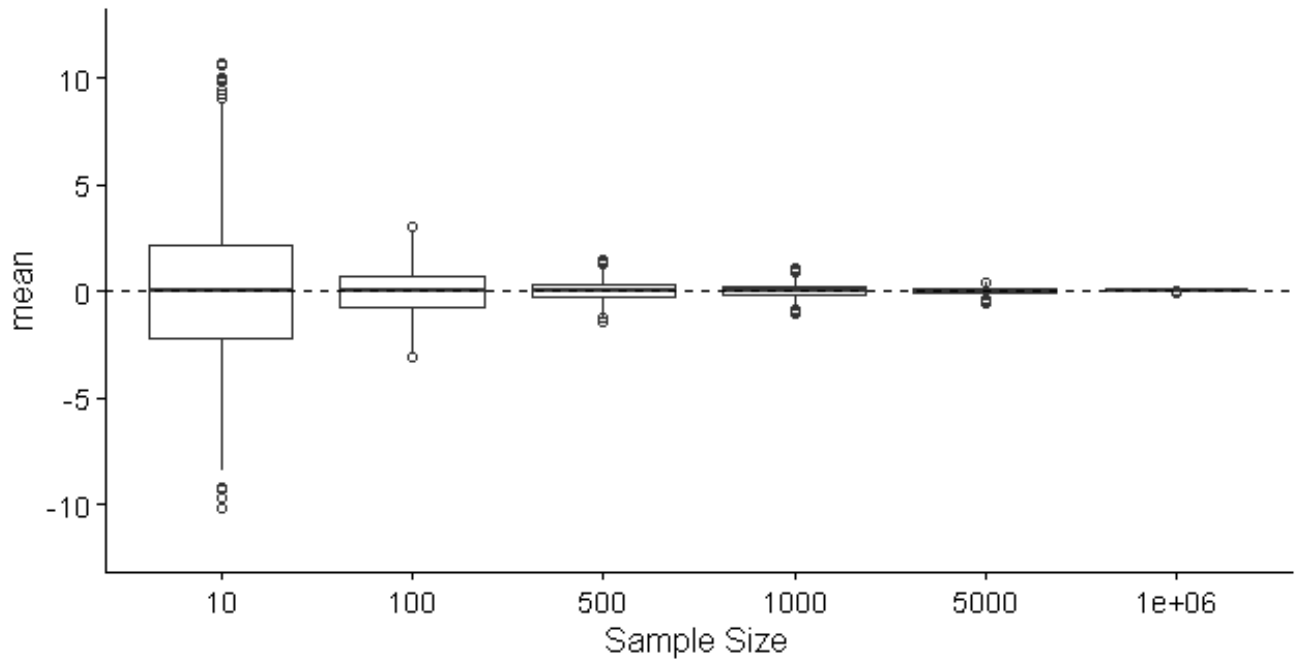
- Bias and variance are exact properties of estimators, i.e., they hold regardless of the sample size
- In practice, many useful estimators are biased, but this bias shrinks as the sample size increases
- Asymptotic properties are defined based on how estimators perform as the sample size increases (as $n \to \infty$)
- Examples of asymptotic properties are
  - Asymptotic unbiasedness
  - Consistency
  - Asymptotic normality

Properties of Estimators: Asymptotic Unbiasedness

- Bias tends to zero as $n \to \infty$
- The variance does not necessarily shrink as $n \to \infty$
- An estimator can be asymptotically unbiased even if it produces estimates far from the true parameter value as long as the average of these estimates is close to the true parameter value as the sample size increases

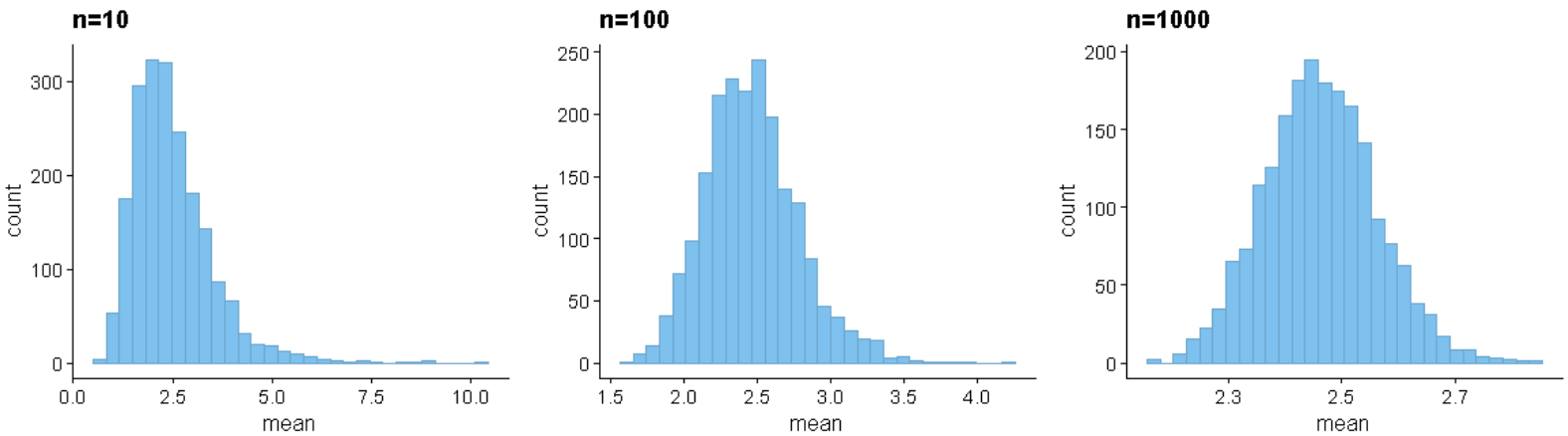
Properties of Estimators: Consistency

- Estimator converges in probability to the estimand ($\hat{\theta} \xrightarrow{p} \theta$)
- High probability estimator close to the estimand for large n
- May exhibit bias in small samples
- Examples of biased (but consistent) estimators
  - Odds ratio
  - Hazard ratio
  - Hajek estimator



Properties of Estimators: Asymptotic Normality

- As the sample size grows, the distribution of some estimators converges to a Normal distribution
- This does not imply that the distribution of the **outcome** in the sample is or becomes normal for large samples
- The distribution of the **estimator** across potential samples becomes more normal as the sample sizes grows
- Useful property for constructing confidence intervals (CIs)
- Commonly-used Wald-type asymptotic CI has the form: $\hat{\theta} \pm z_{1-\alpha/2}\widehat{SE}(\hat{\theta})$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution



What about M-estimators?

- Assume our data O_i are independent and identically distributed (iid) for $i = 1, ..., n$
- Let $\hat{\theta}$ be the solution to an unbiased estimating equation vector $\psi(O_i, \theta)$, i.e., $E\{\psi(O_i, \theta)\} = 0$
- Under suitable regularity conditions (Stefanski and Boos, 2002), $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V(\theta))$
  - $V(\theta) = B(\theta)^{-1} F(\theta) \{B(\theta)^{-1}\}^T$
  - $B(\theta) = E(\psi'(O_i, \theta))$
  - $F(\theta) = E\{\psi(O_i, \theta)\psi(O_i, \theta)^T\}$
- $V(\theta)$ can be consistently estimated with the empirical sandwich variance estimator

What about M-estimators?

Take-aways

- M-estimators have good large-sample properties
  - **Consistency**: M-estimators are expected to be close to the estimand for large n
  - **Asymptotic normality**: We can use the empirical sandwich variance estimator to construct Wald-type CIs for θ
- To demonstrate that our estimator is consistent and asymptotically normal, we need only demonstrate that it is the solution to an unbiased estimating equation vector (for the true θ) under our assumptions! (see example in Appendix)

One more advantage of M-estimation... (an aside)

Double Robust Estimators

- Common in causal inference
- Consistently estimate the average causal effect if either the propensity model or the outcome model is correctly specified (but not necessarily both)
- One common doubly robust estimator is a weighted regression estimator, which combines the two standardization-based methods presented in Part II
- The outcome model is weighted by the estimated inverse probability of treatment weight (IPTW) for each participant
- **Challenge**: While the point estimator is doubly robust, the commonly used influence function based variance estimator is only consistent if **both** models are correctly specified (Daniel, 2014)

Double Robust Estimators as M-estimators

- The weighted regression estimator can be cast as an M-estimator (as can other double robust estimators)
- Its estimating equation vector is:

$$\sum_{i=1}^{n} \psi(O_i;\theta) = \sum_{i=1}^{n} \begin{bmatrix} \{X_i - \text{expit}(W_i^T\alpha)\}W_i \\ IPTW_i\{Y_i - \phi^{-1}(W_i^T\beta)\}W_i \\ b_1(W_i,\beta) - \mu^1 \\ b_0(W_i,\beta) - \mu^0 \\ \mu^1 - \mu^0 - DR \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where $\theta = [\alpha, \beta, \mu^1, \mu^0, DR]$, ϕ is the link function for the outcome model, and $b_x(W_i,\beta) = E(Y_i \mid W_i, X_i = x)$ for $x \in \{0,1\}$

Double Robust Variance Estimation

$$\sum_{i=1}^{n} \psi(O_i;\theta) = \sum_{i=1}^{n} \begin{bmatrix} \{X_i - \text{expit}(W_i^T\alpha)\}W_i \\ IPTW_i\{Y_i - \phi^{-1}(W_i^T\beta)\}W_i \\ b_1(W_i,\beta) - \mu^1 \\ b_0(W_i,\beta) - \mu^0 \\ \mu^1 - \mu^0 - DR \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

- Because these estimating equations are unbiased if either model is correctly specified (but not necessarily both) (Gabriel et al., 2023), M-estimation allows for **double robust variance estimation** (Shook-Sa et al., 2024a)

Alternative Approaches

Alternative 1: "GEE Trick"

- Commonly used for estimating the average causal effect from marginal structural models
- Fit a propensity model and use it to estimate IPTWs
- Use standard regression software to fit the marginal structural model, treating IPTWs as known (i.e., not accounting for the fact that they were estimated in the prior stage)
- Use the "robust" Huber-White variance to estimate the variance of the average causal effect estimator

Alternative 1: "GEE Trick" - Limitations

- This is a **conservative** estimator of the variance of the average causal effect estimator
- That is, it is likely too large, resulting in confidence intervals that are too wide
- The "GEE trick" is not valid in all settings
  - e.g., estimation of the average treatment effect in the treated (ATT)
  - See Reifeis and Hudgens (2022) for more details
- It only applies to IPTW estimators, not in the other settings we have considered
- M-estimation is not conservative, is applicable for estimation of the ATT, and in other settings where the "GEE trick" does not apply

Alternative 2: Nonparametric Bootstrap

- Take B simple random samples (with replacement) of size n from our observed sample
- Estimate the parameter of interest θ by applying our estimator $\hat{\theta}$ to each bootstrap sample
- This results in $\hat{\theta}_b$, $b = 1, 2, ..., B$
- The distribution of the bootstrap estimates $\hat{\theta}_b$ mimics the sampling distribution of $\hat{\theta}$
- Estimated standard error of $\hat{\theta}$ is equal to the standard deviation of $\hat{\theta}_b$
- Appropriately propagates error when estimating parameters in earlier models because variation is captured by sequentially fitting models in bootstrap samples

Alternative 2: Nonparametric Bootstrap - Limitations

- Much more computationally intensive than M-estimation
- Example: Assume we are estimating 20 parameters across all models
  - For M-estimation, we only estimate those 20 parameters once
  - For the nonparametric bootstrap, we estimate them B times, which is likely on the order of estimating 20 x 1000(+) = 20,000(+) parameters across all resamples
- This can be intractable for some datasets or settings (e.g., simulation studies)
- Requires choices about the appropriate B, and which variation of the bootstrap technique to apply

Cautions with M-estimators

M-estimation is a large sample method

- As we have seen, M-estimation is based in large sample theory
- The empirical sandwich variance estimator is known to underestimate the variance of $\hat{\theta}$ in small samples (Fay and Graubard, 2001)
- Finite sample corrections can be made to the empirical sandwich variance estimator (see, e.g., Saul and Hudgens (2020))

What if data are not iid?

- So far we have assumed data are independent and identically distributed
- Adjustments can be made to account for non-independent data (see, e.g., Saul and Hudgens (2020); Stefanski and Boos (2002))
- Additional assumptions may be needed for non identically distributed data (see, e.g., Yuan and Jennrich (1998)); directly applies to fusion settings

Are there other settings where M-estimation does not apply?

- M-estimation assumes that the parameter θ is finite dimensional
- M-estimation does not directly apply in settings where the parameter is infinite dimensional (e.g., for nonparametric approaches in survival analysis)
- With standard M-estimation, the estimating function $\psi(O_i, \theta)$ cannot depend on the values of any other observations. This poses challenges in certain settings (e.g., the Cox model) and requires adaptations, (e.g., Lin and Wei (1989)).

Additional Resources

Where can I learn more about M-estimation?

- Theory of M-estimation
  - Boos, D.D. and Stefanski, L.A., 2013. Essential Statistical Inference: Theory and Methods. New York: Springer.
  - Stefanski, L.A. and Boos, D.D., 2002. The Calculus of M-estimation. The American Statistician, 56(1), pp.29-38.
- Computing M-estimators with software
  - R: geex (see Saul and Hudgens (2020))
  - Python: Delicatessen (see Zivich et al. (2022))
- Applied examples
  - Applications on Slide 6
  - Ross, R.K., Zivich, P.N., Stringer, J.S. and Cole, S.R., 2024. M-estimation for common epidemiological measures: introduction and applied examples. International Journal of Epidemiology, 53(2), p.dyae030.

Questions?

References I

Casella, G. and Berger, R. (2024). *Statistical inference*. CRC Press.

Cole, S. R., Edwards, J. K., Breskin, A., Rosin, S., Zivich, P. N., Shook-Sa, B. E., and Hudgens, M. G. (2023). Illustration of 2 fusion designs and estimators. *American Journal of Epidemiology*, 192(3):467–474.

Cole, S. R., Shook-Sa, B. E., Zivich, P. N., Edwards, J. K., Richardson, D. B., and Hudgens, M. G. (2024). Higher-order evidence. *European Journal of Epidemiology*, 39(1):1–11.

Dahabreh, I. J., Robins, J. M., Haneuse, S. J.-P., Saeed, I., Robertson, S. E., Stuart, E. A., and Hernán, M. A. (2023). Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *Statistics in Medicine*, 42(13):2029–2043.

Daniel, R. M. (2014). Double robustness. *Wiley StatsRef: Statistics Reference Online*, pages 1–14.

DeMonte, J. B., Shook-Sa, B. E., and Hudgens, M. G. (2024). Assessing COVID-19 vaccine effectiveness in observational studies via nested trial emulation. *arXiv: 2403.18115*.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206.

References II

Gabriel, E. E., Sachs, M. C., Martinussen, T., Waernbaum, I., Goetghebeur, E., Vansteelandt, S., and Sjölander, A. (2023). Inverse probability of treatment weighting with generalized linear outcome models for doubly robust estimation. *Statistics in Medicine*, 10.1002/sim.9969:1–14.

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408):1074–1078.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Reifeis, S. A. and Hudgens, M. G. (2022). On variance of the treatment effect in the treated when estimated by inverse probability weighting. *American Journal of Epidemiology*, 191(6):1092–1097.

Saul, B. and Hudgens, M. (2020). The Calculus of M-estimation in R with geex. *Journal of Statistical Software, Articles*, 92(2):1–15.

Shook-Sa, B. E., Zivich, P. N., Lee, C., Xue, K., Ross, R. K., Edwards, J. K., Stringer, J. S., and Cole, S. R. (2024a). Double robust variance estimation. *arXiv: 2404.16166*.

References III

Shook-Sa, B. E., Zivich, P. N., Rosin, S. P., Edwards, J. K., Adimora, A. A., Hudgens, M. G., and Cole, S. R. (2024b). Fusing trial data for treatment comparisons: Single vs multi-span bridging. *Statistics in Medicine*, 43(4):793–815.

Stefanski, L. A. and Boos, D. D. (2002). The Calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Tchetgen Tchetgen, E. J., Park, C., and Richardson, D. B. (2024). Universal difference-in-differences for causal inference in epidemiology. *Epidemiology*, 35(1):16–22.

Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2):245–260.

Zivich, P. N., Klose, M., Cole, S. R., Edwards, J. K., and Shook-Sa, B. E. (2022). Delicatessen: M-estimation in python. *arXiv preprint arXiv:2203.11300*.

Zivich, P. N., Ross, R. K., Shook-Sa, B. E., Cole, S. R., and Edwards, J. K. (2023). Empirical sandwich variance estimator for iterated conditional expectation g-computation. *arXiv: 2306.10976*.

Appendix: Demonstrating Consistency and Asymptotic Normality

In Part II, we presented the following IPTW estimator:

$$\sum_{i=1}^{n} \psi(O_i, \theta) = \sum_{i=1}^{n} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \\ \psi_5 \\ \psi_6 \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} X_i - \text{expit}(\alpha_0 + \alpha_1 W_i) \\ \{X_i - \text{expit}(\alpha_0 + \alpha_1 W_i)\} W_i \\ e_i^{-1} X_i Y_i - \mu_1 \\ (1 - e_i)^{-1}(1 - X_i) Y_i - \mu_2 \\ (\mu_1 - \mu_2) - \delta_1 \\ \text{ln}(\mu_1/\mu_2) - \delta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where $e_i = \text{expit}(\alpha_0 + \alpha_1 W_i)$. To demonstrate consistency and asymptotic normality of the IPTW estimators δ_1 and δ_2, we demonstrate that this estimating equation vector is unbiased.

Appendix: Demonstrating Consistency and Asymptotic Normality

- First note that if the propensity model is correctly specified, ψ_1 and ψ_2 are unbiased based on maximum likelihood theory
- Consider ψ_3 (and drop subscripts):

$$
\begin{aligned}
E(\psi_3) &= E(e^{-1}XY - \mu_1) \\
&= E_W\left\{E_{X,Y^1|W}(e^{-1}XY^1 - \mu_1)\right\} \\
&\qquad \text{(by iterated expectation and causal consistency)} \\
&= E_W\left\{e^{-1}E_{X,Y^1|W}(XY^1)\right\} - \mu_1 \\
&\qquad \text{(because } e^{-1} \text{ is constant given } W\text{)} \\
&= E_W\left\{e^{-1}E(X \mid W)E(Y^1 \mid W)\right\} - \mu_1 \\
&\text{(by conditional exchangeability of } X \text{ and } Y^1 \text{ given } W\text{)} \\
&= E_W\left\{E(Y^1 \mid W)\right\} - \mu_1 \qquad \text{(by definition of } e^{-1}\text{)} \\
&= E(Y^1) - \mu_1 = 0 \\
&\qquad \text{(by iterated expectation and the definition of } \mu_1\text{)}
\end{aligned}
$$

Appendix: Demonstrating Consistency and Asymptotic Normality

- $E(\psi_4) = 0$ analogously
- ψ_5 and ψ_6 are delta method transformations of μ_1 and μ_2
- Thus, $\psi(O_i, \theta)$ is an unbiased estimating equation vector under correct specification of the propensity model
- It follows under suitable regularity conditions (Stefanski and Boos, 2002) that: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V(\theta))$
- $V(\theta) = B(\theta)^{-1} F(\theta) \{B(\theta)^{-1}\}^T$
- $B(\theta) = E(\psi'(O_i, \theta))$
- $F(\theta) = E\{\psi(O_i, \theta)\psi(O_i, \theta)^T\}$