Workshop on M-estimation: Part 3

Stephen R. Cole
cole@unc.edu

Acknowledgements: Thanks to Drs. Jess Edwards, Michael Hudgens, Rachael Ross, Bonnie Shook-Sa, and Paul Zivich. This work was supported in part by NIH grant R01AI157758. Errors are mine.

Disclaimers: I receive salary from UNC, salary and research support from NIH, and honoraria from Oxford University Press for editorial duties at the AJE.

One More Example: Higher Order Evidence

In settings where a large main study depends on an error-prone outcome measure and appropriate validation data are available, epidemiologists are likely to account for the potential misclassification of the outcome.

The "higher order" evidence about sensitivity and specificity is typically assumed to be free of measurement error but may not be.[1]

Let's illustrate!

[1] People used to assume there is no sampling error in the validation data, but that is rare now.

Main Study

Imagine a clinical trial with 400 participants, with 200 randomly allocated to standard and novel treatment arms.

There are 30 self-reported cases of the study outcome in the novel arm and 45 in the standard arm.

TABLE 0: Risk by Treatment Arm

	Case, $W = 1$	Participants
Treatment arm:		
Novel	30	200
Standard	45	200
Total	75	400

So, the naïve estimated risk difference is -0.075.

Second Order Evidence

Second-order evidence is provided by an auxiliary data sample where the self-report outcome has been validated by medical record adjudication.

TABLE 1: Self-Report with Medical Adjudication as Gold Standard

	Medical Adjudication	
Self-Report:	Case, $V = 1$	Non-case, $V = 0$
Case, $W = 1$	80	5
Non-case, $W = 0$	20	95
Total	100	100

However, this adjudicated outcome is itself subject to measurement error.

Third Order Evidence

Third-order evidence is also available from an auxiliary data sample where the medical record adjudication outcome has been validated by biopsy pathology which is assumed to be error-free (i.e., a gold standard).

TABLE 2: Medical Adjudication with Pathology as Gold Standard

	Pathology	
Medical Adjudication:	Case, $Y = 1$	Non-case, $Y = 0$
Case, $V = 1$	40	5
Non-case, $V = 0$	10	45
Total	50	50

Statistical Methods

The interest parameter is the difference in risk of outcome $E(Y^1) - E(Y^0)$, where Y^a is a potential outcome under treatment $A = a$ (5).

We propose[2] a twice-corrected estimator $\hat{\gamma}_a$, where we apply a Rogan-Gladen correction to the Rogan-Gladen estimator (6),

$$\hat{\gamma}_a = \frac{\hat{\beta}_a + \widehat{sp}_2 - 1}{\widehat{se}_2 + \widehat{sp}_2 - 1},$$

where $\widehat{se}_2$ and $\widehat{sp}_2$ are the estimated sensitivity and specificity of medical adjudication with pathology as the gold standard, and $\hat{\beta}_a = (\hat{\alpha}_a + \widehat{sp}_1 - 1)/(\widehat{se}_1 + \widehat{sp}_1 - 1)$, where $\hat{\alpha}_a$ is, for treatment $A = a$, the naïve estimator of the risk, and $\widehat{se}_1$ and $\widehat{sp}_1$ are the estimated sensitivity and specificity of self-report with medical adjudication as the gold standard.

[2] We skip the proof that the estimating function is unbiased, and simulations.

Notation

Data given in Tables 0-2 are organized into 6 separate data sets each of containing n_j records for $j = 1, \ldots, 6$ and $n = n_1 + \cdots + n_6$.

$n_1 = 200$ main trial where $A = 0$
$n_2 = 200$ main trial where $A = 1$
$n_3 = 100$ sensitivity of W among $V = 1$
$n_4 = 100$ specificity of W among $V = 0$
$n_5 = 50$ sensitivity of V among $Y = 1$ and
$n_6 = 50$ specificity of V among $Y = 0$.

Let $R = r$ indicate membership in sample r = {1,…,6}.

S indicates main study, i.e., $S = 1$ if $R \in \{1,2\}$, $S = 0$ otherwise.

Observed data is $X = \{R, I(R < 5)W, I(R > 4)V\}$.

More Notation, Sorry

$$\theta = (\alpha_0, \alpha_1, RD_0, se_1, sp_1, \beta_0, \beta_1, RD_1, se_2, sp_2, \gamma_0, \gamma_1, RD_2,),$$

where:

$\alpha_a = E(W^a|S = 1)$,
$RD_0 = \alpha_1 - \alpha_0$,
$se_1 = P(W = 1|V = 1)$,
$sp_1 = P(W = 0|V = 0)$,
$\beta_a = E(V^a|S = 1)$,
$RD_1 = \beta_1 - \beta_0$,
$se_2 = P(V = 1|Y = 1)$,
$sp_2 = P(V = 0|Y = 0)$,
$\gamma_a = E(Y^a|S = 1)$, and
$RD_2 = \gamma_1 - \gamma_0$ is the average treatment effect on biopsy pathology outcomes.

Stacked Estimating Function

$$g(X,\theta) = \begin{pmatrix} g_1(X,\theta) = I(R=1)(W-\alpha_0) \\ g_2(X,\theta) = I(R=2)(W-\alpha_1) \\ g_3(\theta) = (\alpha_1-\alpha_0) - RD_0 \\ g_4(X,\theta) = I(R=3)(W-se_1) \\ g_5(X,\theta) = I(R=4)(W-(1-sp_1)) \\ g_6(\theta) = \beta_0(se_1+sp_1-1) - (\alpha_0+sp_1-1) \\ g_7(\theta) = \beta_1(se_1+sp_1-1) - (\alpha_1+sp_1-1) \\ g_8(\theta) = (\beta_1-\beta_0) - RD_1 \\ g_9(X,\theta) = I(R=5)(V-se_2) \\ g_{10}(X,\theta) = I(R=6)(V-(1-sp_2)) \\ g_{11}(\theta) = \gamma_0(se_2+sp_2-1) - (\beta_0+sp_2-1) \\ g_{12}(\theta) = \gamma_1(se_2+sp_2-1) - (\beta_1+sp_2-1) \\ g_{13}(\theta) = (\gamma_1-\gamma_0) - RD_2 \end{pmatrix}$$

Results

TABLE 3: Estimated Risk Differences, N = 400

	Risk A=1	Risk A=0	RD	95% CI	SE
Naïve	0.150	0.225	-0.075	-0.151, 0.001	0.039
Rogan-Gladen [a]	0.133	0.233	-0.100	-0.202, 0.002	0.052
Double Rogan-Gladen [b]	0.048	0.190	-0.143	-0.292, 0.006	0.076

[a] Sensitivity = 0.8 and specificity = 0.95 of self-report with medical adjudication as gold standard.

[b] Sensitivity = 0.8 and specificity = 0.9 of medical adjudication with pathology as gold standard.

Here we just did the nondifferential setting to keep things simpler.

Aside: Learning Functions

Let $\theta = (\alpha, \beta)$ now be parameters describing a system, where α and β can be classified as interest and nuisance parameters, respectively, n be combined number of study units, and X be p-dimensional data vector.

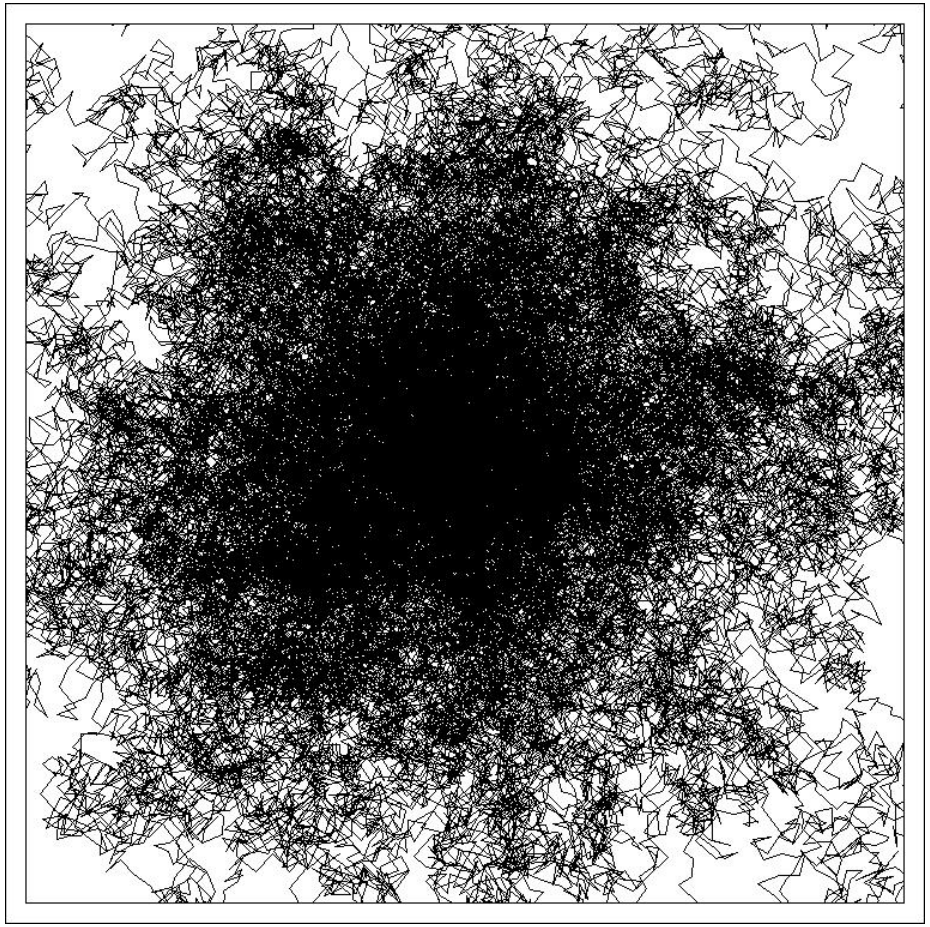
Let $g(X, \theta)$ be an estimating function.

Let S be a set of identification conditions.

Then $h\{g(X, \theta), S\}$ is a learning function, which formalizes the currently implicit dependence on S of inference based on the estimating function g.

The learning function $h\{g(X, \theta), S\} = h_g(X, \theta, S)$ can be used to construct consistent estimators of θ by solving $\sum_{i=1}^{n} g(X_i, \theta) = 0$ for any function g such that $E\{g(X, \theta)\} = 0$ <u>if S holds</u>.

Use of learning functions may make it harder to overlook identification issues.



Theory

Theory provides justification for the use of tools, like m-estimation.

Statistical theory can be broadly categorized as large sample or exact.

Large sample (or asymptotic) theory focuses on the behavior of parameter estimators in the limit as the sample size tends towards infinity, assuming that the population of interest is infinite. [3]

See (Lehmann 1999).

[3] All talk of infinity is just shorthand for delta-epsilon tolerance games.

Identification

A parameter is identified when it can be written as a function of the observed data distribution, otherwise we say it is not identified.[4]

A parameter is nonparametrically identified when there are no restrictions placed on the observed data distribution, otherwise we say it is locally identified.

We can consistently estimate identified parameters.[5] We cannot consistently estimate unidentified parameters.

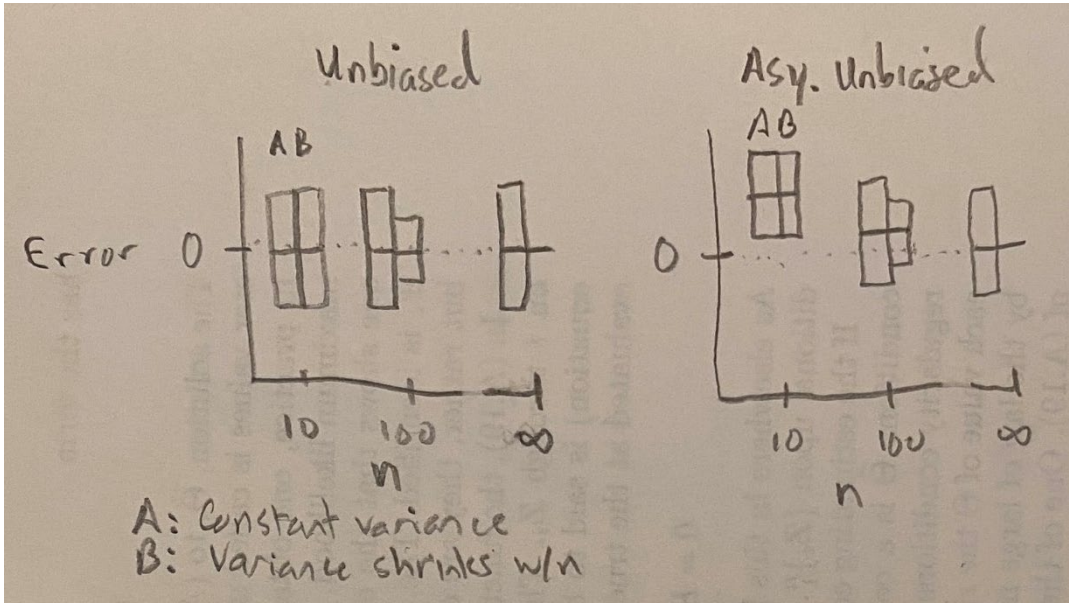
[4] Here we avoid partial identification (perhaps we shouldn't).

[5] Not all identified parameters are estimable. See Aronow et al Identification is not enough arXiv 2021.

(Statistical) Bias

Unbiased estimators equal the parameter value on average, at any sample size.

Asymptotically unbiased estimators have bias that tends towards zero as the sample size tends to infinity, but their variance does not necessarily tend to zero.



All unbiased estimators are asymptotically unbiased but not vice-versa.

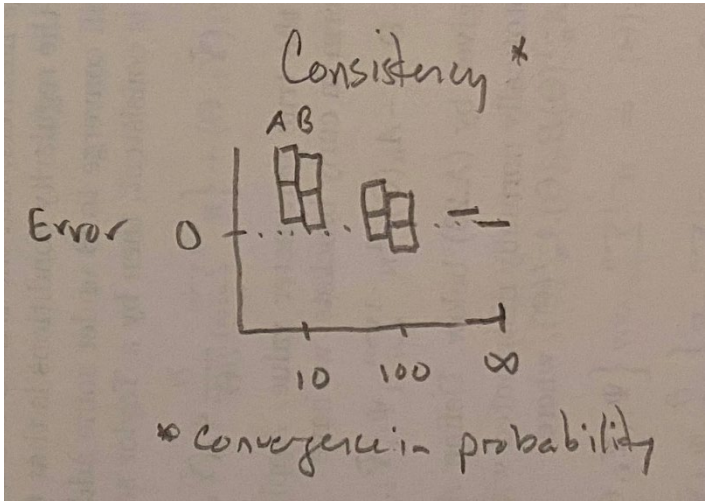
Many nonlinear estimators we use are asymptotically unbiased, e.g., RR, OR, HR, but have nontrivial finite sample bias.

Unbiasedness is crucial because biased estimators give the wrong answer.

(Statistical) Consistency

Consistent estimators converge in probability to the parameter of interest.

Asymptotically unbiased estimators that have variances that also shrink with increasing sample size (as most do) are consistent.



Note that not all unbiased estimators are consistent. Consistency is a stronger property than asymptotic unbiasedness.

Consistent estimators may exhibit bias in small samples but perform well in large samples.

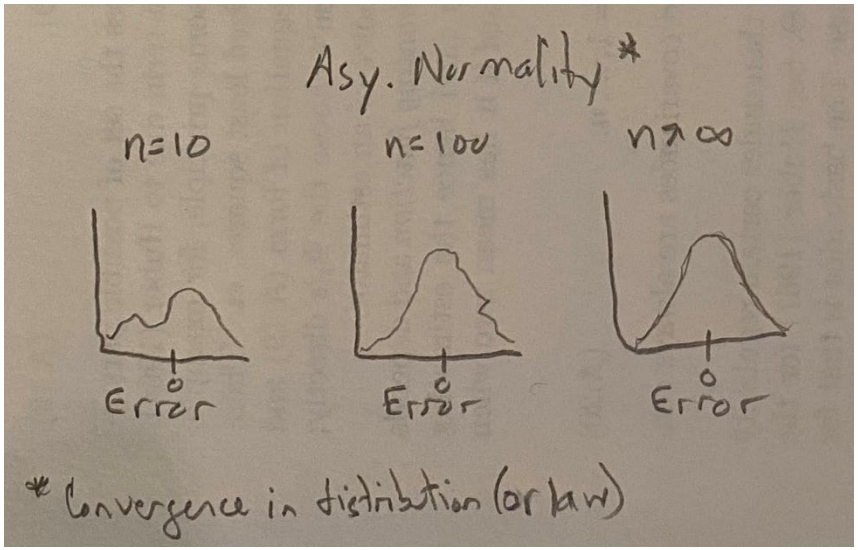
On right, A is consistent but biased and B is consistent and unbiased (same as B from prior slide).[6]

[6] You should be getting bias and consistency confused.

Asymptotic Normality

As the sample size goes to infinity, an asymptotically normal estimator converges <u>in distribution</u> to a normal density.

Wald-type confidence intervals are justified for consistent and asymptotically normal (CAN) estimators.[7]



[7] Honest Wald-type intervals require a stronger consistency condition, uniform consistency, but…

Proving M-estimators are Consistent and Asymptotically Normal

If we can prove an estimating function $g(X, \theta)$ is (asymptotically) unbiased,

(which is just showing that the estimating function, averaged over n, and set equal to 0, yields a $\hat{\theta}$ that equals the population parameter)

and we assume some regularity conditions, like those given by (Boos and Stefanski 2013) in chapter 7, [8]

then that m-estimator of θ is both consistent and asymptotically normal.

[8] Cox's British Regularity conditions were "Whatever is required to get the correct answer."

Next Steps [9]

1. Read and work through the 2 examples in (Cole, Edwards et al. 2023) and the example 3 in the forthcoming Rejoinder.

2. Read Chapter 7 of (Boos and Stefanski 2013).

3. Apply m-estimation to a problem you have!

[9] Understanding M-estimation might be sped up by first understanding MLE (see Cole et al AJE 2014).

What Problems Can M-estimation Address?

Any estimation problem that can be cast as an average taken over observed units, which is broad.

Can be used for semiparametric estimators.

Can be used for Bayes or semibayes estimators (Godambe 1997).

For the sandwich variance estimator, the estimating function must have findable derivatives (even if form is unknown).

M-estimation is under the hood of marginal structural models.

Examples of M-estimation at SER 2023

Mark Klose poster **904**, Revisiting the Population Attributable Fraction, Session 1, Tuesday June 13

Ning Zhang poster **871**, Accounting for outcome misclassification under outcome dependent sampling, Session 2 Wednesday June 14

Steve Cole talk on Higher Order Evidence, Symposium Friday June 16

Some Questions

Can m-estimation be applied in the setting of weighting for missingness or censoring? If so, how? Or is it just for IPTW?

Yes, a key strength of m-estimation is that it can be used to simultaneously account for multiple nuisance models by stacking the estimating functions.

Does m-estimation work in the setting of multiple imputation? How?

Yes, there are various approaches. One could obtain an m-estimator for each imputation, and then combine using Rubin's rule. [10]

What are the next steps in continuing the use of m-estimation in other areas of research? E.g., survival analyses.

Parametric survival analyses are straightforward with M-estimation, but...

[10] If we had to bootstrap rather than use m-estimation, then the imputations and bootstraps quickly become computationally overwhelming.

More Questions?

References

Boos, D. D. and L. A. Stefanski (2013). Essential Statistical Inference. New York, Springer.

Cole, S. R., J. K. Edwards, A. Breskin, S. Rosin, P. N. Zivich, B. E. Shook-Sa and M. G. Hudgens (2023). "Illustration of 2 Fusion Designs and Estimators." Am J Epidemiol **192**(3): 467-474.

Godambe, V. P. (1997). "Estimating functions: a synthesis of least squares and maximum likelihood." Institute of Mathematical Statistics Lecture Notes - Monograph Series **32**: 5-16.

Lehmann, E. L. (1999). Elements of large-sample theory. New York, Springer-Verlag.