

---

# Causal Inference

a summary

---

# Contents

<b>1 Prerequisites</b>	<b>3</b>	Sequential Identifiability . . . . .	8
1.1 Real Analysis . . . . .	3	Treatment-Confounder Feedback . . . . .	8
Real Numbers . . . . .	3	Causal Mediation . . . . .	8
Sequences . . . . .	3		
Series . . . . .	3	<b>3 Models</b>	<b>9</b>
Sets . . . . .	3	Modeling . . . . .	9
Functional Limits and Continuity . . . . .	3	Variable Selection . . . . .	9
Derivation . . . . .	4	Super Learning . . . . .	9
Functional Sequences . . . . .	4	Marginal Structural Models . . . . .	9
Functional Series . . . . .	4	3.1 Traditional Methods . . . . .	9
1.2 Measure Theory . . . . .	4	Stratification . . . . .	9
Riemann Integral . . . . .	4	Outcome Regression . . . . .	9
Measure . . . . .	4	Propensity Score Methods . . . . .	9
		Instrumental Variable Estimation . . . . .	9
<b>2 General</b>	<b>6</b>	Causal Survival Analysis . . . . .	9
Ladder Of Causation . . . . .	6	3.2 G-Methods . . . . .	10
Causal Roadmap . . . . .	6	G-Methods . . . . .	10
Average Causal Effect . . . . .	6	Standardization . . . . .	10
Target Trial . . . . .	6	IP Weighting . . . . .	10
Identifiability Conditions . . . . .	6	G-Estimation . . . . .	11
Effect Modification . . . . .	6	3.3 Doubly Robust Methods . . . . .	11
Interaction . . . . .	7	Double-Robustness . . . . .	11
NPSEM . . . . .	7	Machine Learning . . . . .	11
Causal DAG . . . . .	7	Advantages . . . . .	11
Noncausal DAGs . . . . .	7	Influence Curve . . . . .	11
SWIGs . . . . .	7	Delta Method . . . . .	11
Confounding . . . . .	7	Simple Plug-In Estimator . . . . .	12
Selection Bias . . . . .	7	Augmented IPTW . . . . .	12
Measurement Bias . . . . .	7	TMLE . . . . .	12
Random Variabilty . . . . .	7	LMTP . . . . .	12
Time-Varying Treatments . . . . .	8	Methods for continuous $A$ . . . . .	13

# 1 Prerequisites

## 1.1 Real Analysis (Abbott, 2015)

**Real Numbers** *triangle inequality*  $|a + b| \leq |a| + |b|$   
*Density of  $\mathbb{Q}$  in  $\mathbb{R}$ :*  $\forall a, b \in \mathbb{R} : \exists r \in \mathbb{Q} : a < r < b$   
*Archimedian Property:*  $\forall x \in \mathbb{R} \exists n \in \mathbb{N} : x < n$  &  $\forall y > 0 \exists n \in \mathbb{N} : \frac{1}{n} < y$   
**Bounds of  $A \subseteq \mathbb{R}$**  upper:  $\exists b \in \mathbb{R}$  s.t.  $a \leq b \forall a \in A$  (lower:  $\geq$ )  
*least upper bound (supremum)*  $s \in \mathbb{R}$  s.t.  $s$  is upper bound &  $\forall$  upper bounds  $b$ :  $s \leq b$ ; greatest lower (infimum) analogous  
**Cardinality:**  $A \sim B$ , if  $\exists f : A \rightarrow B$ , where  $f$  bijective (in+sur)  
*function  $f : A \rightarrow B$  mapping  $f(x) = \dots$ , domain =  $A$ , range  $\subseteq B$*   
*in/1-1:*  $a_1 \neq a_2 \Rightarrow f(a_1) \neq f(a_2)$ ; *sur/onto:* if  $\forall b \in B \exists a \in A : f(a) = b$   
**Axiom of Completeness:** every nonempty set of real numbers that is bounded above has a least upper bound; AoC, NIP, BW, CC, MCT are equivalent: if one is assumed the others follow  
**Nested Interval Property:** if  $I_n = [a_n, b_n] = \{x \in \mathbb{R} : a_n \leq x \leq b_n\}$ , where  $n \in \mathbb{N}$  and  $I_1 \supseteq I_2 \supseteq I_3 \dots$ , then  $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$

**Sequences** are functions with domain  $\mathbb{N}$   
**Convergence:**  $(a_n)$  converges to  $a \in \mathbb{R}$  if  $\forall \epsilon > 0 \exists N \in \mathbb{N}$  s.t.  $n \geq N \Rightarrow |a_n - a| < \epsilon$ ; written as  $\lim a_n = a$  or  $(a_n) \rightarrow a$   
*Cauchy Criterion:* sequence converges  $\Leftrightarrow$  is Cauchy sequence  
*Cauchy sequence:*  $\forall \epsilon > 0 \exists N \in \mathbb{N}$  s.t.  $m, n \geq N \Rightarrow |a_n - a_m| < \epsilon$   
**Boundedness:**  $(x_n)$  is bounded if  $\exists M > 0$  s.t.  $|x_n| \leq M \forall n \in \mathbb{N}$   
**Algebraic Limit Theorem:** if  $(a_n) \rightarrow a$ ,  $(b_n) \rightarrow b$ , then  $(ca_n) \rightarrow ca$ ,  $(a_n + b_n) \rightarrow a + b$ ,  $(a_n b_n) \rightarrow ab$ ,  $(a_n/b_n) \rightarrow a/b$  for  $b \neq 0$   
**Order Limit Theorem:**  $a_n \geq 0 \forall n \in \mathbb{N} \Rightarrow a \geq 0$  ( $\leq$  analogous),  $\exists c \in \mathbb{R}$  s.t.  $c \leq b_n \forall n \in \mathbb{N} \Rightarrow c \leq b$  ( $\geq$  analogous)  
**Monotone Convergence Theorem:** bounded & monotone (*increasing  $a_n \leq a_{n+1}$  or decreasing  $a_n \geq a_{n+1}$* ) sequences converge  
**Bolzano-Weierstrass Theorem:**  
all bounded sequences have a convergent subsequence  
subsequences of a convergent sequence converge to the same limit

**Series** infinite series are sums over sequences:  $\sum_{n=1}^{\infty} b_n$   
*harmonic series  $\sum_{n=1}^{\infty} \frac{1}{n}$ , geometric series  $\sum_{k=0}^{\infty} ar^k \stackrel{|r| \leq 1}{=} \frac{a}{1-r}$*   
**Convergence:** to  $B$ , if  $(s_m) \rightarrow B$ , partial sums  $s_m = \sum_{n=1}^m b_n$   
*Cauchy Criterion:*  $\sum_{k=1}^{\infty} a_k$  converges  $\Leftrightarrow$   
 $\forall \epsilon > 0 \exists N \in \mathbb{N}$  s.t.  $n > m \geq N \Rightarrow |a_{m+1} + a_{m+2} + \dots + a_n| < \epsilon$ ;  
that implies if  $\sum_{k=1}^{\infty} a_k$  converges then  $(a_k) \rightarrow 0$   
**Algebraic Limit Theorem:** if  $\sum_{k=1}^{\infty} a_k = A$  and  $\sum_{k=1}^{\infty} b_k = B$   
then  $\sum_{k=1}^{\infty} ca_k = cA$  and  $\sum_{k=1}^{\infty} a_k + b_k = A + B$   
**Cauchy Condensation Test:** if  $(b_n)$  is decreasing and  $b_n \geq 0 \forall n \in \mathbb{N}$  then:  $\sum_{n=1}^{\infty} b_n$  converges  $\Leftrightarrow \sum_{n=0}^{\infty} 2^n b_{2^n}$  converges  
**Comparison Test:** if  $0 \leq a_k \leq b_k \forall k \in \mathbb{N}$ , then  $\sum_{k=1}^{\infty} b_k$  converges  $\Rightarrow \sum_{k=1}^{\infty} a_k$  too &  $\sum_{k=1}^{\infty} a_k$  diverges  $\Rightarrow \sum_{k=1}^{\infty} b_k$  too  
**Absolute Convergence Test:**  $\sum_{n=1}^{\infty} |a_n|$  conv  $\Rightarrow \sum_{n=1}^{\infty} a_n$  too  
**Alternating Series Test:** if  $(a_n)$  is decreasing and converges, then  $\sum_{n=1}^{\infty} (-1)^{n+1} a_n$  converges  
**Absolute Convergence:** if  $\sum_{n=1}^{\infty} |a_n|$  converges then  $\sum_{n=1}^{\infty} a_n$  converges *absolutely*, if only the latter, then *conditionally*  
**Rearrangements:** if  $\sum_{k=1}^{\infty} a_k$  converges absolutely, then any rearrangement converges to the same limit  
**Double Series:** if  $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |a_{ij}|$  converges  $\Rightarrow \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij} = \lim_{n \rightarrow \infty} s_{nn}$ , where  $s_{nn} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}$

**Sets** *Cantor Set:*  $C = \bigcap_{n=0}^{\infty} C_n$ , with  $C_n$  removing the middle third of all intervals, e.g.  $C_1 = C_0 \setminus (\frac{1}{3}, \frac{2}{3}) = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$   
**Open Sets:**  $\forall a \in O \exists V_{\epsilon}(a) \subseteq O$ , with  $\epsilon$ -neighborhood of  $a$

$V_{\epsilon}(a) = \{x \in \mathbb{R} : |x - a| < \epsilon\}$ ; union of open sets is open, the finite intersection of open sets is open  
**Closed Sets:** contain their limit points  $\Leftrightarrow$  every Cauchy sequence has a limit that lies within the set  
 $x$  is *limit point* of  $A \Leftrightarrow \forall \epsilon > 0 V_{\epsilon}(x) \cap A$  includes other points than  $x \Leftrightarrow x = \lim(a_n)$  for some  $(a_n) \in A$  with  $a_n \neq x \forall n \in \mathbb{N}$ ; all non limit point  $a \in A$  are *isolated points*; the finite union of closed sets is closed, the intersection of closed sets is closed  
**Closure:**  $\bar{A} = A \cup L$ , with  $L$  the set of  $A$ 's limit points; the closure is the smallest closed set containing  $A$   
**Complement:**  $A^c = \{x \in \mathbb{R} : x \notin A\}$ ;  $A$  closed  $\Leftrightarrow A^c$  open  
**Compact  $\Leftrightarrow$  Bounded and Closed  $\Leftrightarrow \exists$  Finite Subcover**

- Compactness:**  $K$  compact  $\Leftrightarrow$  every sequence has a subsequence that converges in  $K$ ; the intersection of a sequence of nested nonempty compact sets is not empty
- Boundedness:**  $\exists M > 0$  s.t.  $|a| \leq M \forall a \in A$
- Any open cover for  $A$  has a finite subcover:** An open cover is a set of open sets  $\{O_{\lambda} : \lambda \in \Lambda\}$  whose union contains  $A$ ; a finite subcover is a finite subset that still covers  $A$

**Perfection:** closed and no isolated points; a nonempty perfect set is uncountable; the Cantor set is perfect  
**Separation:** of  $A$  and  $B$  if  $\bar{A} \cap B = \emptyset$  &  $A \cap \bar{B} = \emptyset$   
**Disconnection:** if  $A = B \cup C$ , with  $B, C$  nonempty & separated;  $E$  is *connected*  $\Leftrightarrow$  all nonempty disjoint sets  $B, C$  s.t.  $E = B \cup C$  have a convergent sequence with a limit in the other set  $\Leftrightarrow$  whenever  $a < c < b$  with  $a, b \in E$ , then  $c \in E$   
**Baire's Theorem:**  $\mathbb{R}$  cannot be written as the countable union of nowhere-dense sets;  $E$  is *nowhere-dense* if  $\bar{E}$  contains no nonempty open intervals

## Functional Limits and Continuity

**Functional Limit:** let  $f : A \rightarrow \mathbb{R}$  and  $c$  limit point of  $A$  if  $\forall \epsilon > 0 \exists \delta > 0$  s.t.  $0 < |x - c| < \delta$  (and  $x \in A$ ) it follows  $|f(x) - L| < \epsilon$ , then  $\lim_{x \rightarrow c} f(x) = L$   
*Sequential Criterion:*  $\lim_{x \rightarrow c} f(x) = L \Leftrightarrow$  for all sequences  $(x_n) \subseteq A$ , with  $x \neq c$  and  $(x_n) \rightarrow c$  follows  $f(x_n) \rightarrow L$   
**Algebraic Limit Theorem:** if  $\lim_{x \rightarrow c} f(x) = L$  and  $\lim_{x \rightarrow c} g(x) = M$  then  $\lim_{x \rightarrow c} kf(x) = kL$ ;  $\lim_{x \rightarrow c} [f(x) + g(x)] = L + M$ ;  $\lim_{x \rightarrow c} [f(x)g(x)] = LM$ ;  $\lim_{x \rightarrow c} f(x)/g(x) = L/M$  if  $M \neq 0$   
**Divergence Criterion:** if  $(x_n)$  and  $(y_n)$  with  $x_n \neq c \neq y_n$  and  $\lim x_n = \lim y_n = c$  but  $\lim f(x_n) \neq \lim f(y_n)$  then  $\nexists \lim_{x \rightarrow c} f(x)$   
**Continuity at  $c$ :**  $\forall \epsilon > 0 \exists \delta > 0$  s.t. whenever  $|x - c| < \delta$  (and  $x \in A$ ) then  $|f(x) - f(c)| < \epsilon$ , can also be expressed as:  $(x_n) \rightarrow c$  (with  $x_n \in A$ )  $\Rightarrow f(x_n) \rightarrow f(c)$   
**Algebraic Continuity Theorem:** if  $f, g$  continuous at  $c$  then these are too:  $kf(x)$ ,  $f(x) + g(x)$ ,  $f(x)g(x)$ ,  $f(x)/g(x)$  (if  $g(x) \neq 0$ )  
**Compositions:**  $f$  continuous at  $c$  and  $g$  is continuous at  $f(c) \Rightarrow g \circ f$  is continuous at  $c$  (if  $g \circ f(x)$  well-defined)  
**Boundedness:**  $f$  is bounded on its domain  $A \Leftrightarrow f(A)$  is bounded;  $f$  is bounded on  $B \subseteq A \Leftrightarrow f(B)$  is bounded  
**Preservation of Compact Sets:**  $K$  compact  $\Rightarrow f(K)$  is too; if  $f$  is continuous on a compact set,  $f$  attains min/max values  
**Uniform Continuity:**  $\forall \epsilon > 0 \exists \delta > 0$  s.t. whenever  $|x - y| < \delta$  then  $|f(x) - f(y)| < \epsilon$  (i.e. difference between them is bounded);  $f$  continuous on compact set  $K \Rightarrow f$  uniformly continuous on  $K$   
*Sequential Criterion for Nonuniformity:*  $\exists \epsilon_0 > 0$  and  $(x_n), (y_n)$

in  $A$  s.t.  $|x_n - y_n| \rightarrow 0$  but  $|f(x_n) - f(y_n)| > \epsilon_0$

**Intermediate Value Theorem:**  $f : [a, b] \rightarrow \mathbb{R}$  continuous then  $f(a) < L < f(b) \Rightarrow \exists c \in (a, b)$ , where  $f(c) = L$

alternatively: *Preservation of Connectedness:*  $f : A \rightarrow \mathbb{R}$  continuous,  $E \subseteq A$  connected  $\Rightarrow f(E)$  connected

*Intermediate Value Property* (converse of IVT):  $f$  has IVP on  $[a, b]$  if  $\forall x < y$  &  $L$  s.t.  $f(x) < L < f(y) \exists c \in (x, y)$ , where  $f(c) = L$  (implies continuity if  $f$  is monotone)

**Discontinuity:** • *removable:* if  $\lim_{x \rightarrow c} f(x)$  exists but  $\neq f(c)$   
• *jump:*  $\lim_{x \rightarrow c^+} f(x) \neq \lim_{x \rightarrow c^-} f(x)$   
• *essential:* not continuous for another reason

**The Set of Discontinuous Points**  $D_f$  can be written as the countable union of closed sets ( $=: F_\sigma$ )

**Derivation**  $g'(c) = \lim_{x \rightarrow c} \frac{g(x) - g(c)}{x - c}$   
 $(f+g)'(c) = f'(c) + g'(c)$ ;  $(fg)'(c) = f'(c)g(c) + f(c)g'(c)$ ;  
 $(kf)'(c) = kf'(c)$ ;  $(f/g)'(c) = \frac{g(c)f'(c) - f(c)g'(c)}{g^2(c)}$ , for  $g(c) \neq 0$ ;  
 $(g \circ f)'(c) = g'(f(c)) \cdot f'(c)$

**Differentiability:** Differentiability at  $c$  implies continuity at  $c$

**Interior Extremum Theorem:** let  $f$  differentiable on  $(a, b)$ ; if  $f$  has a maximum or minimum at  $f(c)$ , then  $f'(c) = 0$

**Darboux's Theorem:** if  $f$  differentiable on  $[a, b]$  and  $f'(a) < \alpha < f'(b)$  then  $\exists c \in (a, b)$ , where  $f'(c) = \alpha$

**Rolle's Theorem:**  $f(a) = f(b) \Rightarrow \exists c \in (a, b)$  with  $f'(c) = 0$

**Mean Value Theorem:** if  $f : [a, b] \rightarrow \mathbb{R}$  continuous on  $[a, b]$  and differentiable on  $(a, b)$  then  $\exists c \in (a, b)$  with  $f'(c) = \frac{f(b) - f(a)}{b - a}$

**Generalized Mean Value Theorem:**  $f, g$  continuous on  $[a, b]$  and differentiable on  $(a, b) \Rightarrow \exists c \in (a, b)$  with  $[f(b) - f(a)]g'(c) = [g(b) - g(a)]f'(c)$ ; if  $g \neq 0$ :  $\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}$

**L'Hopital's Rules:**

**0/0:** let  $a \in I$ ,  $f, g$  continuous on  $I$ , differentiable on  $I \setminus a$ :

if  $f(a) = 0 = g(a)$  then  $\lim_{x \rightarrow a} \frac{f'(x)}{g'(x)} = L \Rightarrow \lim_{x \rightarrow a} \frac{f(x)}{g(x)} = L$

**$\infty/\infty$ :** let  $f, g$  differentiable on  $(a, b)$ :

if  $\lim_{x \rightarrow a} g(x) = \infty$  then  $\lim_{x \rightarrow a} \frac{f'(x)}{g'(x)} = L \Rightarrow \lim_{x \rightarrow a} \frac{f(x)}{g(x)} = L$

## Functional Sequences

**Convergence:** let  $f_n$  be defined on  $A \subseteq \mathbb{R}$

*Pointwise:* if  $\lim_{n \rightarrow \infty} f_n(x) = f(x) \forall x \in A$  then  $f_n \rightarrow f$

*Uniform:*  $\forall \epsilon > 0 \exists N \in \mathbb{N}$  s.t.  $|f_n(x) - f(x)| < \epsilon \forall x \in A, n \geq N$

*Cauchy Criterion:*  $(f_n)$  converges uniformly if and only if  $\forall \epsilon > 0 \exists N \in \mathbb{N}$  s.t.  $|f_n(x) - f_m(x)| < \epsilon \forall x \in A, m, n \geq N$

**Continuity:** let  $(f_n)$  converge uniformly to  $f$ , if all  $f_n$  are continuous at  $c \in A$ , then  $f$  is continuous at  $c$

**Differentiability:** let  $(f_n)$  differentiable on  $[a, b]$  and  $(f'_n) \rightarrow g$  uniformly on  $[a, b]$ ; if  $\exists x_0 \in [a, b]$  where  $f_n(x_0)$  convergent, then (1)  $(f_n)$  converges uniformly (2)  $f = \lim f_n$  differentiable (3)  $f' = g$

## Functional Series $\sum_{n=1}^{\infty} f_n(x)$

**Convergence:** converges *pointwise* (uniformly) on  $A$  to  $f(x)$  if sequence of partial sums converges pointwise (uniformly) to  $f(x)$

*Cauchy Criterion:*  $\forall \epsilon > 0 \exists N \in \mathbb{N}$  s.t.

$\forall n > m \geq N : |f_{m+1}(x) + f_{m+2}(x) + \dots + f_n(x)| < \epsilon \forall x \in A$

*Weierstrass M-Test:* let  $(f_n)$  defined on  $A \subseteq \mathbb{R}$ , let  $M_n > 0$

satisfy  $|f_n(x)| \leq M_n \forall x \in A$ ; if  $\sum_{n=1}^{\infty} M_n$  converges, then  $\sum_{n=1}^{\infty} f_n$  converges uniformly on  $A$

**Power Series:**  $f(x) = \sum_{n=0}^{\infty} a_n x^n$

*Convergence:*  $f(x)$  converges for  $x_0 \in \mathbb{R} \Rightarrow$  converges absolutely for any  $x$  with  $|x| \leq |x_0|$  (the set of convergence is  $([-R, R])$  with  $R \in \mathbb{R}_0^+ \cup \infty$ ,  $R$  is called the radius of convergence)

*Abel's Theorem:* let  $f(x)$  converges at  $x = R > 0$ ; then  $f(x)$

converges uniformly on  $[0, R]$  (similar for  $x = -R$ );

*Uniform Convergence:* if a power series converges pointwise on  $A \subseteq \mathbb{R}$ , then it converges uniformly on any compact set  $K \subseteq A$

*Differentiability:* if  $f(x)$  converges on interval  $A \subseteq \mathbb{R}$ , then

•  $f$  continuous on  $A$  and differentiable on any  $(-R, R) \subseteq A$

• the derivative is  $f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$

•  $f$  is infinitely differentiable on  $(-R, R)$

**Taylor Series:**  $f(x) = \frac{f^{(n)}(a)}{n!} (x - a)^n$  (here  $a = 0$ )

*Lagrange's Remainder Theorem:* let  $f$  infinitely differentiable on  $(-R, R)$ , define  $a_n = f^{(n)}(0)/n!$ , let  $S_N$  partial sums to  $N$ , then for  $x \neq 0 \exists c : |c| < |x|$ , where error  $f(x) - S_N(x) = \frac{f^{(N+1)}(c)}{(N+1)!} x^{N+1}$

## 1.2 Measure Theory (Capiński and Kopp, 2004)

**Riemann Integral** coinciding upper  $U$  and lower  $L$  sums for partitions  $P$ :  $U = \sum_{i=1}^n \sup_{p_i} \text{len}(p_i)$ ;  $L = \sum_{i=1}^n \inf_{p_i} \text{len}(p_i)$

**Riemann's Criterion:**  $f$  integrable iff  $\forall \epsilon > 0 \exists P_\epsilon$  s.t.  $U - L < \epsilon$

**Fundamental Theorem of Calculus:** if  $f : [a, b] \rightarrow \mathbb{R}$  continuous and  $F' = f$ , then  $F(b) - F(a) = \int_a^b f(x) dx$

**Problems:** why do we need a Lebesgue integral?

- scope: many results need continuous  $f$  and bounded intervals
- dependence on intervals: otherwise often not defined
- lack of completeness:  $\int_a^b f_n(x) dx \rightarrow \int_a^b f(x) dx$  doesn't hold

### Measure

**Null Sets**  $\forall \epsilon > 0 \exists$  a sequence of intervals  $(I_n)$  s.t.

$A \subseteq \bigcup_{n=1}^{\infty} I_n$  and  $\sum_{n=1}^{\infty} l(I_n) < \epsilon$  ( $= \exists$  arbitrarily small cover); the union of a sequence of null sets is also null

**Lebesgue Outer Measure** of  $A$ :  $m^*(A) = \inf Z_A$ , with  $Z_A = \{\sum_{n=1}^{\infty} l(I_n) : I_n \text{ are intervals, } A \subseteq \bigcup_{n=1}^{\infty} I_n\}$

- null sets have outer measure zero
- monotonicity:  $A \subset B \Rightarrow m^*(A) \leq m^*(B)$
- the outer measure of an interval equals its length
- countable subadditivity:  $m^*(\bigcup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} m^*(E_n)$
- translation invariance:  $m^*(A) = m^*(A + t)$

**Lebesgue Measurability:** of  $E \subseteq \mathbb{R}$  (write  $E \in \mathcal{M}$ ) if  $\forall A \subseteq \mathbb{R} : m^*(A) = m^*(A \cap E) + m^*(A \cap E^c)$  (additivity)

$\mathcal{M}$  is a  $\sigma$ -field: and countably additive ( $=$ : measure)

•  $\mathbb{R} \in \mathcal{M}$  (btw so are all null sets and intervals)

•  $E \in \mathcal{M} \Rightarrow E^c \in \mathcal{M}$

•  $E_n \in \mathcal{M} \forall n \Rightarrow \bigcup_{n=1}^{\infty} E_n \in \mathcal{M}$ , with countable additivity:

$m^*(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} m^*(E_n)$  for  $E_j \cap E_k = \emptyset$  if  $j \neq k$

(it follows that  $\mathcal{M}$  is closed under countable intersections as well) the intersection of  $\sigma$ -fields is a  $\sigma$ -field

*Open Sets:* every open set in  $\mathbb{R}$  can be expressed as a union of countable open intervals  $\Rightarrow$  all open sets are in  $\mathcal{M}$

**Lebesgue Measure**  $m$ :  $f : \mathcal{M} \rightarrow [0, \infty]$ , countably additive  $m^*$

• let  $A_n \in \mathcal{M} : A_n \subset A_{n+1} \forall n \Rightarrow m(\bigcup_n A_n) = \lim_{n \rightarrow \infty} m(A_n)$

let  $m(A_n) < \infty : A_n \supset A_{n+1} \forall n \Rightarrow m(\bigcap_n A_n) = \lim_{n \rightarrow \infty} m(A_n)$

•  $m$  is continuous at  $\emptyset$ , i.e.  $(B_n)$  decrease to  $\emptyset \Rightarrow m(B_n) \rightarrow 0$

**Borel Sets:**  $\sigma$ -field generated by a family of sets

$\mathcal{B} = \bigcap \{\mathcal{F} : \mathcal{F} \text{ is a } \sigma\text{-field containing all intervals}\}$

$\mathcal{B}$  is also generated by intervals of a particular type

*Completion:* not all null sets are in  $\mathcal{B}$

$\mathcal{M}$  is the completion of  $\mathcal{B}$ : a measure space  $(X, \mathcal{F}, \mu)$  is complete if  $\forall F \in \mathcal{F}$  with  $\mu(F) = 0$ :  $N \subset F$  is in  $\mathcal{F}$  and  $\mu(N) = 0$

<p><b>Borel Regular Measure:</b> <math>\mu(B) = \inf\{\mu(O) : O \text{ open}, O \supset B\} =</math>  <math>\sup\{\mu(F) : F \text{ closed}, F \subset B\}</math>; <i>Approximations:</i> of <math>m</math> for <math>E \in \mathcal{M}</math></p>	$\left  \begin{array}{l} \text{above: } \forall \epsilon > 0 \exists \text{ open set } O \text{ s.t. } A \subset O, m(O) \leq m^*(A) + \epsilon \\ \text{below: if } E \in \mathcal{M} \text{ then } \epsilon > 0 \exists \text{ closed set } F \subset E \text{ s.t. } m(E \setminus F) < \epsilon \end{array} \right.$
---	--

## 2 General

### Ladder Of Causation (Pearl, 2019)

1. rung: **association**  $Pr[y|x]$  observation *What is?*
  2. rung: **intervention**  $Pr[y|do(x), z]$  experiment *What if?*
  3. rung: **counterfactuals**  $Pr[y^x|x', y']$  retrospection *Why?*
- if a tool can answer rung  $i$  questions, it can also answer rung  $j < i$

### Causal Roadmap (Petersen and van der Laan, 2014)

systematic approach linking causality to statistical procedures

**1. Specifying Knowledge.** structural causal model (unifying counterfactual language, structural equations, & causal graphs): a set of possible data-generating processes, expresses background knowledge and its limits

**2. Linking Data.** specifying measured variables and sampling specifics (latter can be incorporated into the model)

**3. Specifying Target.** define hypothetical experiment: decide

1. variables to intervene on: one (point treatment), multiple (longitudinal, censoring/missing, (in)direct effects)
2. intervention scheme: static, dynamic, stochastic
3. counterfactual summary of interest: absolute or relative, marginal structural models, interaction, effect modification
4. population of interest: whole, subset, different population

**4. Assessing Identifiability.** are knowledge and data sufficient to derive estimand and if not, what else is needed?

**5. Select Estimand.** current best answer: knowledge-based assumptions + which minimal convenience-based assumptions (transparency) gets as close as possible

**6. Estimate.** choose estimator by statistical properties, nothing causal here

**7. Interpret.** hierarchy: statistical, counterfactual, feasible intervention, randomized trial

### Average Causal Effect $E[Y^{a=1}] \neq E[Y^{a=0}]$

$$E[Y^a] = \sum_y y p_{Y^a}(y) \quad (\text{discrete})$$

$$= \int y f_{Y^a}(y) dy \quad (\text{continuous})$$

individual causal effect  $Y_i^{a=1} \neq Y_i^{a=0}$  generally unidentifiable

*null hypothesis:* no average causal effect

*sharp null hypothesis:* no causal effect for any individual

**notation**  $A, Y$ : random variables (differ for individuals);  $a, y$ : particular values; counterfactual  $Y^{a=1}$ :  $Y$  under treatment  $a = 1$

**stable unit treatment value assumption (SUTVA)**  $Y_i^a$  is well-defined: no interference between individuals, no multiple versions of treatment (weaker: treatment variation irrelevance)

**causal effect measures** typically based on means

*risk difference:*  $Pr[Y^{a=1} = 1] - Pr[Y^{a=0} = 1]$

*risk ratio:*  $\frac{Pr[Y^{a=1}=1]}{Pr[Y^{a=0}=1]}$

*odds ratio:*  $\frac{Pr[Y^{a=1}=1]/Pr[Y^{a=1}=0]}{Pr[Y^{a=0}=1]/Pr[Y^{a=0}=0]}$

*number needed to treat (NNT)* to save 1 life:  $-1/\text{risk difference}$

**sources of random error:** sampling variability (use consistent estimators), nondeterministic counterfactuals

**association** compares  $E[Y|A = 1]$  and  $E[Y|A = 0]$ , **causation** compares  $E[Y^{a=1}]$  and  $E[Y^{a=0}]$  (whole population)

### Target Trial

emulating an ideal randomized experiment

explicitly formulate target trial & show how it is emulated  $\rightarrow$

less vague causal question, helps spot issues

**missing data problem** unknown counterfactuals

*randomized experiments:* missing completely at random  $\rightarrow$

exchangeability (= exogeneity as treatment is exogenous)

*ideal randomized experiment:* no censoring, double-blind,

well-defined treatment, & adherence  $\rightarrow$  association is causation

*pragmatic trial:* no placebo/blindness, realistic monitoring

**PICO** (population, intervention, comparator, outcome): some components of target trial

### three types of causal effects:

*intention-to-treat effect* (effect of treatment assignment)

*per-protocol effect* (usually dynamic when toxicity arises)

*other intervention effect* (strategy changed during follow-up)

**controlled direct effects:** effect of  $A$  on  $Y$  not through  $B$

*natural direct effect*  $A$  on  $Y$  if  $B^{a=0}$  (cross-world quantity)

*principal stratum effect*  $A$  on  $Y$  for subset with  $B^{a=0} = B^{a=1}$

**crossover experiment:** sequential treatment & outcome  $t=0, 1$  individual causal effect  $Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$  only identifiable if: no carryover effect, effect  $\perp$  time, outcome  $\perp$  time

**time zero** if eligibility at multiple  $t$  (observational data):

earliest, random  $t$ , all  $t$  (adjust variance with bootstrapping)

**grace periods:** usually treatment starts  $x$  months after first

eligible, if death before: randomly assign strategy/copy into both

### Identifiability Conditions hold in ideal experiments

**consistency** counterfactuals correspond to data  $Y = Y^A$ :

if  $A = a$ , then  $Y^a = Y$  for each individual

- precise definition of  $Y^a$  via specifying  $a$  (sufficiently well-defined  $a$  maybe impossible (effect of DNA before it was discovered), relies on expert consensus)
- linkage of counterfactuals to data ( $a$  must be seen in data)

**positivity**  $Pr[A = a|L = l] > 0 \forall l$  with  $Pr[L = l] > 0$ ;

$$f_L(l) \neq 0 \Rightarrow f_{A|L}(a|l) > 0 \forall a, l$$

- structural violations (inference not on full population)
- random variability (smooth over with parametric models)

can sometimes be empirically verified (if all is seen in data)

**exchangeability** unverifiable without randomization

- *marginal:*  $Y^a \perp\!\!\!\perp A \hat{=}$  randomized experiment, counterfactuals are missing completely at random (MCAR)
- *conditional:*  $Y^a \perp\!\!\!\perp A|L \hat{=}$  conditionally randomized, counterfactuals are missing at random (MAR)

alternative definition:  $Pr[A = 1|Y^{a=0}, L] = Pr[A = 1|L]$

### additional conditions:

*correct measurement* mismeasurement of  $A, Y, L$  results in bias

*correct model specification* models  $\xrightarrow{\text{may}}$  misspecification bias

### Effect Modification $A$ on $Y$ varies across levels of $V$

null average causal effect  $\neq$  null causal effect per subgroup

**population characteristics:** causal effect measure is actually "effect in a population with a particular mix of effect modifiers"

**transportability:** extrapolation of effect to another population (issues: effect modification, versions of treatment, interference)

effects conditional on  $V$  may be more transportable

**types:** additive/multiplicative scale, qualitative (effect in

opposite directions)/quantitative, surrogate/causal

### calculation:

- *stratify* by  $V$  then standardize/IP weight for  $L$ ,
- $L$  as *matching* factor (ensures positivity, difficult if high-dimensional  $L$ )

**collapsibility:** causal risk difference and ratio are weighted

averages of stratum-specific risks, can not be done for odds ratio

**Interaction** effects of joint interventions  $A$  and  $E$

$$\Pr[Y^{1,1}=1] - \Pr[Y^{0,1}=1] \neq \Pr[Y^{1,0}=1] - \Pr[Y^{0,0}=1]$$

$A$  and  $E$  have equal status and could also be considered a combined treatment  $AE$ , exchangeability for both is needed  
*additive scale* (above): “>” superadditive and “<” subadditive;  
*multiplicative scale*: “>” super- and “<” submultiplicative

**difference to effect modification**: if  $E$  is randomly assigned methods coincide, but  $V$  can not be intervened on as  $E$  can  
**monotonicity** effect is either nonnegative or nonpositive  $\forall i$   
**sufficient component-cause framework** pedagogic model  
*response types* for binary  $A$ : helped, immune, hurt, doomed;  
for binary  $A$  and  $E$ : 16 types

(minimal) sufficient causes:

- (minimal)  $U_1$  together with  $A = 1$  ensure  $Y = 1$
- (minimal)  $U_2$  together with  $A = 0$  ensure  $Y = 1$

*sufficient cause interaction*:  $A$  and  $E$  appear together in a minimal sufficient cause

**NPSEM** nonparametric structural equation model

$$V_m = f_m(pa_m, \epsilon_m)$$

counterfactuals are obtained recursively, e.g.  $V_3^{v_1} = V_3^{v_1, V_2^{v_1}}$   
implies any variable can be intervened on

aka finest causally interpreted structural tree graph (FCISTG)

**additional assumption**  $\cap$  FCISTG  $\Rightarrow$  causal Markov condition:

- independent errors (NPSEM-IE): all  $\epsilon_m$  mutually independent
- fully randomized (FFRCISTG):  $V_m^{\bar{v}_{m-1}} \perp\!\!\!\perp V_j^{\bar{v}_{j-1}}$  if  $\bar{v}_{j-1}$  subvector of  $\bar{v}_{m-1}$

NPSEM-IE  $\Rightarrow$  FFRCISTG (assume DAGs represent latter)

NPSEM-IE assume crossworld independencies  $\rightarrow$  unverifiable

**Causal DAG** draw assumptions before conclusions

*rules*: arrow means direct causal effect for at least one  $i$ , absence

means sharp null holds, all common causes are on the graph

*neglects*: direction of cause (harmful/protective), interactions

*convention*: time flows from left to right

**causal Markov assumption**: any variable ( $v$ ) | its direct causes ( $pa_j$ )  $\perp\!\!\!\perp$  its non-descendants ( $\neg v_j$ )  $\Leftrightarrow$  Markov factorization

$$f(v) = \prod_{j=1}^M f(v_j | pa_j)$$

**d-separation** (d for directional): a pathway in a DAG is ...

- blocked if collider or conditioned on non-collider
- opened if conditioned on collider or descendent of collider

2 variables are d-separated if all connecting paths are blocked

under causal Markov: d-separation  $\Rightarrow$  independence

under faithfulness: independence  $\Rightarrow$  d-separation

**faithfulness**: effects don't cancel out perfectly

*discovery*: process of learning the causal structure; requires faithfulness, but even with it is often impossible

**Noncausal DAGs** (Hernán and Robins, 2023)  $Y^a$  has to

be well-defined (identifiability), what about  $Y^l$  (if  $L \rightarrow Y$ )?

if  $Y^l$  is not well-defined, but  $L \rightarrow Y$ , then the graph is not causal

**statistical interpretation**: only  $A \rightarrow Y$  is causal, the rest simply encodes conditional independencies, *but* why should a DAG corresponding to the study variables even exist then?

**hidden factor**:  $L$  is only a surrogate for  $H$ , with  $Y^h$

well-defined, however,  $L$  being a surrogate can introduce bias

**pragmatic approach**: “cause” as a primary concept which does not need explanation in terms of well-defined interventions (approach is in need of mathematical theory)

**SWIGs** single world intervention graphs

**counterfactual graphic approach**:  $A$  turns into  $A|a$ , the left (right) side inherits incoming (outgoing) arrows (intervention with  $A = a$ ); all outcomes of  $A$  get a superscript  $a$ , e.g.  $Y^a$ ; more than one intervention possible, dynamic strategies require additional arrows from  $L$  to  $a$

$A$  and  $Y^a$  are d-separated for  $L \rightarrow Y^a \perp\!\!\!\perp A|L$  (for FFRCISTG)

**Confounding** bias due to common cause of  $A$  &  $Y$  *not in*  $L$   
randomization prevents confounding

**backdoor path**: noncausal path  $A$  to  $Y$  with arrow into  $A$

**backdoor criterion**: all backdoor paths are blocked by  $L$  & no descendants of  $A$  in  $L \Rightarrow$  conditional exchangeability

$Y^a \perp\!\!\!\perp A|L \Rightarrow L$  fulfills backdoor criterion if faithful (FFRCISTG)

**confounders in observational studies**: occupational factors (*healthy worker bias*), clinical decisions (*confounding by indication/channeling*), lifestyle, genetic factors (*population stratification*), social factors, environmental exposures

given a DAG, confounding is an absolute, confounder is relative  
surrogate confounders in  $L$  may reduce confounding bias

**negative outcome controls**: if  $A$  and  $Y$  share a common cause  $U$ : measure effect for  $Y_0$  (before treatment) and  $Y_1$  (after), subtract (assumption of additive equi-confounding)

**front door criterion** using the full mediator  $M$ :  $\Pr[Y^a = 1] = \sum_m \Pr[M = m|A = a] \sum_{a'} \Pr[Y = 1|M = m, A = a'] \Pr[A = a']$

**Selection Bias** bias due to common effect of  $A$  &  $Y$  *in*  $L$   
 $=$  conditioning on collider (can't be fixed by randomization)

**examples**: informative censoring, nonresponse bias, healthy worker bias, volunteer bias; often M-bias ( $A \leftarrow U_1 \rightarrow L \leftarrow U_2 \rightarrow Y$ )

**solution**: target  $Y^{A,C}$ ,  $AC$  fulfills identifiability conditions, if competing events, interventions may not be well-defined

**multiplicative survival model**:  $\Pr[Y=0|E=e, A=a] = g(e)h(a)$

$\rightarrow$  no interaction between  $E$  and  $A$  on the multiplicative scale;

if  $Y = 0$  is conditionally independent, then  $Y = 1$  can't be as  $\Pr[Y=1|E=e, A=a] = 1 - g(e)h(a) \rightarrow$  conditioning on a collider could be unbiased if restricted to certain levels ( $Y = 0$ )

**Measurement Bias** aka information bias

measurements  $X^*$  of variables  $X$  can be included in DAG

**independent** errors  $U$  if  $f(U_A, U_Y) = f(U_A)f(U_Y)$

**nondifferential**  $A$ : if  $f(U_A|Y) = f(U_A)$ ;  $Y$ :  $f(U_Y|A) = f(U_Y)$

mismeasurement  $\rightarrow$  bias, if:  $A \rightarrow Y$  or dependent or differential

**reverse causation bias** caused by e.g. recall bias: independent but differential  $A$  (caused by  $Y \rightarrow U_A$ )

**misclassified treatment**: assignment  $Z$  does not determine  $A$

*exclusion restriction*: ensure  $Z \not\rightarrow Y$ , e.g. via double-blinding

- **per-protocol effect**: either as-treated ( $\rightarrow$  confounded) or restricted to protocol adhering individuals ( $\rightarrow$  selection bias)
- **intention-to-treat effect** ( $\rightarrow$  measurement bias): advantages:  $Z$  is randomized, preserves null (if exclusion restriction holds), = underpowered  $\alpha$ -level test of the null (only if monotonicity; underpowered may be problematic if treatment safety is tested)

sometimes mismeasurement doesn't matter as the measurement itself is of interest (Hernán and Robins, 2023)

**Random Variability** quantify uncertainty due to small  $n$

**CI**: e.g. Wald CI =  $\hat{\theta} \pm 1.96 \times se(\hat{\theta})$ , *calibrated* if it contains 95 % of estimands (>: *conservative*, <: *anticonservative*)

*large sample* CI: converge to 95 % vs. *small-sample*: always valid

*honest*:  $\exists n$  where coverage  $\geq 95\%$ , *valid*: large-sample & honest  
**inference**: either restrict inference to sample (randomization-based inference) or inference on super-population  
**super-population**: generally a fiction, but  $\rightarrow$  simple statistical properties (where does the variability of the distribution come from: assumption population is sampled from super-population)  
**conditionality principle**: inference should be performed conditional on ancillary statistics (e.g. L-A association) as

$$\mathcal{L}(Y) = f(Y|A, L)f(A|L)f(L)$$

*exactly ancillary*  $A, L$ :  $f(Y|A, L)$  depends on parameter of interest, but  $f(A, L)$  does not share parameters with  $f(Y|A, L)$   
*approximately ancillary*: ... does not share **all** parameters ...  
 continuity principle: also condition on approximate ancillaries  
**curse of dimensionality**: difficult to do conditionality principle

**Time-Varying Treatments** compare 2 treatments  
 treatment history up to  $k$ :  $\bar{A}_k = (A_0, A_1, \dots, A_k)$   
 shorthand: always treated  $\bar{A} = \bar{1}$ , never treated  $\bar{A} = (\bar{0})$   
**static strategy**:  $g = [g_0(\bar{a}_{-1}), \dots, g_K(\bar{a}_{K-1})]$   
**dynamic strategy**:  $g = [g_0(\bar{l}_0), \dots, g_K(\bar{l}_K)]$   
**stochastic strategy**: non-deterministic  $g$   
 optimal strategy is where  $E[Y^g]$  is maximized (if high is good)

**Sequential Identifiability** sequential versions of  
**exchangability**:  $Y^g \perp\!\!\!\perp A_k | \bar{A}_{k-1} \quad \forall g, k = 0, 1, \dots, K$   
*conditional exchangeability*:  
 $(Y^g, L_{k+1}^g) \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{L}_k), \bar{L}^k \quad \forall g, k = 0, 1, \dots, K$   
**positivity**:  $f_{\bar{A}_{k-1}, \bar{L}_k}(\bar{a}_{k-1}, \bar{l}_k) \neq 0 \Rightarrow$   
 $f_{\bar{A}_k | \bar{A}_{k-1}, \bar{L}_k}(a_k | \bar{a}_{k-1}, \bar{l}_k) > 0 \quad \forall (\bar{a}_{k-1}, \bar{l}_k)$   
**consistency**:  
 $Y^{\bar{a}} = Y^{\bar{a}^*} \quad \text{if } \bar{a} = \bar{a}^*; \quad Y^{\bar{a}} = Y \quad \text{if } \bar{A} = \bar{a};$   
 $\bar{L}_k^{\bar{a}} = \bar{L}_k^{\bar{a}^*} \quad \text{if } \bar{a}_{k-1} = \bar{a}_{k-1}^*; \quad \bar{L}_k^{\bar{a}} = \bar{L}_k \quad \text{if } \bar{A}_{k-1} = \bar{a}_{k-1}$

**generalized backdoor criterion** (static strategy): all backdoors into  $A_k$  (except through future treatment) are blocked  $\forall k$   
**static sequential exchangeability for  $Y^{\bar{a}}$**  (weaker version)

$$Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1}, \bar{L}_k \quad \text{for } k = 0, 1, \dots, K$$

sufficient to identify mean counterfactual outcome for static strategies and can be checked on SWIGS via d-separation  
**time-varying confounding**  $E[Y^{\bar{a}} | L_0] \neq E[Y | A = \bar{a}, L_0]$

**Treatment-Confounder Feedback**  $A_0 \rightarrow L_1 \rightarrow A_1$ :  
 an unmeasured  $U$  influencing  $L_1$  and  $Y$  turns  $L_1$  into a collider;  
 traditional adjustment (e.g. stratification) biased: use g-methods  
**g-null test** sequential exchangeability & sharp null true  $\Rightarrow$   
 $Y^g = Y \quad \forall g \Rightarrow Y \perp\!\!\!\perp A_0 | L_0 \quad \& \quad Y \perp\!\!\!\perp A_1 | A_0, L_0, L_1$ ; therefore:  
 if last two independences don't hold, one assumption is violated  
**g-null theorem**:  $E[Y^g] = E[Y]$ , if the two independences hold  
 ( $\Rightarrow$  sharp null: only if strong faithfulness (no effect cancelling))

**Causal Mediation** (Hernán and Robins, 2023)  
 $A \xrightarrow{\quad} M \xrightarrow{\quad} Y$  seen as longitudinal with  $k_0$ :  $A$  and  $k_1$ :  $M$   
**decompose**  $E[Y^{a=1}] - E[Y^{a=0}]$  into cross-world quantities  

- pure (aka natural) direct effect (upper path)  
 $E[Y^{a=1, M^{a=0}}] - E[Y^{a=0, M^{a=0}}]$
- total (aka natural) indirect effect (lower path)  
 $E[Y^{a=1, M^{a=1}}] - E[Y^{a=1, M^{a=0}}]$

**mediation formula** under NPSEM-IE (requires  $Y^{a=1, m} \perp\!\!\!\perp M^{a=0}$  cross-world independence)  
 $E[Y^{a=1, M^{a=0}}] = \sum_m E[Y | A = 1, M = m] \Pr[M = m | A = 0]$   
**interventional interpretation** advocating NPSEM-IE assumption:  
 $A \xrightarrow{\quad} N \xrightarrow{\quad} M \xrightarrow{\quad} Y$  (thick arrows are deterministic)  
 no controlled direct effects: no  $N \rightarrow Y$  and no  $O \rightarrow M$   
 FFRCISTG point of view: intervention on  $N$  and  $O$  separately  
 if decomposable (can be verified in a randomized trial), g-formula for  $N$  and  $O$  reduces to mediation formula for  $A$



### 3 Models

**Modeling** data are a sample from the target population

*estimand*: quantity of interest, e.g.  $E[Y|A = a]$   
*estimator*: function to use, e.g.  $\hat{E}[Y|A = a]$   
*estimate*: apply function to data, e.g. 4.1

**model**: a priori restriction of joint distribution/dose-response curve; *assumption*: no model misspecification (usually wrong)

**non-parametric estimator**: no restriction (saturated model) = *Fisher consistent estimator* (entire population data  $\rightarrow$  true value)

**parsimonious model**: few parameters estimate many quantities

**bias-variance trade-off**:

wiggleness  $\uparrow \rightarrow$  misspecification bias  $\downarrow$ , CI width  $\uparrow$

**Variable Selection** can induce bias if  $L$  includes:

(descendant of) collider: *selection bias under the null*

noncollider effect of  $A$ : *selection bias under the alternative*

mediator: *overadjustment for mediators*

temporal ordering is not enough to conclude anything

**bias amplification**: e.g. by adjusting for an instrument  $Z$  (can also reduce bias)

**Super Learning** (van der Laan et al., 2007, 2011)

**oracle selector**: select best estimator of set of learners  $Z_i$

**discrete super learner**: select algorithm with smallest cross-validated error (converges to oracle for large sample size)

**super learner**: improves asymptotically on discrete version

$\text{logit}(Y = 1|Z) = \sum_i \alpha_i Z_i$ , with  $0 < \alpha_i < 1$  and  $\sum \alpha_i = 1$

weights  $\alpha_i$  are determined inside the cross-validation; for the prediction,  $Z_i$  trained on the full data set are used

can be cross-validated itself to check for overfitting (unlikely)

**Marginal Structural Models** association is causation in the IP weighted pseudo-population

associational model  $E[Y|A] =$  causal model  $E[Y^a]$

*step 1*: estimate/model  $f[A|L]$  (and  $f[A]$ )  $\rightarrow$  get  $(S)W^A$

*step 2*: estimate regression parameters for pseudo-population

**effect modification** variables  $V$  can be included (e.g.

$\beta_0 + \beta_1 a + \beta_2 V a + \beta_3 V$ ; technically not marginal anymore),

$SW^A(V) = \frac{f[A|V]}{f[A|L]}$  more efficient than  $SW^A$

### 3.1 Traditional Methods

**Stratification** calculate risk for each stratum of  $L$   
 only feasible if enough data per stratum

**Outcome Regression** often assume no effect modification

$$E[Y^{a,c=0}|L] = \beta_0 + \beta_1 a + \beta_2 aL + \beta_3 L = E[Y|A, C = 0, L]$$

faux marginal structural model as no IP weighting/ $SW^A(L) = 1$   
 for ATE only  $\beta_1, \beta_2$  of interest, the rest are *nuisance parameters*

**Propensity Score Methods**  $\Pr[A = 1|L] =: \pi(L)$

$\Rightarrow A \perp\!\!\!\perp L|\pi(L)$  (definition of a balancing score); can be modelled

- **stratification**: create strata with similar  $\pi(L)$  (e.g. deciles), but the average  $\pi(L)$  might still be different in some strata
- **standardization**: use  $\pi(L)$  instead of  $L$  to standardize
- **matching**: find close ( $\rightarrow$  bias-variance trade-off) values of  $\pi(L)$ , positivity issues arise often

propensity models don't need to predict well, just ensure exchangeability (good prediction leads to positivity problems)

**Instrumental Variable Estimation**  $L$  unmeasured surrogate/proxy instruments can be used

**instrumental conditions**:

1. **relevance condition**:  $Z \not\perp\!\!\!\perp A$ , meaning  $Z$  is associated with  $A$  (weak association (F-statistic  $< 10$ )  $\rightarrow$  weak instrument)
2. **exclusion restriction**:  $Z$  affects  $Y$  at most through  $A$ 
  - (a) population level:  $E[Y^{z,a}] = E[Y^{z',a}]$  (sometimes enough)
  - (b) **individual level**:  $Y_i^{z,a} = Y_i^{z',a} = Y_i^a$
3. **exchangeability**:  $Z$  and  $Y$  have no shared causes
  - (a) **marginal**:  $Y^{a,z} \perp\!\!\!\perp Z$  (typically enough)
  - (b) joint:  $\{Y^{z,a}; a \in [0, 1], z \in [0, 1]\} \perp\!\!\!\perp Z$
4. (not needed for an instrument, just the IV estimand below)
  - (a) **effect homogeneity**: (i) constant effect  $A \rightarrow Y \forall i$  (ii) constant average effect  $A \rightarrow Y \forall A$  (iii) no additive effect modifiers (iv) additive Z-A association is constant across  $L$
  - (b) **monotonicity**:  $A^{z=1} \geq A^{z=0} \forall i$  (more credible than 4a)

**common instruments**: (physician's) general preference, access to/price of  $A$ , genetic factors (Mendelian randomization)

**bounds**: binary outcome ATE  $[-1, 1]$  (width 2)  $\xrightarrow{\text{data}}$  (width 1) *natural bounds* need 2a,3a (width  $\Pr[A=1|Z=0] + \Pr[A=0|Z=1]$ ) *sharp bounds* require 2a,3b (narrower than natural bounds)

**IV estimand ATE**: intention-to-treat  $\div$  measure of compliance

(1,2b,3a,4a): ATE; (1,2b,3a,4b): ATE in compliers

binary  $Z$ :  $\frac{E[Y|Z=1] - E[Y|Z=0]}{E[A|Z=1] - E[A|Z=0]}$ , continuous  $Z$ :  $\frac{Cov(Y, Z)}{Cov(A, Z)}$ ;

can be calculated as *two-stage-least-squares estimator*:

1.  $E[A|Z]$  2.  $E[Y|Z] = \beta_0 + \beta_1 \hat{E}[A|Z]$  3.  $\hat{\beta}_1$  is IV estimate

**disadvantages**: often leads to wide CI, small violations of conditions can lead to large biases

**regression discontinuity design**: if threshold in  $L$  exists which determines  $A$  perfectly + assumption of continuity in  $L \rightarrow$  jump in  $Y$  at threshold is the causal effect (if no effect modification by  $L$ ); a fuzzy variant also exists (Hernán and Robins, 2023)

**Causal Survival Analysis** time-to-event data

additional censoring due to administrative end of follow-up

**competing events** (often death): censoring (assume population with death abolished) or not (after death, chance of event is zero, but what is the effect of  $A$ ?)  $\rightarrow$  create composite event

**survival quantities**  $k$  is a time point,  $T$  is time of event

- **survival** at  $k$ :  $\Pr[T > k] =: \Pr[D_k = 0]$
- **risk** at  $k$ :  $1 - \Pr[T > k] = \Pr[T \leq k] = \Pr[D_k = 1]$
- **hazard** at  $k$ :  $\Pr[T = k|T > k-1] = \Pr[D_k = 1|D_{k-1} = 0]$ ,  
*hazard ratio* is paradoxical due to in-built selection bias

**modeling**: some options

• **Kaplan-Meier** aka product limit formula (nonparametric):

$$\Pr[D_k = 0] = \prod_{m=1}^k \Pr[D_m = 0|D_{m-1} = 0]$$

• parametric e.g. log hazards model:

– use **IP weights**  $SW^A$  in structural marginal model

$$\text{logit } \Pr[D_{k+1}^{a,c=0} = 0|D_k^{a,c=0} = 0] = \beta_{0,k} + \beta_1 a + \beta_2 a k$$

– **standardize** ( $\prod_k 1 -$ ) parametric hazards model

$$\Pr[D_{k+1} = 1|D_k = 0, C_k = 0, L, A] \text{ weighting across } L$$

- **structural nested cumulative failure time model (CFT):**  
 $\frac{\Pr[D_k^a=1|L,A]}{\Pr[D_k^a=0=1|L,A]} = \exp[\gamma_k(L, A; \psi)]$  (log-linear has no upper limit  $1 \rightarrow$  rare failure  $\uparrow$ ; if  $\downarrow$ , use a survival model (CST)), use g-estimation like with AFT
- **accelerated failure time model (AFT)** with g-estimation:  
 $T_i^a/T_i^{a=0} = \exp(-\psi_1 a - \psi_2 a L_i)$ , exchangeability for  $C$  is guaranteed via artificial censoring (include only individuals who would not have been censored either way)

## 3.2 G-Methods

**G-Methods** generalized treatment contrasts: adjust for  $L$

- **standardization:** two types of g-formula
- **IP weighting:** (in theory) also g-formula
- **g-estimation:** not needed unless longitudinal

**standardization and IP weighting** are equivalent, **but** if modeled, different “no misspecification” assumptions: outcome model (standardization), treatment model (IP weighting)

**big g-formula** not all methods use (sequential) exchangeability

- **problem:** DAG is known, but unmeasured variables exist
- **solution:** include un- & measured variables in big g-formula  $\rightarrow$  derive alternative effect identification methods using only d-separation (e.g. front door formula)

it can always be determined, if the DAG allows for identification with the big g-formula (Hernán and Robins, 2023)

**censoring:** measure joint effect of  $A$  and  $C$  with  $E[Y^{a,c=0}]$   
**standardization**  $E[Y|A=a] = \int E[Y|L=l, A=a, C=0] dF_L[l]$

**IP weights**  $W^{A,C} = W^A \times W^C$  (uses  $n$ ) or  
 $SW^{A,C} = SW^A \times SW^C$  (uses  $n^{c=0}$ )

**g-estimation** only adjusts for confounding  $\rightarrow$  use IP weights

**time-varying censoring  $\bar{C}$ :** monotonic type of missing data

**standardization:**  $\int f(y|\bar{a}, \bar{c}=\bar{0}, \bar{l}) \prod_{k=0}^K dF(l_k|\bar{a}_{k-1}, c_{k-1}=0, \bar{l}_{k-1})$

**IP weighting:**

$$SW^{\bar{C}} = \prod_{k=1}^{K+1} \frac{1 \cdot \Pr(C_k = 0|\bar{A}_{k-1}, C_{k-1} = 0)}{\Pr(C_k = 0|\bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k)}$$

**Standardization** plug-in (parametric if so) g-formula

$$E[Y^a] = \overbrace{E[E[Y|A=a, L=l]]}^{\text{conditional expectation}} = \overbrace{\int E[Y|A=a, L=l] f_L[l] dl}^{\text{joint density estimator}}$$

weighted average of stratum-specific risks; unknowns can be estimated non-parametrically or modeled

**no need to estimate  $f_L[l]$ /integrate** as empirical distribution

can be used: estimate outcome model  $\rightarrow$  predict counterfactuals

on whole dataset  $\rightarrow$  average the results ( $\rightarrow$  CI by bootstrapping)

**for discrete  $L$**   $E[Y|A=a]$  is  $\sum_l E[Y|A=a, L=l] \Pr[L=l]$

**time-varying** standardize over all possible  $\bar{l}$ -histories  
simulates joint distribution of counterfactuals  $(Y^{\bar{a}}, \bar{L}^{\bar{a}})$  for  $\bar{a}$   
**joint density estimator (jde)**

$$\text{discrete: } E[Y^{\bar{a}}] = \sum_{\bar{l}} E[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}] \prod_{k=0}^K f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1})$$

$$\text{continuous: } \int f(y|\bar{a}, \bar{l}) \prod_{k=0}^K f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}) dl$$

for **stochastic strategies** multiply with  $\prod_{k=0}^K f^{int}(a_k|\bar{a}_{k-1}, \bar{l}_k)$

**time-varying** two options based on g-methods as examples

**standardization** (plug-in estimate): risk is  $\Pr[D_{k+1}^{\bar{a}, \bar{c}=\bar{0}} = 1] =$

$$\sum_{\bar{l}_k} \sum_{j=0}^k \Pr[D_{j+1} = 0|\bar{A}_j = \bar{a}_j, \bar{L}_j = \bar{l}_j, \bar{D}_j = 0] \times \prod_{s=0}^j \left\{ \Pr[D_s = 0|\bar{A}_{s-1} = \bar{a}_{s-1}, \bar{L}_{s-1} = \bar{l}_{s-1}, \bar{D}_{s-1} = 0] \times f(l_s|\bar{a}_{s-1}, \bar{l}_{s-1}, D_s = 0) \right\}$$

**IP weighting:** fit a pooled logistic hazard model with time-varying weights  $W_k^{\bar{A}} = \prod_{m=0}^k \frac{1}{f(A_m|\bar{A}_{m-1}, \bar{L}_m)}$

**estimation** (Young et al., 2011; Schomaker et al., 2019)

1. model  $f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1})$  and  $E[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}]$
2. simulate data forward in time:  
at  $k=0$ : use empirical distribution of  $L_0$  (observed data)  
at  $k>0$ : set  $\bar{A}=\bar{a}$ , draw from models estimated in 1.
3. calculate mean of  $\hat{Y}_{K,i}^{\bar{a}}$  (bootstrap for CI)

**iterated conditional expectation (ice)**

$$E[Y_T^{\bar{a}}] = E[E[E[...E[Y_T|\bar{A}_{T-1}=\bar{a}_{T-1}, \bar{L}_T] ...|\bar{A}_0=a_0, L_1]|L_0]]$$

**estimation** (Schomaker et al., 2019)

1. model inside out:  $Q_T = E[Y_T|\bar{A}_{T-1}, \bar{L}_T]$  to  $Q_0 = E[Q_1|\bar{L}_0]$ , predict  $Q_t$  with  $\bar{A}=\bar{a}$  in each step
2. calculate mean of  $\hat{Q}_{0,i}^{\bar{a}}$  (bootstrap for CI)

**g-null paradox** even if the sharp null holds, model misspecification can lead to it being falsely rejected

Proof: for  $L_0 \rightarrow A_0 \rightarrow Y_0 \rightarrow L_1 \rightarrow A_1 \rightarrow Y_1$ ,  $\bar{a} = (a_0, a_1)$

$$E[Y_1^{\bar{a}}] \stackrel{\text{CE}}{=} E[E[Y_1^{\bar{a}}|A_0=a_0, L_0]]$$

$$(\text{ice}) \stackrel{\text{CE}^*}{=} E[E[E[Y_1|\bar{L}, \bar{A}=\bar{a}, Y_0]|A_0=a_0, L_0]]$$

$$\stackrel{\text{LIE}}{=} E\left[\sum_{l_1} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0, y_0]\right]$$

$$\stackrel{\text{LIE}}{=} \sum_{l_0} \left[ \sum_{l_1} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0, y_0] \right] \Pr[l_0]$$

$$(\text{jde}) \stackrel{\text{sum}}{=} \sum_{\bar{l}} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0] \Pr[l_0]$$

CE: conditional expectation; \*: exchangeability;

LIE: law of iterated expectation

**IP Weighting** inverse probability of treatment (g-formula)

$$E[Y^a] = E\left[\frac{I(A=a)Y}{f(A|L)}\right]; W^A = \frac{1}{f(A|L)}; SW^A = \frac{f(A)}{f(A|L)}$$

unknowns can be estimated non-parametrically or modeled

**pseudo-population:** everyone is treated & untreated ( $L \not\rightarrow A$ )

**FRCISTG** (fully randomized causally interpreted structured

graph): probability tree for  $L \rightarrow A \rightarrow Y$ , can be used to

calculate/visualize simulation of values for  $A$

**for discrete  $A, L$ :**  $f[a|l] = \Pr[A=a, L=l]$

**estimators:** Horvitz-Thompson; Hajek (modified version)

**stabilized weights  $SW^A$**  should have an average of 1 (check!)

$\rightarrow$  pseudo-population same size  $\rightarrow$  (if non-saturated) CI width  $\downarrow$

**time-varying**

$$W^{\bar{A}} = \prod_{k=0}^K \frac{1}{f(A_k|\bar{A}_{k-1}, \bar{L}_k)}; SW^{\bar{A}} = \prod_{k=0}^K \frac{f(A_k|\bar{A}_{k-1}, \bar{L}_k)}{f(A_k|\bar{A}_{k-1}, \bar{L}_k)}$$

**G-Estimation** (additive) structural nested models

$$\text{logit Pr} [A = 1 | H(\psi^\dagger), L] = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 L$$

$$H(\psi^\dagger) = Y - \psi_1 A$$

find  $\psi^\dagger$  which renders  $\alpha_1 = 0$ ; 95 %-CI: all  $\psi^\dagger$  for which  $p > 0.05$   
closed-form solution for linear models

**derivation:**  $H(\psi^\dagger) = Y^{a=0}$

$$\text{logit Pr} [A = 1 | Y^{a=0}, L] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$$

$Y^{a=0}$  unknown, but because of exchangeability  $\alpha_1$  should be zero

$$Y^{a=0} = Y^a - \psi_1 a$$

equivalent to  $Y^{a=0} = Y^{a=1} - \psi_1$ , but using no counterfactuals

**structural nested mean model**

$$\text{additive: } E[Y^a - Y^{a=0} | A = a, L] = \beta_1 a (+\beta_2 a L)$$

$$\text{multiplicative: } \log \left( \frac{E[Y^a | A = a, L]}{E[Y^{a=0} | A = a, L]} \right) = \beta_1 a (+\beta_2 a L)$$

multiplicative is preferred if  $Y$  always positive, but does not extend to longitudinal case

semi-parametric: agnostic about  $\beta_0$  and effect of  $L \rightarrow$  robust  $\uparrow$

**no time-varying:** no nesting; model equals marginal structural models with missing  $\beta_0, \beta_3$  (unspecified “no treatment”)

**sensitivity analysis:** unmeasured confounding ( $\alpha_1 \neq 0$ ) can be examined: do procedure for different values of  $\alpha_1 \rightarrow$  plot  $\alpha_1$  vs.  $\psi^\dagger \rightarrow$  how sensitive is estimate to unmeasured confounding?

**effect modification:** add  $V$  in both g-estimation equations

**doubly robust estimators** exist

**time-varying nested equations:** for each time  $k$

**structural nested mean models** separate effect of each  $a_k$

$$E[Y^{\bar{a}_{k-1}, a_k, \bar{a}_{k+1}} - Y^{\bar{a}_{k-1}, \bar{a}_{k+1}} | \bar{L}^{\bar{a}_{k-1}} = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}] =$$

$$a_k \gamma_k (\bar{a}_{k-1}, \bar{l}_k, \beta)$$

**calculations**

$$H_k(\psi^\dagger) = Y - \sum_{j=k}^K A_j \gamma_j (\bar{A}_{j-1}, \bar{L}_j, \psi^\dagger)$$

function  $\gamma_j$  can be, e.g. constant ( $\psi_1$ ), time-varying only ( $\psi_1 + \psi_2 k$ ), or dependent on treatment/covariate history

$$\text{logit Pr} [A_k = 1 | H_k(\psi^\dagger), \bar{L}_k, \bar{A}_{k-1}] =$$

$$\alpha_0 + \alpha_1 H_k(\psi^\dagger) + \alpha_2 w_k (\bar{L}_k, \bar{A}_{k-1})$$

find  $\alpha_1$  that is closest to zero

a closed form estimator exists for the linear case

### 3.3 Doubly Robust Methods

**Double-Robustness** (Hernán and Robins, 2023)

g-formula: *either* treatment model  $f(L)$  *or* outcome model  $b(L)$   
*or* appropriately combine both: “two chances to get it right”

**all doubly robust estimators**

- involve a correction of outcome  $\hat{b}(L)$  using the treatment  $\hat{f}(L)$
- have a bias depending on a product of the errors  $\frac{1}{\pi(l)} - \frac{1}{\pi(l)}$  and  $b(l) - \hat{b}(l)$  known as second order bias

**time-varying:** multiple robustness for  $k = 0, 1, \dots, K$

$K+2$  robustness: consistent, if  $\hat{f}_0$  to  $\hat{f}_I$  and  $\hat{b}_{I+1}$  to  $\hat{b}_K$  are

$2^{K+1}$  robustness: consistent, if for each  $k$ , either  $\hat{f}_k$  or  $\hat{b}_k$  are

**Machine Learning**  $L$  is high-dimensional

one could use lasso or ML for IP weighting/standardization

**but:** ML does not guarantee elimination of confounding and has largely unknown statistical properties: how to get CI?

**sample splitting:** train estimators on training sample  $T_r$ , use resulting estimators for doubly robust method on estimation sample (CIs on estimation sample are valid, but  $n$  halved)

**cross-fitting:** do again the other way round, average the two estimates, get CI via bootstrapping [*alternatively:* split into  $M$  samples, use one sample for estimation and  $M-1$  for training  $\rightarrow$  improved finite sample behavior (Hernán and Robins, 2023)]

**asymptotic behavior** for valid (Wald) CI we need:

- a bias much smaller than  $c \cdot 1/\sqrt{n}$ , which is how the  $se$  typically scales (use doubly robust methods for small bias)
- asymptotic normality (for Wald CI)
- for a doubly robust estimator  $\psi_{dr}$ , we need sample splitting, otherwise  $\hat{b}(l)$  and  $\hat{f}(l)$  are correlated with  $\psi_{dr}$

if  $\hat{b}(l)$  and  $\hat{f}(l)$  are consistent and  $E[\hat{\psi} - \psi | T_r] / se(\hat{\psi})$  converges to  $0 \rightarrow \hat{\psi}$  with sample splitting is asymptotically normal and unbiased  $\rightarrow$  CI is calibrated (Hernán and Robins, 2023)

**problems:** unclear choice of algorithm, is bias small enough?

**Advantages** (van der Laan et al., 2011)

**consistent** if either  $\bar{Q}_0$  or  $g_n$  are consistent (*doubly robust*):

$$\forall \epsilon > 0, P \in \mathcal{M} : \Pr_P [\hat{\theta}_n - \theta(P) > \epsilon] \rightarrow 0 \text{ as } n \rightarrow \infty$$

**collaboratively doubly robust:**  $g_n$  only needs predictors of  $Y$ , as it does not try to fit  $g_0$  well, but improve the fit of  $\bar{Q}_n^*$

**asymptotic unbiasedness** if either  $\bar{Q}_0$  or  $g_0$  are consistent, super learning makes  $\bar{Q}_0$  and  $g_n$  max. asymptotically unbiased

**asymptotic efficiency** if both  $\bar{Q}_0$  and  $g_n$  are consistent:

achieves Cramer-Rao bound of minimum possible asymptotic variance (requires asymptotic unbiasedness)

**asymptotic linearity** if either  $\bar{Q}_0$  or  $g_n$  are consistent:

means estimator behaves like empirical mean

- bias converges to zero at rate smaller than  $1/\sqrt{n}$
- for large  $n$  estimator is approximately normally distributed

**Influence Curve** how robust is an estimator?

$$IC_{T, P_n}(O) = \lim_{\epsilon \rightarrow 0} \frac{T[(1-\epsilon)P_n + \epsilon\delta_O] - T(P_n)}{\epsilon}$$

for estimator  $T$  and distribution  $P_n$  with  $0 < \epsilon < 1$

can also be rewritten as a **directional derivative** at  $P_n$

$$IC_{T, P_n} = \frac{d}{d\epsilon} T[(1-\epsilon)P_n + \epsilon\delta_O] = \frac{d}{dP_n} T(\delta_O - P_n)$$

in direction  $(\delta_O - P_n)$ , where  $P_n$  empirical probability measure that puts mass  $1/n$  on  $O_i$  (Hampel, 1974)

**special cases** (van der Laan et al., 2011)

- $\bar{IC}(P_0) = 0$  and  $\text{Var}(IC(P_0))$  asymptotic variance of the standard estimator  $\sqrt{n}(\psi_n - \psi_0) \rightarrow \text{Var}(\hat{\Psi}(P_n)) = \frac{\text{Var}_{IC}}{n}$
- efficient IC: an estimator is asymptotically efficient  $\Leftrightarrow$  its influence curve is the efficient influence curve  $IC(O) = D^*(O)$

**Delta Method** (Zepeda-Tello et al., 2022) estimand is a

function of  $\theta$ , i.e.  $\psi := \phi(\theta)$ ,  $\text{Var}(\hat{\theta})$  known, but what is  $\text{Var}(\hat{\psi})$ ?

**Taylor's approximation** requirements:

- univariate  $\phi$ : differentiable at  $\theta$
- multivariate  $\phi$ :  $\exists \partial_v \phi(\theta)$  (directional derivative)
- functional  $\phi$  (function of functions):  $\exists \partial_v \phi(\theta)$  & coincides with one-sided directional (Hadamard) derivatives ( $\stackrel{*}{=} \nabla \phi(\theta)^T v$ )

first order Taylor (rearranged $^\dagger$ ):  $\phi(\hat{\theta}_n) \approx \phi(\theta) + \partial_{v:=\hat{\theta}-\theta} \phi(\theta)$

**classical delta method:** if  $\{r_n\}_{n=1}^\infty$  with  $\lim_{n \rightarrow \infty} r_n = \infty$ ,

where  $r_n(\hat{\theta}_n - \theta)$  converges to  $Z \sim N(0, 1)$  (e.g.  $r_n = \sqrt{n/\sigma^2}$ ), then

$$r_n \left( \phi(\hat{\theta}_n) - \phi(\theta) \right) \stackrel{\dagger}{\approx} \nabla \phi(\theta)^T r_n(\hat{\theta}_n - \theta) \stackrel{d}{\rightarrow} \nabla \phi(\theta)^T Z$$

$$\Rightarrow \text{Var} \left[ \phi(\hat{\theta}_n) - \phi(\theta) \right] = \text{Var} \left[ \phi(\hat{\theta}_n) \right] \approx \frac{1}{r_n^2} \text{Var} \left[ \nabla \phi(\theta)^T Z \right]$$

**functional delta:**  $r_n(\hat{\theta}_n - \theta) \stackrel{d}{\rightarrow} Z \Rightarrow r_n \left( \phi(\hat{\theta}_n) - \phi(\theta) \right) \stackrel{d}{\rightarrow} \partial_Z \phi(\theta)$

**influence function:**  $\psi = \phi(\mathbb{P}_X)$  is a functional

estimations rate of change for  $\mathbb{P}_X$  to  $Q$ , where  $Q = \mathbb{1}_{\{Y\}}$

$$\text{IF}_{\phi, \mathbb{P}_X}(Y) := \partial_{Q - \mathbb{P}_X} \phi(\mathbb{P}_X) = \lim_{h \downarrow 0} \frac{\phi((1-h)\mathbb{P}_X + hQ) - \phi(\mathbb{P}_X)}{h},$$

*interpretation:* rate of change if distribution deviates from  $\mathbb{P}_X$  to  $Q = \text{one observation } Y$ , assigns probability 1 to  $X$  taking value  $Y$

*use delta:*  $\phi(\hat{\mathbb{P}}_X) \approx \phi(\mathbb{P}_X) + \text{IF}_{\phi, \mathbb{P}_X}(Y)$ , if  $(\hat{\theta}_n - \theta) \stackrel{n \rightarrow \infty}{\sim} N(., .)$

$$\hat{\psi}_n - \psi = \phi(\hat{\theta}_n) - \phi(\theta) \stackrel{\text{approx}}{\sim} N(0, \text{Var}[\text{IF}_{\phi, \mathbb{P}_X}(Y)]),$$

where  $\widehat{\text{Var}}[\text{IF}_{\phi, \mathbb{P}_X}(Y)] = \frac{1}{n} \sum_{i=1}^n (\text{IF}_{\phi, \mathbb{P}_X}(X_i))^2$ , which is the classical  $S^2$  estimator since the mean is known ( $= 0$ )

**using the delta method (general case)**

1. determine asymptotic distribution of  $v := r_n(\hat{\theta}_n - \theta)$
2. define  $\phi$  and compute Hadamard derivative
3. multiply asymptotic distribution with Hadamard derivative, then estimate the variance

### Simple Plug-In Estimator proto-TMLE

1. fit outcome regression with variable  $R = \begin{cases} +W^A & \text{if } A=1 \\ -W^A & \text{if } A=0 \end{cases}$
2. standardize by averaging

**time-varying  $K+2$  robust estimator (related to TMLE)**

1. estimate  $\hat{f}(A_m | \bar{A}_{m-1}, \bar{L}_m)$  (e.g. logistic model), use it to calculate at each time  $m$ :  $\bar{W}^{\bar{A}_m} = \prod_{k=0}^m \frac{1}{\hat{f}(A_k | \bar{A}_{k-1}, \bar{L}_k)}$  and modified IP weights at  $m$ :  $\bar{W}^{\bar{A}_{m-1}, a_m} = \frac{\bar{W}^{\bar{A}_m}}{\hat{f}(a_m | \bar{A}_{m-1}, \bar{L}_m)}$
2. with  $\hat{T}_{K+1} := Y$ , recursively for  $m = K, K-1, \dots, 0$ :
  - (a) fit outcome regression on  $\hat{T}_{m+1}$  with variable  $\bar{W}^{\bar{A}_m}$
  - (b) calculate  $\hat{T}_m$  using the outcome model with  $\bar{W}^{\bar{A}_{m-1}, a_m}$
3. calculate standardized mean outcome  $\hat{E}[Y^a] = E[\hat{T}_0]$

### Augmented IPTW (Hernán and Robins, 2023)

$$\hat{E}[Y^a] = \frac{1}{n} \sum_{i=1}^n \left[ \frac{\mathbb{1}(A=a)Y}{\hat{f}(A|L)} - \left( \frac{\mathbb{1}(A=a)}{\hat{f}(A|L)} - 1 \right) \hat{b}(a, L) \right]$$

**disadvantages:** ignores global constraints  $\rightarrow$  often unstable if sparsity, sometimes not well-defined (van der Laan et al., 2011)

Relationship between AIPTW and TMLE for causal effect:

$$\hat{\psi}_{1, \text{AIPTW}} - \hat{\psi}_{0, \text{AIPTW}} = P_n \left[ \hat{b}(1, L) \right] - P_n \left[ \hat{b}(0, L) \right]$$

$$- P_n \left[ \frac{\{ \mathbb{1}(A=1) - \mathbb{1}(A=0) \} (Y - \hat{b}(A, L))}{\hat{f}(A|L)} \right]^\dagger$$

using the IRLS estimate for  $b(A, L; \beta, \theta) = \phi \left[ m(A, L; \beta) + \theta \left\{ \frac{\mathbb{1}(A=1) - \mathbb{1}(A=0)}{\hat{f}(A|L)} \right\} \right]$  with canonical link  $\phi$  sets the last part<sup>†</sup> to zero (as the score equation for  $\theta$ )

**TMLE** (van der Laan and Rubin, 2006; van der Laan et al., 2011) *targeted maximum likelihood estimation*: an ML-based substitution estimator of the g-formula

$$O = (W, A, Y) \sim P_0; \quad \mathcal{L}(O) = \overbrace{\Pr(Y|A, W)}^Q \overbrace{\Pr(A|W)}^g \overbrace{\Pr(W)}^{Q_W}$$

target  $\Psi(P_0) = \Psi(\bar{Q}_0, Q_{W,0}) = \psi_0$ , *ATE*:  $\bar{Q}_0 = E_0(Y|A, W)$

**first step:** outcome model  $\bar{Q}_n^0(A, W)$  estimating  $\bar{Q}_0$  (part of  $P_0$ )

- super learning is often used here, but leads to a biased estimate
- not all of  $P_0$  is estimated, just relevant portion  $\bar{Q}_0 \rightarrow$  efficiency

**second step:** update  $\bar{Q}_n^0(A, W)$  to  $\bar{Q}_n^1(A, W)$  using treatment model  $g_n$  estimating  $g_0 = P_0(A|W)$ , e.g. for binary  $A$ :

1. model  $g_n$ , super learning is a popular choice here, too
2. calculate  $n$  clever covariates:  $H_n^*(A, W) = \begin{cases} \frac{1}{g_n(1|W)} & \text{if } A_i=1 \\ \frac{-1}{g_n(0|W)} & \text{if } A_i=0 \end{cases}$
3. update  $\bar{Q}_n^0$ , by estimating  $\epsilon_n$  with offset logistic regression:  $\text{logit } \bar{Q}_n^1(A, W) = \text{logit } \bar{Q}_n^0(A, W) + \epsilon_n H_n^*(A, W)$  (converges after first update), then calculate counterfactuals

- goal: bias reduction, get optimal bias-variance trade-off

- removes all asymptotic bias, if consistent estimator is used here

**third step:** use empirical distribution for  $Q_{W,0}$  in a substitution estimator, e.g.:  $\psi_n^{TMLE} = \frac{1}{n} \sum_{i=1}^n [\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)]$

**advantages:** loss-based (does not only solve efficient influence curve estimating equation, but also uses a loss and working model preserving global constraints), well-defined (as a loss-based learner), substitution estimator (respects global constraints  $\rightarrow$  more robust to outliers and sparsity)  $\rightarrow$  good finite sample performance

**closed form inference based on the influence curve**, e.g.:

$$IC_n^*(O_i) = \overbrace{\left[ \frac{\mathbb{1}(A_i=1)}{g_n(1, W_i)} - \frac{\mathbb{1}(A_i=0)}{g_n(0, W_i)} \right]}^a [Y - \bar{Q}_n^1(A_i, W_i)]$$

$$+ \overbrace{\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i) - \psi_{TMLE, n}}^b$$

TMLE sets the mean of the IC,  $\overline{IC}_n$ , to zero ( $b$  has already mean zero, see third step, MLE sets the sum of  $a$  to zero, if  $H_n^*(A, W)$  is chosen correctly  $\rightarrow$  the first part of  $a$  is the clever covariate)

*sample variance* is then:  $S^2(IC_n) = \frac{1}{n} \sum_{i=1}^n (IC_n(O_i) - \overline{IC}_n)^2$

*standard error* of estimator:  $\sigma_n = \sqrt{\frac{S^2(IC_n)}{n}}$

95% CI:  $\psi_{TMLE, n} \pm z_{0.975} \frac{\sigma_n}{\sqrt{n}}$ ; p-value:  $2 \left[ 1 - \Phi \left( \left| \frac{\psi_{TMLE, n}}{\sigma_n / \sqrt{n}} \right| \right) \right]$

**time-varying LTMLE** (Schomaker et al., 2019; van der Laan and Gruber, 2012) *longitudinal TMLE*: based on ice g-formula for  $t = T, \dots, 1$ :

1. model  $\hat{E}(Y_t | \bar{A}_{t-1}, \bar{L}_t)$  (for individuals observed at  $t-1$ )
2. plug in  $\bar{A}_{t-1} = \bar{d}_{t-1}$ ; use regression from step 1 to predict outcome at time  $t$ , i.e.  $\bar{Y}_t^{\bar{d}_t}$
3. update estimate with  $\bar{Y}_{t, \text{new}}^{\bar{d}_t} = \text{offset}(\bar{Y}_t^{\bar{d}_t}) + \epsilon \hat{H}(\bar{A}, \bar{C}, \bar{L})_{t-1}$ : update  $\bar{Y}_t^{\bar{d}_t}$  (or regress  $\text{offset}(\bar{Y}_t^{\bar{d}_t}) + \epsilon 1$  with weights  $\hat{H}(\bar{A}, \bar{C}, \bar{L})_{t-1}$ , with clever covariate (without censoring):  $\hat{H}(\bar{A}, \bar{L})_{t-1} = \prod_{s=0}^{t-1} \frac{\mathbb{1}(\bar{A}_s = \bar{d}_s)}{\widehat{\Pr}(\bar{A}_s = \bar{d}_s | \bar{A}_{s-1} = \bar{d}_{s-1}, \bar{L}_s = \bar{l}_s)}$
4.  $\hat{\psi}_T = \text{mean of } \bar{Y}_1^{\bar{d}_1}$ , get CI using influence curve result is a  $K+2$  multiply robust estimator (Díaz et al., 2021)

### targeted minimum loss-based estimation

target parameter  $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ , with  $\mathcal{M}$  the statistical model used

1. compute  $\Psi$ 's pathwise derivative at  $P$  and its corresponding canonical gradient  $D^*(P)$  (efficient influence curve)
2. define a loss  $L()$  s.t.  $P \rightarrow E_0 L(P)$  is minimized at true  $P_0$
3. for a  $P$  in model  $\mathcal{M}$  define a parametric working model  $\{P(\epsilon) : \epsilon\}$  s.t.  $P(\epsilon=0) = P$  and a "score"  $\frac{d}{d\epsilon} L(P(\epsilon))$  s.t. it (or linear combination of its components) equals  $D^*(P)$  at  $P$
4. compute  $\epsilon_n^0 = \arg \min_{\epsilon} \sum_{i=1}^n L(P_n^0(\epsilon))(O_i)$ , with initial estimate  $P_n^0$ , then first iteration  $P_n^1 = P_n^0(\epsilon_n^0)$ , repeat until  $\epsilon_n^k = 0$
5. get TMLE estimate  $\psi_0$  by plugging  $P_n^*$  into  $\Psi$  (substitution)
6. TMLE solves the efficient influence curve equation  $\sum_{i=1}^n D^*(P_n^*)(O_i) = 0 \rightarrow$  asymptotic linearity and efficiency can also be carried out for a relevant part  $Q$  instead of all of  $P$

**LMTP** (Díaz et al., 2021) *modified treatment policies*

**problems** for (longitudinal) continuous or multi-valued  $A$ :

- fixed value counterfactuals unrealistic

- infinite-dimensional dose-response curve needs parametric assumptions or is not  $n^{1/2}$  consistent
- positivity is often violated

**solution:** longitudinal MTP  $A_t^d = d(A_t(\bar{A}_{t-1}^d), H_t(\bar{A}_{t-1}^d))$ , e.g. threshold ( $\max(c, a_t)$ ), shift ( $a_t + \delta$  if positivity else  $a_t$ ), stochastic (draw from  $F(d(A_t, H_t)|H_t)$ ; randomizer  $\perp\!\!\!\perp U, P$ ), shifted propensity score (only for binary  $A$ )

**identification** for a given NPSEM, assumptions:

- *positivity* if  $(a_t, h_t)$  in  $\text{supp}\{A_t, H_t\}$  then  $(d(a_t, h_t), h_t)$  too
- *sequential randomization*:
  - standard  $U_{A,t} \perp\!\!\!\perp \underline{U}_{L,t+1}|H_t$  (for stochastic LMTP)
  - strong  $U_{A,t} \perp\!\!\!\perp (\underline{U}_{L,t+1}, \underline{U}_{A,t+1})|H_t$  (for other LMTP)

iterative process: set  $m_{\tau+1} := Y$ , for  $t = \tau, \dots, 1$ :

$$m_t : (a_t, h_t) \mapsto E[m_{t+1}(A_{t+1}^d, H_{t+1})|A_t = a_t, H_t = h_t]$$

$$\text{solve } \theta = E[m_1(A_1^d, L_1)]$$

**optimality limitations:** threshold LMTPs can't be  $n^{1/2}$

consistent as parameter not pathwise differentiable, continuous  $A$  can only be considered, if  $d(\cdot, h_t)$  *piecewise smooth invertible* efficient influence curve (assumes  $d \perp\!\!\!\perp P$ ):

$$EIF\left(E\left[m_1(A^d, L_1)\right]\right) = \phi_1(Z) - \theta$$

with  $r_t(a_t, h_t) = \frac{g_t^d(a_t|h_t)}{g_t(a_t|h_t)}$  and  $\phi_t : z \mapsto \sum_{s=t}^{\tau} (\prod_{k=t}^s r_k(a_k, h_k)) \{m_{s+1}(a_{s+1}^d, h_{s+1}) - m_s(a_s, h_s)\} + m_t(a_t^d, h_t)$

**estimation** use Super Learner for  $\hat{r}_t$  and  $\hat{m}_t$

- ***g-methods:*** asymptotically linear and  $n^{1/2}$  consistent if models correctly specified, asymptotic distribution generally unknown

*substitution (standardization):*  $\hat{\theta}_{\text{sub}} = \frac{1}{n} \sum_{i=1}^n \hat{m}_1(A_{1,i}^d, L_{1,i})$

*IPTW:*  $\hat{\theta}_{\text{iptw}} = \frac{1}{n} \sum_{i=1}^n (\prod_{t=1}^{\tau} \hat{r}_t(A_{t,i}, H_{t,i})) Y_i$

- ***TMLE:*** use sample splitting and cross-fitting with sets  $\mathcal{T}_j$ , TMLE sets cross-validated EIF  $P_n\{\phi_1(\cdot, \tilde{\eta}_j(\cdot)) - \hat{\theta}_{\text{tmle}}\}$  to zero  $\tau+1$  multiply robust &  $n^{1/2}$  consistent (if nuisance consistent)

*step 1:* initialize  $\tilde{\eta} = \hat{\eta}$  and  $\tilde{m}_{\tau+1,j(i)}(A_{\tau+1,i}^d, H_{\tau+1,i}) = Y_i$

*step 2:* compute  $\tau$  weights  $\omega_{s,i} = \prod_{k=1}^s \hat{r}_{k,j(i)}(A_{k,i}, H_{k,i})$

*step 3:* for  $t = \tau, \dots, 1$ : fit generalized linear tilting model

$$\text{link } \tilde{m}_t^\epsilon(A_{t,i}, H_{t,i}) = \epsilon + \text{link } \tilde{m}_{t,j(i)}(A_{t,i}, H_{t,i})$$

with the canonical link and use  $\hat{\epsilon}$  to update  $\tilde{m}_{t,j(i)}^{\hat{\epsilon}}$

*step 4:*  $\hat{\theta}_{\text{tmle}} = \frac{1}{n} \sum_{i=1}^n \tilde{m}_{1,j(i)}(A_{1,i}^d, L_{1,i})$

- ***SDR:***  $2^\tau$  multiply robust (sequentially double robust) and same rate of  $n^{1/2}$  consistency as TMLE, better finite sample behavior than TMLE but estimate is not guaranteed to be in support

*step 0:* cross-fit estimates  $\hat{r}_{1,j(i)}, \dots, \hat{r}_{\tau,j(i)}$

*step 1:*  $\phi_{\tau+1}(Z_i; \tilde{\eta}_{\tau,j(i)}) = Y_i$

*step 2:* for  $t = \tau, \dots, 1$ :

- compute pseudo-outcome  $\check{Y}_{t+1,i} = \phi_{t+1}(Z_i; \tilde{\eta}_{\tau,j(i)})$

- for  $j = 1, \dots, J$ : regress  $\check{Y}_{t+1,i}$  on  $(A_{t,i}, H_{t,i})$  only using  $i \in \mathcal{T}_j$ , with  $\tilde{m}_{t,j}$  output, update  $\tilde{\eta}_{t,j} = (\hat{r}_{t,j}, \check{m}_{t,j}, \dots, \hat{r}_{\tau,j}, \check{m}_{\tau,j})$

*step 3:*  $\hat{\theta}_{\text{sdr}} = \frac{1}{n} \sum_{i=1}^n \phi_1(Z_i, \tilde{\eta}_{j(i)})$

- \* ***estimate density ratio  $r_t$ :*** duplicate dataset, where

duplicates get assigned  $A_t^d$  with indicator  $\Lambda \in \{0, 1\}$

$$r_t(a_t, h_t) \stackrel{1}{=} \frac{p^\lambda(a_t, h_t|\Lambda=1)}{p^\lambda(a_t, h_t|\Lambda=0)} \stackrel{2}{=} \frac{P^\lambda(\Lambda=1|A_t=a_t, H_t=h_t)}{P^\lambda(\Lambda=0|A_t=a_t, H_t=h_t)} \stackrel{3}{=} \frac{u_t^\lambda(a_t, h_t)}{1 - u_t^\lambda(a_t, h_t)}$$

with 1 definition of  $r_t$ , 2 Bayes rule, and 3 by definition

$\Rightarrow$  any classification method can be used (e.g. Super Learning), cross-fitting should be used

## Methods for continuous $A$ (Kennedy et al., 2017)

doubly robust methods possible for continuous  $A$  for *parametric* effect curves otherwise a  $\sqrt{n}$  consistent estimator can not exist

**procedure:** found double robust  $\xi$  using efficient influence curve in  $\mathbb{E}\{\xi(Z; \pi, \mu)|A = a\} = \theta(a)$ , with data  $Z$  and nuisance  $\pi, \mu$

*step 1:* estimate nuisance  $\pi, \mu$  and predict

*step 2:* construct pseudo-outcome  $\hat{\xi}(Z; \hat{\pi}, \hat{\mu})$  and regress on  $A$  (e.g. using local linear kernel regression)

**consistent if:** either  $\bar{\pi} = \pi$  or  $\bar{\mu} = \mu$ ,  $\theta(a)$  twice continuously differentiable (and two other items are continuous, assumptions on the kernel part and the function class of nuisance)

**asymptotic normality** if at least one nuisance is fast enough

**TMLE version** a clever covariate is given by the authors

# References

*If no citation is given, the information is taken from the book (Hernán and Robins, 2020)*

- Abbott, S. (2015). *Understanding analysis*. Springer.
- Capiński, M. and Kopp, P. E. (2004). *Measure, integral and probability*. Springer.
- Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2021). Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, pages 1–16.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.
- Hernán, M. A. and Robins, J. M. (2020). *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.
- Hernán, M. A. and Robins, J. M. (2023). *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.
- Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245.
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60.
- Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418–426.
- Schomaker, M., Luque-Fernandez, M. A., Leroy, V., and Davies, M.-A. (2019). Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Statistics in medicine*, 38(24):4888–4911. ISBN: 0277-6715 Publisher: Wiley Online Library.
- van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1).
- van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1). Article 24.
- van der Laan, M. J., Rose, S., et al. (2011). *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer.
- van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(11).
- Young, J. G., Cain, L. E., Robins, J. M., O’Reilly, E. J., and Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in biosciences*, 3:119–143.
- Zepeda-Tello, R., Schomaker, M., Maringe, C., Smith, M. J., Belot, A., Rachet, B., Schnitzer, M. E., and Luque-Fernandez, M. A. (2022). The delta-method and influence function in medical statistics: a reproducible tutorial. *arXiv preprint arXiv:2206.15310*.