Causal Inference

a summary

Table of Contents

1 Prerequisites

1.1 Real Analysis (Abbott, 2015)

Real Numbers
triangle inequality $|a+b| \leq |a|+|b|$
Density of Q in 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 $\exists f : A \to B$, where f bijective (in+sur)
function $f : A \to B$ mapping $f(x) = ...$, 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 $I_1 \supseteq I_2 \supseteq I_3 ...$, then $\cap_{n=1}^{\infty} I_n \neq \emptyset$

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 $(a_n) \to a$
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 $(a_n){\to}a$, $(b_n){\to}b$, then $(ca_n){\to}ca$, $(a_n{+}b_n){\to}a{+}b$, $(a_nb_n){\to}ab$, $(a_n/b_n){\to}a/b$ 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* $\sum_{k=0}^{\infty} ar^k \overset{|r|<1}{=} \frac{a}{1-r}$
Convergence: to B, if $(s_m){\to}B$, 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. $n > m \geq N \Rightarrow |a_{m+1} + a_{m+2} + ... + a_n| < \epsilon$; that implies if $\sum_{k=1}^{\infty} a_k$ converges then $(a_k) \to 0$
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: $\sum_{n=}^{\infty} b_n$ 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 $\Rightarrow \sum_{i=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}$ $= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij} = \lim_{n\to\infty} s_{nn}$, where $s_{nn} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$

Sets
Cantor Set: $C = \cap_{n=0}^{\infty} C_n$, with C_n removing the middle third of all intervals, e.g. $C_1 = C_0 \setminus (\frac{1}{3}, \frac{2}{3}) = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]$
Open Sets: $\forall a \in O\,\exists V_\epsilon(a) \subseteq O$, with ϵ-neighborhood of a $V_\epsilon(a) = \{x \in \mathbb{R} : |x-a| < \epsilon\}$; union of open sets is open, the finite intersecton 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 $f : A \to \mathbb{R}$ 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 $\lim_{x\to c} f(x) = L$
Sequential Criterion: $\lim_{x\to c} f(x) = L \Leftrightarrow$ for all sequences $(x_n) \subseteq A$, with $x \neq c$ and $(x_n) \to c$ follows $f(x_n) \to L$
Algebraic Limit Theorem: if $\lim\limits_{x\to c} f(x) = L$ and $\lim\limits_{x\to c} g(x) = M$ then $\lim\limits_{x\to c} kf(x) = kL$; $\lim\limits_{x\to c} [f(x) + g(x)] = L + M$; $\lim\limits_{x\to c} [f(x)g(x)] = LM$; $\lim\limits_{x\to c} f(x)/g(x) = L/M$ 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 $\lim f(x_n){\neq}f(y_n)$ then $\nexists \lim_{x\to c} f(c)$
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: $(x_n) \to c$ (with $x_n \in A$) $\Rightarrow f(x_n) \to f(c)$
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. $|x_n - y_n| \to 0$ but $|f(x_n) - f(y_n)| > \epsilon_0$
Intermediate Value Theorem: $f : [a, b] \to \mathbb{R}$ continuous then $f(a) < L < f(b) \Rightarrow \exists c \in (a, b)$, where $f(c) = L$
alternatively: *Preservation of Connectedness:* $f : A \to \mathbb{R}$ continuous, $E \subseteq A$ connected $\Rightarrow f(E)$ connected
Intermediate Value Property (converse of IVM): 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 $\lim_{x\to c} f(x)$ exists but $\neq f(c)$
- *jump*: $\lim_{x\to c^+} f(x) \neq \lim_{x\to c^-} f(x)$
- *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 $\quad g'(c) = \lim_{x\to c} \frac{g(x)-g(c)}{x-c}$

$(f+g)'(c) = f'(c)+g'(c)$; $\quad (fg)'(c) = f'(c)g(c)+f(c)g'(c)$;
$(kf)'(c) = kf'(c)$; $\quad (f/g)'(c) = \frac{g(c)f'(c)-f(c)g'(c)}{g^2(c)}$, 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 $f : [a, b] \to \mathbb{R}$ 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 $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L \Rightarrow \lim_{x\to a} \frac{f(x)}{g(x)} = L$
∞/∞: let f, g differentiable on (a, b): if $\lim_{x\to a} g(x) = \infty$ then $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L \Rightarrow \lim_{x\to a} \frac{f(x)}{g(x)} = L$

Functional Sequences

Convergence: let f_n be defined on $A \subseteq \mathbb{R}$
Pointwise: if $\lim_{n\to\infty} f_n(x) = f(x)$ $\forall x \in A$ then $f_n \to f$
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 in continuous at c
Differentiability: let (f_n) differentiable on $[a, b]$ and $(f_n') \to g$ 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 $\quad \sum_{n=1}^{\infty} f_n(x)$

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.
$\forall n > m \geq N : |f_{m+1}(x) + f_{m+2}(x) + ...f_n(x)| < \epsilon \forall x \in A$
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 $f : [a, b] \to \mathbb{R}$ continuous and $F' = f$, then $F(b) - F(a) = \int_a^b f(x)\mathrm{d}x$
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: $\int_a^b f_n(x)\mathrm{d}x \to \int_a^b f(x)\mathrm{d}x$ doesn't hold

Measure $\quad \mathcal{B} \subset \mathcal{M} \subset \mathcal{P}(\mathbb{R})$

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 $Z_A = \left\{\sum_{n=1}^{\infty} l(I_n) : I_n \text{ are intervals, } A \subseteq \bigcup_{n=1}^{\infty} I_n\right\}$
- 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: $m^* \left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} m^*(E_n)$
- 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 σ-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 σ-fields is a σ-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**:** $f : \mathcal{M} \to [0, \infty]$, countably additive m^*
- let $A_n \in \mathcal{M} : A_n \subset A_{n+1} \forall n \Rightarrow m(\bigcup_n A_n) = \lim_{n\to\infty} m(A_n)$
  let $m(A_n) < \infty : A_n \supset A_{n+1} \forall n \Rightarrow m(\bigcap_n A_n) = \lim_{n\to\infty} m(A_n)$
- m is continuous at $\emptyset$, i.e. (B_n) decrease to $\emptyset \Rightarrow m(B_n) \to 0$

Borel Sets: σ-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 $\mu(N) = 0$

Borel Regular Measure: $\mu(B) = \inf\{\mu(O) : O \text{ open}, B \subset O\} = \sup\{\mu(F) : F \text{ closed}, F \subset B\}$; *Approximations:* of m for $E \in \mathcal{M}$
above: $\forall \epsilon > 0 \, \exists$ open set O s.t. $A \subset O$, $m(O) \le m^*(A) + \epsilon$
below: if $E \in \mathcal{M}$ then $\epsilon > 0 \, \exists$ closed set $F \subset E$ s.t. $m(E \setminus F) < \epsilon$
Probability restriction to $\mathcal{M}_B = \{A \cap B : A \in \mathcal{M}\}$
Probability Space: $(\Omega, \mathcal{F}, P)$, with Ω an arbitrary set, $\mathcal{F}$ a σ-field of subsets of Ω called events, and P measure on $\mathcal{F}$ s.t. $P(\Omega) = 1$;
Independence:
- events $E = \{A_1, ...A_n\}$: $\forall K \subseteq E : P(\bigcap_K A_i) = \prod_K P(A_i)$
- σ-fields $S = \{\mathcal{F}_1, ...\mathcal{F}_n\}$ on $(\Omega, \mathcal{F}, P)$:
  $\forall S_s \subseteq S : F_i \in \mathcal{F}_j \in S_s \Rightarrow P(\bigcap_{S_s} F_i) = \prod_{S_s} P(F_i)$

Measurable Functions define f only up to null sets (a.e.)

Extended $\mathbb{R}$: $\bar{\mathbb{R}} = [-\infty, \infty]$, with $0 \cdot \infty = 0$ and avoid $\infty - \infty$
(Lebesgue) Measurability: target range (Riemann: domain)
E is measurable, $f : E \to \mathbb{R}$ is measurable if $\forall I \subseteq \mathbb{R}$:
$f^{-1}(I) \in \mathcal{M}$ (if $\in \mathcal{B}$, then Borel(-measurable) function)
Equivalences: f is measurable $\Leftrightarrow \forall a : f^{-1}$ is measurable
for x in $\{(a, \infty), [a, \infty), (-\infty, a), (-\infty, a]\}$
Measurable f: constant, continuous (due to open sets), monotone
Properties: m'able $f : E \to \mathbb{R}$: vector space, i.e. $f+g$, fg m'able
$F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ continuous and f, g m'able $\Rightarrow F(f(x), g(x))$ m'able
f^-, f^+ m'able $\Leftrightarrow f$ m'able $\Rightarrow |f|$ m'able
(f_n) m'able $\Rightarrow \max_{n \le k} f_n, \sup_{n \in N} f_n, \lim \sup_{n \to \infty} f_n$ m'able (also min etc.)
$f, g : E \to \mathbb{R}$, $E \in \mathcal{M}$, f m'able, $\{x : f(x) = g(x)\}$ null $\Rightarrow g$ m'able
(f_n) m'able, $f_n(x) \to f(x)$ a.e. for $x \in E \Rightarrow f$ m'able
Essential Supremum: $\operatorname{ess\,sup} f := \inf\{z : f \le z \text{ a.e.}\}$ (also inf)
Probability *random variable* m'able $X : \Omega \to \mathbb{R}$
σ-field gen. by X: $X^{-1}(\mathcal{B}) = \{S \in \mathcal{F} : S = X^{-1}(B) \text{ for some } B \in \mathcal{B}\}$
probability distribution: $P_X(B) = P(X^{-1}(B))$ is count. add.
Dirac measure: $\delta_a(B) = \mathbb{1}(a \in B)$
X, Y indep.: $\forall B, C$ in $\mathbb{R} : X^{-1}(B) \perp\!\!\!\perp Y^{-1}(C)$

Integral over $E \in \mathcal{M}$ with measure m: $\int_E f \, \mathrm{d}m$

on *simple function* φ (range is finite set of reals $a_i > 0$, with corresp. domain m'able A_i): $\int_E \varphi \, \mathrm{d}m = \sum_{i=1}^n a_i m(A_i \cap E)$
on m'able function $f \ge 0$: $\int_E f \, \mathrm{d}m = \sup Y(E, f)$, where
$Y(E, f) = \{\int_E \varphi \, \mathrm{d}m : 0 \le \varphi \le f, \varphi \text{ simple}\}$; $f=0$ a.e. $\Leftrightarrow \int f \, \mathrm{d}m = 0$
Monotone Convergence Theorem if $\{f_n\} \ge 0$, m'able, &
$f_n \to f$ pointwise then: $\lim_{n \to \infty} \int_E f_n(x) \, \mathrm{d}m = \int_E f \, \mathrm{d}m$
(if limit is only a.e. then E has to be m'able for result to hold)
on m'able function f: $\int_E f \, \mathrm{d}m = \int_E f^+ \, \mathrm{d}m - \int_E f^- \, \mathrm{d}m$
(f is only integrable if both components are finite)
set of all functions that are integrable over E is called $\boldsymbol{\mathcal{L}^1(E)}$
Properties: let f, g integrable, $c \in \mathbb{R}$, A, E measurable then
$f \le g \Rightarrow \int f \, \mathrm{d}m \le \int g \, \mathrm{d}m$; $\int (cf) \, \mathrm{d}m = c \int_E f \, \mathrm{d}m$,
$f+g$ int'able, $\mathcal{L}^1$ vector space; $\int_E (f+g) \, \mathrm{d}m = \int_E f \, \mathrm{d}m + \int_E g \, \mathrm{d}m$;
$\int_A f \, \mathrm{d}m \le \int g \, \mathrm{d}m \forall A \Rightarrow f \le g$ a.e.; int'able f is finite a.e.;
$m(A) \inf_A f \le \int_A f \, \mathrm{d}m \le m(A) \sup_A f$; $|\int f \, \mathrm{d}m| \le \int |f| \, \mathrm{d}m$;
let $f \ge 0$: (1) $\int f \, \mathrm{d}m = 0 \Rightarrow f = 0$ a.e. (2) $A \to \int_A f \, \mathrm{d}m$ is measure
Dominated Convergence Theorem: (f_n) measurable, $|f_n| \le g$ a.e. on $E \in \mathcal{M}$, g E-int'able $\Rightarrow \lim_{n \to \infty} \int_E f_n(x) \, \mathrm{d}m = \int_E f \, \mathrm{d}m$
Beppo-Levi: $\sum_{n=1}^\infty \int |f_n| \, \mathrm{d}m$ finite $\Rightarrow \sum_{n=1}^\infty f_n(x)$ converges a.e., is integrable and $\int \sum_{n=1}^\infty f_n \, \mathrm{d}m = \sum_{n=1}^\infty \int f_n \, \mathrm{d}m$
Relation to Riemann integrals: let $[a, b] \to \mathbb{R}$ bounded
f continuous a.e. $\Leftrightarrow f$ is R-int'able $\Rightarrow f$ is L-int'able ($F_R = F_L$)
L-int's also equal improper R-int's (e.g. $\lim_{a \to -\infty, b \to \infty} \int_a^b f(x) dx$)
Approximating m'able functions:
f on $[a, b]$ bounded: $\forall \epsilon > 0 \, \exists$ step function h s.t. $\int_a^b |f - h| \, \mathrm{d}m \le \epsilon$
$f \in \mathcal{L}^1$: $\forall \epsilon > 0 \, \exists g$ (cont's, zero outside int.) s.t. $\int |f - g| \, \mathrm{d}m \le \epsilon$

Riemann-Lebesgue: $f \in \mathcal{L}^1$: $s_n = \int_{-\infty}^\infty f(x) \overset{(\cos)}{\sin} (kx) \, \mathrm{d}x$ go to 0
Probability: for r.v. X: $\int_\Omega g(X(\omega)) \, \mathrm{d}P(\omega) = \int_\mathbb{R} g(x) \, \mathrm{d}P_X(x)$
absolutely continuous: int'able $f \ge 0$, where measure
$A \to P(A) = \int_A f \, \mathrm{d}m$; f is called a *density* with $\int f \, \mathrm{d}m = 1$;
cdf: defined as $F(y) = P_X((-\infty, y]) \overset{a.c.}{=} \int_{-\infty}^y f(x) \, \mathrm{d}x$, continuity of F does not imply f exists (see Lebesgue function),
F_X is 1 non-decreasing, 2 right continuous and 3 goes from 0 to 1,
$g : \mathbb{R} \to \mathbb{R}$ increasing & diff'able: $f_{g(X)}(y) = f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$
Expectation: $\mathbb{E}(X) = \int_\Omega X \, \mathrm{d}P = \int x \, \mathrm{d}P_X(x) \overset{a.c.}{=} \int x f_X(x) \, \mathrm{d}x$
$\mathbb{C}$: $\int_E u + \mathrm{i}v \, \mathrm{d}m = \int_E u \, \mathrm{d}m + \mathrm{i} \int_E v \, \mathrm{d}m$, $|\int_E f \, \mathrm{d}m| \le \int_E |f| \, \mathrm{d}m$
Characteristic Function: $\varphi_X(t) = \mathbb{E}(\mathrm{e}^{\mathrm{i}tX}) = \int \mathrm{e}^{\mathrm{i}tx} \, \mathrm{d}P_X(x)$ for r.v. X, with $t \in \mathbb{R}$; $\varphi_X(0) = 1$, $-1 \le \varphi_X(t) \le 1$

Spaces Of Integrable f ∞-dim. normed vector spaces

$\boldsymbol{L^1}$ (Banach space := complete normed space):
norm: $x \to \|x\|$, with $x \in$ vector space: $\|x\| \ge 0$,
$\|x\| = 0 \Leftrightarrow x = 0$, $\|\alpha x\| = |\alpha| \|x\|$ ($\alpha \in \mathbb{R}/\mathbb{C}$), triangle equality
metric: $d : X \times X \to \mathbb{R}$ (X set), with $d(x, y) \ge 0$, symmetric,
$d(x, y) = 0 \Leftrightarrow x = y$, triangle equality; e.g.: $d(x, y) = \|x - y\|$
consider vector space $L^1(E) = \mathcal{L}^1(E)/_\sim$, where $f \sim g \Leftrightarrow$
$f(x) = g(x)$ a.e. with $\|f\|_1 = \int_E |f| \, \mathrm{d}m$, $L^1(E)$ is complete
$\boldsymbol{L^2}$ (Hilbert space := Banach space + $(.,.)$ inducing $\|.\|$):
$\mathcal{L}^2$ is set of m'able functions where norm $\|f\|_2 = \left(\int_E |f|^2 \, \mathrm{d}m\right)^{\frac{1}{2}}$
is finite, with L^2 its set of equivalence classes
norm fulfills Δ-eq. due *Schwarz Inequality:* $f, g \in L^2(E, \mathbb{C}) \Rightarrow$
$|\int_E f \bar{g} \, \mathrm{d}m| \le \|fg\|_1 \le \|f\|_2 \|g\|_2$ ($\bar{g}$ complex conjugate: $a - bi$)
$L^1 \not\subset\not\supset L^2$, but if for set D: $m(D) < \infty$ then $L^2(D) \subset L^1(D)$
inner product: $(f, g) := \int f \bar{g} \, \mathrm{d}m$ (induces $\|.\|_2$)
inner product: $(.,.) : H \times H \to \mathbb{C}$, which is (with induced norm)
- linear (1st .): $(f+g, h) = (f, h) + (g, h)$, $(cf, h) = c(f, h)$
- conjugate symmetric: $(f, g) = \overline{(g, f)}$
- positive definite: $(f, f) \ge 0$, $(f, f) = 0 \Leftrightarrow f = 0$
- (if $\mathbb{C}$: conjugate) linear (2nd .): $(f, cg + h) = \bar{c}(f, g) + (f, h)$

$\to$ *parallelogram law:* $\|f+g\|^2 + \|f-g\|^2 = 2\|f\|^2 + 2\|g\|^2$
$\to$ *polar.:* $4(f, g) = \|f+g\|^2 - \|f-g\|^2 + \mathrm{i}\left(\|f+\mathrm{i}g\|^2 - \|f-\mathrm{i}g\|^2\right)$
orthogonal: $(f, g) = 0$, *angle:* $\cos\theta = \frac{(f,g)}{\|f\|\|h\|}$ = correlation
$\|h - h'\| = \inf\{\|h - k\| : k \in K\}$
**Decomposition:* $\forall h \in H : h = h' + h''$, where h' orthogonal projection on K (complete subspace of H) and h'' orthogonal to K.
Therefore $H = K \oplus K^\perp$ (all $k \in K$ orthogonal to all $k' \in K^\perp$).
$\boldsymbol{L^p}$ (Banach spaces): E Lebesgue-finite, $p \le q \Rightarrow L^q(E) \subseteq L^p(E)$
for $1 \le p \le \infty$: $\|f\|_p = \left(\int_E |f|^p \, \mathrm{d}m\right)^{\frac{1}{p}}$; $\|f\|_\infty = \operatorname{ess\,sup} f$
Hölder's Inequality (needed for Δ eq.; generalizes Schwarz):
let $\frac{1}{p} + \frac{1}{q} = 1$, $f \in L^p$, $g \in L^q \Rightarrow fg \in L^1$ and $\|fg\|_1 \le \|f\|_p \|g\|_q$
Minkowski's Inequality: $f, g \in L^p \Rightarrow \|f+g\|_p \le \|f\|_p + \|g\|_p$
Probability: *nth moment:* $\mathbb{E}(X^n)$, *central moment:* $\mathbb{E}(X - \mu)^n$
$\mathbb{E}(X^n)$ (in)finite $\Rightarrow \mathbb{E}(X^k)$ (in)finite ($k \le n$); moments (c.m.) of order n are determined by c.m. (moments) of order k for $k \le n$
$\mathbb{E}(X^k) = \frac{1}{i^k} \frac{\mathrm{d}^k}{\mathrm{d}t^k} \varphi_X(0)$ (if kth moment finite, φ_X is $k\times$ diff'able)
independent: $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$, f, g bounded $\mathcal{B}$-m
$\mathbb{E}(XY) = (X, Y)$, it follows: $\mathrm{Cov}(X, Y) = (X_c, Y_c) =$
$\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$, where $X_c = X - \mathbb{E}(X)$ centered r.v.
$\mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$; $\varphi_{X_1 + \ldots + X_n}(t) = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t)$
Conditional Expectation: let $(\Omega, \mathcal{F}, P)$ where $\mathcal{G}$ is sub-σ-field of $\mathcal{F}$, if $X \in \mathcal{L}^1(\mathcal{F})$ $\exists Y \in \mathcal{L}^1(\mathcal{G})$ s.t. $\int_G Y \, \mathrm{d}P = \int_G X \, \mathrm{d}P \quad \forall G \in \mathcal{G}$
$Y = \mathbb{E}(X|\mathcal{G})$ is uniquely defined up to P-null sets
construct Y when $X \in \mathcal{L}^2$: Y as orthogonal projection to
$K = L^2(\mathcal{G})$ (see *) $\Rightarrow (X - Y, \mathbb{1}_G) = 0$ as $\mathbb{1}_G \in L^2(\mathcal{G}) \; \forall G \in \mathcal{G}$
$\to$ can be seen as minimising distance between X and $\mathcal{L}^2(\mathcal{G})$
$\to$ Y is 'best predictor' of X in $\mathcal{L}^2(\mathcal{G})$

Product Measures

$\mathbb{R}^n$: cubes $\boldsymbol{I} = I_1 \times ... \times I_n$ with length $l(\boldsymbol{I}) = l(I_1) \times ... \times l(I_n)$, where hyperplanes are null sets

Product $\boldsymbol{\sigma}$-fields: from measure spaces $(\Omega_i, \mathcal{F}_i, P_i)$: $(\Omega, \mathcal{F}, P)$ with $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{F}$ generated by rectangles $\mathcal{R} = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$, write $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$

Measure: $P{=}P_1{\times}P_2$, with σ-finite P_i, i.e. $\exists (A_n) : \bigcup_{n=1}^{\infty} A_n{=}\Omega_i$, with $P_i(A_n)$ finite; $P(A) = \int_{\Omega_2} P_1(A_{\omega_2})\,\mathrm{d}P_2(\omega_2)$ with *section* $A_{\omega_2} = \{\omega_1 \in \Omega_1 : (\omega_1, \omega_2) \in A\} \subset \Omega_1$ and measurable function $\omega_2 \to P_1(A_{\omega_2}) = \mathbb{1}_{\omega_2 \in A} P(A_1)$

fulfills $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$ and P is countably additive

Fubini's Theorem: integrate $f \in L^1(\Omega_1 \times \Omega_2)$ over sections $\omega_1 \to \int_{\Omega_2} f(\omega_1, \omega_2)\mathrm{d}P_2(\omega_2)$ (ω_2 analogous)

$$\int_{\Omega_1 \times \Omega_2} f\,\mathrm{d}(P_1 \times P_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f\,\mathrm{d}P_2 \right) \mathrm{d}P_1$$

Probability random vector $(X, Y) : \Omega \to \mathbb{R}$

joint density: if $\exists f_{(X,Y)} : P_{(X,Y)}(B) = \int_B f_{(X,Y)}(x, y)\,\mathrm{d}m_2(x, y)$

marginal distributions: if X, Y have a joint density, they are absolutely continuous with $f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x, y)\,\mathrm{d}y$

independence: iff $P_{(X,Y)} = P_X \times P_Y$ or if $\exists$ joint density: $f_{(X,Y)}(x, y) = f_X(x) \times f_Y(y)$

$X{+}Y$: $f_{X+Y}(z) = \int_{\mathbb{R}} f_{X,Y}(x, z{-}x)\,\mathrm{d}x \overset{\perp\!\!\!\perp}{=} \int_{\mathbb{R}} f_X(x) f_Y(z{-}x)\,\mathrm{d}x$

characteristic function: if F_X continuous at $a, b \in \mathbb{R}$: $F_X(b) - F_X(a) = \lim_{c \to \infty} \frac{1}{2\pi} \int_{-c}^{c} \frac{\mathrm{e}^{-\mathrm{i}ua} - \mathrm{e}^{\mathrm{i}ub}}{\mathrm{i}u} \varphi_X(u)\,\mathrm{d}u$;

if φ_X integrable: $f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}ux} \varphi_X(u)\,\mathrm{d}u$

Radon-Nikodym

$(\Omega, \mathcal{F})$ is a *measurable space*

absolute continuity ($\boldsymbol{\nu \ll \mu}$): if $\mu(A){=}0 \Rightarrow \nu(A){=}0\ \forall A{\in}\mathcal{F}$, $\nu \ll \mu$ iff: $\forall \epsilon > 0\ \exists\, \delta > 0$ s.t. $\mu(F) < \delta \Rightarrow \nu(F) < \epsilon\ \forall F{\in}\mathcal{F}$

Radon-Nikodym: ν, μ σ-finite, $\nu \ll \mu$: $\exists$ m'able function $h \geq 0$ $h : \Omega \to \mathbb{R}$ s.t. $\nu(F){=}\int_F h\,\mathrm{d}\mu\ \ \forall F{\in}\mathcal{F}$, h is unique up to μ-null sets

Radon-Nikodym derivative: $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} = h$ with $\lambda \ll \mu$, $\nu \ll \mu$ implying $\phi = \lambda{+}\nu \Rightarrow \frac{\mathrm{d}\phi}{\mathrm{d}\mu} = \frac{\mathrm{d}\lambda}{\mathrm{d}\mu} + \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ a.s. and $\lambda \ll \nu \Rightarrow \frac{\mathrm{d}\lambda}{\mathrm{d}\mu} = \frac{\mathrm{d}\lambda}{\mathrm{d}\nu}\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ a.s.

mutual singularity ($\boldsymbol{\mu \perp \nu}$): concentrated on disjoint $A_i{\subset}\Omega$

λ concentrated on $E{\in}\mathcal{F}$: $\lambda(F) = \lambda(E \cap F)\ \forall F{\in}\mathcal{F}$

Lebesgue decomposition: $\lambda{=}\lambda_a{+}\lambda_s$, where $\lambda_a \ll \mu$ and $\lambda_s{\perp}\mu$ $\lambda(E) = \int_E h\,\mathrm{d}\mu + \lambda_s(E)$ "linear combination": μ acting as "basis"

Lebesgue-Stieltjes measure: distribution function F has only jump discontinuities & is right continuous, i.e. $F(x^+) = F(x)$

F-outer measure m_F^*: m^* using l_F on intervals, restricted to $\mathcal{M}_F$: complete, contains $\mathcal{B}$ but $\neq \mathcal{M}$, if $F(x) = x$ then $m_F = m$

absolute continuity of real function $\boldsymbol{F}$ on $[a, b]$ if $\forall \epsilon{>}0\ \exists\, \delta{>}0$ s.t. $\forall$ sets of n disjoint $J_k = (x_k, y_k)$ in $[a, b]$ with $\sum_{k=1}^n (y_k - x_k) < \delta$ follows $\sum_{k=1}^n |F(x_k) - F(y_k)| < \epsilon$;

F monotone increasing & absolutely continuous $\Rightarrow$ every Lebesgue-measurable set is m_F-measurable and $m_F \ll m$

bounded variation ($\boldsymbol{F \in BV[a, b]}$) if $T_F[a, b] < \infty$ where $T_F[a, x] = \sup\{\sum_{k=1}^n |F(x_k) - F(x_{k-1})|\} \Leftrightarrow$ is the difference of two monotone increasing real functions on $[a, b]$;

absolutely continuous functions are $\in BV[a, b]$ as well as the components of the minimal decomposition $F = F_1 - F_2$,

Lebesgue-Stieltjes *signed measure of F* $m_F = m_{F_1} - m_{F_2}$ on $\mathcal{B}$

signed measure: $\nu : \mathcal{F} \to (-\infty, +\infty]$ with $\nu(\emptyset) = 0$ and countably additive, ν monotone increasing $\Rightarrow$ is measure;

total variation $|\nu|$ of a bounded signed measure is a measure, $\nu^{+(-)} = \frac{1}{2}(|\nu|{+}({-})\nu)$; with μ measure: unique $\nu = \nu_a + \nu_s$ with $\nu_a \ll \mu$ & $\nu_s \perp\!\!\!\perp \mu$ and $\nu_a(F) = \int_F h\,\mathrm{d}\mu\ \forall F{\in}\mathcal{F}$ with $h \in \mathcal{L}^1(\mu)$;

ν bounded signed measure and $F(x) = \nu((-\infty, x])$ then: F diff'able with $F'(a) = L \Leftrightarrow \forall \epsilon{>}0\ \exists\, \delta{>}0$ s.t. $|\frac{\nu(J)}{\mu(J)} - L| < \epsilon$ if open interval J around a and $l(J) < \epsilon$

Fundamental Theorem of Calculus: F absolutely continuous on $[a, b]$ then F is diff'able m-a.e., $\frac{\mathrm{d}m_F}{\mathrm{d}m} = F'$, and $F(x) - F(a) = m_F[a, x] = \int_a^x F'(t)\,\mathrm{d}t\ \ \forall x \in [a, b]$

Hahn-Jordan decomposition: ν bounded signed measure then disjoint $A \cup B = \Omega$ with $\nu^{+(-)}(F) = \nu(B(A) \cap F)\ \forall F{\in}\mathcal{F}$, if $\nu = \lambda_1 - \lambda_2$ then $\lambda_1 \geq \nu^+$, $\lambda_2 \geq \nu^-$

integrals of signed measures: $\int_F f\,\mathrm{d}\mu = \int_F f\,\mathrm{d}\mu^+ - \int_F f\,\mathrm{d}\mu^-$ (whenever this is not $\pm(\infty - \infty)$)

Probability Radon-Nikodym means $\exists$ cond. exp. for $X \in \mathcal{L}^1(P)$ $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$; X $\mathcal{G}$-measurable $\Rightarrow \mathbb{E}(X|\mathcal{G}) = X$; $X \perp\!\!\!\perp \mathcal{G} \Rightarrow \mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$; Linearity: $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$; Positivity: $X \geq 0 \Rightarrow \mathbb{E}(X|\mathcal{G}) \geq 0$; 'monotone convergence': $\{X_n\} \geq 0$ increase a.s. to $X \Rightarrow \{\mathbb{E}(X_n|\mathcal{G})\}$ increase a.s. to $\mathbb{E}(X|\mathcal{G})$; Y $\mathcal{G}$-m'able, XY integrable $\Rightarrow \mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$ (known factor); tower property: $\mathcal{H}{\subset}\mathcal{G} \Rightarrow \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$

Martingales: past knowledge $\mathcal{F}_k{=}\sigma\{X_i : i{\leq}k\}$ generated by sequence (aka *stochastic process*) (X_n), $\mathcal{F}_0$ typically $= \{\emptyset, \Omega\}$

Filtration: increasing sequence of sub-σ-fields $(\mathcal{F}_n) := \mathbb{F}$, i.e. $\mathcal{F}_0 \subset \mathcal{F}_1 ... \subset \mathcal{F}$. (X_n) is *adapted* to F if X_n is $\mathcal{F}_n$-m'able.

(X_n) is *martingale* in *filtered prob. space* $(\Omega, \mathcal{F}, (\mathcal{F}_n), P)$ w.r.t. $\mathbb{F}$ if • (X_n) adapted to $\mathbb{F}$ • $X_n{\in}\mathcal{L}^1(P)$ • $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$.

$\Rightarrow$ if $m{>}n$ then $\mathbb{E}(X_m|\mathcal{F}_n){=}X_n$, & constant exp.: $\mathbb{E}(X_n){=}\mathbb{E}(X_0)$.

a martingale is only predictable if a. s. constant, processes can be decomposed in a predictable and a martingale component

discrete stochastic integral $c \cdot X$ $I_n(\omega) = \sum_{k=1}^n c_k(\omega)(\Delta X_k(\omega))$, with c bounded predictable process & X mart. $\Rightarrow I_n$ martingale (interpretation: c could be the money that you bet in game X)

stopping time: r.v. $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$

stopped process $X_n^\tau(\omega) = X_{\min(n, \tau(\omega))}(\omega)$ (again a martingale)

Limit Theorems

focus on probability

convergence $f_n{\to}f$: (1)$\Rightarrow$(2)$\Rightarrow$(3); finite measures: (1)$\Rightarrow$(4)

(1) *uniformly on E:* $\forall \epsilon{>}0\ \exists N{=}N(\epsilon)$ s.t. $\forall n \geq N$
$\|f_n - f\|_\infty = \sup_{x \in E}(|f_n(x) - f(x)|) < \epsilon$

(2) *pointwise on E:* $\forall x \in E : \forall \epsilon{>}0\ \exists N{=}N(\epsilon, x)$ s.t. $\forall n \geq N$
$|f_n(x) - f(x)| < \epsilon$

(3) *a. e. on E:* $\exists F \subset E$ null s.t. $f_n \to f$ pointwise on $E \setminus F$

(4) *in L^p-norm (p^{th} mean):* $\|f_n - f\|_p \to 0$ as $n \to \infty$, i.e.
$\forall \epsilon{>}0\ \exists N{=}N(\epsilon)$ s.t. $\forall n \geq N : \left(\int_E |f_n - f|^p dm\right)^{1/p} < \epsilon$

(5) ***convergence in probability:*** sequence $(X_n) \to X$ if $\forall \epsilon{>}0$
$P(|X_n - X| > \epsilon) \overset{n\to\infty}{\to} 0$; (3)$\Rightarrow$(5), (4)$\Rightarrow$(5)

Chebyshev's Ineq. r.v. $Y{\geq}0$, $\epsilon{>}0, 0{<}p{<}\infty \Rightarrow P(Y{\geq}\epsilon) \leq \frac{\mathbb{E}(Y^p)}{\epsilon^p}$ it follows r.v. X with $\mathbb{E}(X) = m$, variance σ^2 and $0 < a < \infty$ then $P(|X - m| \geq a\sigma) \leq a^{-2}$

Weak Law of Large Numbers X_i independent, $\mathbb{E}(X_i) = m$, $\mathrm{Var}(X_i) \leq K < \infty \Rightarrow \frac{S_n}{n} \to m$ in probability

Borel-Cantelli Lemmas $\overset{\text{(lower: } \cup\cap)}{\text{upper limit:}}$ $\limsup\limits_{n\to\infty} A_n = \bigcap\limits_{n=1}^{\infty} \bigcup\limits_{m=n}^{\infty} A_m$

(a) $\sum_{n=1}^\infty P(A_n) < \infty \Rightarrow P(\limsup_{n\to\infty} A_n) = 0$

(b) $\sum_{n=1}^\infty P(A_n) = \infty \Rightarrow P(\limsup_{n\to\infty} A_n) = 1$ (for A_n indep.)

it follows: $X_n \to X$ in probability $\Rightarrow$ subsequence $\to X$ a. s.

Strong Law of Large Numbers $S_n = X_1 + ... + X_n$; versions:

•X_n indep., $\mathbb{E}(X_n){=}m$, $\mathbb{E}(X_n^4){<}K \Rightarrow \frac{S_n}{n}{=}\frac{1}{n}\sum_{k=1}^n X_k \to m$ a. s.

•X_n indep., $\mathbb{E}(X_n){=}0$, $\sum_{n=1}^\infty \frac{1}{n^2}\mathrm{Var}(X_n) < \infty \Rightarrow \frac{S_n}{n} \to 0$ a. s.

•X_n i.i.d., $\mathbb{E}(X_1){=}m < \infty \Rightarrow \frac{S_n}{n} \to m$ a. s. (final form of the law)

(6) ***weak convergence*** "weak" as implied by weakest (5)$\Rightarrow$(6):
sequence P_n of Borel prob. measures on $\mathbb{R}^n \to P$ weakly $\Leftrightarrow$
their CDFs $F_n \to F$ (CDF of P) wherever F continuous

Skorokhod Representation Theorem $P_n{\to}P$ weakly$\Rightarrow \exists$ r.v. X_n, X on $\big([0, 1], \mathcal{B}, m_{[0,1]}\big)$ s.t. $P_n{=}P_{X_n}$, $P{=}P_X$ & $X_n{\to}X$ a. s.

Prokhorov's Theorem $P_n{\in}\mathbb{R}^d$ tight $\exists k_n$ s.t. $P_{k_n} \to P$ weakly with *tight*: $\forall \epsilon{>}0\ \exists M$ s.t. $P_n(\mathbb{R}_d \backslash [-M, M]) < \epsilon\ \forall n$ ("light tails")

Central Limit Theorem normalized r.v. $T_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}}$

X_n indep., $\mathbb{E}(X_n){=}m_n{<}\infty$ & $\mathrm{Var}(X_n){<}\infty$ then $P_{T_n} \overset{weak}{\to} \mathrm{N}(0, 1)$ if $\frac{1}{\mathrm{Var}(S_n)} \sum_{k=1}^n \int_{\{x:|x - m_k| \geq \epsilon\sqrt{\mathrm{Var}(S_n)}\}} (x - m_k)^2 \mathrm{d}P_{X_k} \overset{n\to\infty}{\to} 0$

1.3 Functional Analysis (Kreyszig, 1991)

Metric Spaces (additional notes on) (X, d)

discrete metric $d(x,x) = 0$ and $d(x,y) = 1$ for $x \neq y$
topology (X, d) is a topological space $(X, \mathscr{T})$, satisfying
$\emptyset \in \mathscr{T}, X \in \mathscr{T}$; unions $\in \mathscr{T}$; finite intersections $\in \mathscr{T}$
separability $M \subset X$ is *dense in* X if completion $\bar{M} = X$; X is *separable* if $\exists$ countable subset dense in X (e.g. $\mathbb{R}$, $\mathbb{C}$, l^p for $p{<}\infty$)
completeness X is complete if every Cauchy sequence converges
completion for every $X = (X, d)$ $\exists$ complete $\hat{X} = (\hat{X}, \hat{d})$ with subspace W dense in $\hat{X}$ and isometric with X (unique up to isometries), where $T : X \to \tilde{X}$ is *isometric* if $d(x, y) = \tilde{d}(Tx, Ty)$ and X is *isometric* with $\tilde{X}$ if $\exists$ bijective isometry $X \to \tilde{X}$

Normed Spaces vector space with metric defined by norm

Vector Space over field K: nonempty set of *vectors* $x, y...$ with operations *vector addition* (abelian group) and *scalar* (elements of K) *vector multiplication* $(\alpha(\beta x){=}(\alpha\beta)x)$; $1x{=}x$, distributive)
linear independence $a_1x_1 + ... + a_mx_m = 0$ only for $a_i = 0$
dimensionality max number of vectors in independent set, for $\dim X = n$, a linearly independent n-tuple is called a basis
Normed Space vector space + norm $\|x\|$ (PD, homogeneity, Δ)
if $\exists(e_n)$ in X s.t. $\forall x$ $\|x - (\alpha_1e_1 + ... + \alpha_ne_n)\| \overset{n\to\infty}{\to} 0$ then e_n is called *Schauder basis* with $x{=}\sum_{k=1}^{\infty} \alpha_k e_k$; $\exists(e_n) \Rightarrow X$ separable
Banach Space complete (in metric $\|x - y\|$) normed space
a metric induced norm is translation invariant
subspace of Banach space Y is complete iff it is closed in Y
completion normed space X to Banach space $\hat{X}$ with isometry A s.t. $A(X)$ dense in $\hat{X}$ (unique up to isometries)
Finite Dimensional Normed Space every finite dimensional subspace B of a normed space A is complete and closed in A;
in a finite dimensional vector space X:
- all norms are equivalent: $\exists a, b > 0$ s.t $a\|x\|_* \leq \|x\|. \leq b\|x\|_*$
- any $M \subset X$ is compact iff M closed and bounded

Riesz's Lemma $Y \subset Z \subset X$ with Y closed $\Rightarrow \forall\, \theta{\in}(0, 1)$ $\exists z{\in}Z$ s.t. $\|z\|{=}1$ & $\|z{-}y\| \geq \theta$ $\forall\, y{\in}Y$; the lemma shows: closed unit ball $M = \{x | \|x\| \leq 1\}$ compact $\Rightarrow X$ is finite dimensional
continuous mapping $T{:}X{\to}Y$ (metric spaces), $M{\subset}X$ compact $\Rightarrow T(M)$ compact & $T{:}M{\to}\mathbb{R}$ assumes maximum and minimum
Linear Operator $T : \mathscr{D}(T) \to \mathscr{R}(T)$ (vector spaces over K) & $T(x{+}y) = Tx{+}Ty, T(\alpha x) = \alpha Tx$; $\mathscr{N}(T)$ is null space s.t. $Tx{=}0$
• $\mathscr{R}(T), \mathscr{N}(T)$ vector spaces • $\dim \mathscr{D}(T){=}n{<}\infty \Rightarrow \dim \mathscr{R}(T) \leq n$
inverse: linear operator $\Leftarrow T^{-1}$ exits $\Leftrightarrow Tx{=}0$ only at $x{=}0$
$X \overset{T}{\to} Y \overset{S}{\to} Z$ bijective on vector spaces: $(ST)^{-1} = T^{-1}S^{-1}$
bounded linear $T : \mathscr{D}(T) \to Y$ (X, Y normed spaces):
$\exists c \in \mathbb{R}$ s.t. $\forall x \in \mathscr{D}(T)$ $\|Tx\| \leq c\|x\| \Rightarrow T$ bounded ($\Leftrightarrow \|T\|$ exists)
operator norm $\|T\|{=}\sup_{x\in\mathscr{D}(T), x\neq 0} \frac{\|Tx\|}{\|x\|} {=} \sup_{x\in\mathscr{D}(T), \|x\|=1} \frac{\|Tx\|}{\|x\|}$
in a finite normed space X every linear operator is bounded
T continuous at a single point $\Leftrightarrow T$ continuous $\Leftrightarrow T$ bounded
$\mathscr{N}(T)$ is closed & $x_n \to x$ implies $Tx_n \to Tx$
extension $\tilde{T} : \overline{\mathscr{D}(T)} \to Y$ bounded with norm $\|\tilde{T}\| = \|T\|$
matrix representation for $\dim X{<}\infty$: linear operator = matrix
Functional operator $f : \mathscr{D}(f) \to K$ with $\mathscr{D}(f){\subset}X$ & $K{=}\mathbb{R}/\mathbb{C}$
• *linear:* X vector space over K • *bounded/continuous:* see above
algebraic dual spaces for X vector space: algebraic dual space X^* (values $f(x)$) & second algebraic dual space X^{**} (values $g(f)$)
canonical mapping $X{\to}X^{**}$ is $g(f){=}g_x(f){=}f(x)$ with x fixed
X is ***isomorphic*** to a subspace of X^{**} (*embeddable*), meaning a bijective function exists which preserves the structure (e.g. norm)
dual basis $\dim X{=}n \leq \infty$ with basis $E{=}\{e_1, ..., e_n\} \Rightarrow$ Kronecker delta $\delta_{jk}{=}f_k(e_j){=}\mathbb{1}_{j=k}(e_j)$ is basis for X^* and $\dim X^*{=}n$

Dual Space $\boldsymbol{X'}$: all bounded linear functionals on normed space X
vector space $\boldsymbol{B(X, Y)}$: all bounded linear operators $X \to Y$ (normed spaces) $\to$ normed space with the operator norm;
Y Banach space $\Rightarrow B(X, Y)$ Banach space
isomorphisms: $(\mathbb{R}^n)'{=}\mathbb{R}^n$; $(l^1)'{=}l^\infty$; $(l^p)'{=}l^q$ with $\frac{1}{q}{+}\frac{1}{p}{=}1$

Inner Product Spaces $\langle , \rangle$ + complete = Hilbert

inner product (PD, sesquilinear, $\overline{\text{symmetric}}$) induces norm
$\|x\| = \sqrt{\langle x, x\rangle}$ & metric $\|x - y\|$; *orthogonality* $\perp$: $\langle x, y\rangle = 0$
continuity: $x_n \to x$ & $y_n \to y$ $\Rightarrow$ $\langle x_n, y_n\rangle \to \langle x, y\rangle$
completion: inner product space X to Hilbert space H (unique up to isomorphisms) with isomorphism $A : X \to W$ (dense $W{\subset}H$)
for $Y{\subset}H$: • Y finite $\Rightarrow Y$ complete $\Leftrightarrow Y$ closed in H
• H separable $\Rightarrow Y$ separable (even for incomplete H)
Orthogonal Complement $H{=}Y{\oplus}Y^\perp$ with Y closed subspace
segment all z s.t. $x, y \in X$ (vector space): $\alpha x{+}(1{-}\alpha)y = z{\in}X$ $(0{\leq}\alpha{\leq}1)$; if all $z \in X \forall x, y \in M \Rightarrow M{\subset}X$ is *convex*
minimizing vector: convex, complete M subset X inner product space: $\forall\, x{\in}X$ $\exists$ unique $y{\in}M$ s.t. $\delta = \inf_{\tilde{y}\in M} \|x - \tilde{y}\| = \|x - y\|$
if complete M subspace $Y \Rightarrow z{=}x{-}y$ orthogonal to Y
direct sum: $X{=}Y{\oplus}Z$ meaning $x{=}y{+}z$ $\forall x \in X$; $M{\subset}M^{\perp\perp}$
orthogonal projection $P : H \to Y$*:* orthogonal complement $Y^\perp$ of closed suspace $Y{\subset}H$ is $\mathscr{N}(P)$; $Y^\perp$ closed vector space; $Y{=}Y^{\perp\perp}$
dense set: for any $M \subset H$, span(M) dense in H iff $M^\perp = 0$
Orthonormal Sets $M \subset X$ (inner product space) with pairwise orthogonal elements, called orthonormal if normed to 1, e.g. unit vectors in $\mathbb{R}^n$, sequence (e_n) in l^2 where nth element 1 others 0
for orthonormal (e_k), inner product space X: $x = \sum_{k=1}^{n}\langle x, e_k\rangle e_k$
Bessel inequality: $\sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 \leq \|x\|^2$; $\langle x, e_k\rangle$ Fourier coef.
Fourier Series *periodic* $\exists\, p{\geq}0$ s.t. $f(t{+}p) = f(t)$
trigonometric $a_0{+}\sum_{k=1}^{\infty}(a_k \cos kt{+}b_k \sin kt)$ with *Euler functions*
•$a_0{=}\frac{1}{2\pi}\int_0^{2\pi} x(t)dt$ •$a_k{=}\frac{1}{\pi}\int_0^{2\pi} x(t) \cos kt\, dt$ •$b_k{=}\frac{1}{\pi}\int_0^{2\pi} x(t) \sin kt\, dt$
$x{=}\langle x, e_0\rangle e_0{+}\sum_{k=1}^{\infty}[\langle x, e_k\rangle e_k{+}\langle x, \tilde{e}_k\rangle \tilde{e}_k]$ w/ $e_j{=}\frac{\cos kt}{\|\cos kt\|}$ ($\tilde{e_j}$: sin)
for (e_k) orthonormal in H: $\sum_{k=1}^{\infty} |\alpha_k|$ converges $\Leftrightarrow \sum_{k=1}^{\infty} \alpha_k e_k$ converges $\Rightarrow x{=}\sum_{k=1}^{n}\langle x, e_k\rangle e_k$ (this converges in any H)
inner product space X: at most countable many nonzero $\langle x, e_k\rangle$
Total Orthonormal Sets *total:* span is dense in X
total orthonormal set $M{\subset}H$ exists $\forall H$, all these sets have the same cardinality: *Hilbert dimension* ($\dim H{=}\dim \tilde{H} \Leftrightarrow H{=}\tilde{H}$)
M is total iff Bessel equality (=*Parseval relation*) holds
- H contains total orthonormal sequence $\Rightarrow H$ separable
- H separable $\Rightarrow$ every orthonormal set is countable

Riesz's Theorem every bounded linear functional f on H can be written as $f(x) = \langle x, z\rangle$ with unique z and $\|z\| = \|f\|$
sesquilinear form $h : X \times Y \to K = \mathbb{R}/\mathbb{C}$ (X, Y vector spaces) linear in the first and conjugate linear in the second argument
bounded if $\forall x, y\, \exists\, c{\in}\mathbb{R}$ s.t. $|h(x, y)| \leq c\|x\|\|y\|$ (normed X,Y)
with norm $\|h\|{=}\sup_{x\in X\setminus 0, y\in Y\setminus 0} \frac{|h(x,y)|)}{\|x\|\|y\|}$
Riesz representation $h{:}H_1{\times}H_2{\to}K$ bounded sesquilinear then $h(x,y){=}\langle Sx, y\rangle$ with unique $S{:}H_1{\to}H_2$ bounded linear, $\|S\|{=}\|h\|$
Hilbert-Adjoint Operator T^*: $\forall$ bounded linear $T{:}H_1{\to}H_2$ $\exists$ unique $T^*{:}H_2{\to}H_1$ s.t. $\langle Tx, y\rangle{=}\langle x, T^*y\rangle$; $\|T^*\|{=}\|T\|$
for bounded linear $S, T{:}H_1{\to}H_2$: • $\langle T^*y, x\rangle = \langle y, Tx\rangle$
• $(\alpha T)^* = \bar{\alpha}T^*$ • $(S{+}T)^* = S^*{+}T^*$ • $(ST)^*{=}T^*S^*$ (if $H_1{=}H_2$)
• $(T^*)^* = T$ • $T^*T{=}0 \Leftrightarrow T{=}0$ • $\|T^*T\| = \|TT^*\| = \|T\|^2$
for bounded linear $Q : X{\to}Y$: $Q{=}0 \Leftrightarrow \langle Qx, y\rangle{=}0$ $\forall x, y$
a bounded linear operator $T{:}H{\to}H$ is ***self-ajoint*** if $T^*{=}T$, ***unitary*** if $T^*{=}T^{-1}$ (T bijective), & ***normal*** if $TT^*{=}T^*T$

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 asspumptions (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 $\mathrm{E}\left[Y^{a=1}\right] \neq \mathrm{E}\left[Y^{a=0}\right]$

$$\begin{aligned}\mathrm{E}\left[Y^a\right] &= \sum_y y p_{Y^a}(y) && \text{(discrete)} \\ &= \int y f_{Y^a}(y) dy && \text{(continuous)}\end{aligned}$$

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: $\Pr\left[Y^{a=1}=1\right] - \Pr\left[Y^{a=0}=1\right]$
risk ratio: $\frac{\Pr\left[Y^{a=1}=1\right]}{\Pr\left[Y^{a=0}=1\right]}$
odds ratio: $\frac{\Pr\left[Y^{a=1}=1\right]/\Pr\left[Y^{a=1}=0\right]}{\Pr\left[Y^{a=0}=1\right]/\Pr\left[Y^{a=0}=0\right]}$

number needed to treat (NNT) to save 1 life: $-1/$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 $E\left[Y^{a=1}\right]$ and $E\left[Y^{a=0}\right]$ (whole population)

Target Trial emulating an ideal randomized experiment

explicitly formulate target trial & show how it is emulated → less vague causal question, helps spot issues
missing data problem unknown counterfactuals
randomized experiments: missing completely at random → exchangeability (= exogeneity as treatment is exogenous)
ideal randomized experiment: no censoring, double-blind, well-defined treatment, & adherence → 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 $\perp\!\!\!\perp$ time, outcome $\perp\!\!\!\perp$ 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:* $Y^a \perp\!\!\!\perp A \mathrel{\widehat{=}}$ randomized experiment, counterfactuals are missing completely at random (MCAR)
- *conditional:* $Y^a \perp\!\!\!\perp A|L \mathrel{\widehat{=}}$ conditionally randomized, counterfactuals are missing at random (MAR)

alternative definition: $\Pr\left[A=1|Y^{a=0}, L\right] = \Pr[A=1|L]$
additional conditions:
correct measurement mismeasurement of A, Y, L results in bias
correct model specification models $\overset{\text{may}}{\to}$ 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

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

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
*n*on*p*arametric *s*tructural *e*quation *m*odel

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

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 ϵ_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
*s*ingle *w*orld *i*ntervention *g*raphs
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* differantial
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 α-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 Variabilty
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: $\bar{A}_k = (A_0, A_1, ..., A_k)$
shorthand: always treated $\bar{A} = \bar{1}$, never treated $\bar{A} = (\bar{0})$
static strategy: $g = [g_0(\bar{a}_{-1}), ..., g_K(\bar{a}_{K-1})]$
dynamic strategy: $g = \left[g_0(\bar{l}_0), ..., g_K(\bar{l}_K)\right]$
stochastic strategy: non-deterministic g
optimal strategy is where $\mathrm{E}\left[Y^g\right]$ is maximized (if high is good)

Sequential Identifiability

sequential versions of
exchangability: $Y^g \perp\!\!\!\perp A_k|\bar{A}_{k-1} \;\; \forall g, k = 0, 1, ..., K$
conditional exchangeability:

$$\left(Y^g, L^g_{k+1}\right) \perp\!\!\!\perp A_k|\bar{A}_{k-1}{=}g\left(\bar{L}_k\right), \bar{L}^k \;\; \forall g, k = 0, 1, ..., K$$

positivity: $f_{\bar{A}_{k-1}, \bar{L}_k}(\bar{a}_{k-1}, \bar{l}_k) \neq 0 \;\Rightarrow$

$$f_{A_k|\bar{A}_{k-1}, \bar{L}_k}(a_k|\bar{a}_{k-1}, \bar{l}_k) > 0 \; \forall \left(\bar{a}_{k-1}, \bar{l}_k\right)$$

consistency:

$$Y^{\bar{a}} = Y^{\bar{a}^*} \text{ if } \bar{a} = \bar{a}^*; \qquad Y^{\bar{a}} = Y \text{ if } \bar{A} = \bar{a};$$

$$\bar{L}^{\bar{a}}_k = \bar{L}^{\bar{a}^*}_k \text{ if } \bar{a}_{k-1} = \bar{a}^*_{k-1}; \qquad \bar{L}^{\bar{a}}_k = \bar{L}_k \text{ if } \bar{A}_{k-1} = \bar{a}_{k-1}$$

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

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

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

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 $\Rightarrow$ $Y^g{=}Y \; \forall g \;\;\; \Rightarrow \;\;\; Y \perp\!\!\!\perp A_0|L_0 \;\; \& \;\; Y \perp\!\!\!\perp A_1|A_0, L_0, L_1$; therefore:
if last two independences don't hold, one assumption is violated
g-null theorem: $\mathrm{E}\left[Y^g\right] = \mathrm{E}\left[Y\right]$, if the two independences hold
($\Rightarrow$ sharp null: only if strong faithfulness (no effect cancelling))

Causal Mediation

(Hernán and Robins, 2023)

$$\begin{tikzcd} A \arrow[r] \arrow[rr, bend left] & M \arrow[r] & Y \end{tikzcd}$$

seen as longitudinal with k_0: A and k_1: M
decompose $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ into cross-world quantities

- pure (aka natural) direct effect (upper path)

$$\mathrm{E}\left[Y^{a=1, M^{a=0}}\right] - \mathrm{E}\left[Y^{a=0, M^{a=0}}\right]$$

- total (aka natural) indirect effect (lower path)

$$\mathrm{E}\left[Y^{a=1, M^{a=1}}\right] - \mathrm{E}\left[Y^{a=1, M^{a=0}}\right]$$

mediation formula under NPSEM-IE (requires $Y^{a=1,m} \perp\!\!\!\perp M^{a=0}$ cross-world independence)

$$\mathrm{E}\left[Y^{a=1, M^{a=0}}\right] = \sum_m \mathrm{E}\left[Y|A = 1, M = m\right] \Pr\left[M = m|A = 0\right]$$

interventional interpretation advocating NPSEM-IE assuming:

$$\begin{tikzcd} A \arrow[r, Rightarrow] \arrow[dr, Rightarrow] & N \arrow[r] & M \arrow[r] & Y \\ & O \arrow[urr] & & \end{tikzcd}$$

(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. $\mathrm{E}[Y \mid A=a]$
estimator:	function to use,	e. g. $\widehat{\mathrm{E}}[Y \mid A=a]$
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 → true value)
parsimonious model: few parameters estimate many quantities
bias-variance trade-off:
wiggliness ↑ → misspecification bias ↓, CI width ↑

Variable Selection

can induce bias if L includes:

(decendant 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 $\operatorname{logit}(Y=1 \mid Z)=\sum_i \alpha_i Z_i$, with $0<\alpha_i<1$ and $\sum \alpha_i=1$
weights α_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

$$\text{associational model } \mathrm{E}[Y \mid A]= \text{ causal model } \mathrm{E}[Y^a]$$

step 1: estimate/model $f[A \mid L]$ (and $f[A]$) → 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), $SW^A(V)=\frac{f[A \mid V]}{f[A \mid L]}$ 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

$$\mathrm{E}\left[Y^{a, c=0} \mid L\right]=\beta_0+\beta_1 a+\beta_2 a L+\beta_3 L=\mathrm{E}[Y \mid A, C=0, L]$$

faux marginal structural model as no IP weighting/$SW^A(L)=1$
for ATE only β_1, β_2 of interest, the rest are *nuisance parameters*

Propensity Score Methods

$\operatorname{Pr}[A=1 \mid L]=: \pi(L)$
$\Rightarrow A \perp\!\!\!\perp L \mid \pi(L)$ (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 (→ 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) → weak instrument)
2. ***exclusion restriction:*** Z affects Y at most through A
   (a) population level: $\mathrm{E}\left[Y^{z, a}\right]=\mathrm{E}\left[Y^{z', a}\right]$ (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) $\overset{data}{\rightarrow}$ (width 1)
natural bounds need 2a,3a (width $\operatorname{Pr}[A{=}1 \mid Z{=}0]+\operatorname{Pr}[A{=}0 \mid Z{=}1]$)
sharp bounds require 2a,3b (narrower than natural bounds)
IV estimand ATE: intention-to-treat ÷ measure of compliance (1,2b,3a,4a): ATE; (1,2b,3a,4b): ATE in compliers
binary Z: $\frac{\mathrm{E}[Y \mid Z=1]-\mathrm{E}[Y \mid Z=0]}{\mathrm{E}[A \mid Z=1]-\mathrm{E}[A \mid Z=0]}$, continuous Z: $\frac{Cov(Y, Z)}{Cov(A, Z)}$;
can be calculated as *two-stage-least-squares estimator:*
1. $\mathrm{E}[A \mid Z]$ 2. $\mathrm{E}[Y \mid Z]=\beta_0+\beta_1 \hat{\mathrm{E}}[A \mid Z]$ 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 → 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?) → create composite event
survival quantities k is a time point, T is time of event

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

modeling: some options

- ***Kaplan-Meier*** aka product limit formula (nonparametric): $\operatorname{Pr}[D_k=0]=\prod_{m=1}^k \operatorname{Pr}[D_m=0 \mid D_{m-1}=0]$
- parametric e. g. log hazards model:
  - use ***IP weigths*** SW^A in structural marginal model $\operatorname{logit} \operatorname{Pr}\left[D_{k+1}^{a, \bar{c}=\bar{0}}=0 \mid D_k^{a, \bar{c}=\bar{0}}=0\right]=\beta_{0, k}+\beta_1 a+\beta_2 a k$
  - ***standardize*** ($\prod_k 1-$) parametric hazards model $\operatorname{Pr}[D_{k+1}=1 \mid D_k=0, C_k=0, L, A]$ weighting across L

- ***structural nested cumulative failure time model (CFT):*** $\frac{\Pr[D_k^a=1|L,A]}{\Pr[D_k^{a=0}=1|L,A]} = \exp\left[\gamma_k(L,A;\psi)\right]$ (log-linear has no upper limit 1 → rare failure ↑; if ↓, 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)

time-varying two options based on g-methods as examples
standardization (plug-in estimate): risk is $\Pr\left[D_{k+1}^{\bar{a},\bar{c}=\bar{0}}=1\right] =$

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

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

3.2 G-Methods

G-Methods
*g*eneralized 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 → 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 $\mathrm{E}\left[Y^{a,c=0}\right]$
standardization $\mathrm{E}[Y|A=a] = \int \mathrm{E}[Y|L{=}l,A{=}a,C{=}0]\,dF_L[l]$
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 → use IP weights

time-varying censoring $\bar{C}$: monotonic type of missing data
standardization: $\int f(y|\bar{a},\bar{c}{=}\bar{0},\bar{l})\prod_{k=0}^{K} dF\left(l_k|\bar{a}_{k-1},c_{k-1}{=}0,\bar{l}_{k-1}\right)$
IP weighting:
$$SW^{\bar{C}} = \prod_{k=1}^{K+1}\frac{1\cdot\Pr\left(C_k=0|\bar{A}_{k-1},C_{k-1}=0\right)}{\Pr\left(C_k=0|\bar{A}_{k-1},C_{k-1}=0,\bar{L}_k\right)}$$

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

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

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 → predict counterfactuals on whole dataset → average the results (→ CI by bootstrapping)
for discrete L $\mathrm{E}[Y|A=a]$ is $\sum_l \mathrm{E}[Y|A=a,L=l]\Pr[L=l]$

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

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

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

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

estimation (Young et al., 2011; Schomaker et al., 2019)
1. model $f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right)$ and $\mathrm{E}\left[Y|\bar{A}=\bar{a},\bar{L}=\bar{l}\right]$
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)
$\mathrm{E}\left[Y_T^{\bar{a}}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[...\mathrm{E}\left[Y_T|\bar{A}_{T-1}{=}\bar{a}_{T-1},\bar{L}_T\right]...|\bar{A}_0{=}a_0,L_1\right]|L_0\right]\right]$

estimation (Schomaker et al., 2019)
1. model inside out: $Q_T{=}\mathrm{E}\left[Y_T|\bar{A}_{T-1},\bar{L}_T\right]$ to $Q_0{=}\mathrm{E}\left[Q_1|\bar{L}_0\right]$, 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 $L_0 \to A_0 \to Y_0 \to L_1 \to A_1 \to Y_1$, $\bar{a}=(a_0,a_1)$
$\mathrm{E}\left[Y_1^{\bar{a}}\right] \overset{\text{CE}}{=} \mathrm{E}\left[\mathrm{E}\left[Y_1^{\bar{a}}|A_0{=}a_0,L_0\right]\right]$
(ice) $\overset{\text{CE*}}{=} \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[Y_1|\bar{L},\bar{A}{=}\bar{a},Y_0\right]|A_0{=}a_0,L_0\right]\right]$
$\overset{\text{LIE}}{=} \mathrm{E}\left[\sum_{l_1}\mathrm{E}\left[Y_1|A_0{=}a_0,\bar{L},Y_0\right]\Pr\left[l_1|a_0,l_0,y_0\right]\right]$
$\overset{\text{LIE}}{=} \sum_{l_0}\left[\sum_{l_1}\mathrm{E}\left[Y_1|A_0{=}a_0,\bar{L},Y_0\right]\Pr\left[l_1|a_0,l_0,y_0\right]\right]\Pr[l_0]$
(jde) $\overset{\text{sum}}{=} \sum_{\bar{l}}\mathrm{E}\left[Y_1|A_0{=}a_0,\bar{L},Y_0\right]\Pr\left[l_1|a_0,l_0\right]\Pr[l_0]$
CE: conditional expectation; *: exchangeability;
LIE: law of iterated expectation

IP Weighting
*i*nverse *p*robability of treatment (g-formula)

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

unknowns can be estimated non-parametrically or modeled
pseudo-population: everyone is treated & untreated ($L \not\to A$)
FRCISTG *(fully randomized causally interpreted structured graph)*: probability tree for $L \to A \to Y$, 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!) → pseudo-population same size → (if non-saturated) CI width ↓

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

G-Estimation (additive) structural nested models

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

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

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

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

$Y^{a=0}$ unknown, but because of exchangeability α_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

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

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

multiplicative is preferred if Y always positive, but does not extend to longitudinal case
semi-parametric: agnostic about β_0 and effect of L → robust ↑
no time-varying: no nesting; model equals marginal structural models with missing β_0,β_3 (unspecified "no treatment")

sensitivity analysis: unmeasured confounding ($\alpha_1\neq 0$) can be examined: do procedure for different values of α_1 → plot α_1 vs. $\psi^\dagger$ → 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
strutural nested mean models separate effect of each a_k

$$\mathrm{E}\left[Y^{\bar{a}_{k-1},a_k,\underline{0}_{k+1}}-Y^{\bar{a}_{k-1},\underline{0}_{k+1}}|\bar{L}^{\bar{a}_{k-1}}=\bar{l}_k,\bar{A}_{k-1}=\bar{a}_{k-1}\right]=a_k\gamma_k\left(\bar{a}_{k-1},\bar{l}_k,\beta\right)$$

calculations

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

function γ_j can be, e. g. constant (ψ_1), time-varying only ($\psi_1+\psi_2 k$), or dependent on treatment/covariate history

$$\text{logit}\Pr\left[A_k=1|H_k\left(\psi^\dagger\right),\bar{L}_k,\bar{A}_{k-1}\right]=\alpha_0+\alpha_1 H_k\left(\psi^\dagger\right)+\alpha_2 w_k\left(\bar{L}_k,\bar{A}_{k-1}\right)$$

find α_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 $\frac{1}{\pi(l)}-\frac{1}{\hat{\pi}(l)}$ and $b(l)-\hat{b}(l)$ known as second order bias

time-varying: multiple robustness for $k=0,1,...K$
$K+2$ robustness: consistent, if $\hat{f}_0$ to $\hat{f}_l$ and $\hat{b}_{l+1}$ 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 → 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 ψ_{dr}, we need sample splitting, otherwise $\hat{b}(l)$ and $\hat{f}(l)$ are correlated with ψ_{dr}

if $\hat{b}(l)$ and $\hat{f}(l)$ are consistent and $\mathrm{E}[\hat{\psi}-\psi|T_r]/se(\hat{\psi})$ converges to 0 → $\hat{\psi}$ with sample splitting is asymptotically normal and unbiased → 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)*:

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

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?

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

for estimator T and distribution P_n with $0<\epsilon<1$
can also be rewritten as a ***directional derivative*** at P_n

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

in direction (δ_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)

- $\overline{IC}(P_0)=0$ and $\mathrm{Var}(IC(P_0))$ asymptotic variance of the standard estimator $\sqrt{n}(\psi_n-\psi_0)$, → $Var(\hat{\Psi}(P_n))=\frac{Var_{IC}}{n}$
- efficient IC: an estimator is asymptotically efficient ⇔ its influence curve is the efficient influence curve $IC(O)=D^*(O)$

Delta Method (Zepeda-Tello et al., 2022) estimand is a function of θ, i.e. $\psi:=\phi(\theta)$, $Var(\hat{\theta})$ known, but what is $Var(\hat{\psi})$?
Taylor's approximation requirements:

- univariate ϕ: differentiable at θ
- multivariate ϕ: $\exists\,\partial_v\phi(\theta)$ (directional derivative)
- functional ϕ (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†): $\phi(\hat{\theta}_n)\approx\phi(\theta)\stackrel{\approx}{+}\partial_{v:=\hat{\theta}-\theta}\phi(\theta)$
classical delta method: if $\{r_n\}_{n=1}^\infty$ with $\lim_{n\to\infty}r_n=\infty$,

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

$$r_n\left(\phi(\hat\theta_n) - \phi(\theta)\right) \overset{\dagger *}{\approx} \nabla\phi(\theta)^T r_n(\hat\theta_n - \theta) \overset{d}{\to} \nabla\phi(\theta)^T Z$$

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

functional delta: $r_n(\hat\theta_n - \theta) \overset{d}{\to} Z \Rightarrow r_n\left(\phi(\hat\theta_n) - \phi(\theta)\right) \overset{d}{\to} \partial_Z\phi(\theta)$

influence function: $\psi = \phi(\mathbb{P}_X)$ is a functional estimations rate of change for $\mathbb{P}_X$ to Q, where $Q = \mathbb{1}_{\{Y\}}$

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

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

use delta: $\phi(\hat{\mathbb{P}}_X) \approx \phi(\mathbb{P}_X) + \mathrm{IF}_{\phi,\mathbb{P}_X}(Y)$, if $(\hat\theta_n - \theta) \overset{n\to\infty}{\sim} \mathrm{N}(.,.)$

$$\hat\psi_n - \psi = \phi(\hat\theta_n) - \phi(\theta) \overset{\text{approx}}{\sim} \mathrm{N}\left(0, \mathrm{Var}[\mathrm{IF}_{\phi,\mathbb{P}_X}(Y)]\right),$$

where $\widehat{\mathrm{Var}}[\mathrm{IF}_{\phi,\mathbb{P}_X}(Y)] = \frac{1}{n}\sum_{i=1}^n \left(\mathrm{IF}_{\phi,\mathbb{P}_X}(X_i)\right)^2$, 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 ϕ 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 $\hat f\left(A_m|\bar A_{m-1}, \bar L_m\right)$ (e. g. logistic model), use it to calculate at each time m: $\widehat{W}^{\bar A_m} = \prod_{k=0}^m \frac{1}{\hat f(A_k|\bar A_{k-1}, \bar L_k)}$ and modified IP weights at m: $\widehat{W}^{\bar A_{m-1}, a_m} = \frac{\widehat{W}^{\bar A_{m-1}}}{\hat f(a_m|\bar A_{m-1}, \bar L_m)}$
2. with $\widehat{T}_{K+1} := Y$, recursively for $m = K, K-1, ..., 0$:
   (a) fit outcome regression on $\widehat{T}_{m+1}$ with variable $\widehat{W}^{\bar A_m}$
   (b) calculate $\widehat{T}_m$ using the outcome model with $\widehat{W}^{\bar A_{m-1}, a_m}$
3. calculate standardized mean outcome $\widehat{\mathrm{E}}\left[Y^{\bar a}\right] = \mathrm{E}\left[\widehat{T}_0\right]$

Augmented IPTW (Hernán and Robins, 2023)

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

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

Relationship between AIPTW and TMLE for causal effect:

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

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

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

TMLE (van der Laan and Rubin, 2006; van der Laan et al., 2011)

*t*argeted *m*aximum *l*ikelihood *e*stimation: 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:* $\bar Q_0 = \mathrm{E}_0(Y|A,W)$

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 $\bar Q_0$ → 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 ϵ_n with offset logistic regression: $\operatorname{logit}\bar Q_n^1(A,W) = \operatorname{logit}\bar Q_n^0(A,W) + \epsilon_n H_n^*(A,W)$ (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.*: $\psi_n^{TMLE} = \frac{1}{n}\sum_{i=1}^n\left[\bar Q_n^1(1,W_i) - \bar Q_n^1(0,W_i)\right]$

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), substition estimator (respects global constraints → more robust to outliers and sparsity) → good finite sample performance

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

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

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

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 → the first part of a is the clever covariate)

sample variance is then: $S^2(IC_n) = \frac{1}{n}\sum_{i=1}^n\left(IC_n(o_i) - \bar{IC}_n\right)^2$

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) *l*ongitudinal TMLE: based on ice g-formula for $t = T, ..., 1$:

1. model $\widehat{\mathrm{E}}(Y_t|\bar A_{t-1}, \bar L_t)$ (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 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): $\hat H(\bar A, \bar L)_{t-1} = \prod_{s=0}^{t-1}\frac{\mathbb{1}(\bar A_s = \bar d_s)}{\widehat{\Pr}(A_s = d_s|\bar A_{s-1} = \bar d_{s-1}, \bar L_s = \bar l_s)}$
4. $\hat\psi_T$ = 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)

***t**argeted **m**inimum **l**oss-based **e**stimation*

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

1. compute Ψ's pathwise derivative at P and its corresponding canonical gradient $D^*(P)$ (efficient influence curve)
2. define a loss $L()$ s. t. $P \to E_0 L(P)$ 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 ψ_0 by plugging P_n^* into Ψ (substitution)
6. TMLE solves the efficient influence curve equation $\sum_{i=1}^n D^*(P_n^*)(O_i) = 0$ → 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 $A_t^{\mathbb{d}} = \mathbb{d}\left(A_t(\bar{A}_{t-1}^{\mathbb{d}}), H_t(\bar{A}_{t-1}^{\mathbb{d}})\right)$, e. g. threshold ($\max(c, a_t)$), shift ($a_t + \delta$ if positivity else a_t), stochastic (draw from $F(\mathbb{d}(A_t, H_t)|H_t)$; 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 $\mathrm{supp}\{A_t, H_t\}$ then $(\mathbb{d}(a_t, h_t), h_t)$ 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 $t = \tau, ..., 1$:

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

solve $\theta = \mathrm{E}\left[m_1(A_1^{\mathbb{d}}, L_1)\right]$

optimality *limitations:* threshold LMTPs can't be $n^{1/2}$ consistent as parameter not pathwise differentiable, continuous A can only be considered, if $\mathbb{d}(\cdot, h_t)$ *piecewise smooth invertible*

efficient influence curve (assumes $\mathbb{d} \perp\!\!\!\perp$ P):

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

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

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

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

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

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

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

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

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

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

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

- ***SDR:*** 2^{τ} 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 $\hat{r}_{1,j(i)}, ..., \hat{r}_{\tau,j(i)}$

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

step 2: for $t = \tau, ..., 1$:

- compute pseudo-outcome $\check{Y}_{t+1,i} = \phi_{t+1}(Z_i; \underline{\check{\eta}}_{\tau,j(i)})$
- for $j = 1, ..., J$: regress $\check{Y}_{t+1,i}$ on $(A_{t,i}, H_{t,i})$ only using $i \in \mathcal{T}_j$, with $\check{m}_{t,j}$ output, update $\underline{\check{\eta}}_{t,j} = (\hat{r}_{t,j}, \check{m}_{t,j}, ..., \hat{r}_{\tau,j}, \check{m}_{\tau,j})$

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

∗ ***estimate density ratio r_t:*** duplicate dataset, where duplicates get assigned $A_t^{\mathbb{d}}$ with indicator $\Lambda \in \{0, 1\}$

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

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 ξ using efficient influence curve in $\mathbb{E}\{\xi(Z; \bar{\pi}, \bar{\mu})|A = a\} = \theta(a)$, with data Z and nuisance π, μ

step 1: estimate nuisance π, μ 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.

Kreyszig, E. (1991). *Introductory functional analysis with applications*. John Wiley & Sons.

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*.