0.1 Casual Inference

- Inferring the effects of any treatment/policy/intervention/etc.
  - effect of treatment on a disease
  - effect of social media on health
- Simpson's Paradox: mortality rate table
- Total population vs subgroup by conditions
- Correlation does not imply causation!
- Correlation: linear statistical dependence
- Association is the more correct term for statistical dependence
- It is possible to have large amounts of association with only *some* being casual. Some association and 0 causation is a case of "assocation is not casuation"
- e.g.Wearing shoes and waking up with a headache, common cause of drinking the night before
- This is a "confoundeer", this is a type of *confounding association*
- If association is causation, then causual inference could be solved using traditional statistics and ML
- Even with infinite amounts of data, we sometimes cannot compute casual quantities
- Identification of casual effects
- Intervention vs. observation. If we can intervene/experiment, identification becomes easy. Observational data is challenging because there is often confounding.

0.1.1 Potential Outcomes

- The *potential outcome* $Y(t)$ denotes what your outcome would be if you were to take treatment t
- Potential outcomes aren't always observed, they can be potentially observed
- The one that is actually observed depends on the treatment
- Individual treatment effect (ITE) for the ith individual

$$\tau_i \triangleq Y_i(1) - Y_i(0)$$

- Fundamental problem of causal inference: can't observe all potential outcomes for a given individual, we cannot observe both $Y(1)$ and $Y(0)$
- The outcomes that you can't observe are called *counterfactuals*
- Average treatment effect (ATE)

$$\tau \triangleq \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y(1) - Y(0)]$$

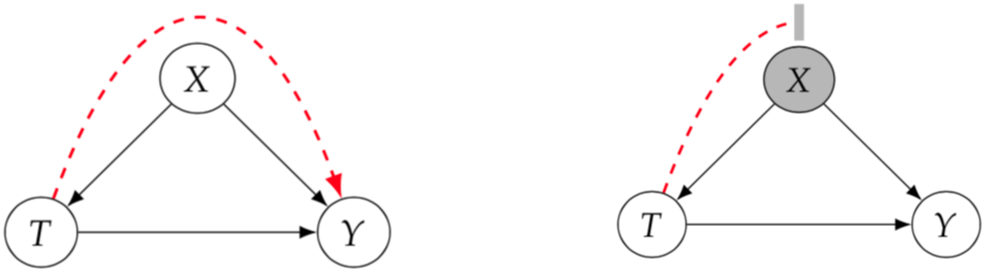
- Association difference is not the same as causal difference due to confounding

$$\mathbb{E}[Y|T=1] - \mathbb{E}[Y|Y=0] \neq \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

0.1.2 Ignorability and Exchangability

- Ignorability - ignoring missing data, remove causal arrow from confounder to treatment
- Ignorability allows us to reduce ATE to associational difference
- Exchangability - treatment groups are exchangable such that if they were swapped, the new treatment group would observe the same outcome as the old treatment group
- Identifying a casual effect is to reduce causal expression to a statistical expression
- Conditional exchangability means if we condition on the covariate X, there is no longer any non-causal association between T and X.

Figure 1: Causal graphical models



(a) X is confounding the effect of T on Y

(b) Conditioning on X leads to no confounding

- Adjustment formula: Given the assumptions of unconfoundedness, positivity, consistency, and no interference, we can identify the ATE:

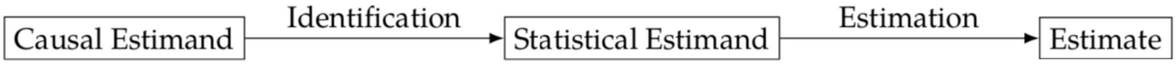
$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]]$$

- Positivity-unconfoundedness tradeoff: conditioning on more covariates can lead to better chance of satisfying unconfoundedness, but it can lead to a higher chance of violating positivity
- No interference: $Y_i(t_1, \cdot, t_n = Y_i(t_i)$ otherwise my outcome is only a function of my own treatment
- Consistency: If treatment is T, then th observed outcome Y is the potential outcome under T. $T = t \rightarrow Y = Y(t)$
- Often we use a model (e.g. linear regression or some ML predictor) in place of conditional expectations $\mathbb{E}[Y|T=t, X=x]$, these models are known as *model-assisted estimators*.

0.1.3 Definitions

- Estimand: quantity that we want to estimate
- Estimate: is an approximation of some estimand
- Estimator: a function that maps a dataset to an estimate of an estimand
- Casual estimand: any estimand that contains a potential outcome
- Statistical estimand: any estimand that does not contain a potential outcome

Figure 2: Identification Flowchart



- Graph Terminology:
  - A **graph** is a collection of **nodes** and **edges** that connect the nodes
  - Undirected graphs: edges don't have any direction
  - Directed graphs: edges go from a *parent* node to a *child* node, parents of node X are $\text{pa}(X)$
  - Two nodes are *adjacent* if they're connected by an edge
  - A *path* is any sequence of adjacent nodes regardless of direction, vs. a *directed path*
  - X is an *ancestor* of Y, and Y is a *descendant* of X
  - Cycle
  - If there are no cycles in a graph, then it is a *directed acyclic graph* (DAG)
- Bayesian Networks
  - An intuitive way to model many variables together in a joint distribution is to only model local dependencies
  - Local Markov Assumption: given its parents in the DAG, a node X is independent of its non-descendants
  - Bayesian Network Factorization: given a probability P and a DAG G, P factorizes according to G if

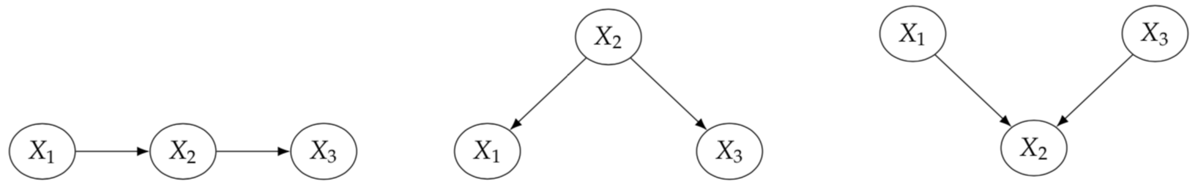
$$P(x_1, \cdot, x_n) = \prod_i P(x_i|pa_i)$$

  - Also known as the *chain rule for Bayesian networks*
- Minimal Assumption: also adds that adjacant nodes in the DAG are dependent, also equivalent to saying that we can't remove any more edges from the graph

0.1.4 Causal Graphs

- A variable X is said to be a cause of variable Y if Y can change in response to changes to X
- In a DAG, every parent is a direct cause of all of its children

Figure 3: Graph building blocks



(a) Chain

(b) Fork

(c) Immorality

- Two unconnected nodes are conditionally independent. $P(x_1, x_2) = P(x_1)P(x_2)$
- X_1 and X_3 are associated in chains and forks because they're commonly associated with X_2
- When we condition on X_2 for forks and chains, it blocks the flow of association because of the local Markov Assumption

Figure 4: Causal graphical models

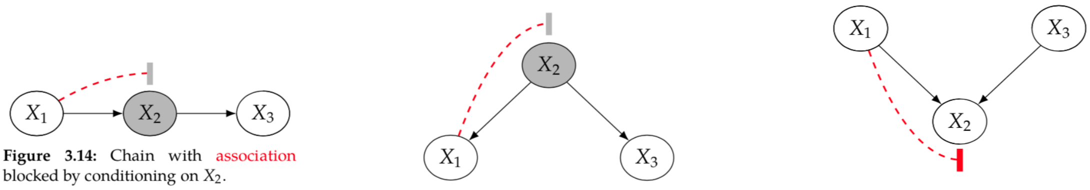


Figure 3.14: Chain with association blocked by conditioning on X_2.

Figure 3.15: Fork with association blocked by conditioning on X_2.

Figure 3.16: Immorality with association blocked by a collider.

- Colliders child of two parents that are not connected by an edge. In a collider, the parents are independent e.g. $X_1 \perp\!\!\!\perp X_3$, this is a *blocked path*
- When we condition on a collider (X_2), its parents X_1 and X_3 become dependent
- Conditioning on a collider can turn a *blocked path* to an *unblocked path*
- This phenomenon is known as *Berkson's paradox*
- Conditioning on descendants of a collider also induces association between parents of the collider
- In causal graphs, the edges have causal meaning

0.1.5 d-separation

- Two sets of nodes X and Y are d-separated by a set of nodes Z if all of the paths between X and Y are blocked by Z

- If all paths between X and Y are blocked, then they are *d-separated*
- D-separation implies conditional independence
- Global markov assumption: $X \perp\!\!\!\perp_G Y|Z \implies X \perp\!\!\!\perp_P Y|Z$
- Conditioning on $T = t$ means we restrict focus to subset of population to those who receive treatment t
- Intervention: take whole population and give everyone treatment t
- Denote intervention using $do(T = t)$ operator
- Interventional distribution: $P(Y|do(t))$ vs observational distribution: $P(Y)$
-

0.1.6 Structural Causal Models (SCMs)

- Structural equation: $B := f(A, U)$
- := gives us casual relation, A causes B
- $\mathcal{U}$ is some unobserved random variable and denotes all the relevant (noisy) background conditions that cause B

$$\begin{aligned} B &:= f_B(A, \mathcal{U}_B) \\ C &:= f_C(A, B, \mathcal{U}_C) \\ D &:= f_D(A, C, \mathcal{U}_D) \end{aligned}$$

- The variables that we write structural equations for are *endogenous* variables, these are the variables whose causal mechanisms we are modeling, $\{B, C, D\}$
- *Exogenous* variables are variables who don't have any parents in the causal graph, $\{A, \mathcal{U}_{\{B,C,D\}}\}$
- A structural casual model is a tuple of:
  - A set of endogenous variables V
  - A set of exogenous variables U
  - A set of functions f to generate each endogenous variable as a function of other variables
- If casual graph has no cycles and noise variables are independent then it is *Markovian*, if noise terms are dependent then it is *semi-Markovian*
- Intervention $do(T = t)$, replace structural equation for T with $T := t$, then we get the *interventional SCM M_t*
- This is by the modularity assumption

0.2 Bayesian Inference

0.3 Variational Inference

0.4 Expectation Maximization

- Parameters θ, evidence X
- Prior: probability of parameters, $p(\theta)$
- Likelihood: probability of evidence given parameters, $p(X|\theta)$
- Posterior: probability of the parametersgiven evidence, $p(\theta|X)$

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)} p(\theta)$$
$$posterior = \frac{likelihood}{constant} prior$$

- Maximum a posteriori estimate (MAP): estimate of an unknown quantity, mode of a posterior distribution
- point estimate of unobservable quantity based on empirical data
- EM: iterative method to find maximum likelihood or maximum a posteriori (MAP) estimate of parameters in a statistical model
- alternate between Expectation step and Maximization step
- E-step: creates a function for expectation of log-likelihood evaluated using current estimates of parameter
- M-step: computes new parameters that maximize expected log-likelihood

0.5 Gaussian Processes

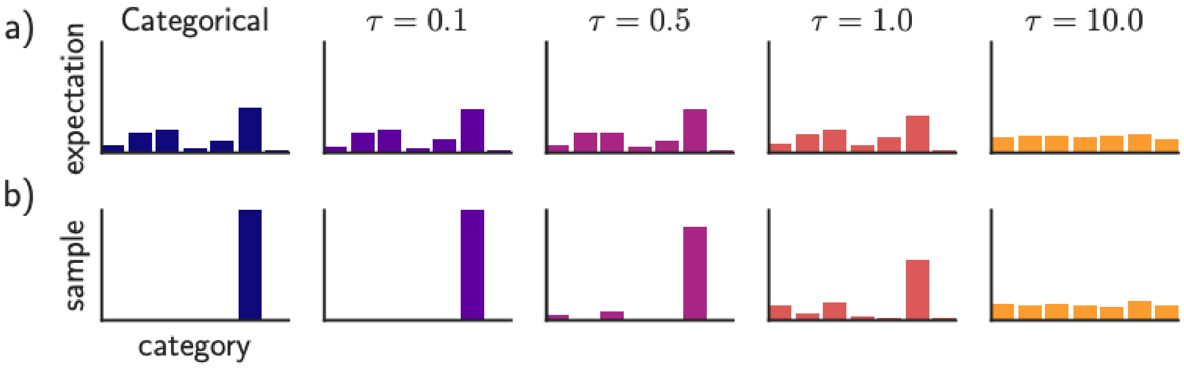
- Multivariate gaussian distributions are defined by a mean vector μ and covariance matrix Σ
- The diagonal of Σ consists of the variances σ_i^2 of the i-th random variable and the off diagonal elements describe correlation between i, j-th random variables
- Gaussian distributions are closed under conditioning and marginalization
- $X|Y \sim \mathcal{N}(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$
- Nice visualizations of these operations
- Simple intuitive explanations
- Kernel, often called the *covariance function*. Kernel computes pairwise similarity between points $k(t, t')$
- Different kernels: rbf kernel, periodic, linear, etc

- GP is a non-parametric approach to regression meaning it finds a distribution over all the functions that are consistent with the observed data
- Starts with a prior distribution and updates this as points are observed
- Non-parametric: infinite parameters
- GP can derive posterior from prior and observations
- We want $p(f_*|x_*, x, f)$ where x is observed data, x_* is test data
- Can sample points from the posterior

0.6 Gumbel Softmax Trick

- Argmax is a non-differentiable function
- Wherever you are as long as you are not on the x1=x2 line, if you move an infinitesimal tiny bit in any direction the output of argmax doesn't change, thus the gradient of argmax is (0,0) almost everywhere
- Gumbel-Softmax is a continuous ditribution that can approximate samples from a categorical distribution
- $z = \text{one\_hot}(argmax_i[g_i + log(\pi_i)])$
- Gumbel-Softmax interpolates between a one-hot encoded categorical distributions and continuous categorical densities
- Low temperature = categorical random variable, high temperature = uniform distribution
- It is smooth for $\tau > 0$ and has well-defined gradient
- Tradeoff during training where smaller temperatures lead to samples closer to one-hot but variance of gradients is large and vice versa for high temperature
- Straight-trhough Gumbel-Softmax allows samples to be sparse, even when temperature is high

Figure 5: Gumbel-Softmax distribution


a)
b)
Categorical
$\tau = 0.1$
$\tau = 0.5$
$\tau = 1.0$
$\tau = 10.0$
expectation
sample
category