Estimators & Estimation (and DAGs too)

Annie Chen

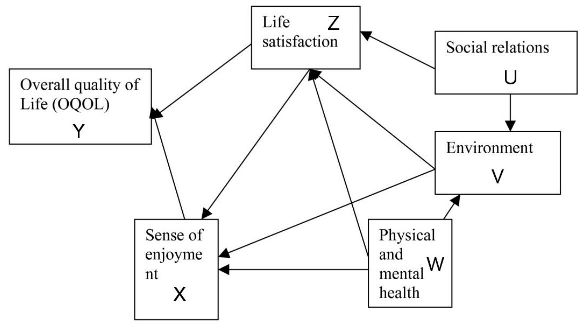
1/10/2020

Some admin

- Push problem sets to GitHub
- Office hours: Tuesdays 2:45pm to 3:45pm
- other admin stuff?

An Alternative Causal Model: Causal Graphs

- The old paradigm: structural equation modeling (SEM) and path analysis



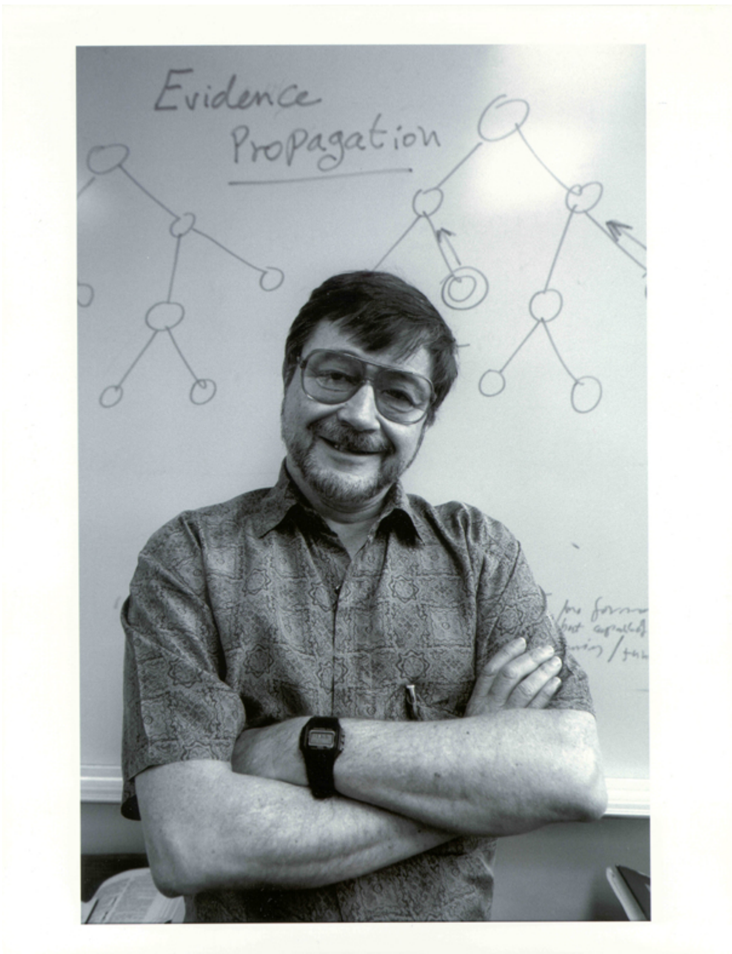
1. Postulate a causal mechanism and draw a corresponding path diagram
2. Translate it into a (typically linear) system of equations:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + \epsilon_\alpha$$
$$X = \beta_0 + \beta_1 Z + \beta_2 W + \beta_3 V + \epsilon_\beta \quad \cdots$$

3. Estimate β, α, etc. typically assuming normality and exogeneity

- However: Strong distributional/functional form assumptions and no language to distinguish causation from association

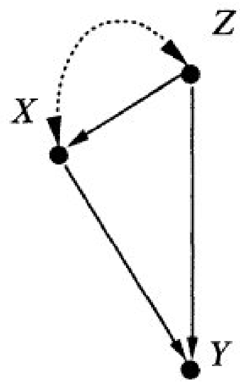
Pearl's Attack

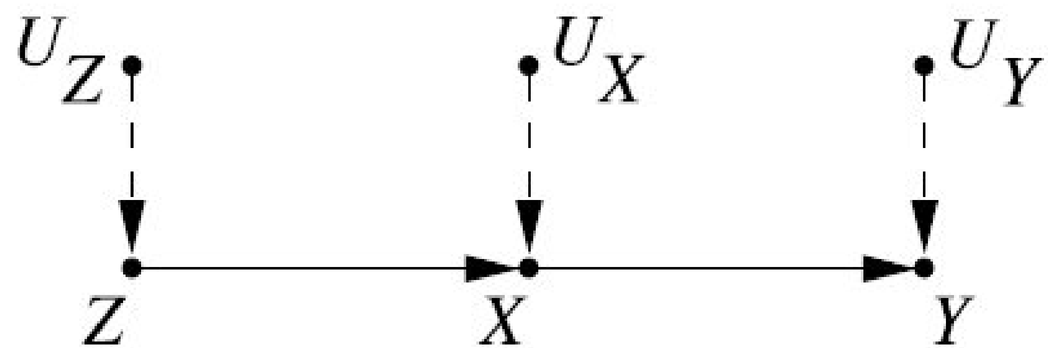


- Judea Pearl (1936–) proposed a new causal inference framework based on nonparametric structural equation modeling (NPSEM)
- Computer scientist, working on Artificial Intelligence
- Causality (2000, Cambridge UP)
- Pearl’s framework builds on SEMs and revives it as a formal language of causality.

Anatomy of a Causal Directed Acyclic Graph (DAG)

- Composed of nodes (• = observed or ∘ = unobserved variables) and directed edges ($\longrightarrow$ = possible causal relationship).
- Exogenous variables that are not explicitly modelled can be omitted from a graph
- It is *missing edges* that encode causal assumptions: missing arrows encode exclusion restrictions

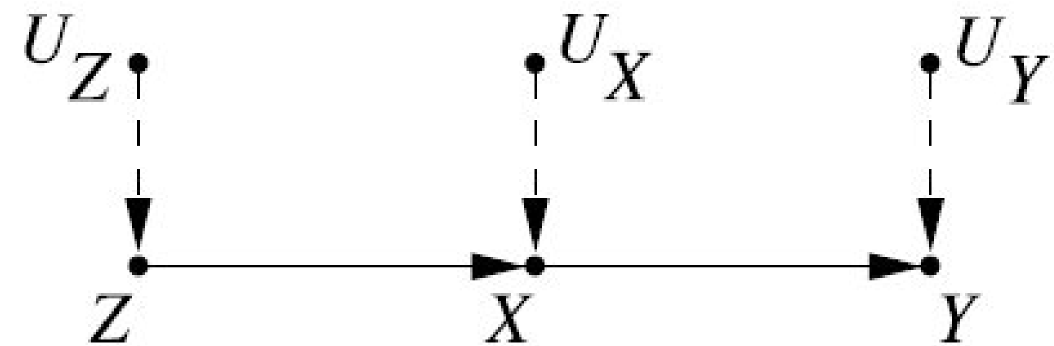




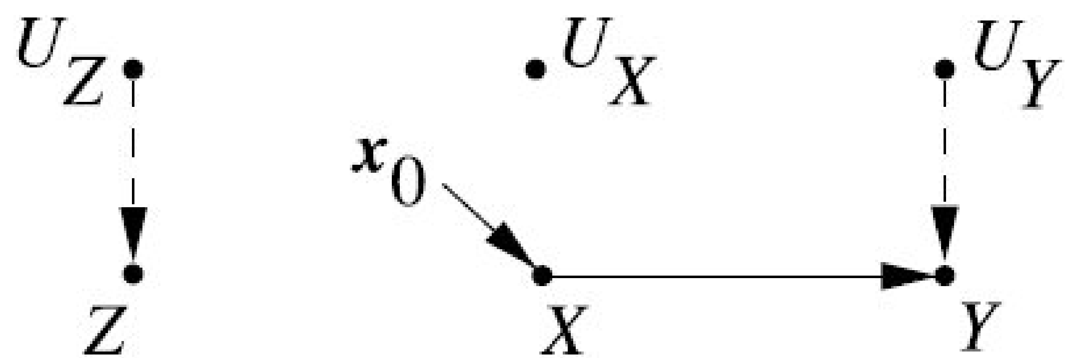
- The corresponding structural equations:

$z = f_z(u_z)$, $x = f_x(z, u_x)$, and $y = f_y(x, u_y)$

How do we get from:

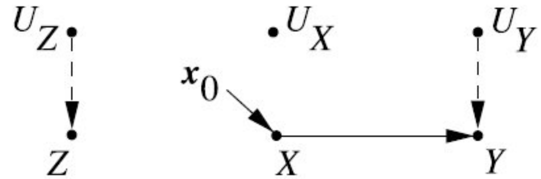


to:



`do()` Operator and Causal Effects

- Treatments (interventions) are represented by the $do()$ operator
- For example, $do(x_0)$ holds X at x_0 exogenously ($x = x_0$)
- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$



- Then, the ATE is defined as $\mathbb{E}[Y \mid do(x_1)] - \mathbb{E}[Y \mid do(x_0)]$
- Identification: Can $P(y \mid do(x))$ be estimated from data governed by the pre-intervention distribution $P(x, y, z)$?

Why DAGs?

- Fundamental equivalence of causal graphs and potential outcomes:

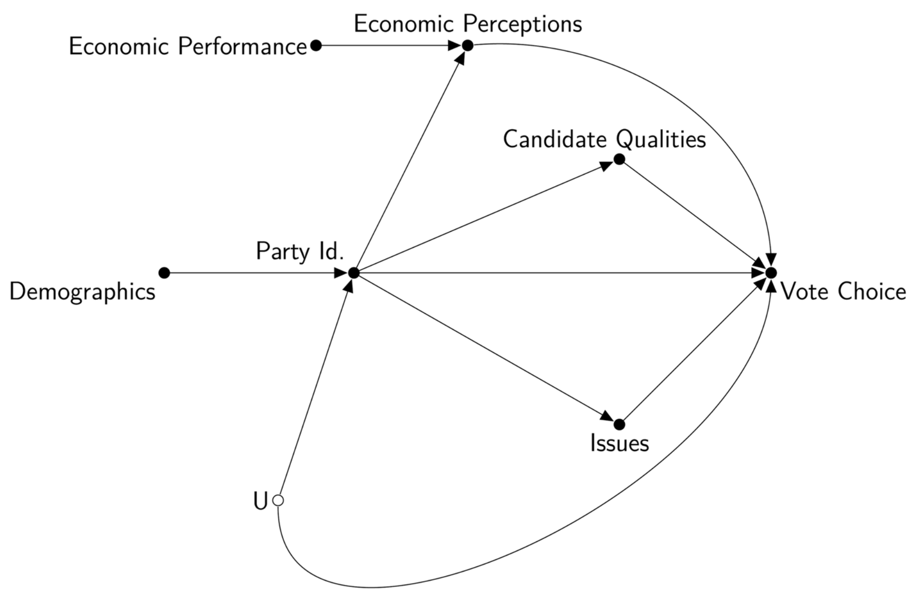
$$z = f_z(u_z),\ x = f_x(z, u_x), \text{ and } y = f_y(x, u_y)$$

$$=$$

$$Z_i,\ X_{Zi}, \text{ and } Y_{Xi}$$

- Potential outcomes framework (Neyman-Rubin model) is the dominant framework for causal inference in the social sciences.

Why DAGs?

- DAGs are a great way of encoding your causal assumptions.
- I.e. Party id as the "unmoved-mover" (e.g., Campbell et al., 1966; Green and Palmquist, 1994; Miller and Shanks, 1996; Green et al., 2004)?
- OR can short-term factors like issue positions [1], candidate characteristics [2], and economic perceptions [3], at least under some conditions, move party id?
- 

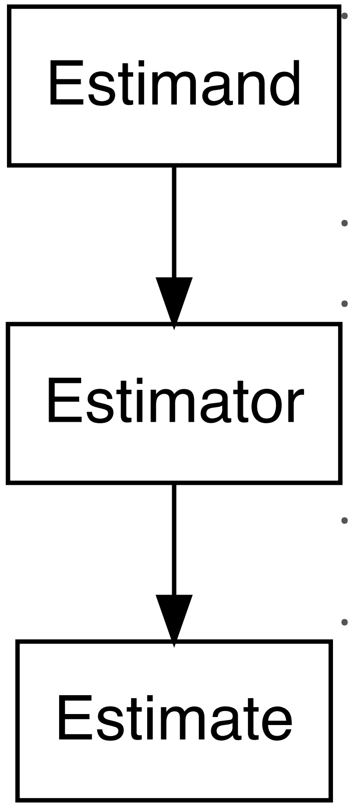
1. Carsey and Layman, 2006; Highton and Kam, 2011
2. Page and Jones, 1979
3. Fiorina, 1981; MacKuen et al., 1989

Estimand, Estimators, and Estimates

What are these?

- $$Y = \beta_0 + \beta_1 X_1 + u$$
- $$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{u}$$
- $$\hat{\beta} = (X'X)^{-1} X'y$$

Estimand, Estimators, and Estimates



- θ: An unknown value that describes a population relationship. The concept we want to estimate.
- $ATE = \mathbb{E}[Y_{1i} - Y_{0i}]$
- $\widehat{\theta}(Y)$: A rule (i.e., function) for combining data to produce a numerical value for a population parameter (estimand);
- $\widehat{ATE} = \frac{1}{n_1} \sum_{i=1}^{n} D_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - D_i) Y_i$
- $\widehat{\theta}(Y = y)$: The numerical value(s) taken on by an estimator for a particular sample of data.

Estimation and Sampling Distribution

- Estimators produce estimates. Under repeated random sampling, they produce many estimates.
- Assuming repeated random sampling, we can calculate the uncertainty around our point estimates.
- Under repeated sampling, we also have a distribution of an estimate: a sampling distribution (the probability distribution of a statistic).

Sampling Distribution of the Mean

-

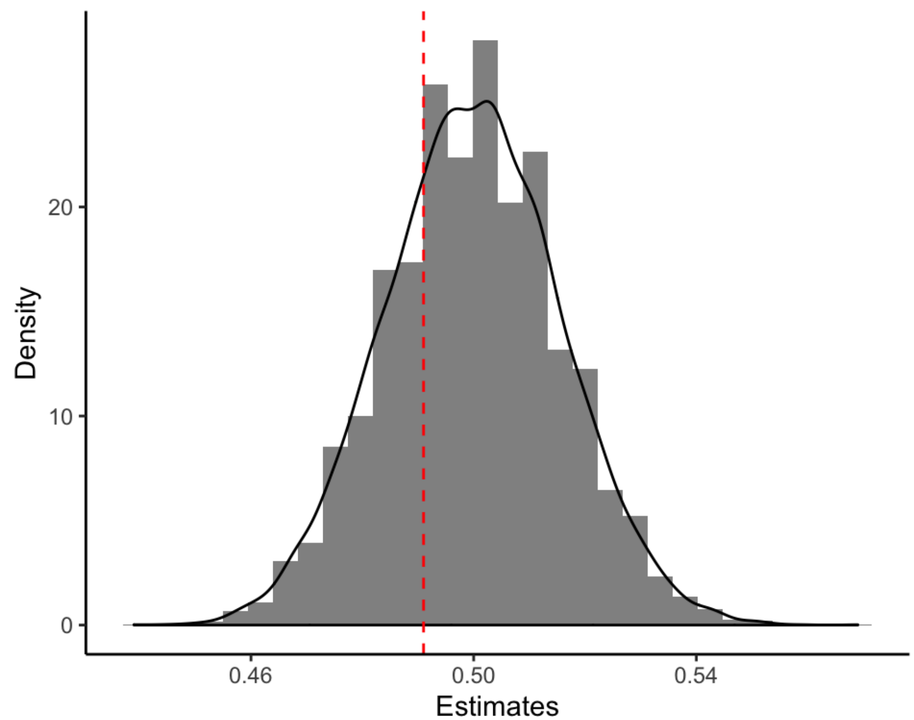
Estimand	Estimator	Sampling Distribution
$\mathbb{E}(X) = \mu$	$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

```
set.seed(01102020)
X <- rbinom(1000, 1, 0.5)

mean.estimator = function(X) {
  # Calculate the size of the sample, n.
  n <- length(X)
  # Apply the estimator to the sample data.
  mu.hat <- (1/n) * sum(X)
  return(mu.hat)
}

# Estimate mean
mean.estimator(X)
```

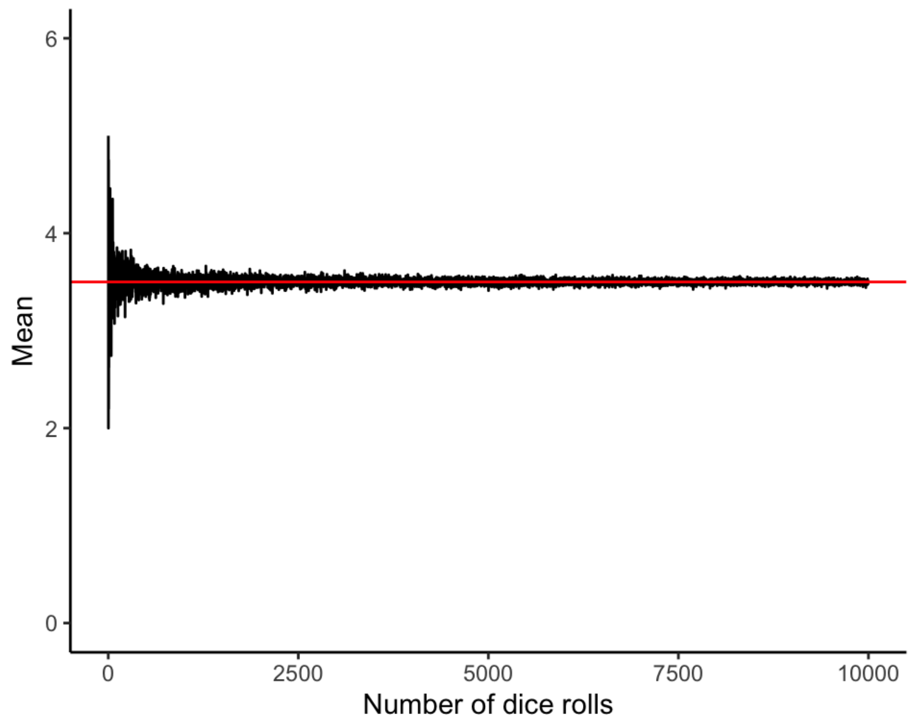
```
## [1] 0.491
```

This is the sampling distribution of our estimator.

```
mu_samp <- replicate(10000, mean.estimator(rbinom(1000, 1, 0.5)))
```

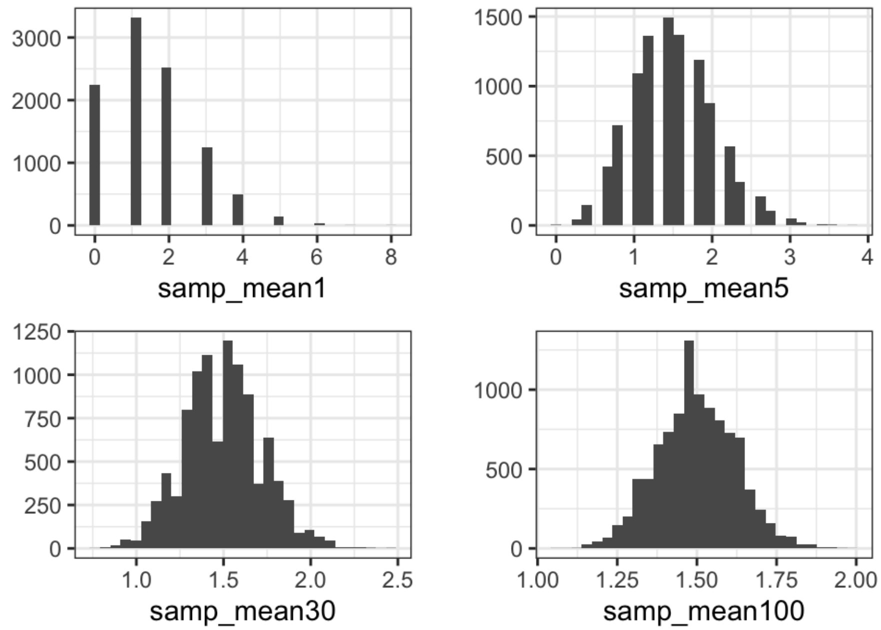


Law of Large Numbers and Central Limit Theorem

- Law of Large Numbers: If $X_1, \ldots, X_n$ are drawn independently from the distribution of X (randomly sampled), then the sample estimator for the mean converges in probability to the population average of X [4]
- 

4. $\bar{X}_n \xrightarrow{p} \mathbb{E}[X]$ as $n \to \infty$, assuming the expected value exists $\mathbb{E}[X] < \infty$

- Central Limit Theorem: The sampling distribution of the sample means approaches a Normal distribution as the sample size increases (no matter what the underlying population distribution is). [5]

- 

5. If $X_1, \ldots, X_N$ are drawn independently drawn from the distribution of X, and $\mathbb{V}[X] < \infty$ then, the difference between the sample average of X and the population average of X (scaled by $\sqrt{N}$) converges in distribution to a Normal distribution. That is, $\sqrt{N}(\bar{X}_N - \mathbb{E}[X]) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$

(Desirable) Properties of Estimators

- Unbiasedness

$$\mathbb{E}[\hat{\mu}] = \mu$$

- (Relative) Efficiency

$$\mathbb{V}[\widehat{\mu_1}] < \mathbb{V}[\widehat{\mu_2}]$$

- Consistency

$$\hat{\mu} \xrightarrow{p} \mu$$

- Asymptotic Normality

$$\hat{\mu} \overset{approx.}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Review of potential outcomes

- D_i: Indicator of treatment intake for unit i, where $i = 1, \ldots, N$; for now, $D_i \in \{0, 1\}$.
- Y_i: Variable of interest, whose value is observed, which may be affected by the treatment.

Review of potential outcomes

Which of these are observed?

	$Y = 1$	$Y = 0$
$D = 1$	$\mathbb{E}[Y_i(1)\vert D_i = 1]$	$\mathbb{E}[Y_i(0)\vert D_i = 1]$
$D = 0$	$\mathbb{E}[Y_i(1)\vert D_i = 0]$	$\mathbb{E}[Y_i(0)\vert D_i = 0]$

Review of potential outcomes

Which of these are observed?

	$Y = 1$	$Y = 0$
$D = 1$	$\mathbb{E}[Y_i(1) \vert D_i = 1]$	$\mathbb{E}[Y_i(0) \vert D_i = 1]$
$D = 0$	$\mathbb{E}[Y_i(1) \vert D_i = 0]$	$\mathbb{E}[Y_i(0) \vert D_i = 0]$

- This problem has a name…
- The Fundamental Problem of Causal Inference!
- Can't we just use this naive estimator? $\mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0]$
- It depends.

Selection Bias

- Data from National Health Interview Survey (NHIS) 2005. Health status measured from 1 (poor health) to 5 (excellent health).

Group	Sample Size	Avg Health Status	Std Error
Hospital	7,774	3.21	0.014
No Hospital	90,049	3.93	0.003

- Going to the hospital makes people sick!? O_o

Quiz yourself!

Are the following equalities always true?

- $Y_{di} = Y_i(d)$
- Yes.
- $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$
- Nope.
- $Y_i(d_i) = Y_i(d_i, d_{i-1}, d_{i+1})$
- Nope.
- $\mathbb{E}[Y_i(1) - Y_i(0)] = Y_i(1) - Y_i(0)$
- Nope again.

GitHub markdown