Bayesian Causal Inference

Ben Goodrich

February 16, 2021

Fisherian / Randomization / Design Inference

- The data are NOT a random sample from some population
- Randomization is between treatment and control status
- If there are N observations, there are $\binom{N}{N/2}$ ways to assign to treatment
- If the treatment has no effect, then $\ldots$

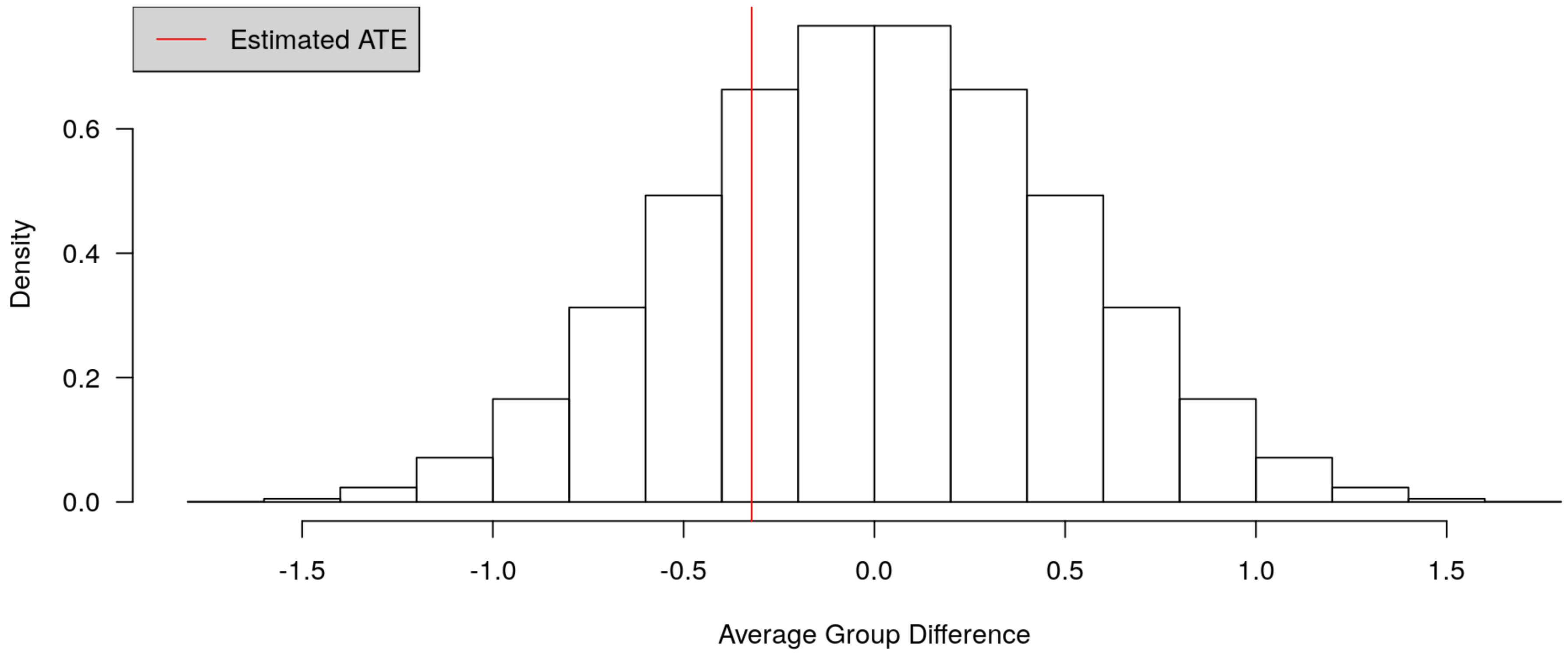
```
N <- 20
treated <- sample(1:N, size = 10, replace = FALSE); control <- (1:N)[-treated]
y <- rnorm(N)
diffs <- combn(N, N / 2, FUN = function(group_1) {
  group_2 <- (1:N)[-group_1]
  return(mean(y[group_1]) - mean(y[group_2]))
})
ATE <- mean(y[treated]) - mean(y[control])
mean(abs(diffs) > abs(ATE)) # p-value
```

```
## [1] 0.5254931
```

```
hist(diffs, prob = TRUE, main = "", xlab = "Average Group Difference")
abline(v = ATE, col = "red")
```

Plot based on Previous Slide

```
hist(diffs, prob = TRUE, main = "", xlab = "Average Group Difference")
abline(v = ATE, col = "red")
legend("topleft", legend = "Estimated ATE", lty = 1, col = 2, bg = "lightgrey")
```



Potential Outcomes Framework

- Each of N observations has TWO functions, Y_{i1} and Y_{i0}, that respectively yield the outcome if i were in the treatment or control group respectively
- Individual Causal Effect (ICE) is defined as $\Delta_i = Y_{i1} - Y_{i0}$
- Observed outcome is $y_i = t_i Y_{i1} + (1 - t_i) Y_{i0}$ where t_i is binary
- Fundamental Problem of Causal Inference: Either Y_{i1} or Y_{i0} is not observed
- Average Causal / Treatment Effect (ACE or ATE) can be estimated as $\widehat{\Delta} = \frac{1}{N/2} \sum_{i:t_i=1} y_i - \frac{1}{N/2} \sum_{i:t_i=0} y_i$ and is unbiased if t_i is conditionally independent of both Y_{i1} and Y_{i0}, given a (possibly empty) set of covariates $\mathbf{x}_i$
- This framework is relatively good for understanding the ACE in experiments

Graphical Causal Models Framework

- Potential Outcomes framework is widespread in economics & political science
- Graphical Causal Models framework is widespread in epidemiology and to a lesser extent, in sociology and psychology but is overkill for estimating ACEs in simple experiments
- A theorem in one framework implies a theorem in the other framework
- Directed Acyclic Graphs can serve multiple purposes:
  1. A language to communicate theories
  2. Identification Analysis: Whether a theory implies the ACE could be calculated in a population
  3. Testable Implications: What observable variables are conditionally independent, which can be investigated with data

Directed Acyclic Graphs (DAGs)

- Three elements to a DAG:
    1. Variables / Nodes
    2. Arrows from an earlier (in time) node to a later node
    3. Absence of arrows between nodes, which implies an ACE is zero
- $A \rightarrow B$ means that if A were experimentally manipulated there MAY be a non-zero ACE that is assumed to be unmediated by any other variable
- Cannot start at any node, follow arrows, and get back to where you started
- DAGs are usually written w/o distributional or functional form assumptions

Three Sources of Association between A and B

1. Direct, $A \to B$, or indirect, $A \to C \to B$, causation
2. Confounding due to common cause(s): $A \leftarrow C \to B$
3. Selection due to conditioning on a “collider”, $A \to \boxed{\mathrm{C}} \leftarrow B$, or a (not necessarily direct) descendant of a collider, $$\begin{tikzcd} A \arrow[r] & C \arrow[d] & B \arrow[l] \\ & \boxed{\mathrm{D}} & \end{tikzcd}$$, where boxes indicate stratification or otherwise perfect conditioning (by researchers)

- Technically, a collider has to be defined with reference to a path. In the above, C is a collider along the path from A to B but C is not a collider on the path from A to D

Three Mistakes in Calculating the ACE of X on Y

1. Intercepting: $X \to \boxed{\text{M}} \to Y$ or $X \to Y \to \boxed{\text{S}}$

2. Failure to condition on a confounder:

$$\begin{tikzcd}
X \arrow[r] & Y \\
C \arrow[u] \arrow[ur] &
\end{tikzcd}$$

3. Endogenous selection due to conditioning on (a descendent of) a collider:

$$\begin{tikzcd}
X \arrow[r] \arrow[d] & Y \arrow[dl] \\
\boxed{\text{C}} &
\end{tikzcd}$$

Paths in DAGs

- A path is a sequence of connected nodes, regardless of the arrows' direction
    1. Causal Path: A path that exclusively follows the arrows
    2. Non-causal Path: Any other path
- What are the paths below and are they causal or not?

$$\begin{tikzcd} X \arrow[r] & M \\ W \arrow[r] \arrow[ur] & Y \end{tikzcd}$$

Blocking / Closing a Path

1. Condition on a noncollider along a path: $A \leftarrow \boxed{\mathrm{C}} \rightarrow B$
2. Refrain from conditioning on a (descendant of a) collider along a path: $A \rightarrow C \leftarrow B$

- If a path is not blocked it is open or unblocked
- Two variables are “d-separated” iff all paths between them are blocked by conditioning on a possibly empty set of variables $\{Z\}$, in which case they are conditionally independent
- Two variables that are not “d-separated” are almost surely not conditionally independent

Adjustment and Backdoor Criteria

- Adjustment Criterion is satisfied iff

  1. All causal paths from X to Y are open
  2. All non-causal paths from X to Y blocked by a (possibly empty) set of variables $\{Z\}$

- If the adjustment criterion is satisfied, ACE of X on Y is identified and can be consistently estimated with a (possibly weighted) difference in means

- Backdoor Criterion is satisfied iff

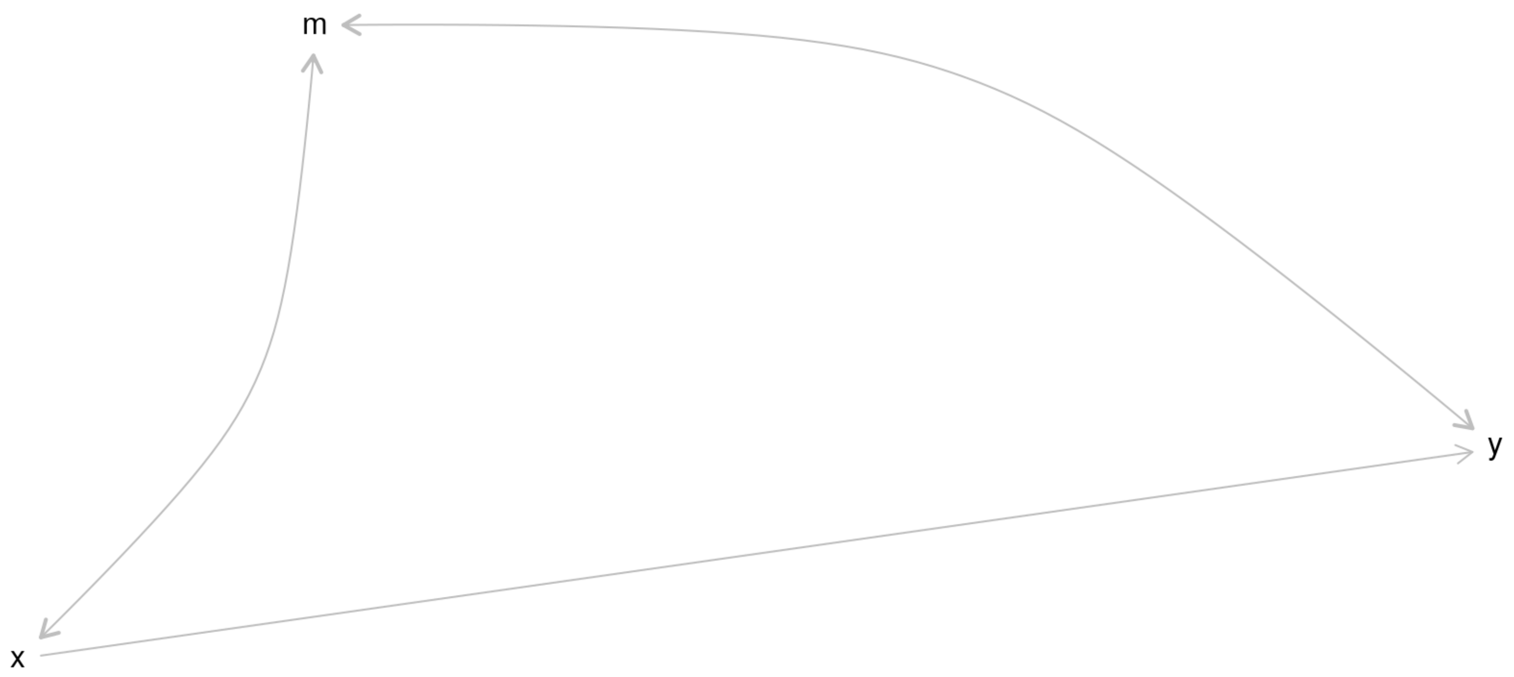
  1. No element of $\{Z\}$ is a descendant of Y
  2. Some element of $\{Z\}$ blocks all “backdoor” paths, i.e. those starting with $\rightarrow X$

- Backdoor Criterion implies Adjustment Criterion but not vice versa

Dagitty

- There is a website, www.dagitty.net , that implements the most common and useful identification strategies for DAGs and a similar R package, dagitty:

```
library(dagitty); g <- dagitty( "dag{ x -> y ; x <-> m <-> y }" ); plot(graphLayout(g))
```



Double-headed Arrows

- The dagitty R package and a lot of authors use $A \leftrightarrow B$ as a shorthand for a $A \leftarrow U \rightarrow B$ with U unobserved or for $\begin{matrix} X & \leftrightarrow & Y \\ \downarrow & \swarrow & \\ \boxed{\text{C}} & & \end{matrix}$ due to conditioning
- Any path involving a $\leftrightarrow$ is a non-causal path

Contrasts with Common Practice

- Supervised learning generally does not utilize DAGs and is not capable of estimating causal effects. Models are scored on how well they predict the outcome (in the testing set), which is better if you condition on descendants of the outcome, mediators, some colliders, etc.
- Regressions generally do not estimate the causal effect of all covariates and rarely estimate the causal effect of any covariate. They implicitly correspond to a DAG like (although the error term is often omitted):

$$\begin{array}{ccccc} & \nearrow & \boxed{\mathrm{X}}_1 & \searrow & \\ C & \rightarrow & \boxed{\mathrm{X}}_2 & \rightarrow & Y \leftarrow \epsilon \\ & \searrow & \boxed{\mathrm{X}}_3 & \nearrow & \end{array}$$

- Non-Bayesians tend to write DAGs without distributions, parameters, & priors

CausalQueries R Package

- All DAGs where all observed nodes are binary have a (multinomial) likelihood function that can be combined with priors on the “type” parameters to perform Bayesian inference
- The CausalQueries R Package takes a DAG, priors, and data and uses Stan to draw from the posterior distribution of the parameters given the data & priors
- Useful for describing your beliefs about (not necessarily average) causal effects in many situations that are not simple experiments

Dataset from Bertrand and Mullainathan (2004)

- Resumes were created for a fictitious person applying for an entry-level job but the name at the top of the resume was randomized to make the company think the applicant was probably black / white / male / female. The outcome is whether the company called the applicant to schedule an interview, etc.

```
data(resume, package = "qss") # may need to install qss from GitHub
library(dplyr)
resume_grouped <- group_by(resume, race, sex, call) %>% summarize(n = n())
resume_grouped
```

```
## # A tibble: 8 x 4
## # Groups:   race, sex [4]
##   race  sex     call     n
##   <chr> <chr>  <int> <int>
## 1 black female     0  1761
## 2 black female     1   125
## 3 black male       0   517
## 4 black male       1    32
## 5 white female     0  1676
## 6 white female     1   184
## 7 white male       0   524
## 8 white male       1    51
```

Basic Frequentist Inference

```
prop.test(matrix(resume_grouped$n, ncol = 2, byrow = TRUE))
```

```
## 
##  4-sample test for equality of proportions without continuity correction
## 
## data:  matrix(resume_grouped$n, ncol = 2, byrow = TRUE)
## X-squared = 17.867, df = 3, p-value = 0.0004686
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4 
## 0.9337222 0.9417122 0.9010753 0.9113043
```

- Many questions prohibited, such as “How sure are you that companies favor white / male over black / female?”

Principal Stratification

- If X and $Y\left(X\right)$ are both binary, then there are $2^2 = 4$ types of observations:

Data	Type
$Y\left(x\right) \neq x$	Adverse
$Y\left(x\right) = x$	Beneficial
$Y\left(x\right) = 0 \forall x$	Chronic
$Y\left(x\right) = 1 \forall x$	Destined

- Chronic and destined types have zero ICE
- The population ACE is the difference between the proportion of beneficial and adverse types
- In complicated models with many nodes, there are a huge number of types

Setup for Basic Bayesian Inference

```
library(CausalQueries)
(model <- make_model("race -> call <- sex"))
```

```
##
## Statement:
## [1] "race -> call <- sex"
##
## DAG:
##   parent children
## 1   race     call
## 2    sex     call
##
##  -----------------------------------------------------------------------------------------
##
## Nodal types:
## $race
## 0  1
##
##   node position display interpretation
## 1 race       NA   race0       race = 0
## 2 race       NA   race1       race = 1
##
## $sex
## 0  1
##
##   node position display interpretation
## 1  sex       NA    sex0        sex = 0
## 2  sex       NA    sex1        sex = 1
##
## $call
## 0000  1000  0100  1100  0010  1010  0110  1110  0001  100
##
##   node position    display            interpretation
## 1 call        1 call[*]*** call | race = 0 & sex = 0
## 2 call        2 call*[*]** call | race = 1 & sex = 0
## 3 call        3 call**[*]* call | race = 0 & sex = 1
## 4 call        4 call***[*] call | race = 1 & sex = 1
##
##
## Number of types by node
## race  sex call
##    2    2   16
##
## Number of unit types:  64
```

Multinomial Distribution

- The multinomial distribution over $\Omega = \{0, 1, \dots, n\}$ has a PMF $\Pr\left(x \middle| \pi_1, \pi_2, \dots, \pi_K\right) = n! \prod_{k=1}^K \frac{\pi_k^{x_k}}{x_k!}$ where the parameters satisfy $\pi_k \geq 0 \forall k$, $\sum_{k=1}^K \pi_k = 1$, and $n = \sum_{k=1}^K x_k$
- The multinomial distribution is a generalization of the binomial distribution to the case that there are K possibilities rather than merely failure vs. success
- Categorical is a special case where $n = 1$
- The multinomial distribution is the count of n independent categorical random variables with the same π_k values
- Draw via `rmultinom(1, size = n, prob = c(pi_1, pi_2, ..., pi_K))`

Dirichlet Distribution

- Dirichlet distribution is over the parameter space of PMFs — i.e. $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$ — and the Dirichlet PDF is $f\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$ where $\alpha_k \geq 0 \, \forall k$ and the multivariate Beta function is $B\left(\boldsymbol{\alpha}\right) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\prod_{k=1}^{K} \alpha_k\right)}$ where $\Gamma\left(z\right) = \int_0^\infty u^{z-1} e^{-u} du$ is the Gamma function
- $\mathbb{E}\pi_i = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k} \, \forall \boldsymbol{i}$ and the mode of π_i is $\frac{\alpha_i - 1}{-1 + \sum_{k=1}^{K} \alpha_k}$ if $\alpha_i > 1$
- Iff $\alpha_k = 1 \, \forall k$, $f\left(\boldsymbol{\pi} \middle| \, \boldsymbol{\alpha} = \mathbf{1}\right)$ is constant over Θ (simplexes)
- Beta distribution is a special case of the Dirichlet where $K = 2$
- Marginal and conditional distributions for subsets of $\boldsymbol{\pi}$ are also Dirichlet
- Dirichlet distribution is conjugate with the multinomial and categorical

Conditioning on the Resume Data

- Default priors on types are flat Dirichlet, which get mapped into simplexes for the multinomial likelihood function

```
model <- update_model(model, keep_transformed = TRUE, seed = 12345,
                      data = mutate(resume, race = race == "white", sex = sex == "female"))
```

```
query_model(model, using = "posteriors", queries =
              c(race_ATE = "call[race = 0] - call[race = 1]",
                sex_ATE  = "call[sex = 0] - call[sex = 1]",
                Pr_fine  = "call[race = 0, sex = 0] >= call[race = 1, sex = 1]"))
```

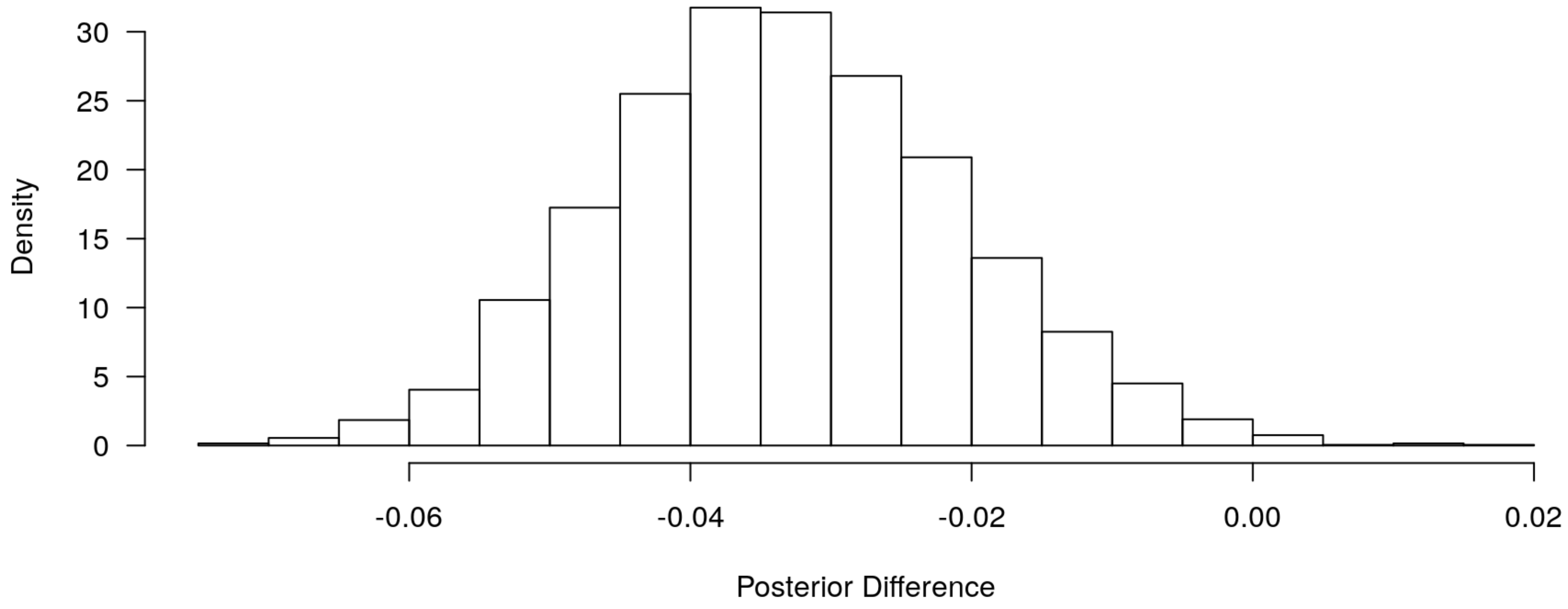
```
##      Query Given      Using Case.estimand         mean          sd
## 1 race_ATE     - posteriors         FALSE -0.030168196 0.007612252
## 2  sex_ATE     - posteriors         FALSE -0.004485754 0.008768055
## 3  Pr_fine     - posteriors         FALSE  0.931494018 0.013185155
```

- The last posterior `mean` value is what people mistake a p-value for

Do Not Limit Yourself to a Posterior Summary

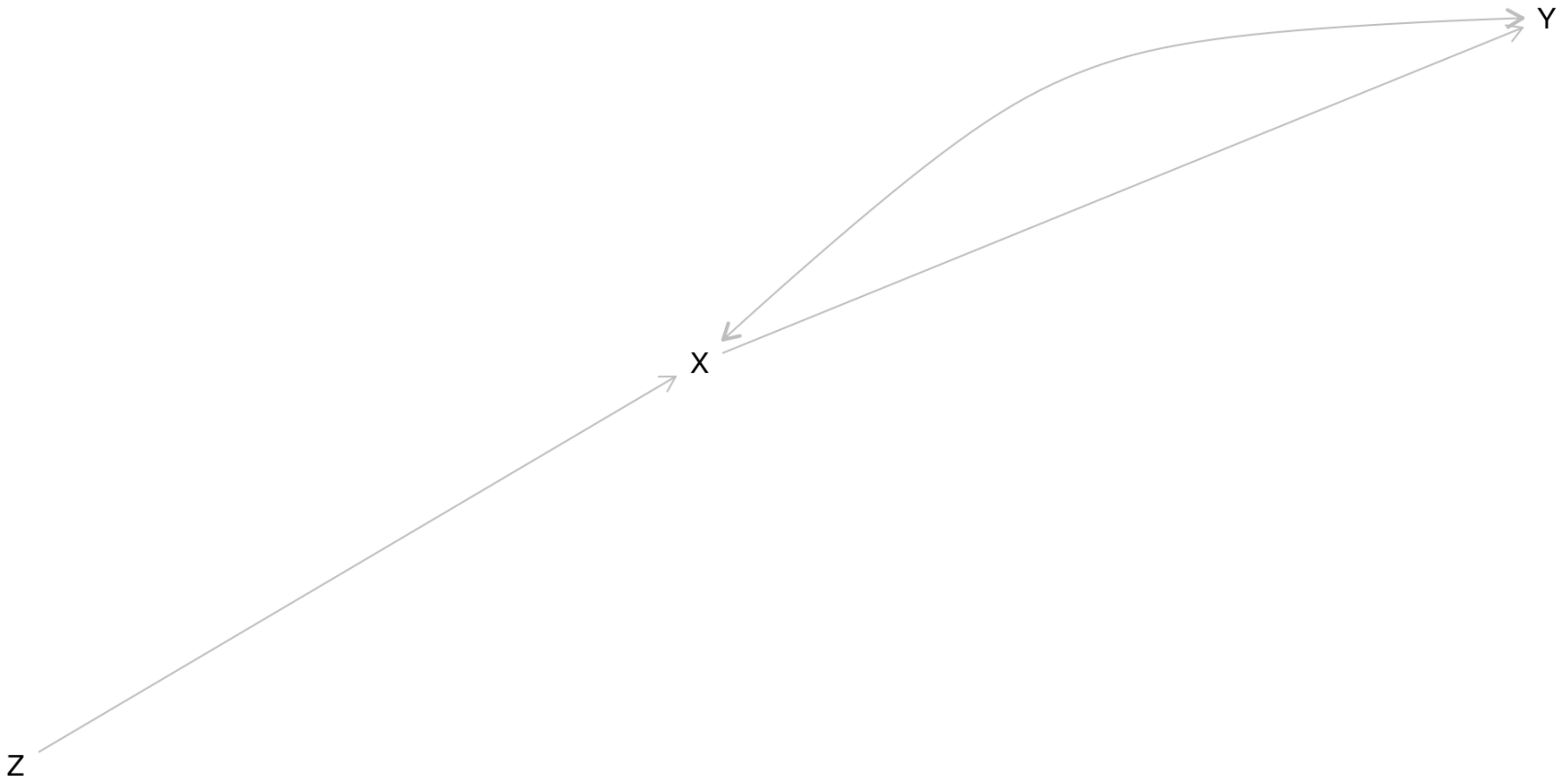
- You can get all the posterior draws (for a single query at a time)

```
d <- c(query_distribution(model, using = "posteriors",
                          query = "call[race = 0, sex = 0] - call[race = 1, sex = 1]"))
hist(d, prob = TRUE, main = "", xlab = "Posterior Difference")
```



Simple DAG for Instrumental Variables

```
plot(make_model("Z -> X -> Y") %>% set_confound(confound = list("X <-> Y")))
```



Frequentist Estimation w/ Instrumental Variables

- By “instrumental variables”, people usually think of Frequentist ESTIMATORS of instrumental variables models but the concept is much more general
- In the previous slide, if all variables are binary, an estimator of the (local) Average Treatment Effect of X on Y is given by $\widehat{\Delta} = \frac{\mathrm{Cov}(Z,Y)}{\mathrm{Cov}(X,Y)}$
- Estimated covariances are asymptotically normal *across datasets* of size N
- HW2 showed a ratio of normals is so heavy-tailed the ratio does not have an $\mathbb{E}$
- Median of estimator *across datasets* of size N is equal to true LATE but individual estimates can be HUGE outliers, especially if the denominator is not mostly bounded away from zero
- Same is true if there are more covariates, continuous variables, etc., in which case Frequentists use an estimator called “two stage least squares”