Bayesian data analysis: Theory & practice

Part 5b: Causal inference & regression modeling

Michael Franke

Causal inference
motivation

- we know: **correlation does not mean causation**
- we want: actionable conclusions
- we use: randomized control trials
- but: **what if we can only observe passively?**

how to disentangle

Simpson’s paradox

Simpson's paradox: Case 1

Gender as a confounder

- 700 patients w/ choice: take drug or not
- variable of interest: recovery rate
- also observed: **gender**

	No drug	Drug
Men	234 / 270 (87%)	81 / 87 (93%)
Women	55 / 80 (68%)	192 / 263 (73%)
Σ	289 / 350 (83%)	273 / 350 (78%)

Case 2

Blood-pressure as a mediator

- same as case 1, but no gender info
- also observed:
  post-treatment blood pressure

	No drug	Drug
Low BP	234 / 270 (87%)	81 / 87 (93%)
High BP	55 / 80 (68%)	192 / 263 (73%)
Σ	289 / 350 (83%)	273 / 350 (78%)

Would you recommend using the drug
in Case 1 and/ or Case 2?

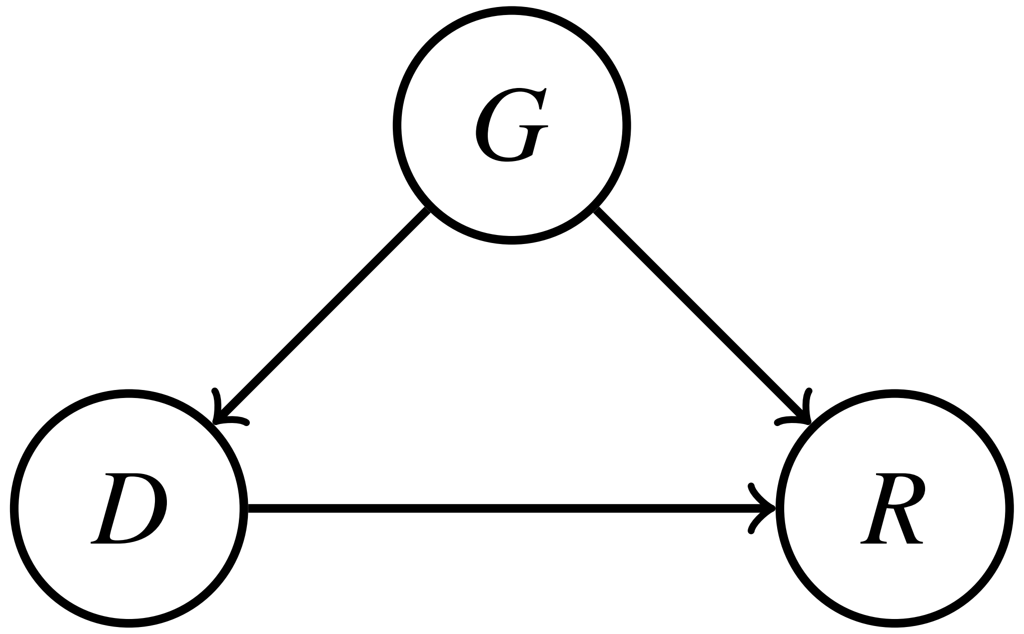
Simpson's paradox: Case 1

Gender as a confounder

- 700 patients w/ choice: take drug or not
- variable of interest: recovery rate
- also observed: **gender**

	No drug	Drug
Men	234 / 270 (87%)	81 / 87 (93%)
Women	55 / 80 (68%)	192 / 263 (73%)
Σ	289 / 350 (83%)	273 / 350 (78%)

causal relation
gender is
a **confound**



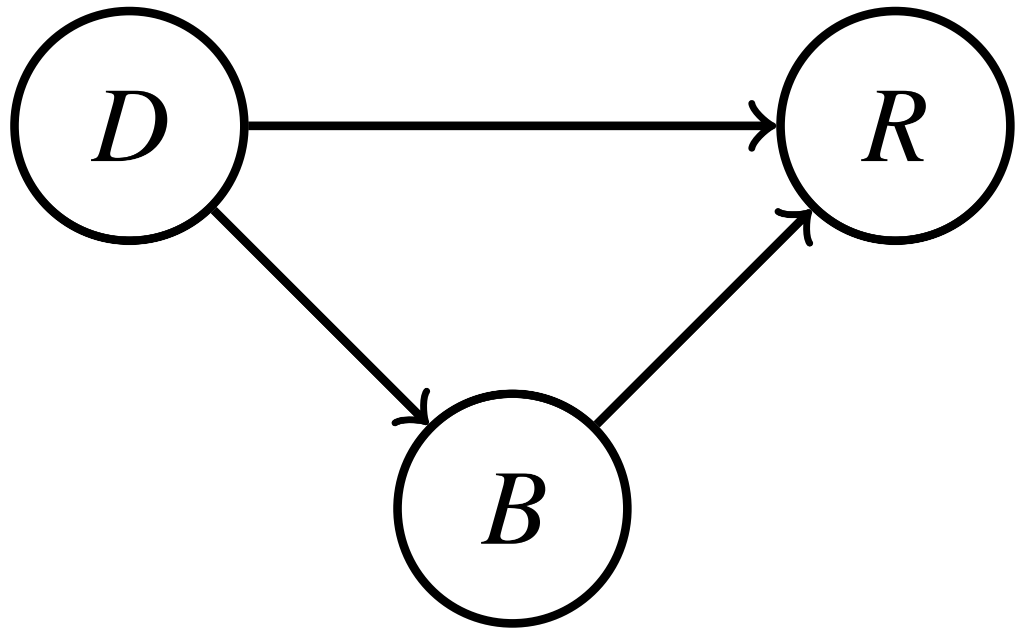
Case 2

Blood-pressure as a mediator

- same as case 1, but no gender info
- also observed:
  post-treatment blood pressure

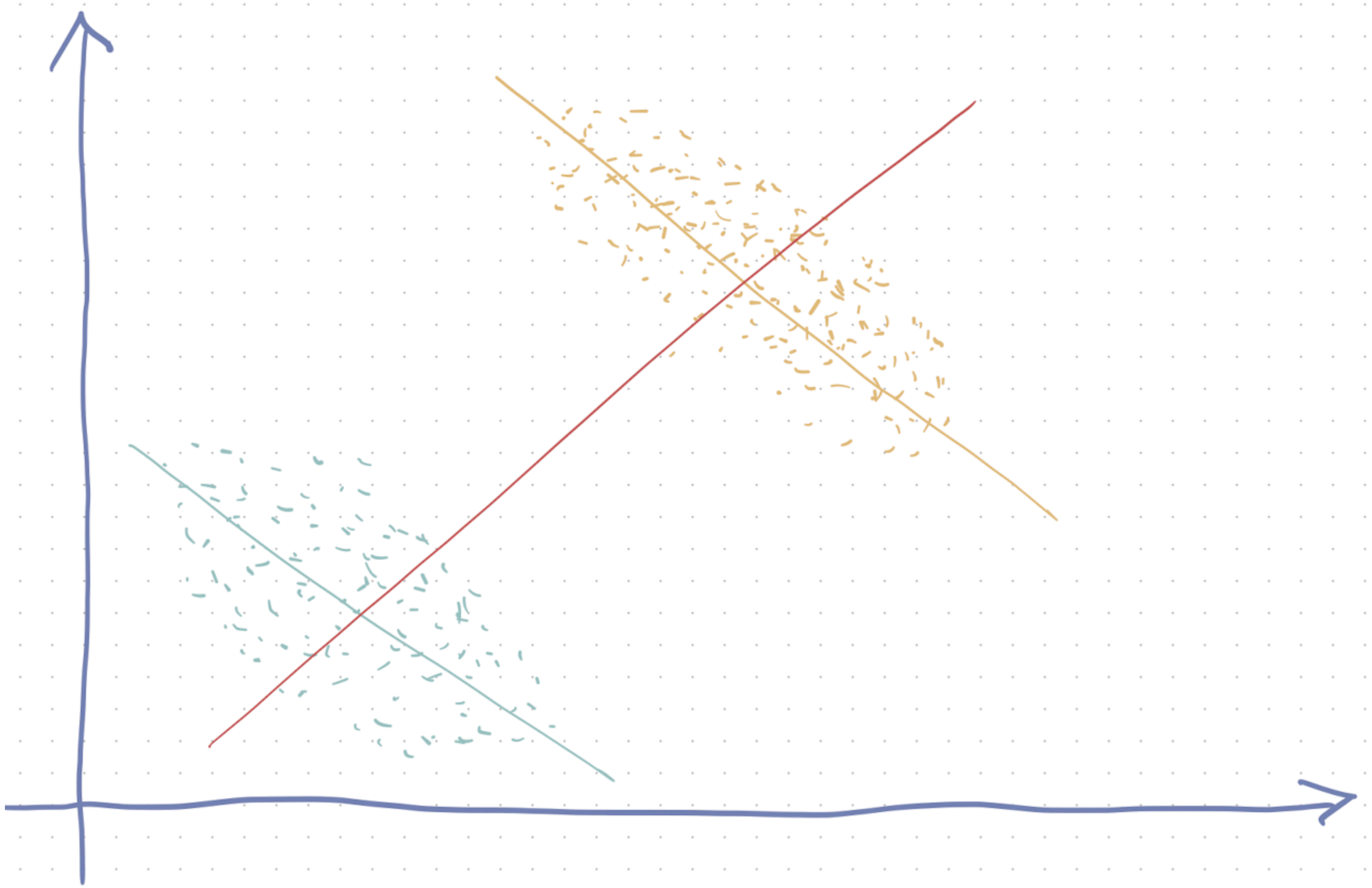
	No drug	Drug
Low BP	234 / 270 (87%)	81 / 87 (93%)
High BP	55 / 80 (68%)	192 / 263 (73%)
Σ	289 / 350 (83%)	273 / 350 (78%)

causal relation
blood pressure is
a **mediator**



Simpson's paradox

- negative correlation in each subgroup, but …
- positive correlation in the whole population



causal models

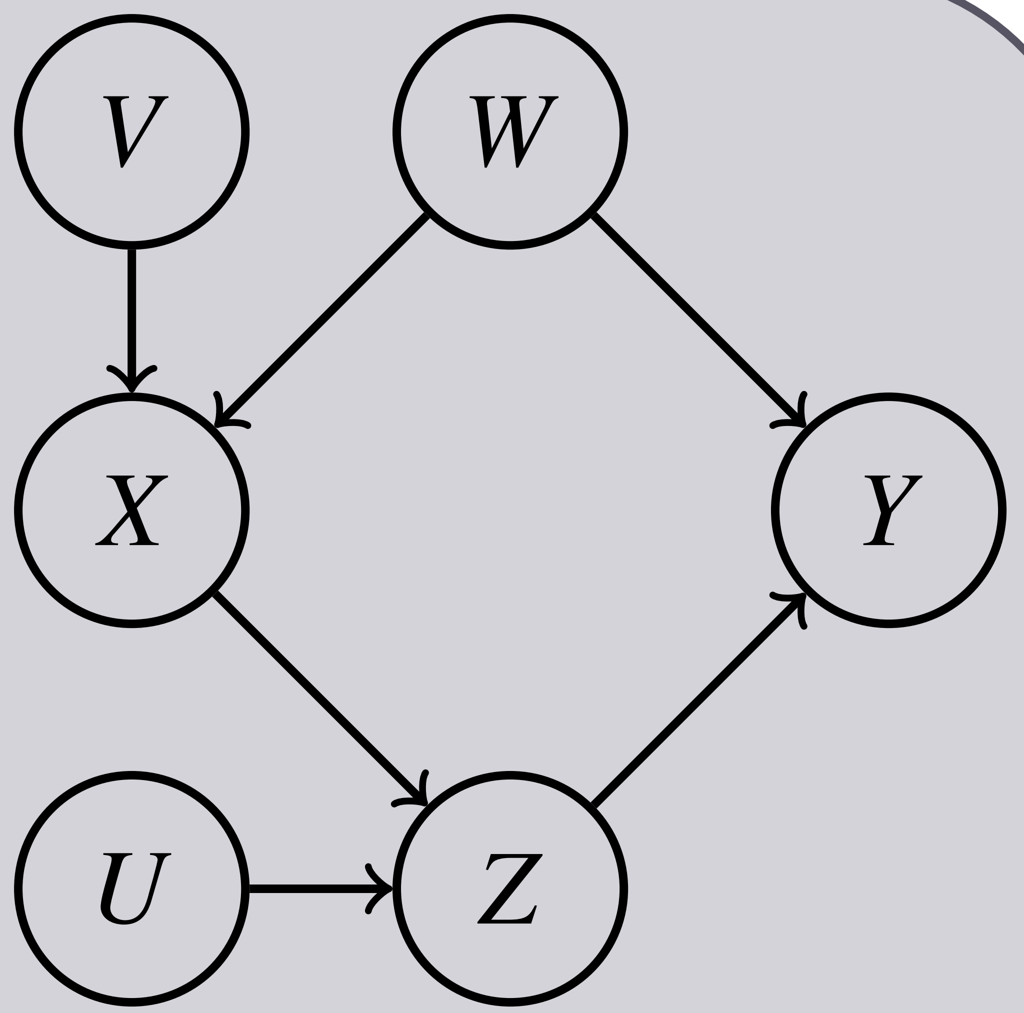
Causal models

intuitive, informal approach

"causal models represent the *mechanism* by which data were generated" (Pearl et. al 2016, p. 36)

- directed acyclic graph (DAG):
  - nodes are variables / events
  - edges indicate direct causal relationship
  - paths indicate (indirect) causal relationship
- beliefs in causal relationships constrain beliefs in stochastic dependency
  - no causal path => stochastic independence
  - single causal path from X to Y via Z => conditional stochastic dependence

example

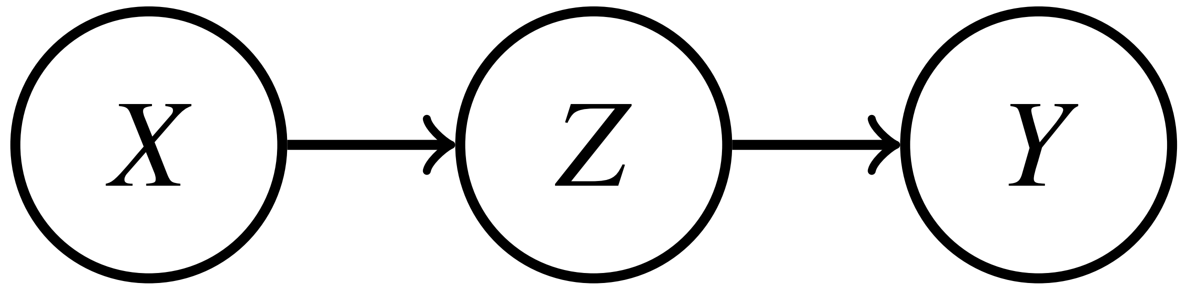


- V and W are independent ($V \perp\!\!\!\perp W$)
- U and Y are independent conditional on Z ($U \perp\!\!\!\perp Y \mid Z$)
- ...

Elementary causal relationships

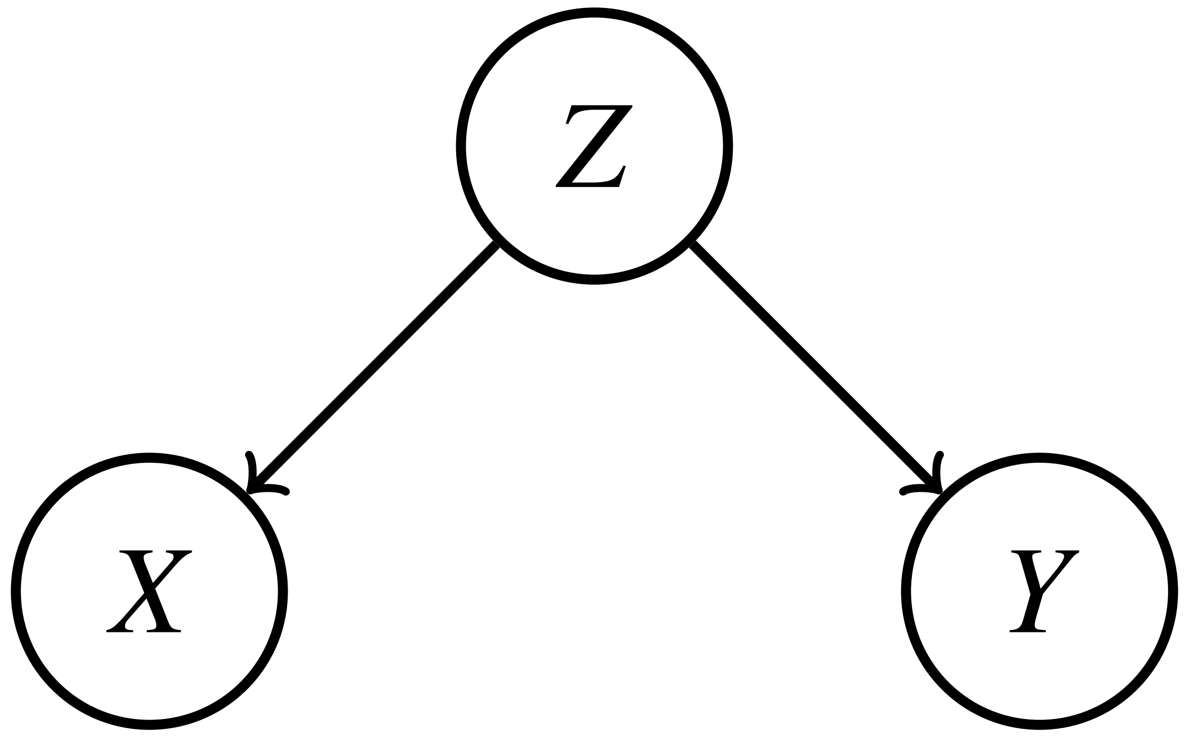
think: “conceptual atoms of complex causal graphs”

Chain



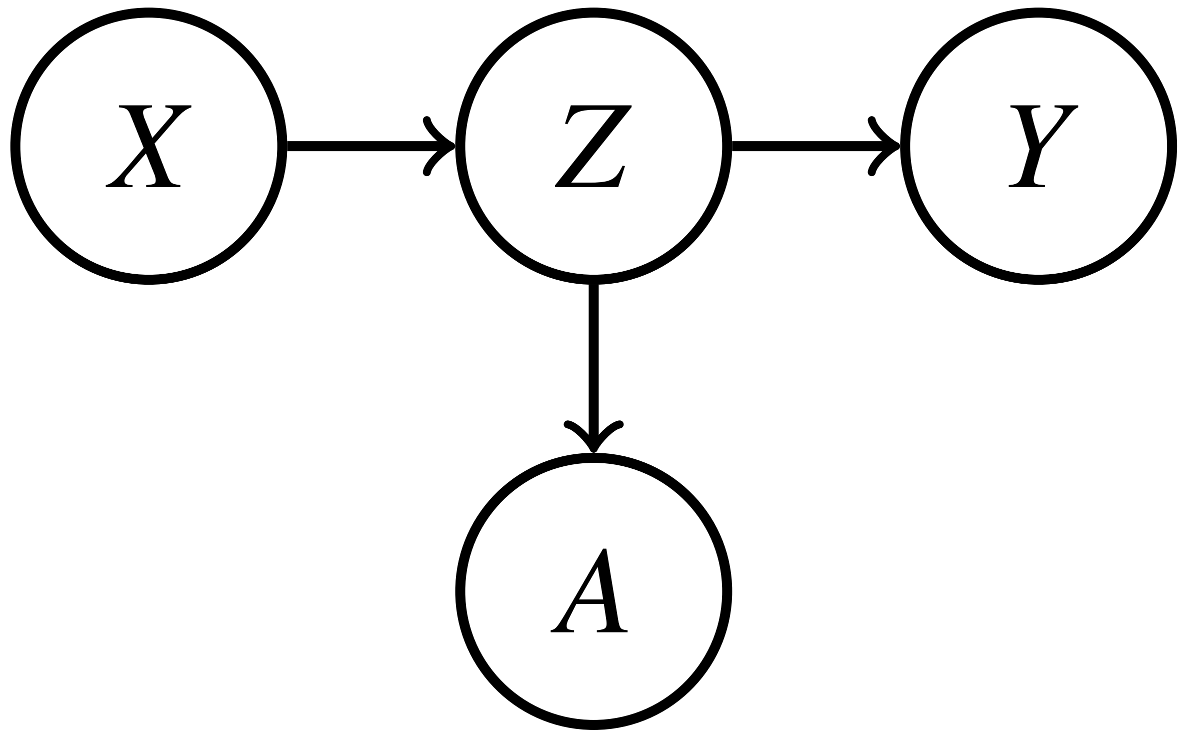
- X & Y independent conditional on Z

Fork



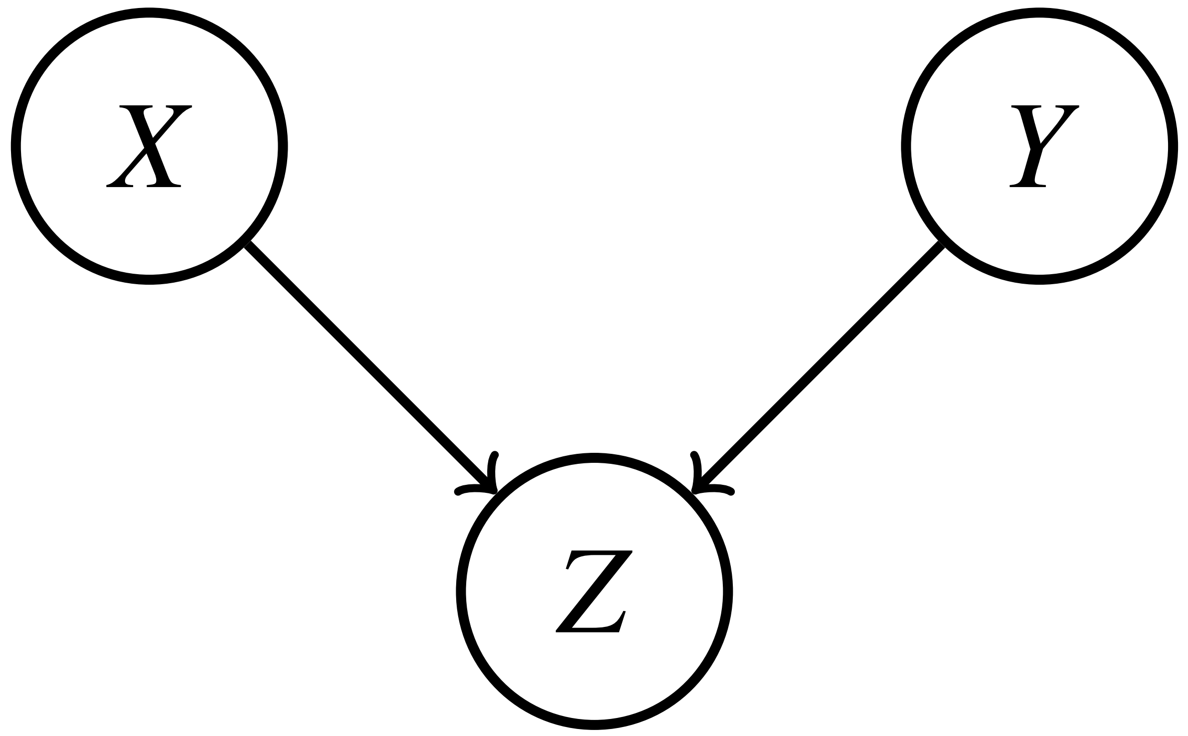
- X & Y stochastically dependent without direct causal relation

Descendant



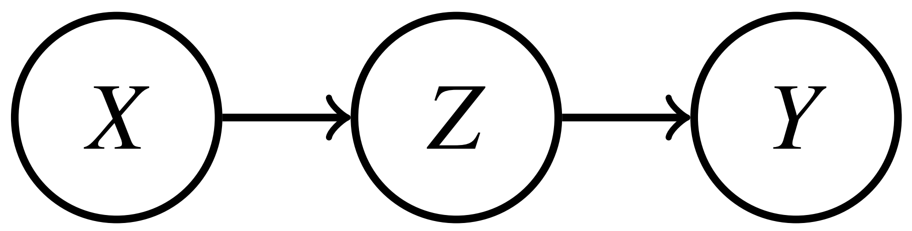
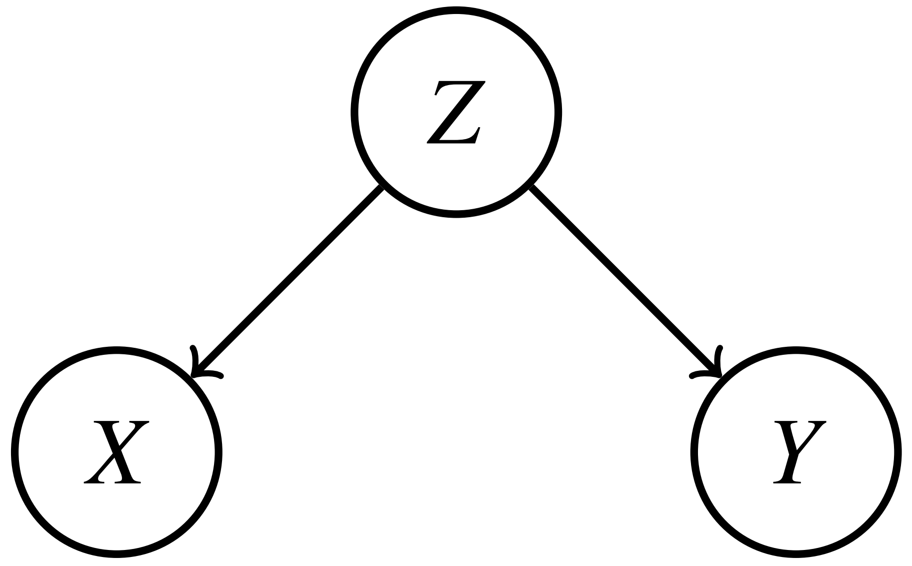
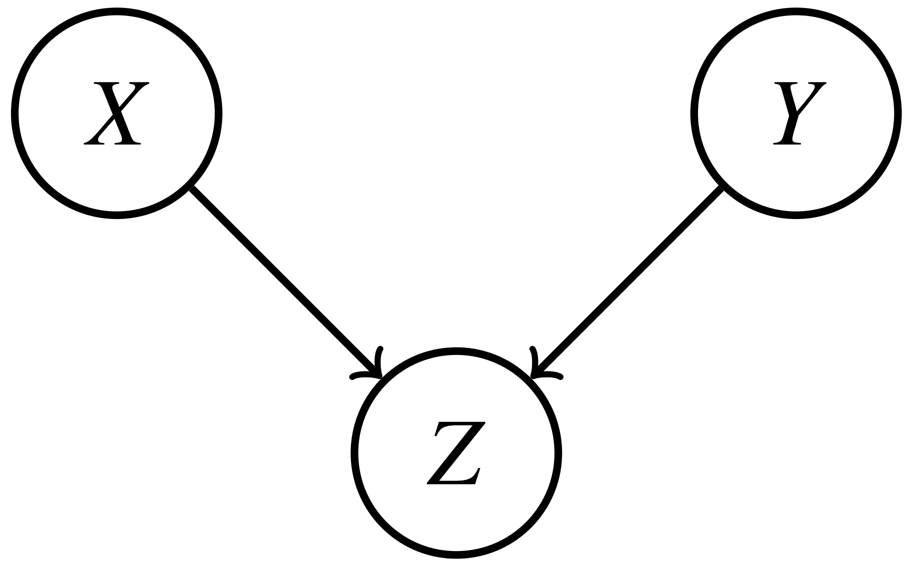
- X & Y independent conditional on A, the more A provides information about Z

Collider



- X & Y independent, but dependent conditional on Z

Causal relation & (conditional) stochastic independence

		$X \perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y \mid Z$
Chain		✗	✓
Fork		✗	✓
Collider		✓	✗

demo

example of using regression to test stochastic (in-)dependence

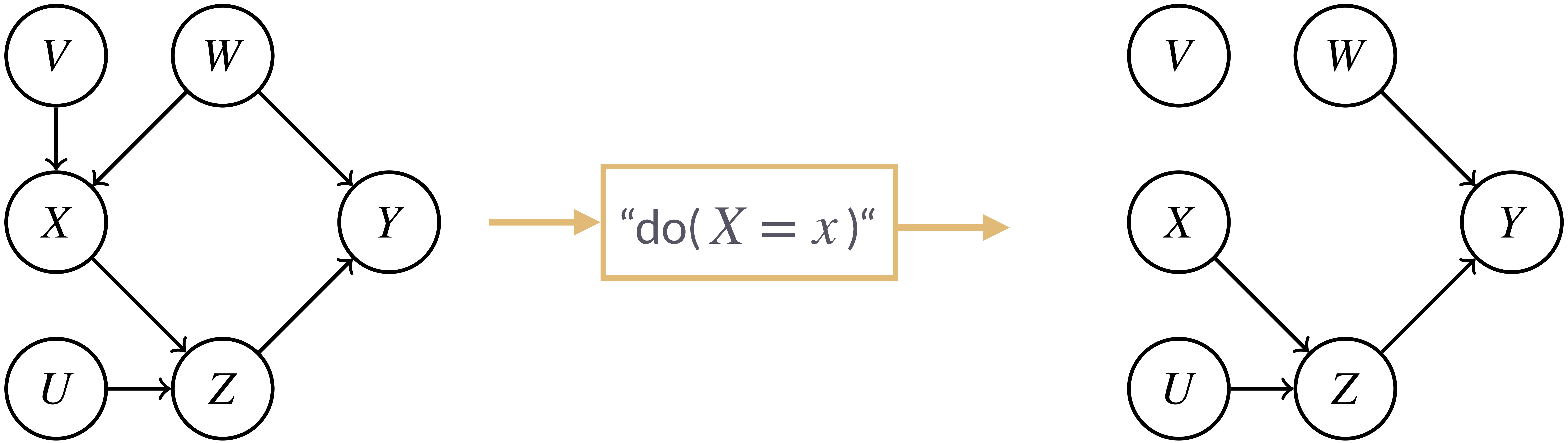
interventions

do-calculus

Formalizing effects of interventions

- new notation $P(Y = y \mid do(X = x))$:
  - probability of $Y = y$ after intervening the causal flow by setting $X = x$
- intervening = pruning:
  - **“doing $X = x$“ => remove all arrows pointing towards X**

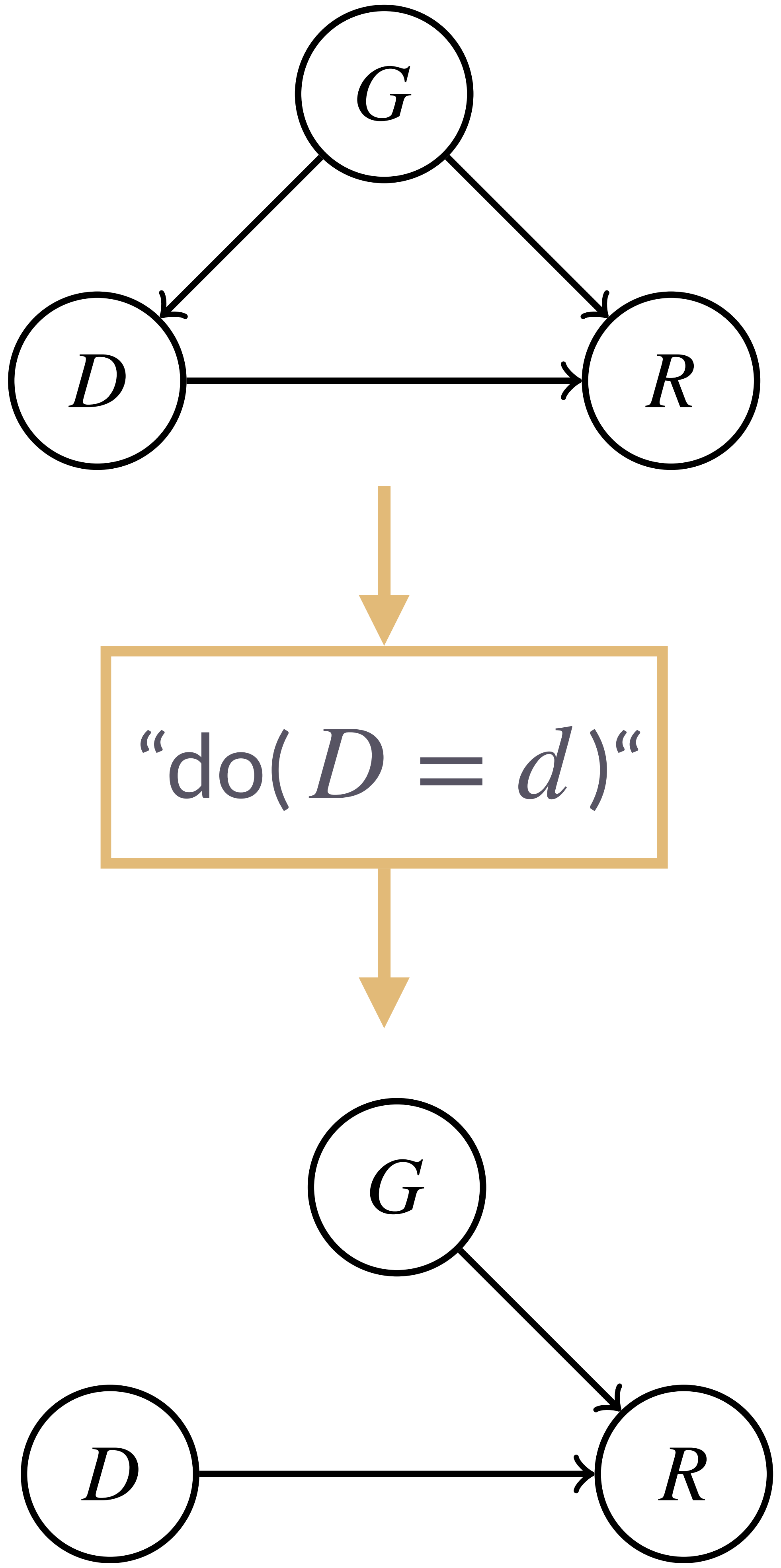
Causal dependencies **Updated graph**



NB: influence of X on Y now passes only via Z, not the confounder W

Example: Simpson's paradox

Case 1: gender as a confounder

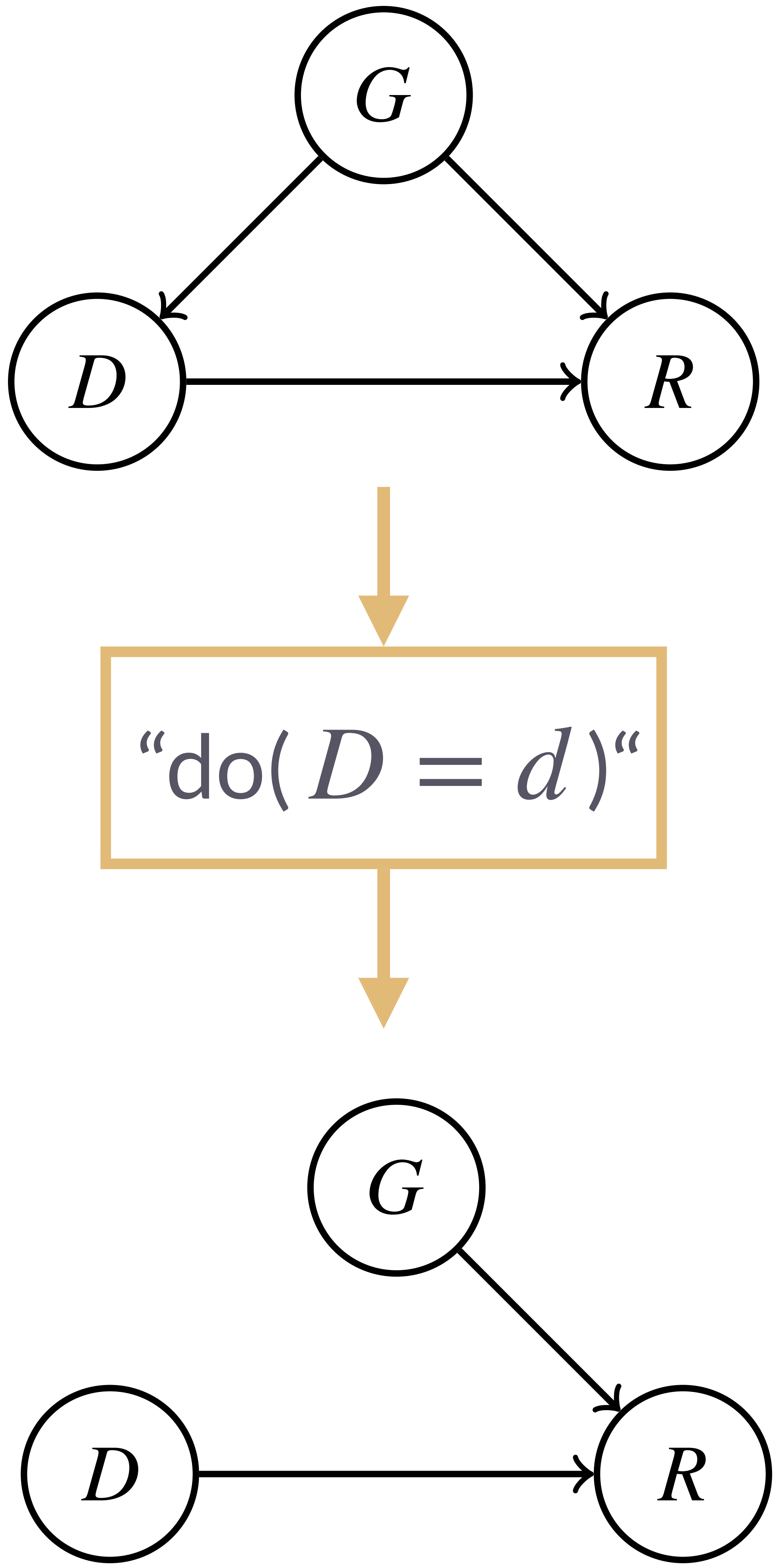


$$
\begin{aligned}
& P\left(R = r \mid do(D = d)\right) \\
&= P^*\left(R = r \mid D = d\right) && \text{[by definition]} \\
&= \sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g \mid D = d\right) && \text{[rules of prob.]} \\
&= \sum_g P^*\left(R = r \mid D = d, G = g\right)\ P^*\left(G = g\right) && \text{[independence]} \\
&= \sum_g P\left(R = r \mid D = d, G = g\right)\ P\left(G = g\right) && \text{[not affected by "do"]}
\end{aligned}
$$

!!! It is possible to express effects of "do"-intervention in terms of observational probabilities alone!

Example: Simpson’s paradox

Case 1: gender as a confounder



$$P\left(R = r \mid do(D = d)\right)$$

$$= \sum_{g} P\left(R = r \mid D = d, G = g\right)\; P\left(G = g\right)$$

	No drug	Drug
Men	234 / 270 (87%)	81 / 87 (93%)
Women	55 / 80 (68%)	192 / 263 (73%)
Σ	289 / 350 (83%)	273 / 350 (78%)

$$P\left(R = 1 \mid do(D = 1)\right) \approx 0.83$$

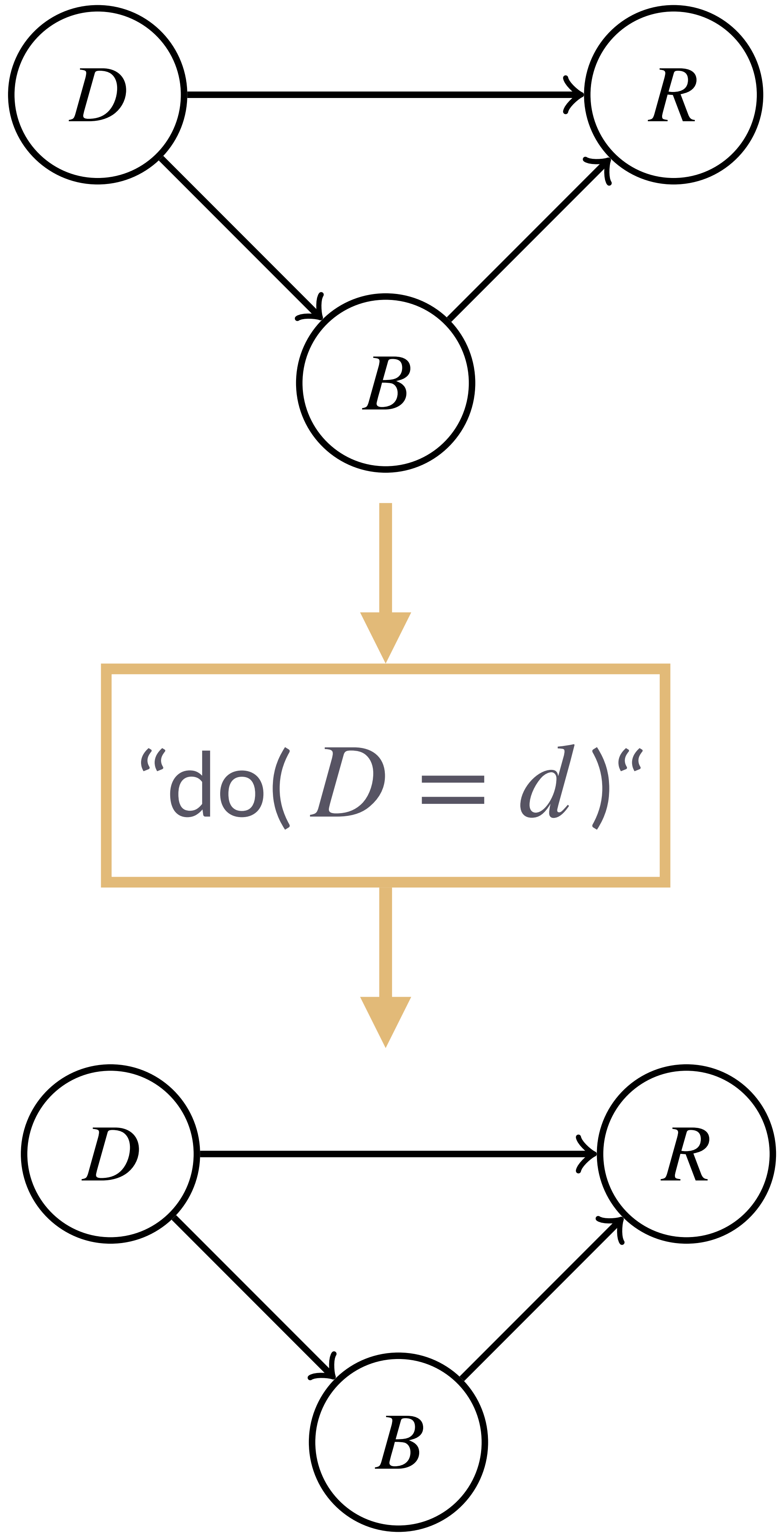
$$P\left(R = 1 \mid do(D = 0)\right) \approx 0.78$$

ML-estimate of **causal effect**

$0.83 - 0.78 = 0.05$

Example: Simpson's paradox

Case 2: blood pressure as a mediator



$$P\left(R = r \mid do(D = d)\right)$$

$$= P^*\left(R = r \mid D = d\right) \quad \text{[by definition]}$$

$$= \sum_b P^*\left(R = r \mid D = d, B = b\right)\ P^*\left(B = b \mid D = d\right) \quad \text{[rules of prob.]}$$

$$\cancel{= \sum_b P^*\left(R = r \mid D = d, B = b\right)\ P^*\left(B = b\right) \quad \text{[independence]}}$$

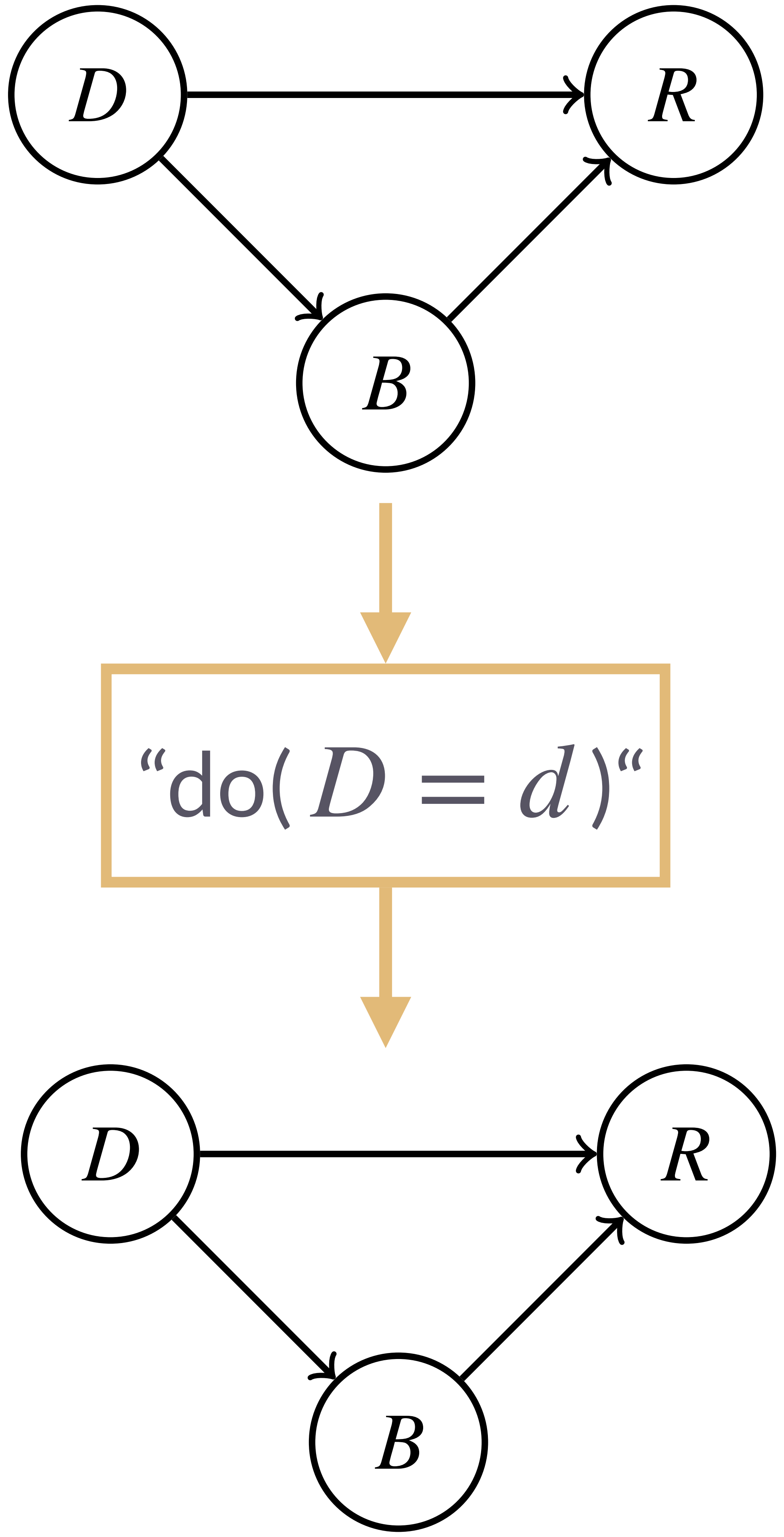
$$= \sum_b P\left(R = r \mid D = d, B = b\right)\ P\left(B = b \mid D = d\right)$$

$$= P\left(R = r \mid D = d\right)$$

Here "do"-intervention does not "create independency".
And it is not necessary at all!

Example: Simpson's paradox

Case 2: blood pressures as a mediator



$$P\left(R = r \mid do(D = d)\right)$$

$$= \sum_b P\left(R = r \mid D = d, G = g\right) \; P\left(G = g \mid B = b\right)$$

	No drug	Drug
Low BP	234 / 270 (87%)	81 / 87 (93%)
High BP	55 / 80 (68%)	192 / 263 (73%)
Σ	289 / 350 (83%)	273 / 350 (78%)

$$P\left(R = 1 \mid do(D = 1)\right) \approx 0.78$$

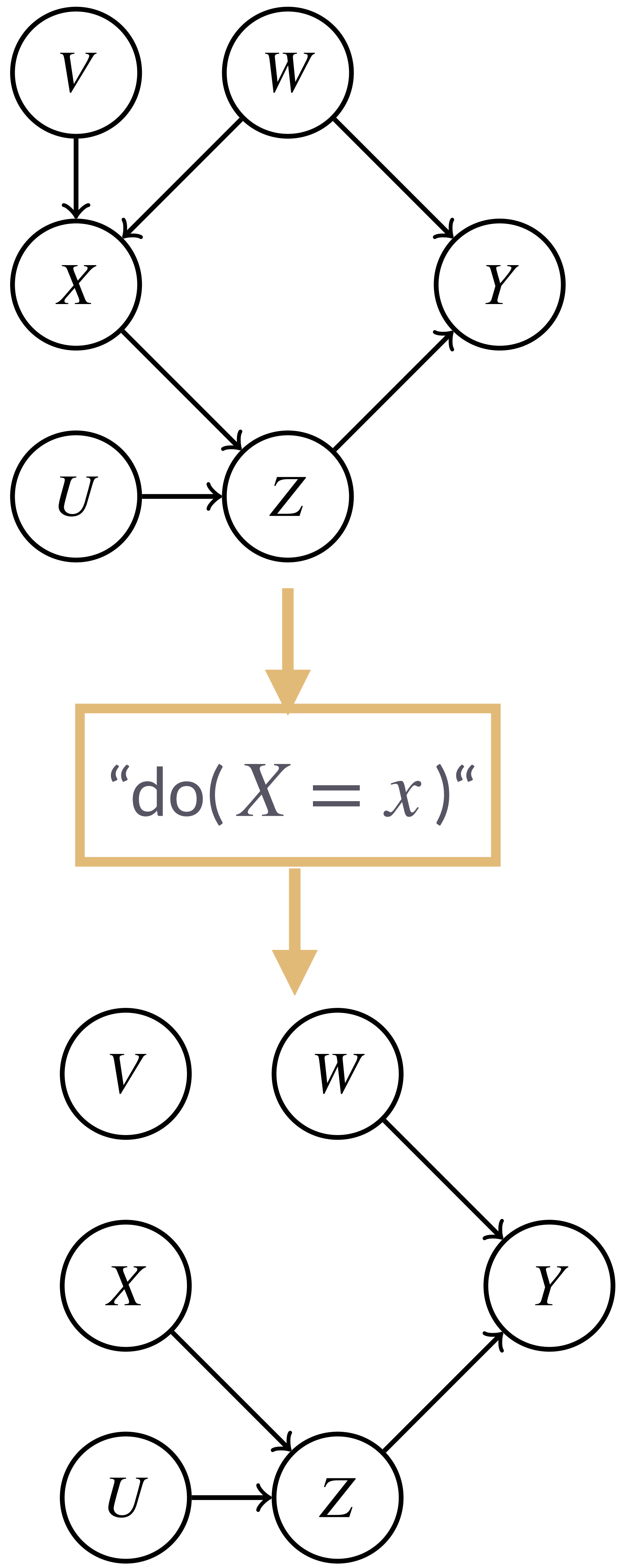
$$P\left(R = 1 \mid do(D = 0)\right) \approx 0.83$$

ML-estimate of
causal effect
$0.78 - 0.83 = -0.05$

Intervention

summary

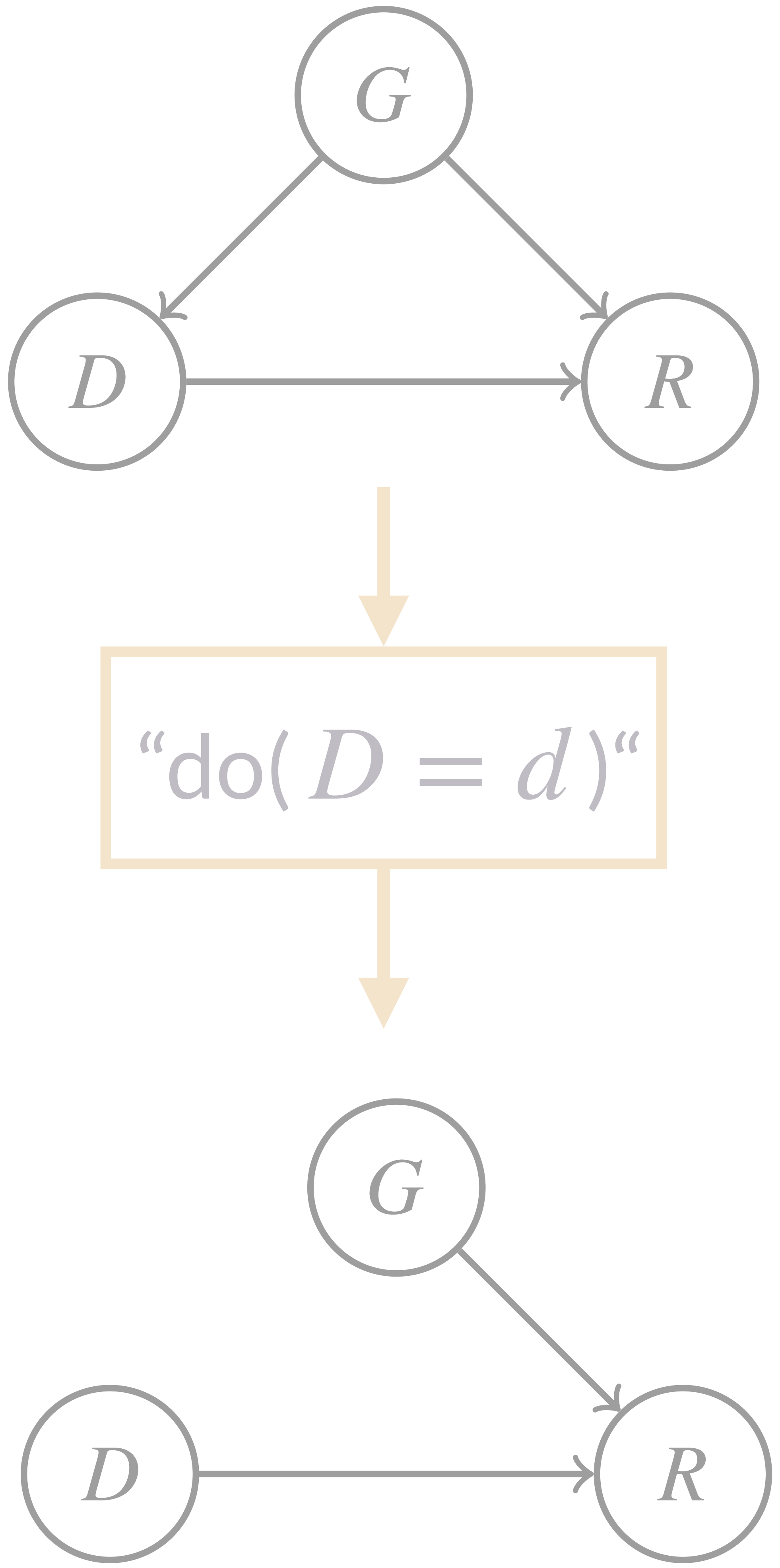
- causal intuitions go beyond stochastic dependence, and imply intuitions about **interventions**
- intervening = pruning:
  - "doing $X = x$" entails removing all arrows pointing towards X
- new formal notion: $P(Y = y \mid do(X = x))$
- sometimes we can express $P(Y = y \mid do(X = x))$ in terms of "normal" conditional probabilities; sometimes we cannot
- follow-up question: when and how can we eliminate "*do(X)*"?



causal effects w/ regression modeling

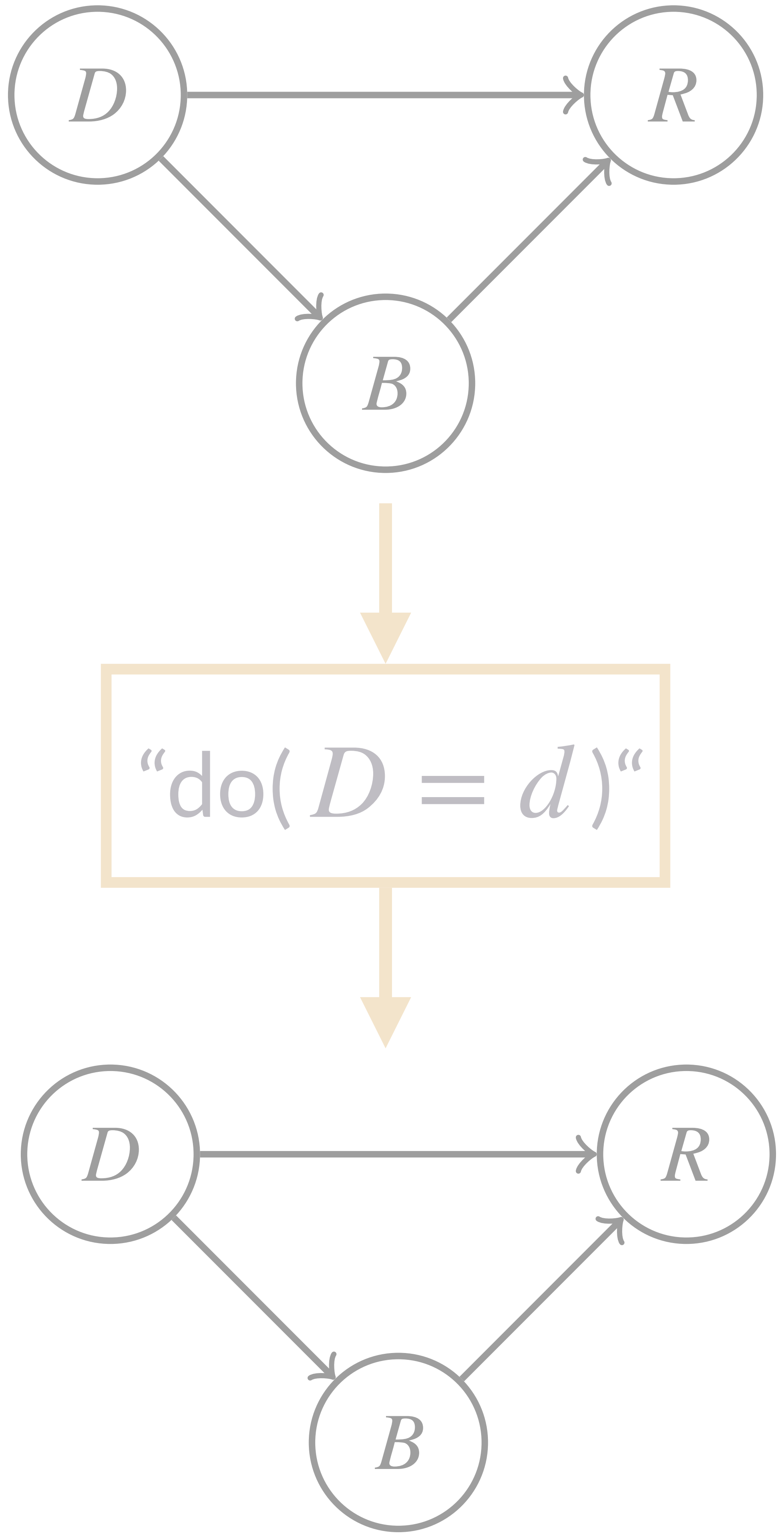
MLE of causal effect

Gender as a confounder



ML-estimate of
causal effect
$0.83 - 0.78 = 0.05$

Blood-pressure as a mediator



ML-estimate of
causal effect
$0.78 - 0.83 = -0.05$

Prepare data

Simpson's paradox

```r
##################################################
# set up the data for SP
##################################################

data_simpsons_paradox <- tibble(
  gender = c("Male", "Male", "Female", "Female"),
  bloodP = c("Low", "Low", "High", "High"),
  drug   = c("Take", "Refuse", "Take", "Refuse"),
  k      = c(81, 234, 192, 55),
  N      = c(87, 270, 263, 80),
  proportion = k/N
)
```

```r
# cast into long format
data_SP_long <- rbind(
  data_simpsons_paradox |> uncount(k) |>
    mutate(recover = TRUE)  |> select(-N, -proportion),
  data_simpsons_paradox |> uncount(N-k) |>
    mutate(recover = FALSE) |> select(-N, -proportion, -k)
)
data_SP_long
```

```
# A tibble: 700 × 4
   gender bloodP drug  recover
   <chr>  <chr>  <chr> <lgl>
 1 Male   Low    Take  TRUE
 2 Male   Low    Take  TRUE
 3 Male   Low    Take  TRUE
 4 Male   Low    Take  TRUE
 5 Male   Low    Take  TRUE
 6 Male   Low    Take  TRUE
 7 Male   Low    Take  TRUE
 8 Male   Low    Take  TRUE
 9 Male   Low    Take  TRUE
10 Male   Low    Take  TRUE
# ℹ 690 more rows
```

Calculating the total causal effect

Case 1: gender as a confound

We want:

$$P(R = 1 \mid do(D = d)) = \sum_{g \in \{0,1\}} P(R = 1 \mid D = d, G = g)\ P(G = g)$$

We do:

1. estimate $P(G)$
   - use intercept-only logistic regression `G ~ 1`
2. estimate $P(R = 1 \mid D = d, G = g)$:
   - use logistic regression model `R ~ D * G`
3. calculate TCE with posterior predictive distributions of these models

Calculating TCE

Case 1: gender as a confound

Step 1: G ~ 1

```
niter = 2000

fit_SP_GonIntercept <- brm(
  formula = gender ~ 1,
  data    = data_SP_long,
  family  = bernoulli(link = "logit"),
  iter    = niter
)
```

Step 2: R ~ D * G

```
fit_SP_RonGD <- brm(
  formula = recover ~ gender * drug,
  data    = data_SP_long,
  family  = bernoulli(link = "logit"),
  iter    = niter
)
```

Step 3:

sample participants

```
postPred_gender <- tidybayes::predicted_draws(
  object  = fit_SP_GonIntercept,
  newdata = tibble(Intercept = 1),
  value   = "gender",
  ndraws  = niter * 2
  ) |>
  ungroup() |>
  mutate(gender = ifelse(gender, "Male", "Female")) |>
  select(gender)
```

do(drug = 1)

```
# posterior predictive samples for D=1
posterior_DrugTaken <- tidybayes::epred_draws(
  object  = fit_SP_RonGD,
  newdata = postPred_gender |> mutate(drug = "Take"),
  value   = "taken",
  ndraws  = niter * 2
) |> ungroup() |>
  select(taken)
```

do(drug = 0)

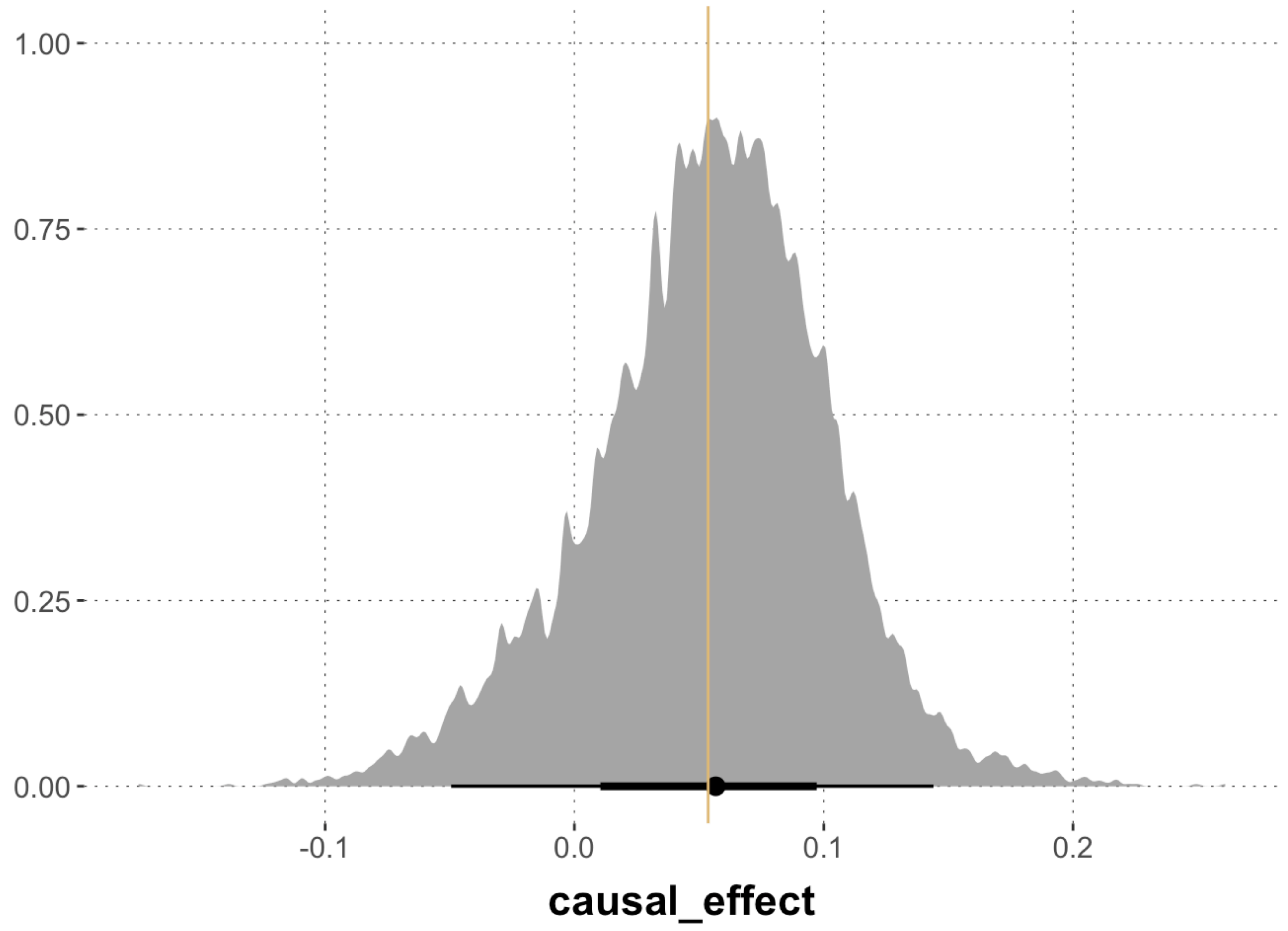
```
# posterior predictive samples for D=0
posterior_DrugRefused <- tidybayes::epred_draws(
  object  = fit_SP_RonGD,
  newdata = postPred_gender |> mutate(drug = "Refuse"),
  value   = "refused",
  ndraws  = niter * 2
) |> ungroup() |>
  select(refused)
```

Calculating TCE

Case 1: gender as a confound

```
# A tibble: 3 × 4
  Parameter      `|95%`   mean `95%|`
  <chr>           <dbl>  <dbl>  <dbl>
1 drug_taken     0.687  0.834   0.972
2 drug_refused   0.614  0.780   0.907
3 causal_effect -0.0505 0.0540  0.142
```

ML-estimate of
causal effect
$0.83 - 0.78 = 0.05$



Calculating TCE

Case 2: blood-pressure as mediator

We want:

$$P\left(R = r \mid do(D = d)\right) = P\left(R = r \mid D = d\right)$$

We do:

```r
fit_SP_RonBD <- brms::brm(
  formula = recover ~ drug,
  data    = data_SP_long,
  family  = bernoulli(link = "logit"),
  iter    = niter
)
```

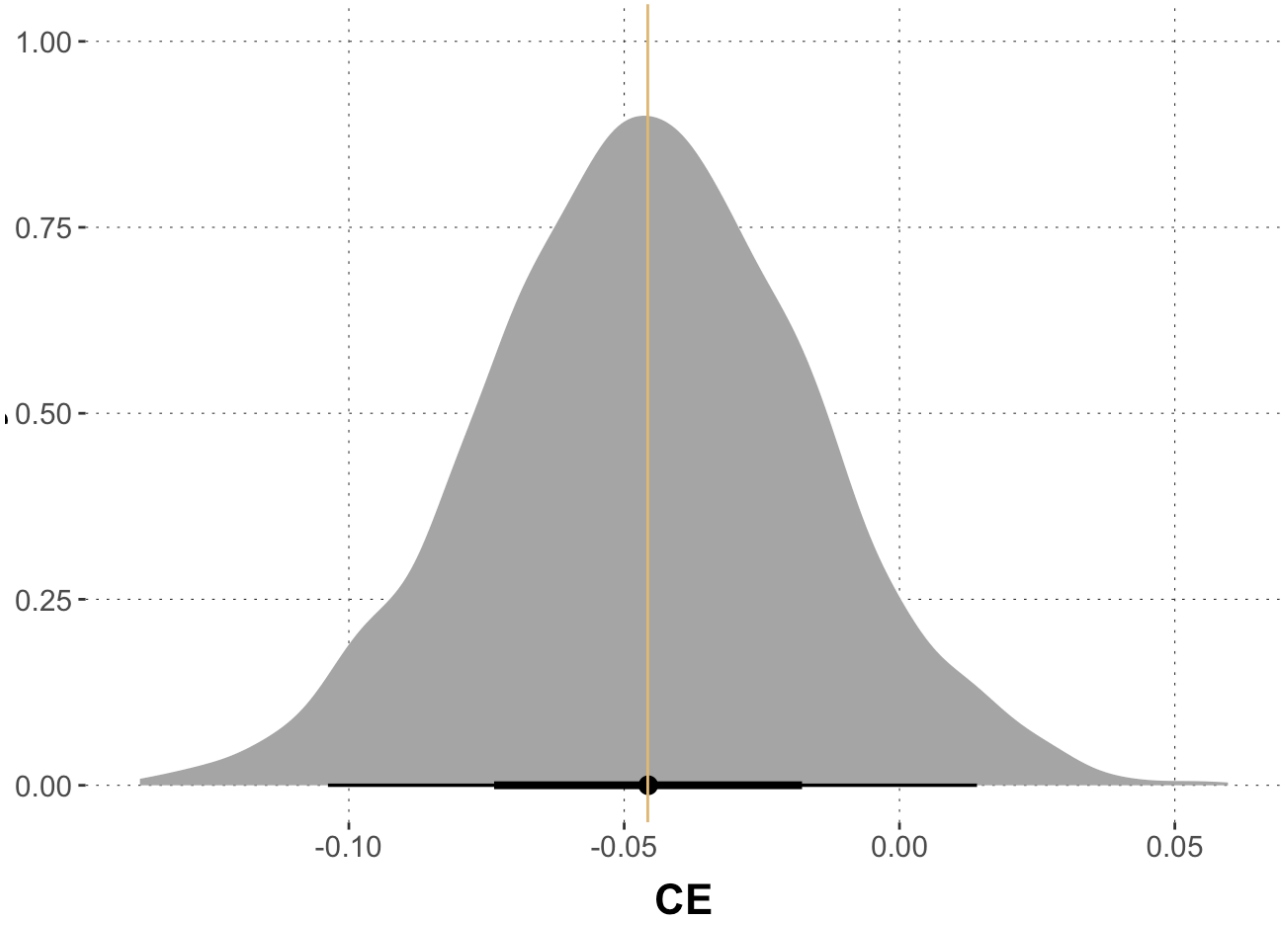
```r
posterior_DrugTaken <-
  faintr::extract_cell_draws(fit_SP_RonBD, drug == "Take") |>
  pull(draws) |>
  logistic()

posterior_DrugRefused <-
  faintr::extract_cell_draws(fit_SP_RonBD, drug == "Refuse") |>
  pull(draws) |>
  logistic()

posterior_causalEffect <-
  posterior_DrugTaken - posterior_DrugRefused
```

Calculating TCE

Case 2: blood-pressure as mediator

```
# A tibble: 3 × 4
  Parameter     `|95%`    mean `95%|`
  <chr>          <dbl>   <dbl>  <dbl>
1 drug_taken     0.734  0.780  0.822
2 drug_refused   0.786  0.825  0.865
3 causal_effect -0.103 -0.0457 0.0146
```

ML-estimate of
causal effect
$$0.78 - 0.83 = -0.05$$



Causal inference w/ Bayesian regression

summary

- do-calculus tells us when and how we can draw "causal conclusions" from observational data
  - we must specify a causal model
  - readily applicable criteria exist: backdoor, front-door
- uncertainty about causal effect is quantifiable using Bayesian regression modeling
  - but (currently) requires manual labor