Statistical Testing

Lauren Vanderlinden, PhD, MS
T15 Postdoctoral Fellow Computational Biology
Division of Rheumatology & Department of Biomedical Informatics
School of Medicine, University of Colorado Anschutz Medical Campus

CPBS 7602 - November 27, 2024

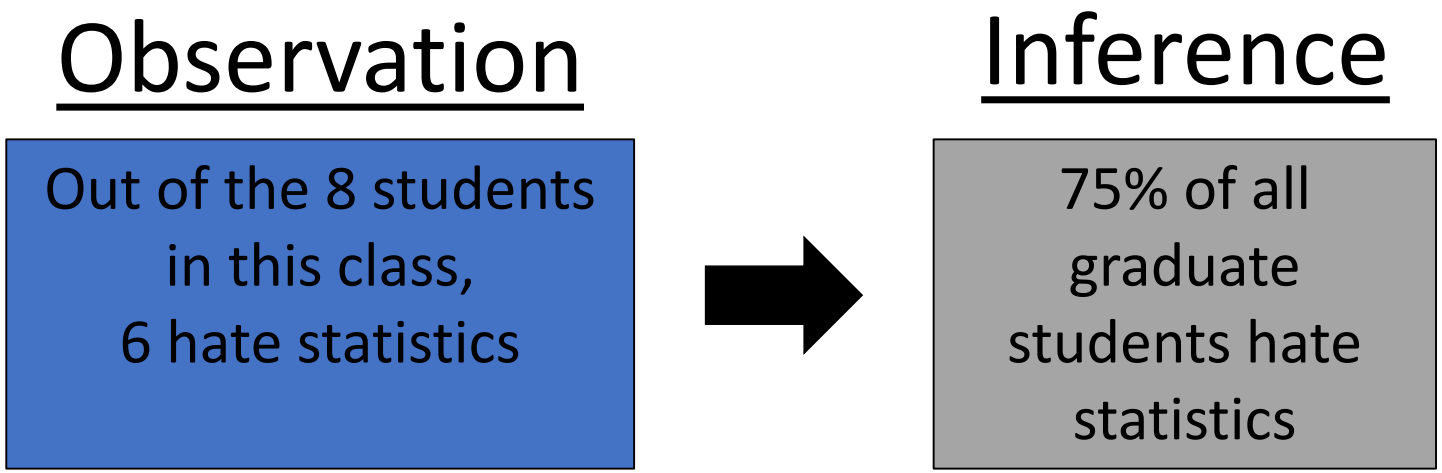
Outline

- Inference
- Populations and sampling
- Study designs
- Directed acyclic graphs (causal diagrams)
- Hypothesis testing
- Effect sizes and p-values
- Power and type II error rate
- Bayesian statistics

Inference

Inference

- An **inference** is a conclusion that patterns in the data (sample) are present in a larger context (population).



Types of Inference

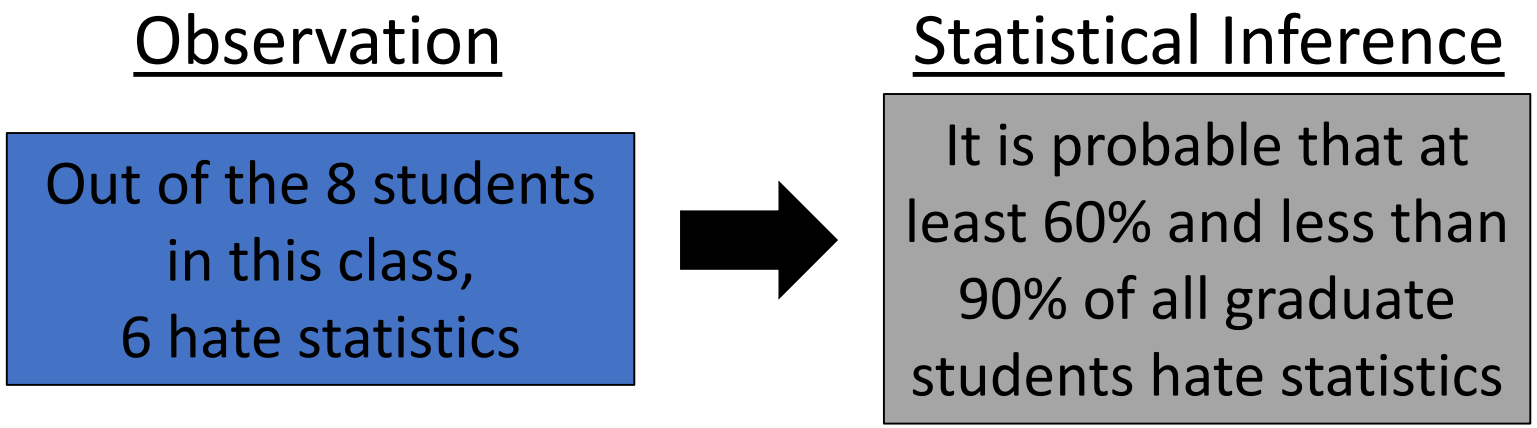
- **Statistical inference** is an inference justified using statistical methods.
- **Scope of inference** is the group of objects to which my conclusion (results) extend.
- **Causal inference** is drawing a cause-and-effect conclusion between an explanatory variable and a response variable.

Statistical Inference

Inference

Statistical Inference

- **Statistical inference** is an inference justified using statistical methods.
  - Statistical inference allows us to *quantify the uncertainty* in our conclusions.



Scope of Inference

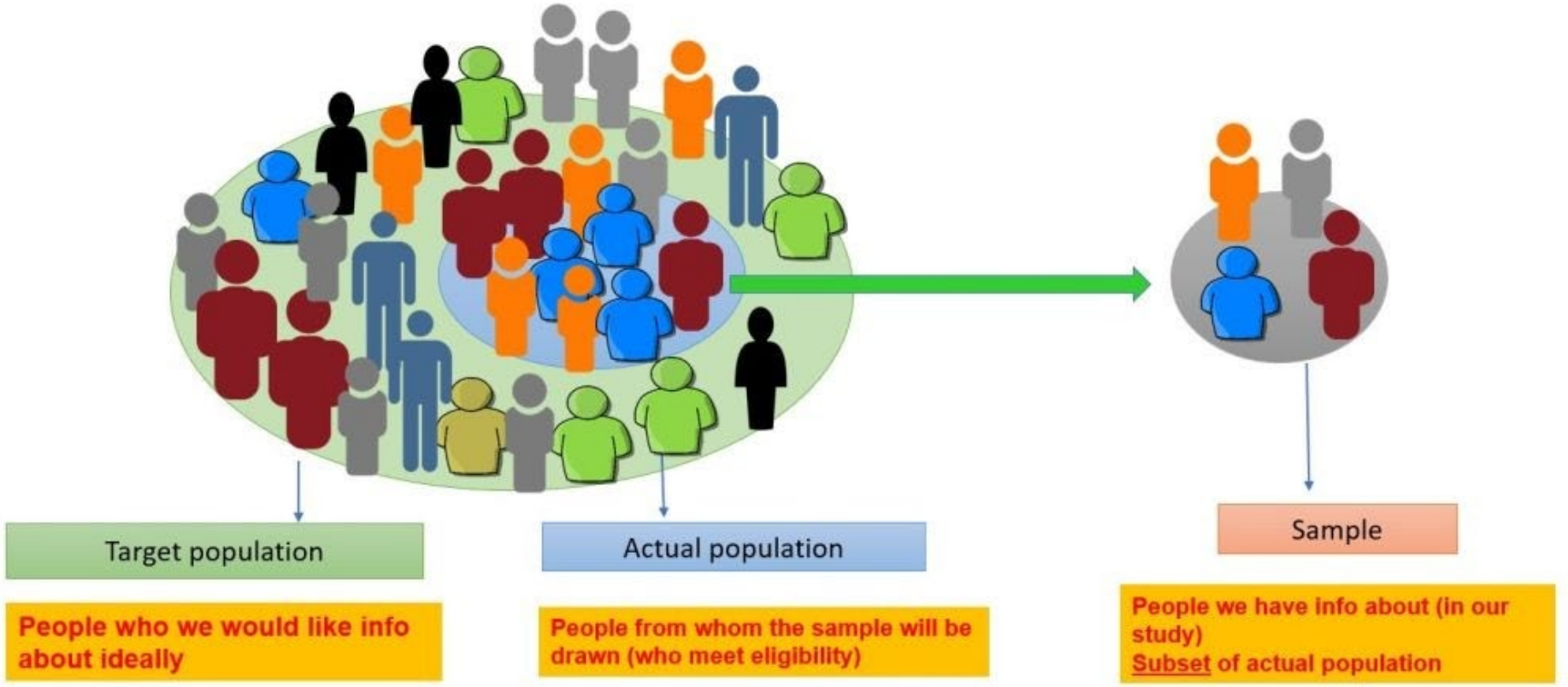
Inference

Populations & Samples

Researchers are typically interested in finding results that apply to an entire population of people or things.

- Population
  - The collection of units (be they people, plankton, plants, cities, suicidal authors, etc.) to which we want to generalize a set of findings or a statistical model.
- Sample
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that population.

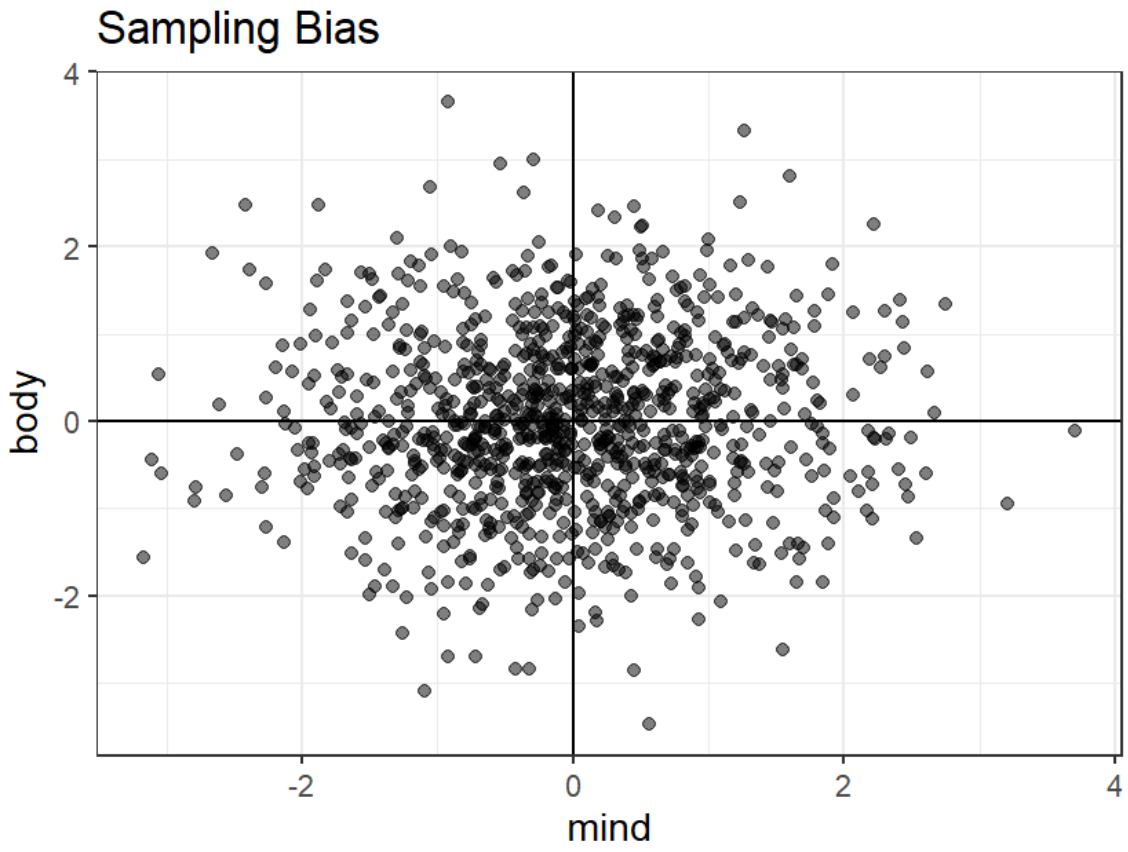
Sampling Figure



https://www.linkedin.com/pulse/population-sample-statistics-alka-pandey/

Sampling Gone Wrong

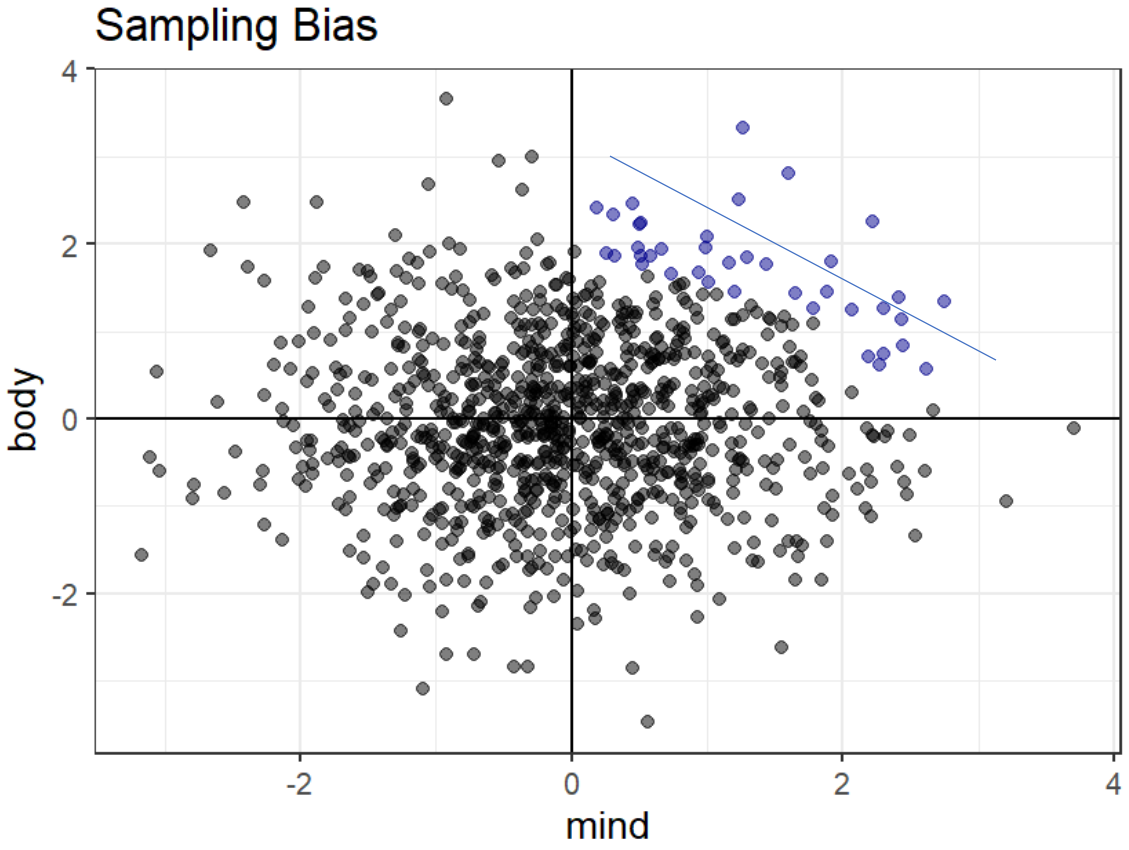
- If good looks and smarts are distributed normally
- If good looks and smarts **have nothing to do with each other**



http://euri.ca/2012/youre-probably-polluting-your-statistics-more-than-you-think/index.html
http://www.theatlantic.com/business/archive/2012/05/when-correlation-is-not-causation-but-something-much-more-screwy/256918/

Sampling Gone Wrong

- If good looks and smarts are distributed normally
- If good looks and smarts **have nothing to do with each other**
- If movie producers want both smarts and looks. Then, by observing **employed actors (blue)** we'll assume that looks and smarts have a **negative correlation**



http://euri.ca/2012/youre-probably-polluting-your-statistics-more-than-you-think/index.html
http://www.theatlantic.com/business/archive/2012/05/when-correlation-is-not-causation-but-something-much-more-screwy/256918/

Selection Bias

- Bias results when different types of people have different probabilities of getting into the study, or staying in the study
- Participants have different probabilities of being included *or retained* in a study based on their **exposure** and/or **disease status**
- The sample obtained is not **representative** of the population to which we would like to generalize the results

Scope of Inference

The **scope of inference** is the group of individuals to whom the statistical conclusions can be extended.

- Inferences to populations can be only drawn from random sampling studies.
  - A **random sampling** study is when units are randomly selected from a well-defined population.
  - Random sampling typically ensures that all subpopulations are represented in the sample in roughly the same proportion as the population.
  - Our statistical procedures take into account that sometimes the sample may not be a very good mix of the population.

Causal Inference

Inference

The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

"Switzerland was the top performer in terms of both the number of Nobel laureates and chocolate consumption. The slope of the regression line allows us to estimate that **it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1**. For the United States, that would amount to 125 million kg per year. The minimally effective chocolate dose seems to hover around 2 kg per year, and the dose–response curve reveals no apparent ceiling on the number of Nobel laureates at the highest chocolate-dose level of 11 kg per year."

Types of Studies

- **Experiments** are studies in which we manipulate one or more explanatory variables (e.g., treatments) to see the effect they have on another variable.
    - In a **randomized experiment** the investigator uses a chance mechanism to assign experimental units to various treatment groups
- **Observational studies** are studies in which the data are measured through observation of the world as it naturally occurs.
    - Grouping (i.e., explanatory variable) occurs naturally and is not assigned

Case reports

Case series

Ecologic studies

Cross-sectional studies

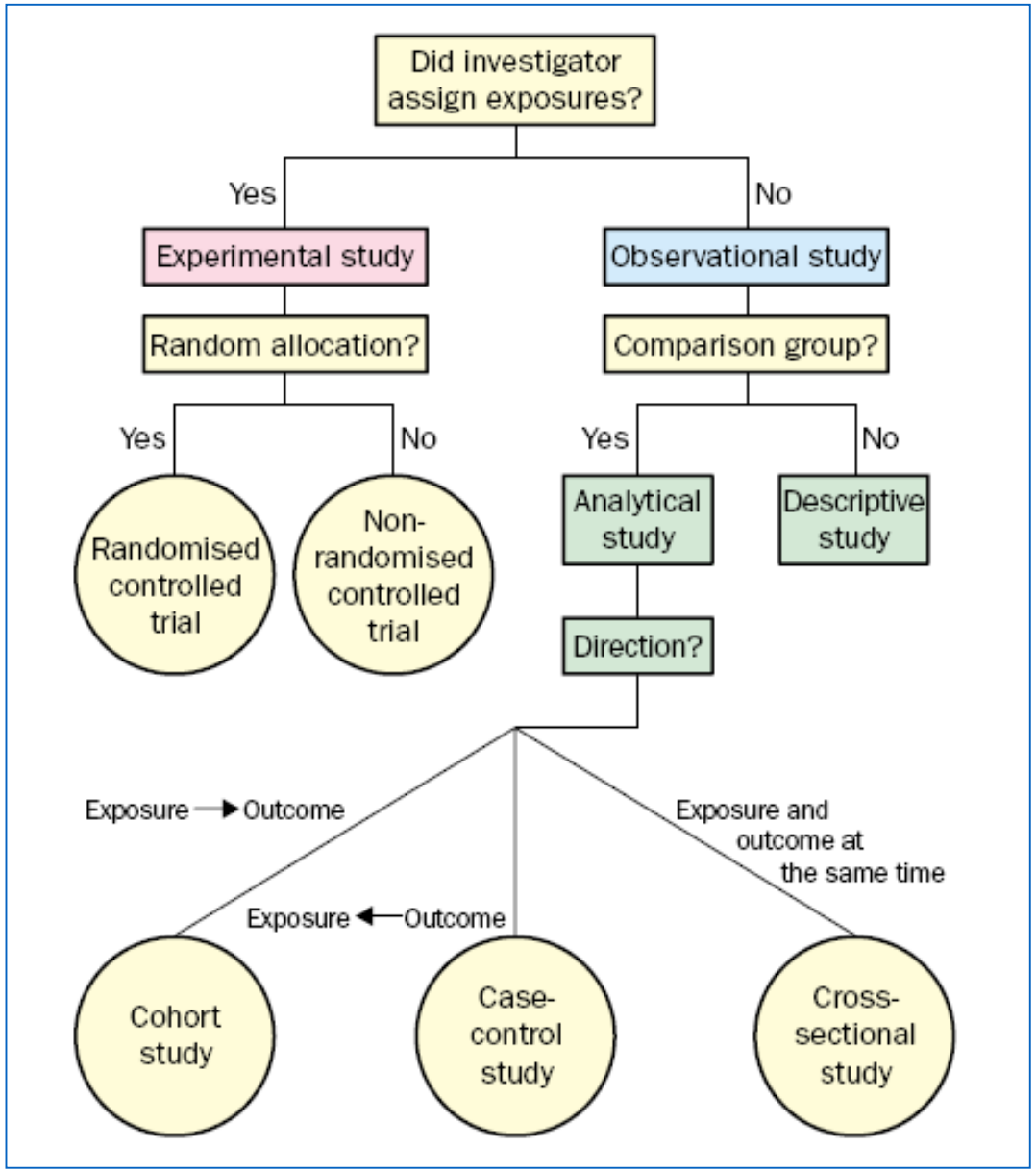
Case-control studies

Cohort studies

Randomized controlled trials

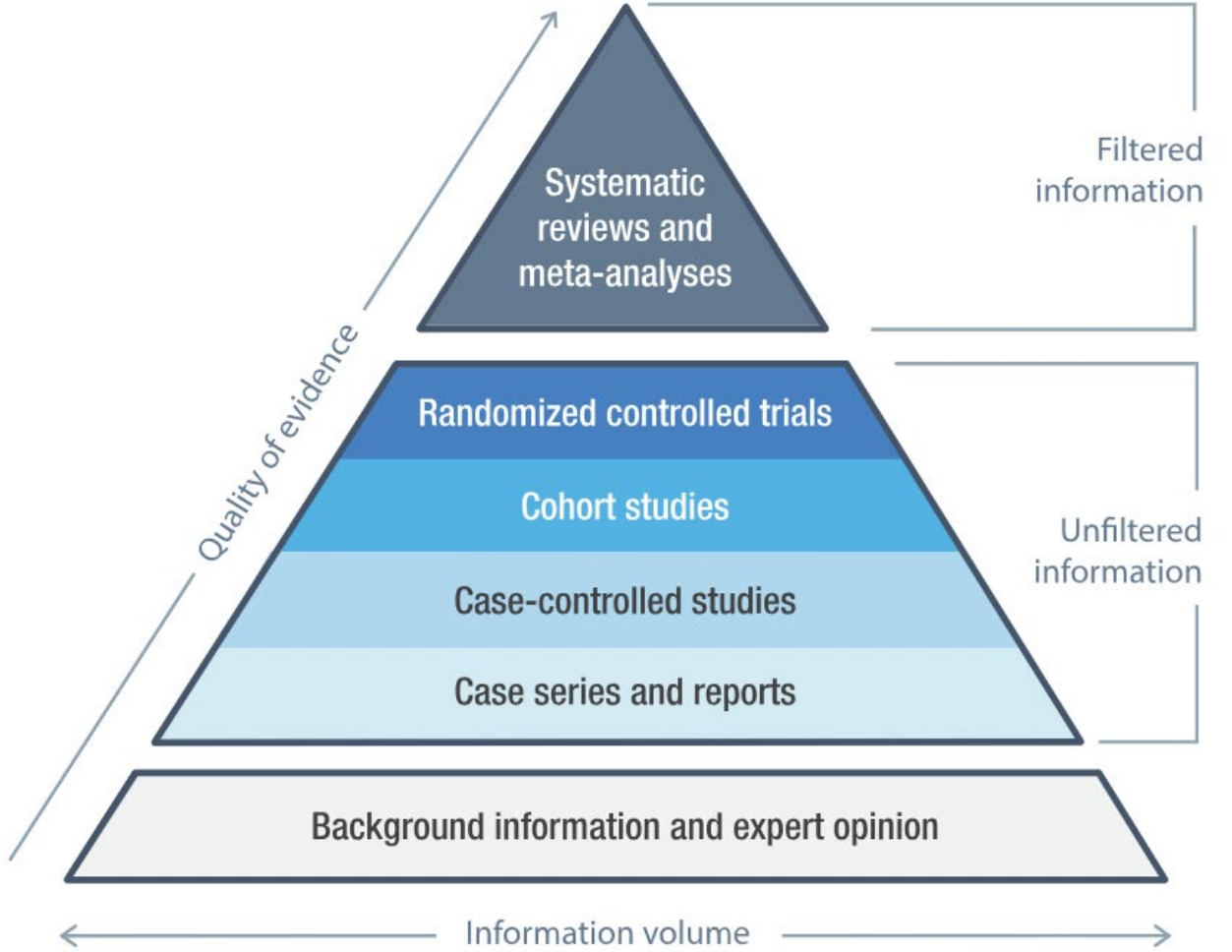
Generate hypotheses

Establish causality



Grimes et al. Lancet 2002;359:57-61

Levels of Evidence



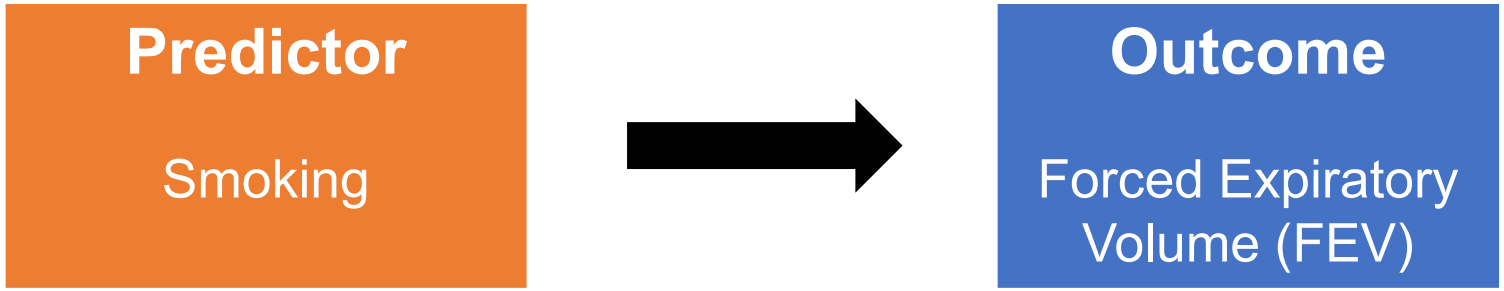
https://openmd.com/guide/levels-of-evidence

Causal Inference & Observational Studies

Causal inference is impossible in observational studies because confounding variables may cause the differences in the behavior of the response variable for different groups.

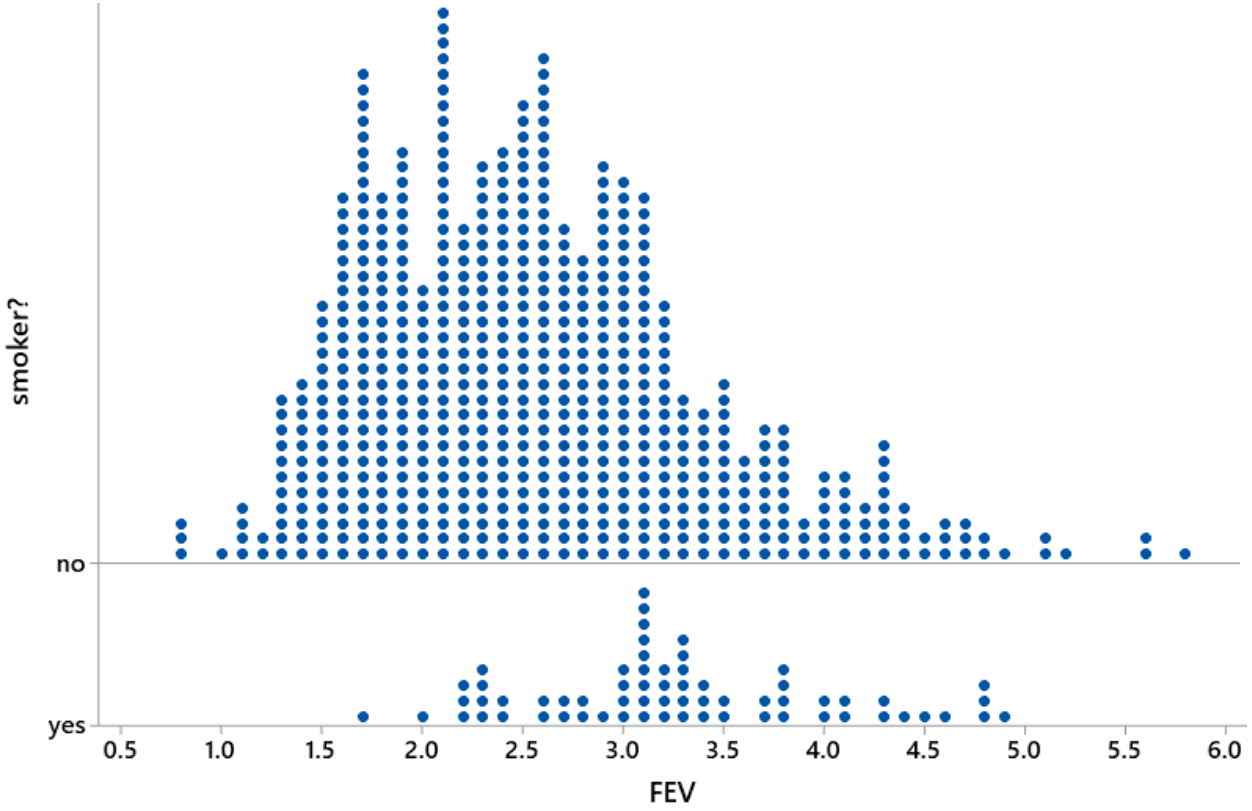
- A **confounding variable** is a variable that explains the group a person (e.g. exposure) is in and also the outcome of interest.
  - e.g., health consciousness is a confounding variable when testing the relationship between takes vitamins and how often a person gets sick

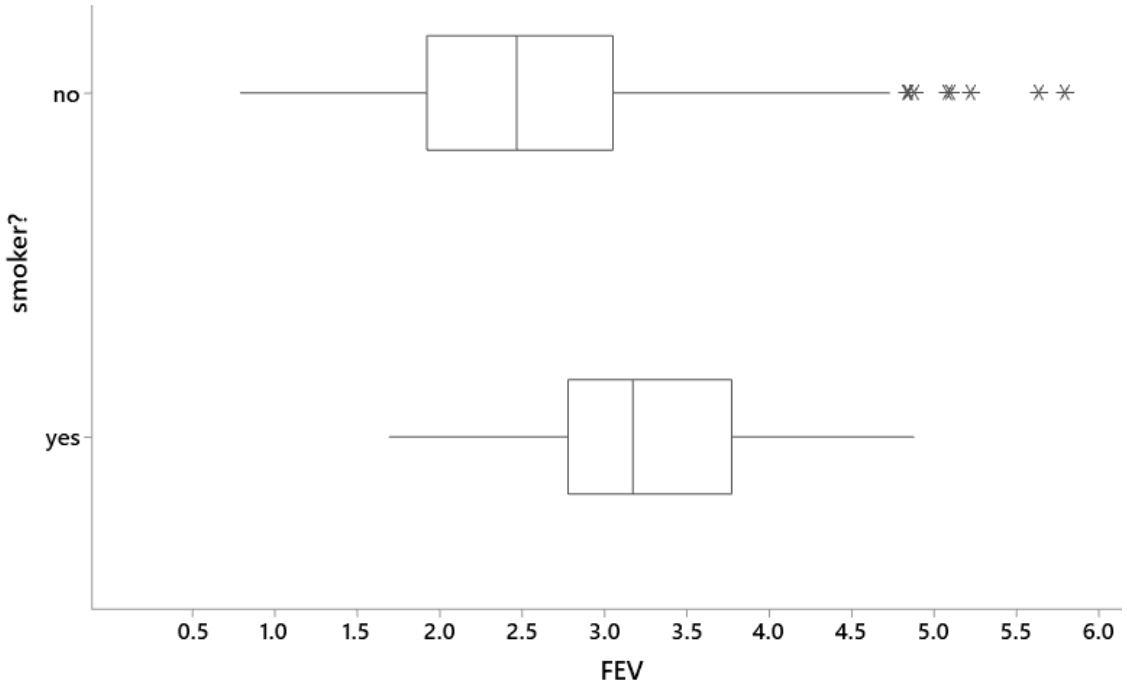
Confounding Example



https://askgoodquestions.blog/2020/03/02/35-statistics-of-illumination-part-4/

Confounding Example

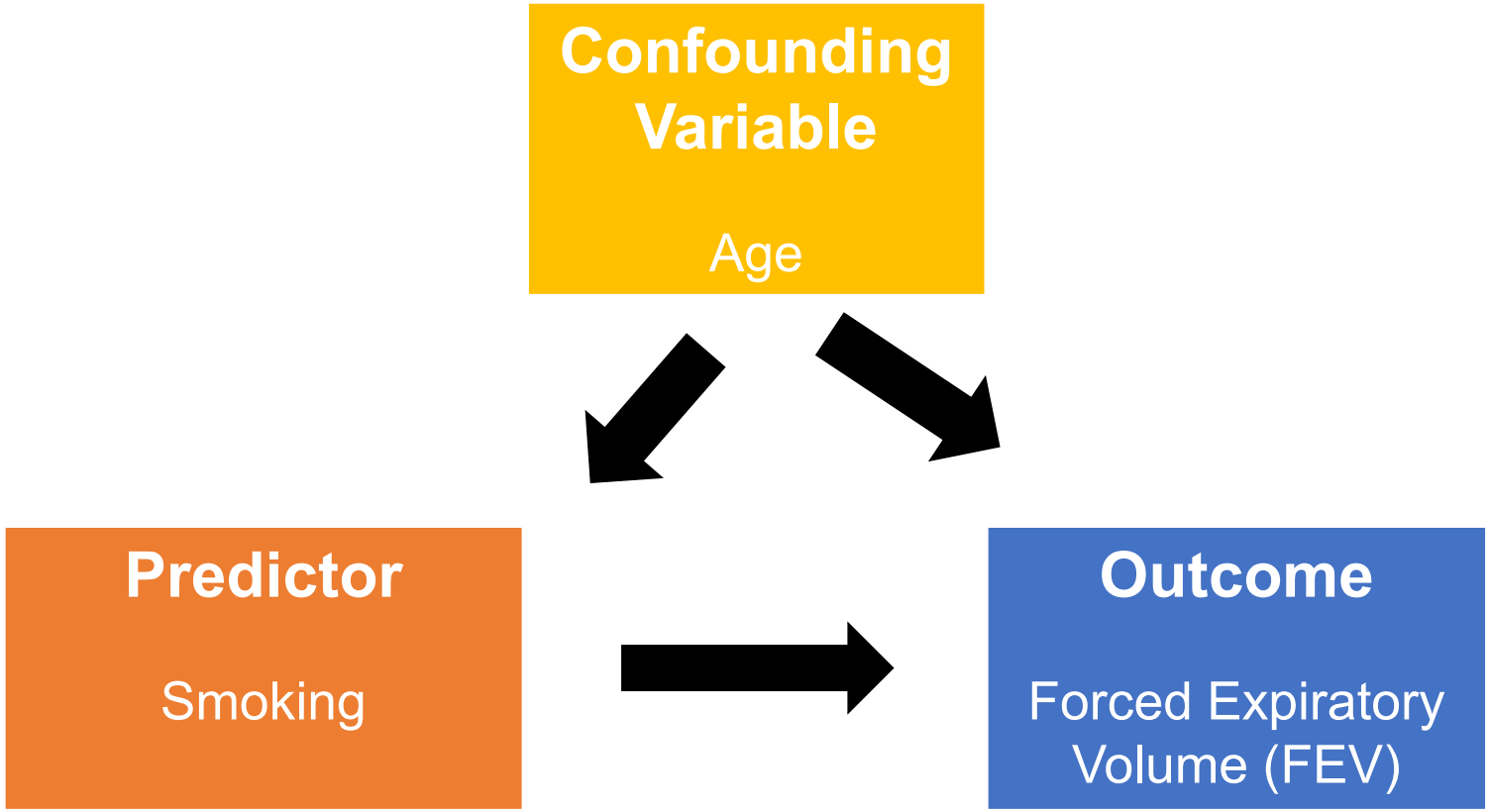




Conclusion: Smokers have Increased Lung Capacity

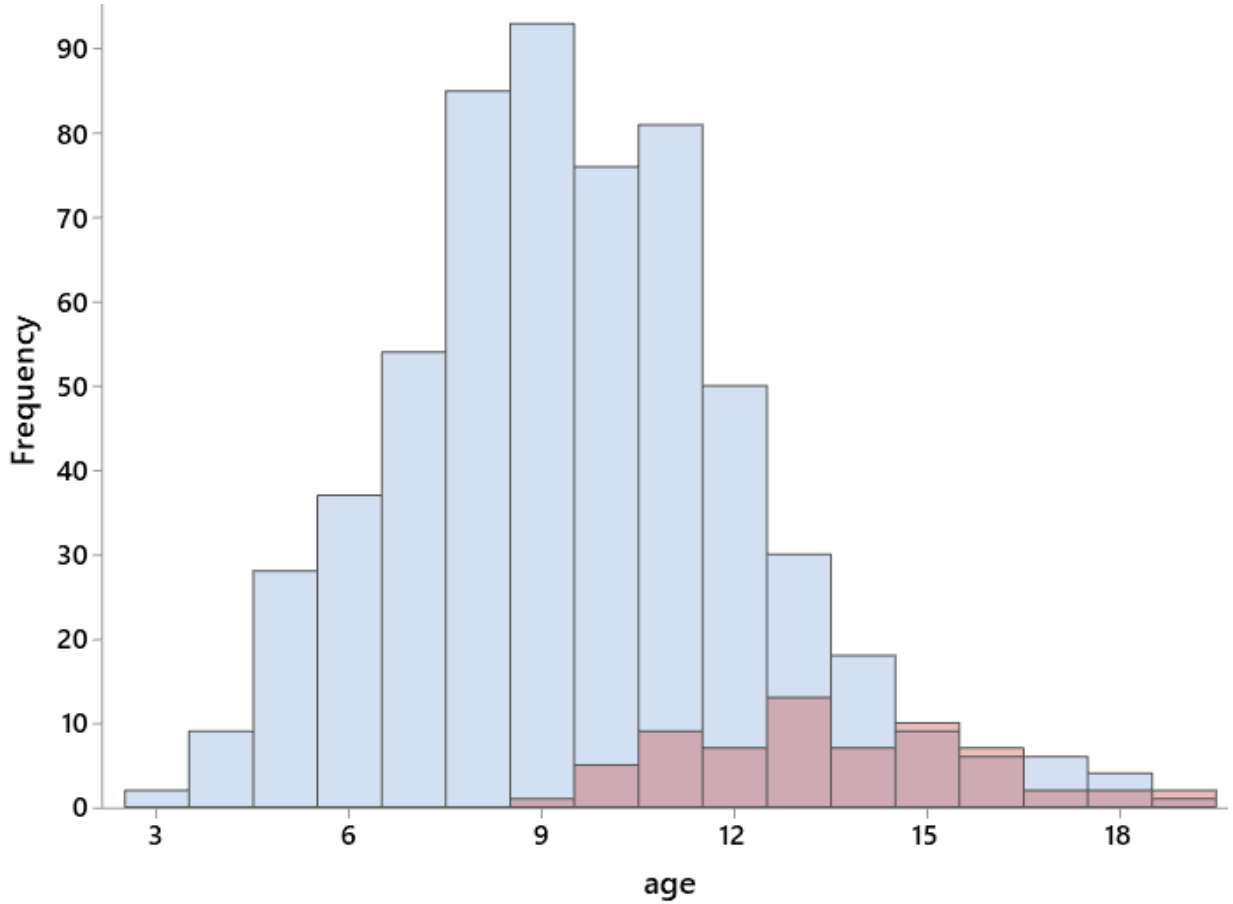
https://askgoodquestions.blog/2020/03/02/35-statistics-of-illumination-part-4/

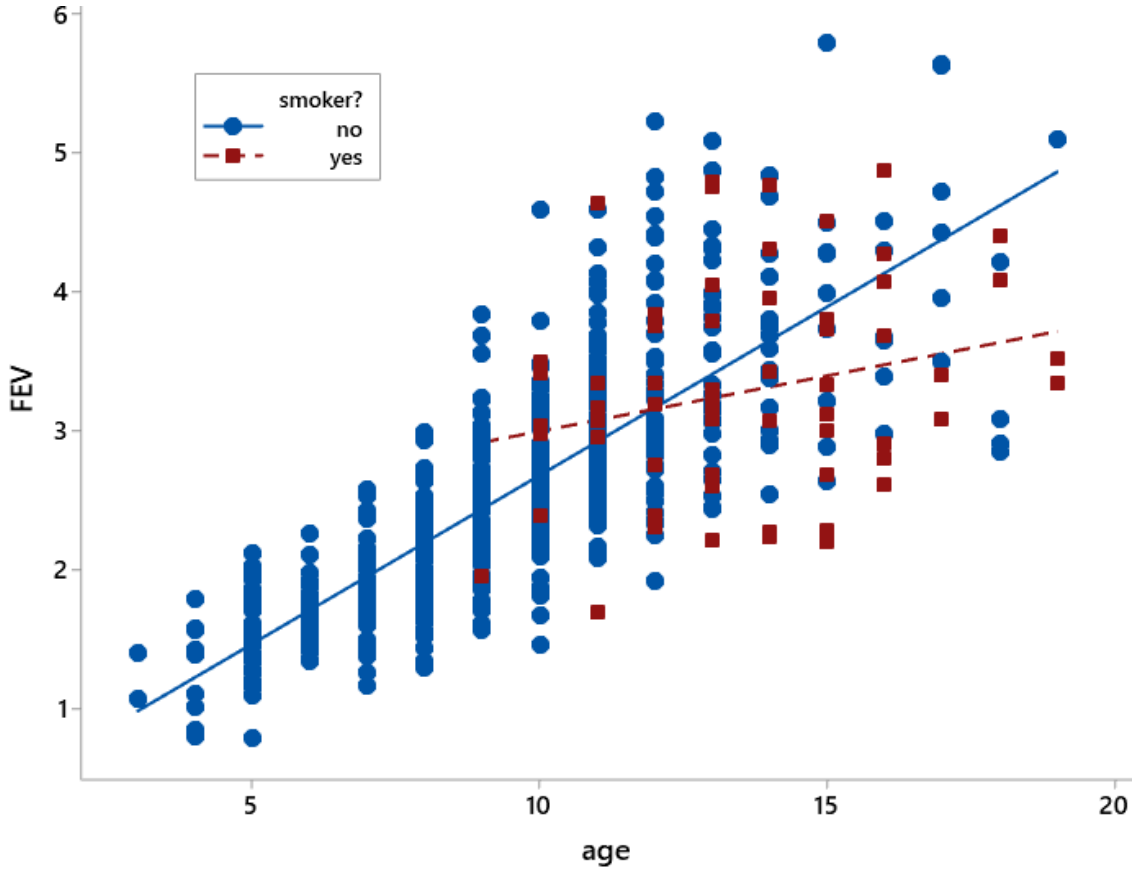
Confounding Example



https://askgoodquestions.blog/2020/03/02/35-statistics-of-illumination-part-4/

Confounding Example





https://askgoodquestions.blog/2020/03/02/35-statistics-of-illumination-part-4/

Causal Inference & Randomized Studies

Causal inference can be made from randomized experiments but not from observational studies.

- Randomization ensures that subjects with different features (i.e., confounding variables) are mixed up evenly among the treatment groups
- Randomized experiments seek to create groups that are totally similar except whether a treatment/exposure is present or absent
- The possibility that the groups may not end up being very “random” (i.e., groups are not mixed very well) is incorporated into the statistical tools used to express our uncertainty

Causal Diagrams

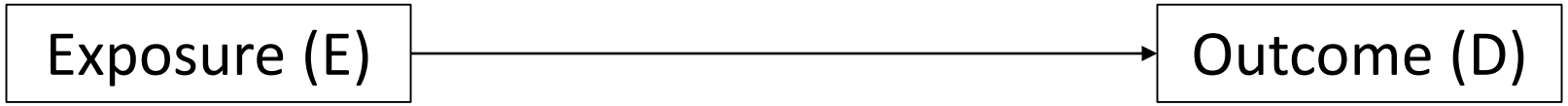
Directed Acyclic Graphs (DAGs)

- Another way to illustrate sources of bias (e.g., confounding) and other threats to validity (e.g., selection bias, measurement error)
- Visualize direction and flow of association (both causal and non-causal)
- Summarizes and informs the analytical approach
- Keeps you anchored to your research question
- If interested in building your own DAG: https://www.dagitty.net/

DAG Components

- Components of a DAG
  - Confounder
  - Precision Covariate
  - Mediator
  - Effect modifier
  - Collider

DAG Backbone

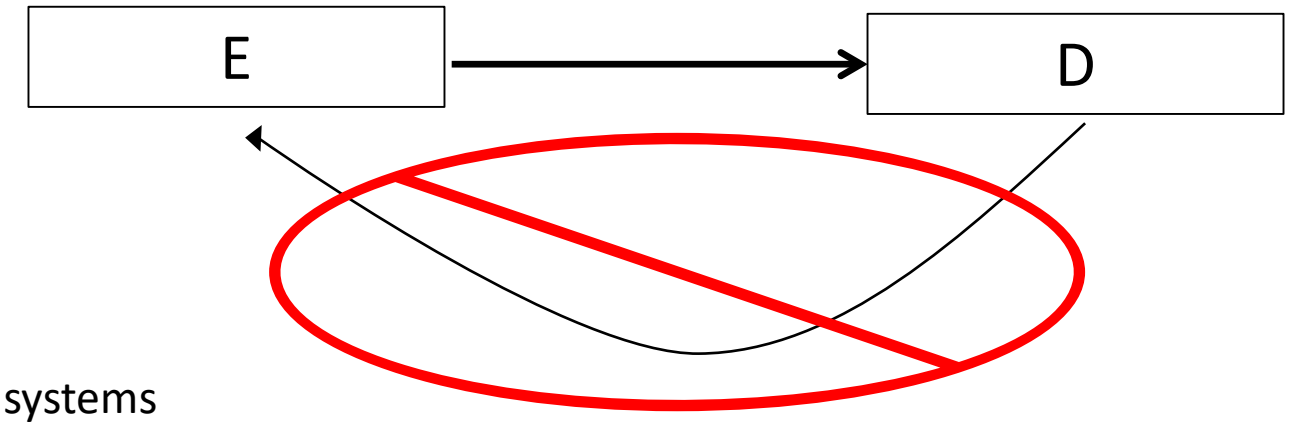


- Cause
- Risk factor
- Antecedent event
- Ancestor
- Treatment
- Independent variable
- Predictor
- Correlate
- Determinant
- Explanatory variable
- X-variable

- Effect
- Disease
- Event
- Result
- Case status
- Dependent variable
- Y-variable

No Feedback Loops

Directed Acyclic Graphs



We know in biological systems there are many feedback loops

Way to think about these variables in DAG form would be to specify a time (e.g. variable at time A, B…)

Basic elements of a DAG

& how they relate to causal inference

- Confounder
- Precision covariate

✓

Accounting for these variables promotes exchangeability between exposed & unexposed groups, and enhances internal validity
Accounting for confounders controls bias; accounting for precision covariates controls for measurement error

- Mediator
- Effect modifier

Maybe

Whether you deal with these types of variables in an analysis depends on your research question

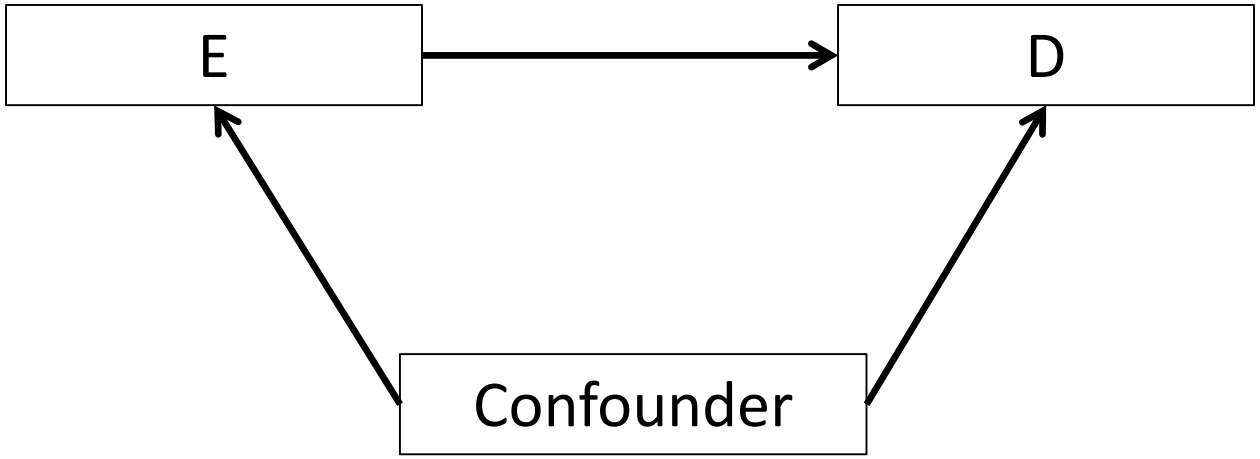
- Collider

✗

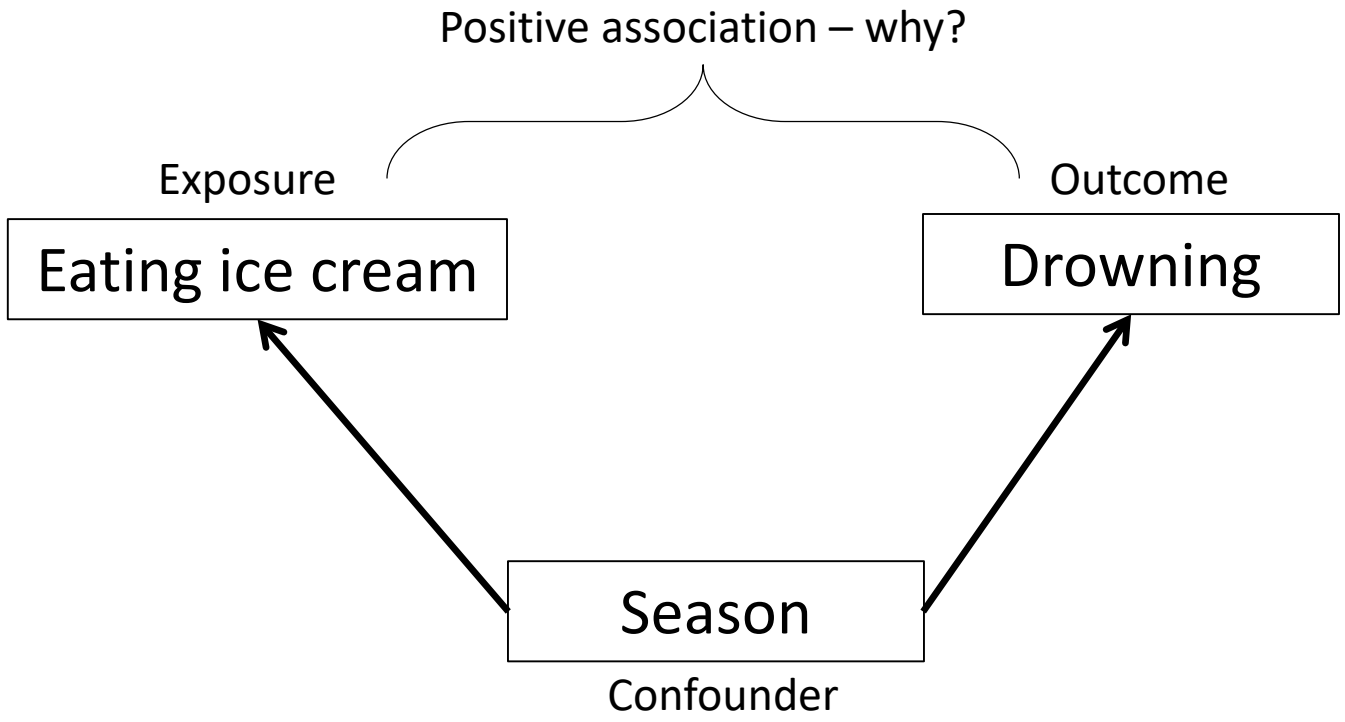
Colliders present a selection bias issue
Conditioning on a collider is a threat to both internal and external validity

Confounder

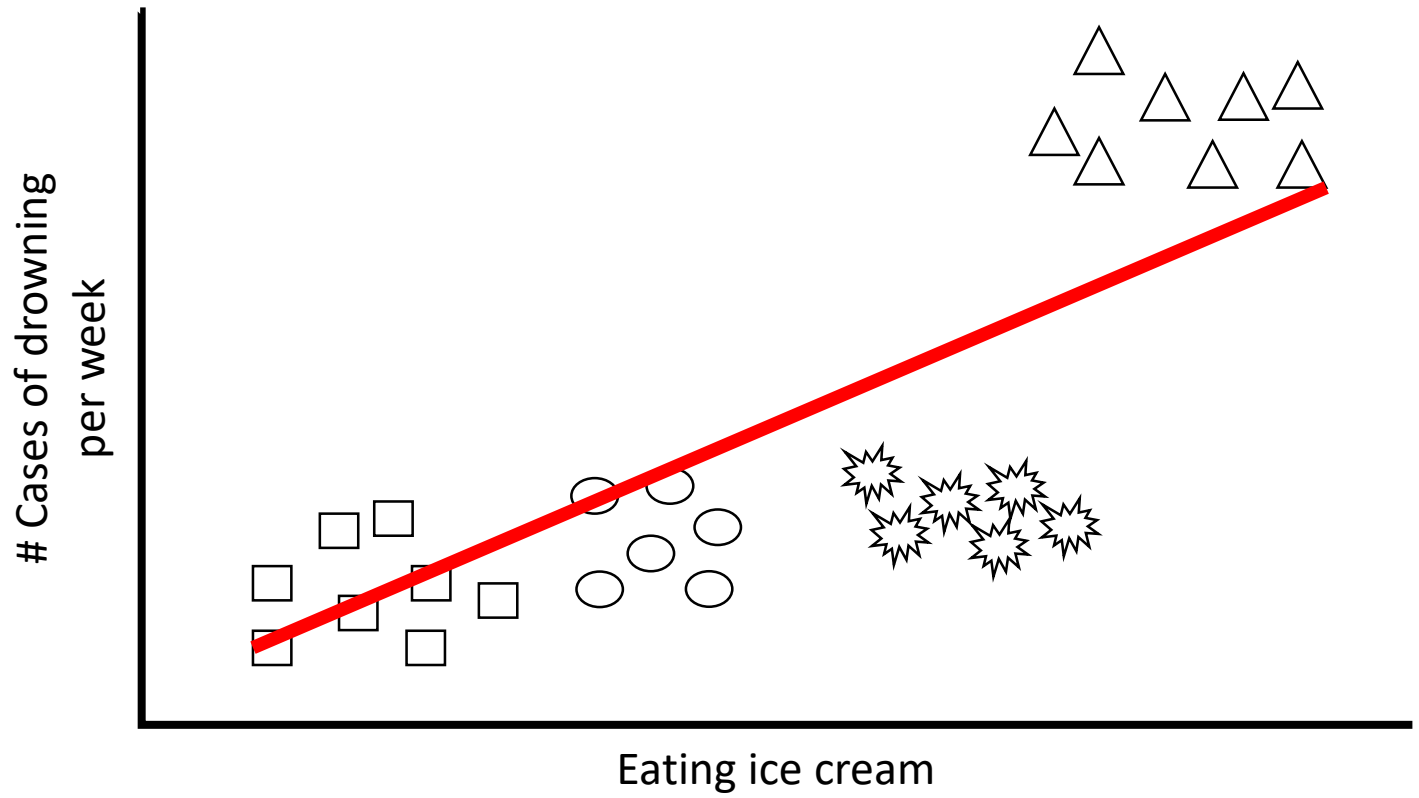
A variable that is associated with the exposure (but not affected by the exposure), and a potential determinant of the outcome.
"A shared common cause of E and D."



Confounder Example

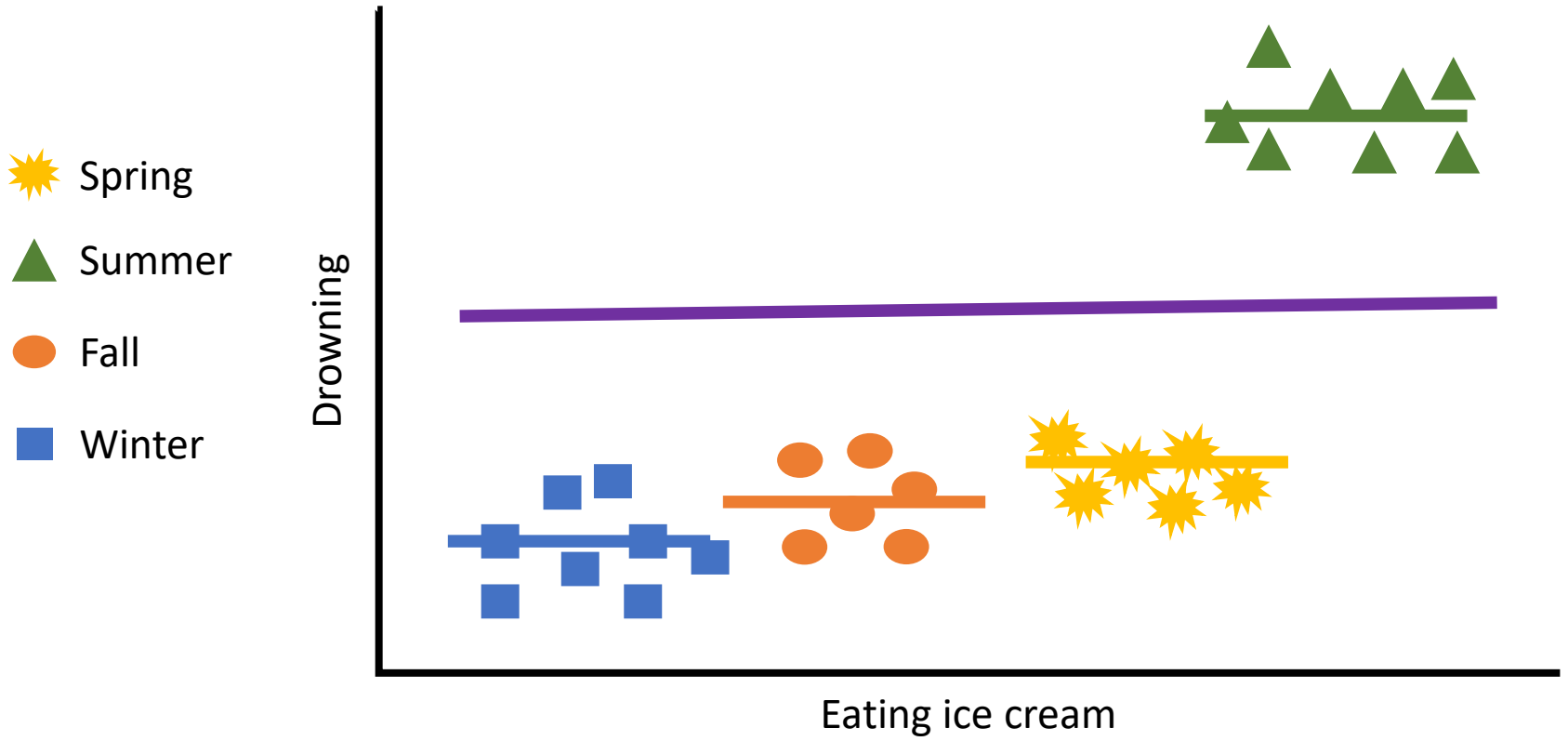


Confounder Example



Positive association between eating ice cream and drowning

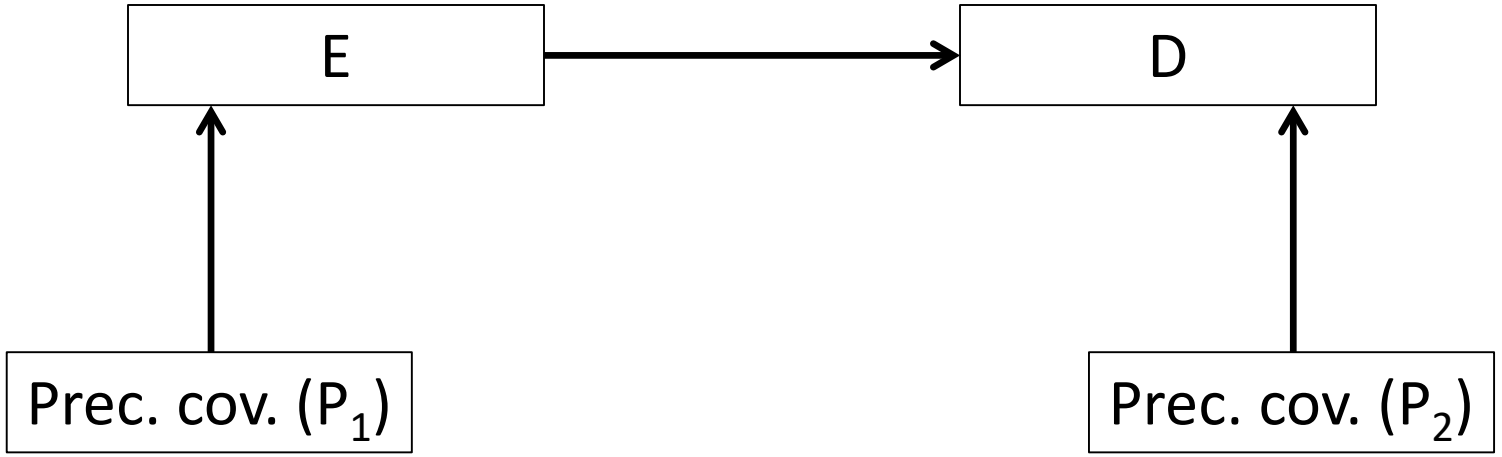
Confounder



No association between eating ice cream and drowning after conditioning on season

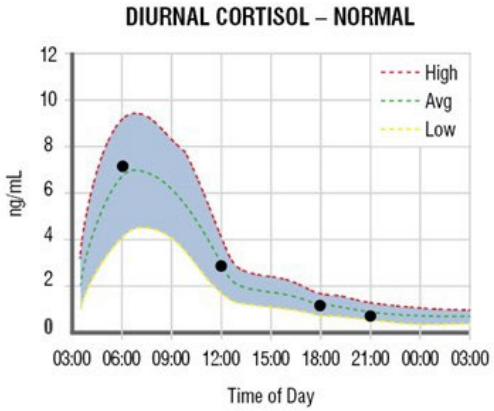
Precision covariate

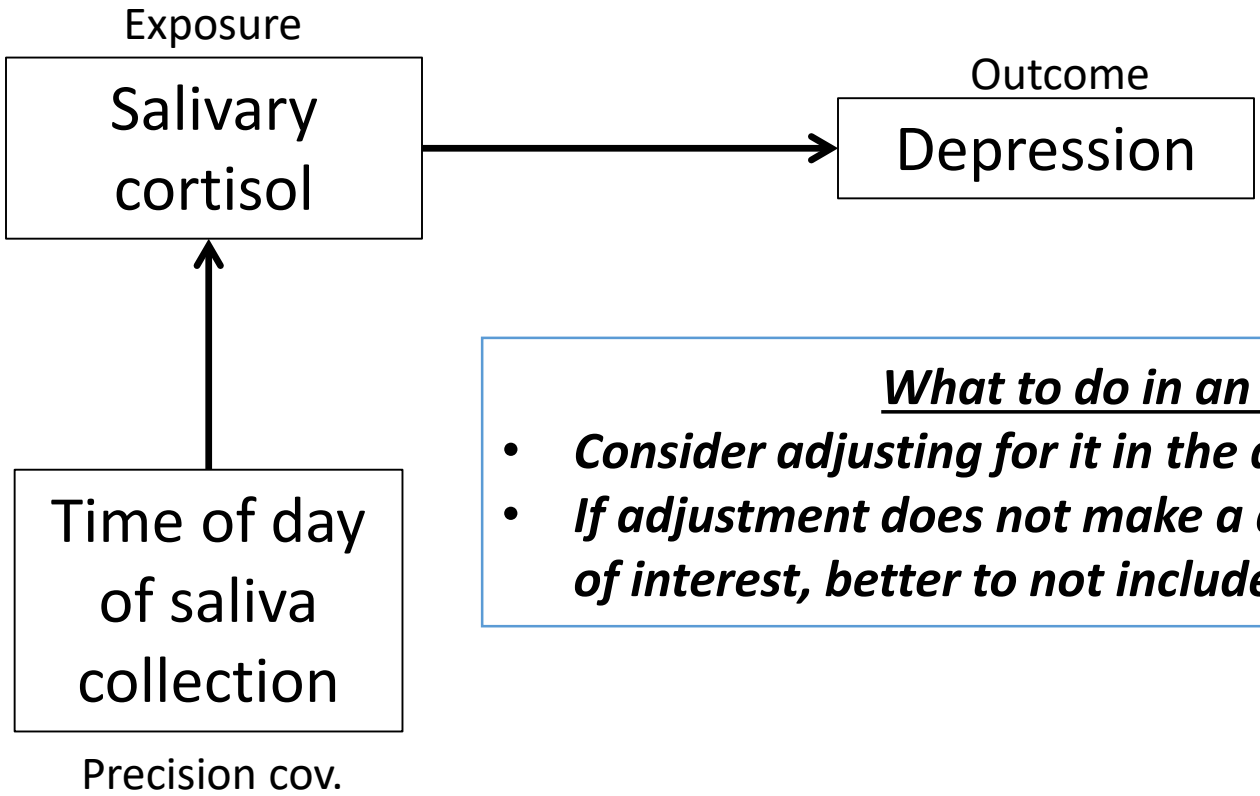
Variable that accounts for variability (i.e., measurement error) in the exposure only, or the outcome only.



Precision covariate

Question: Is higher salivary cortisol associated with depression?



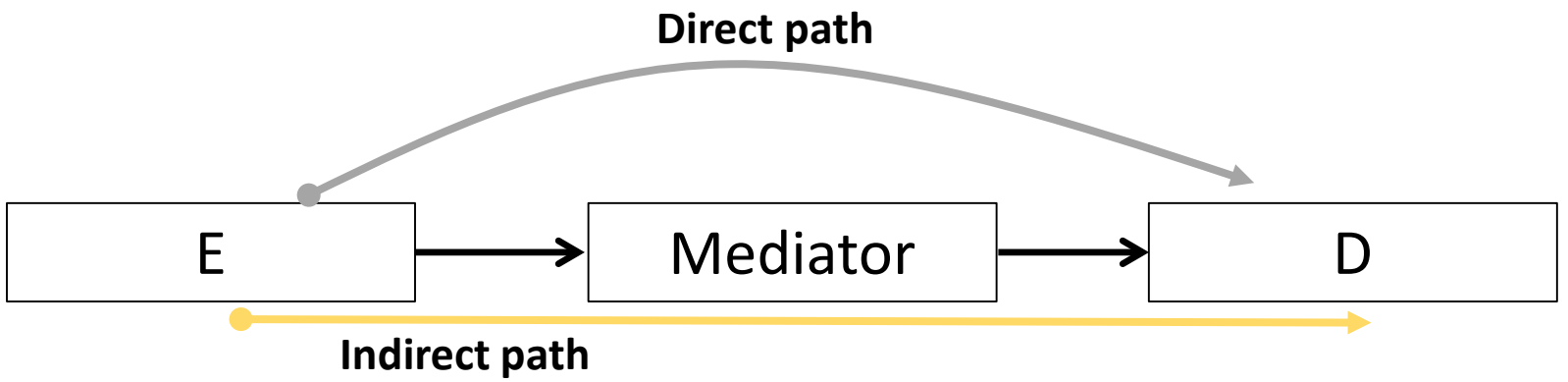


<u>What to do in an analysis?</u>

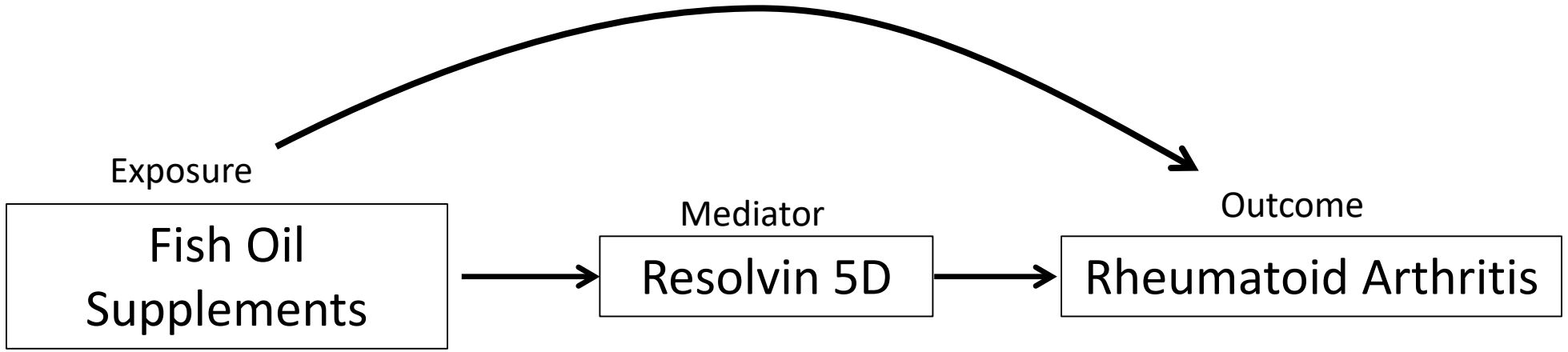
- ***Consider adjusting for it in the analysis***
- ***If adjustment does not make a difference in the estimate of interest, better to not include it in the models***

Mediator

A variable on the causal pathway between the exposure and outcome (i.e., caused by the exposure and a possible determinant of the outcome; also known as an intermediate variable)



Mediator Example



<u>What to do in an analysis?</u>

Depends on research question:

(1) Does the effect of taking fish oil supplements on RA operate through the lipid mediator resolvin 5D anti-inflammatory pathway?

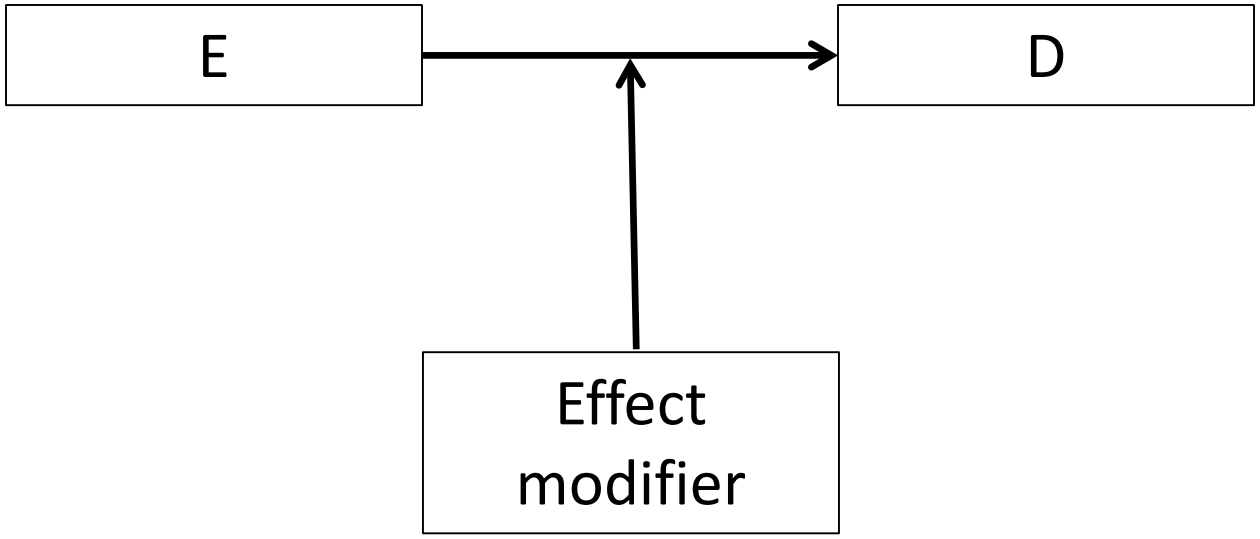
- ***Condition on mediator and compare total vs. direct effect of E on D***

(2) What is the effect of taking fish oil supplements on RA?

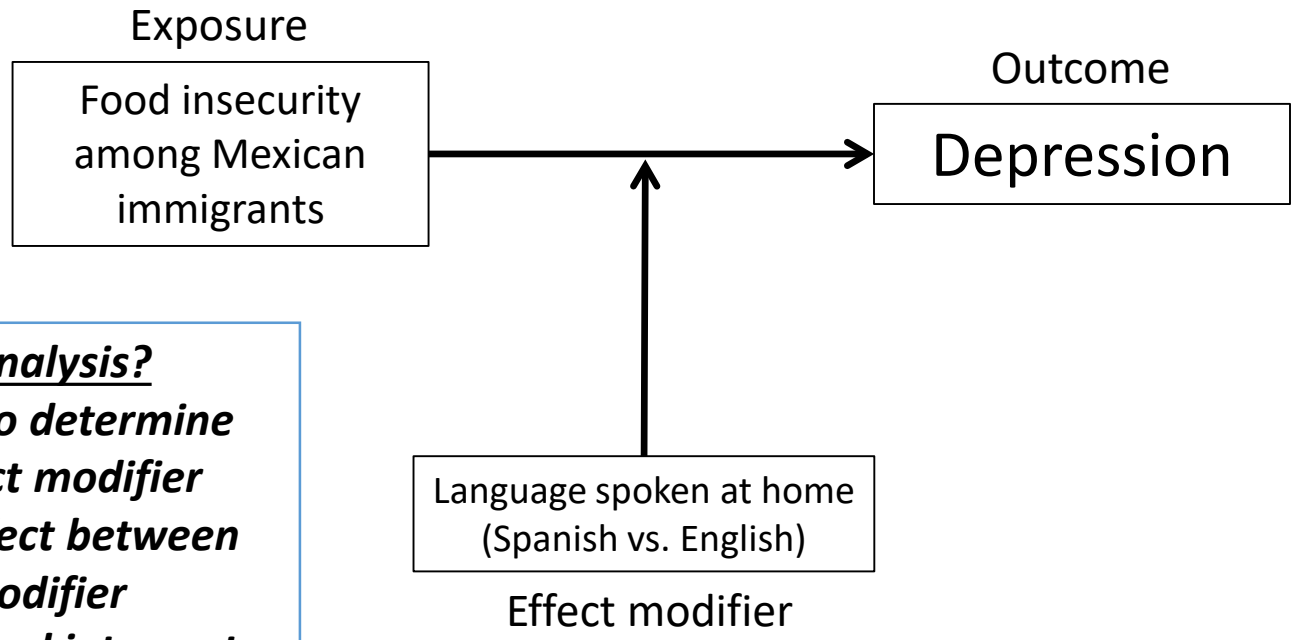
- ***Do not condition on mediator***

Effect Modifier

A variable that changes the nature of the relationship between the exposure and outcome
"Heterogeneity in the effect of E on D across levels of Q."



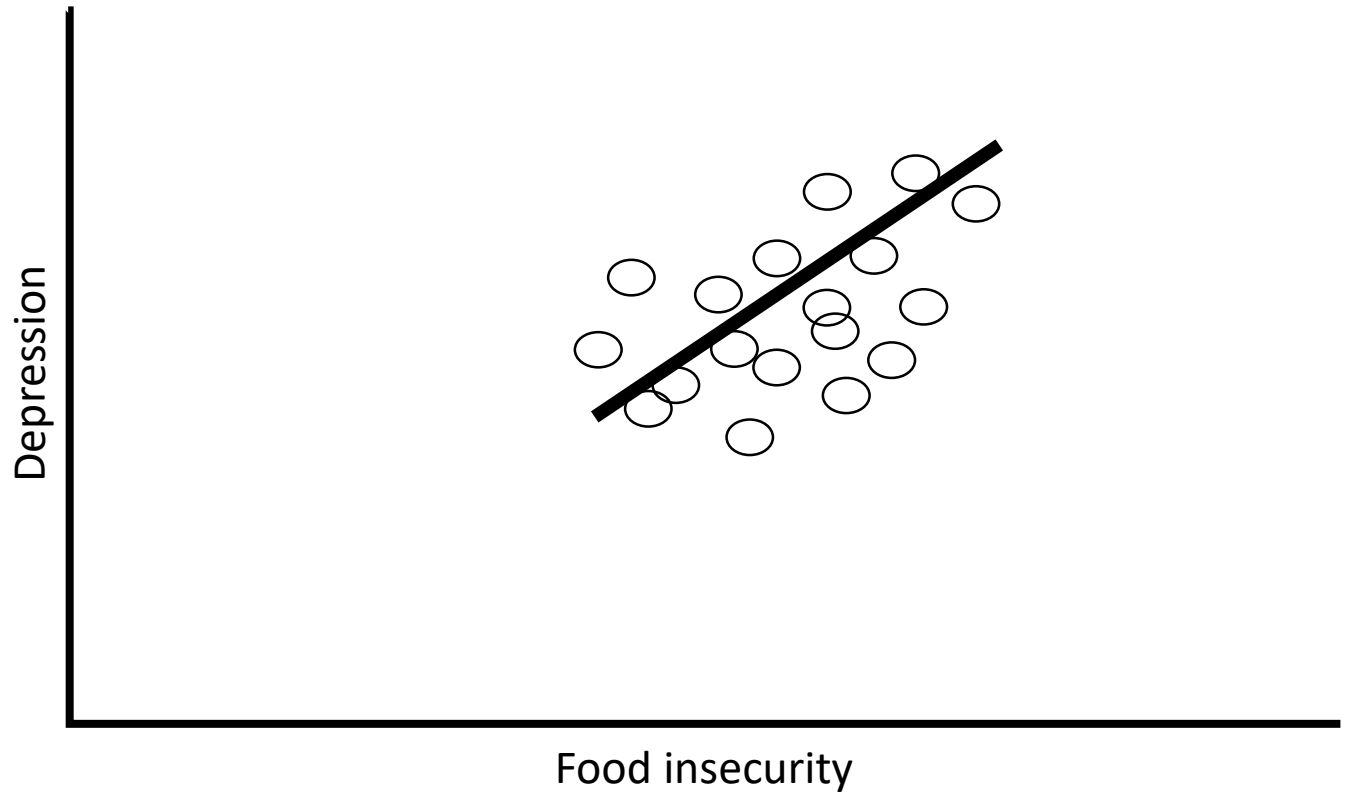
Effect Modifier Example



<u>What to do in an analysis?</u>

- ***Use prior knowledge to determine what could be an effect modifier***
- ***Test for interaction effect between exposure and effect modifier***
- ***Stratify as necessary and interpret stratum-specific estimates***

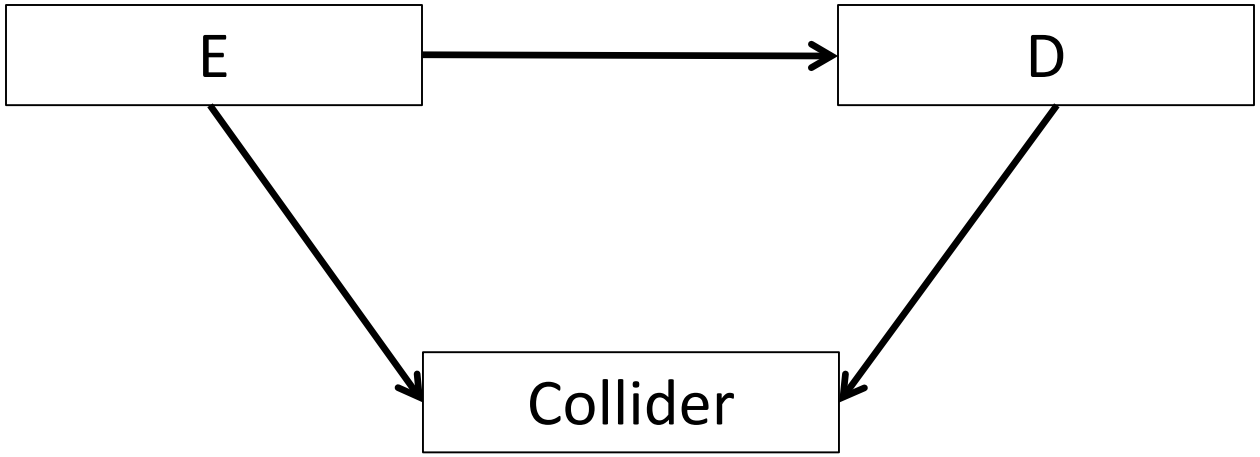
Effect Modifier Example



Effect Modifier Example

Collider

A variable that is caused by the exposure and caused by the outcome



<u>What to do in an analysis?</u>

- ***Identify colliders and leave them alone… do not condition on a collider!***

Collider Example

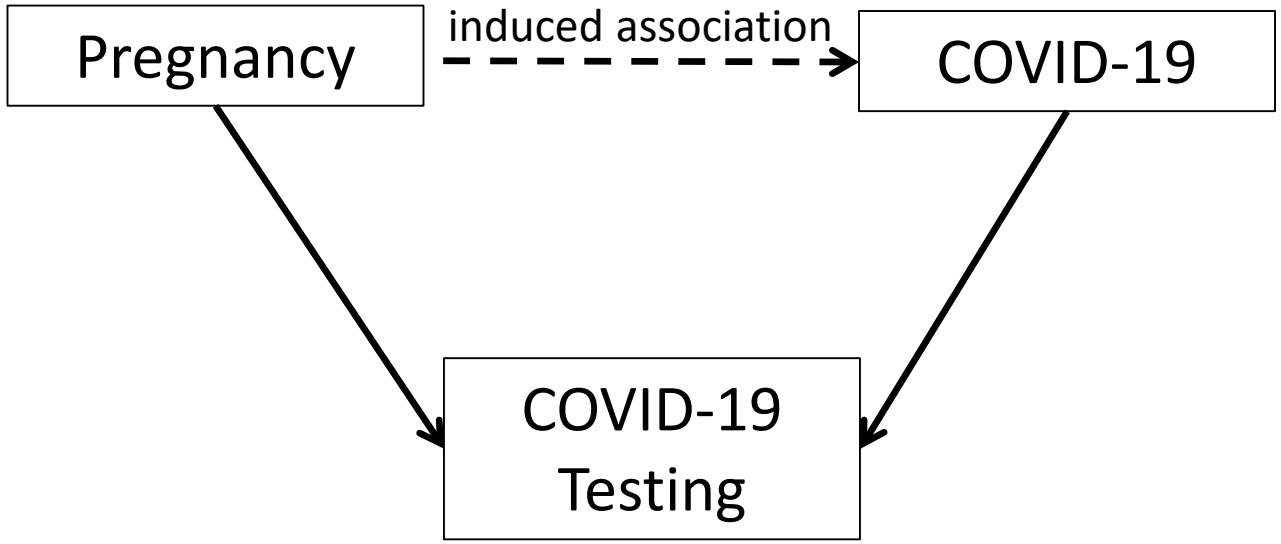
nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19478-2 OPEN

Collider bias undermines our understanding of COVID-19 disease risk and severity

Gareth J. Griffith[1,2,4], Tim T. Morris[1,2,4], Matthew J. Tudball[1,2,4], Annie Herbert[1,2,4], Giulia Mancano[1,2,4], Lindsey Pike[1,2], Gemma C. Sharp[1,2], Jonathan Sterne[2], Tom M. Palmer[1,2], George Davey Smith[1,2], Kate Tilling[1,2], Luisa Zuccolo[1,2], Neil M. Davies[1,2,3] & Gibran Hemani[1,2,4]✉



Hypothesis Testing

Null vs Alternative

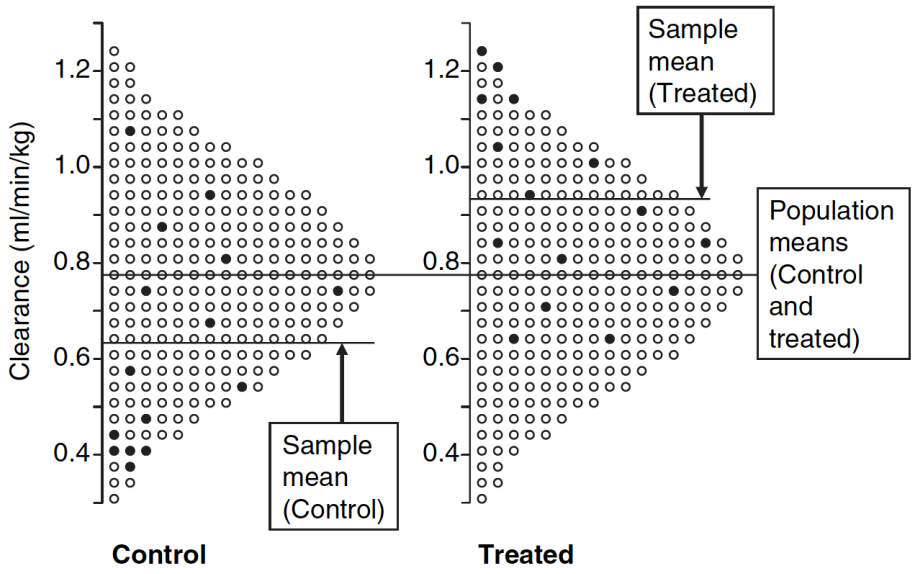
Hypothesis testing is a formal statistical procedure for deciding between two competing claims about a population parameter, in the example of a 2-sample t-test, the difference between two means

Null hypothesis - no real effect; apparent effects arose from random sampling

Alternative hypothesis - real effect; observed effects arose from a true difference

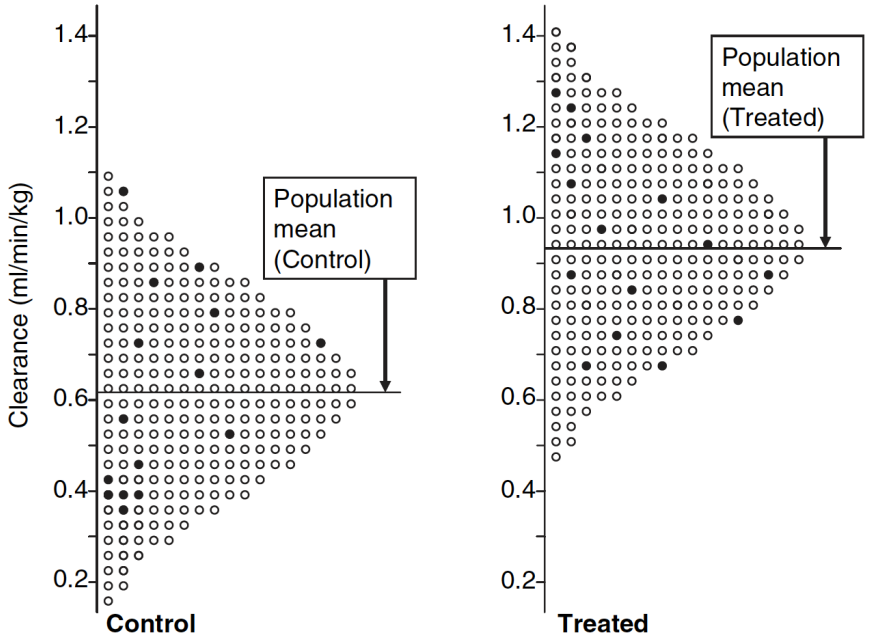
Null Hypothesis

(no difference between population means)



Alternative Hypothesis

(difference between population means)



Two-sample t-test

$\mu_{control}$ = population mean of the control group
$\mu_{treated}$ = population mean of the treated group

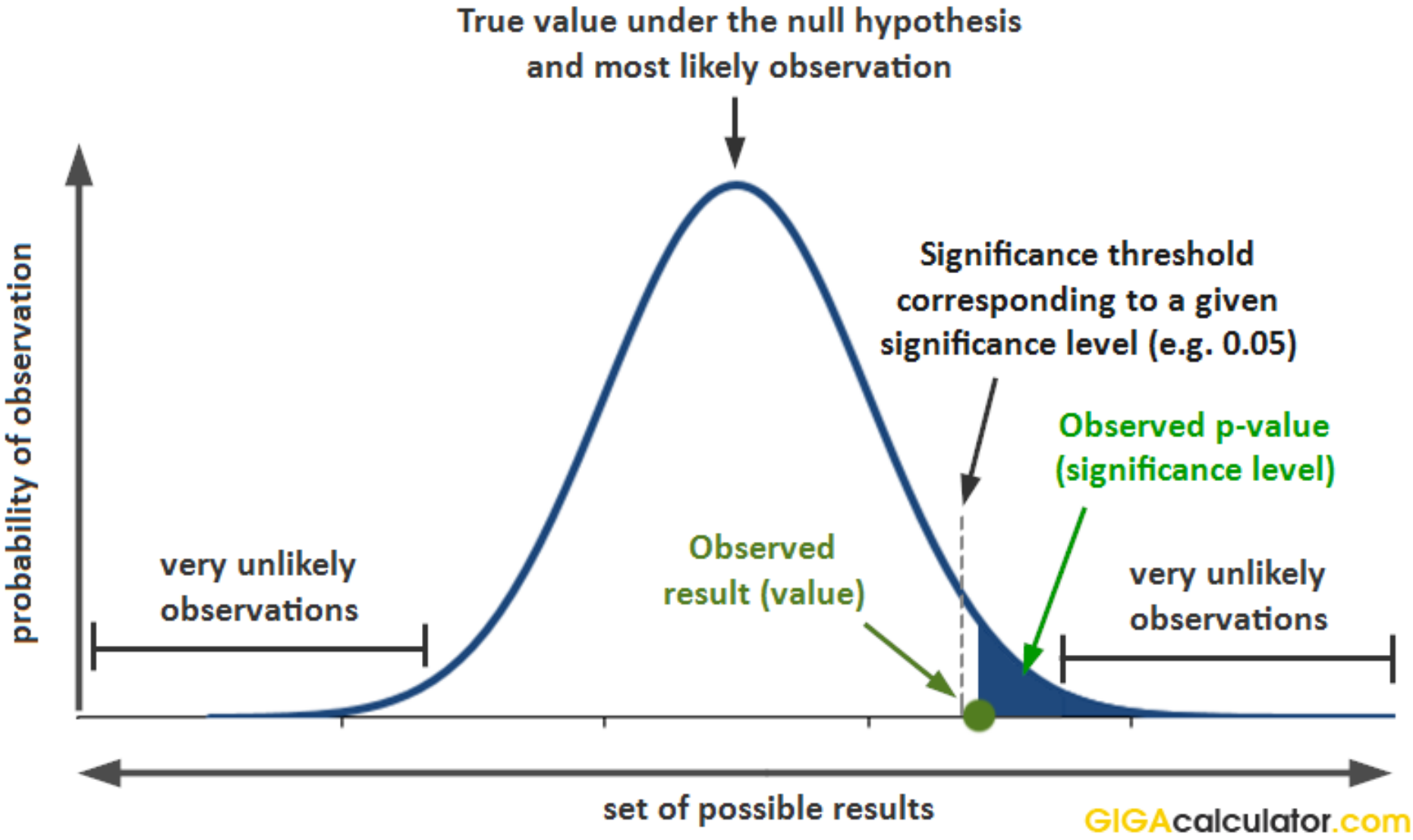
Null Hypothesis: $H_0: \mu_{control} = \mu_{treated}$
Alternative Hypothesis: $H_a: \mu_{control} \neq \mu_{treated}$

p-values

p-value

- **Definition**: The p-value is the probability that with a treatment that has no real effect, we would observe an effect (positive or negative) as extreme as (or more extreme than) the one observed.
- p-value is a quantitative measure of ‘how significant’ or ‘how confident’ you are that there is no effect
- smaller p-value = stronger evidence against ‘no effect’

Null Distribution

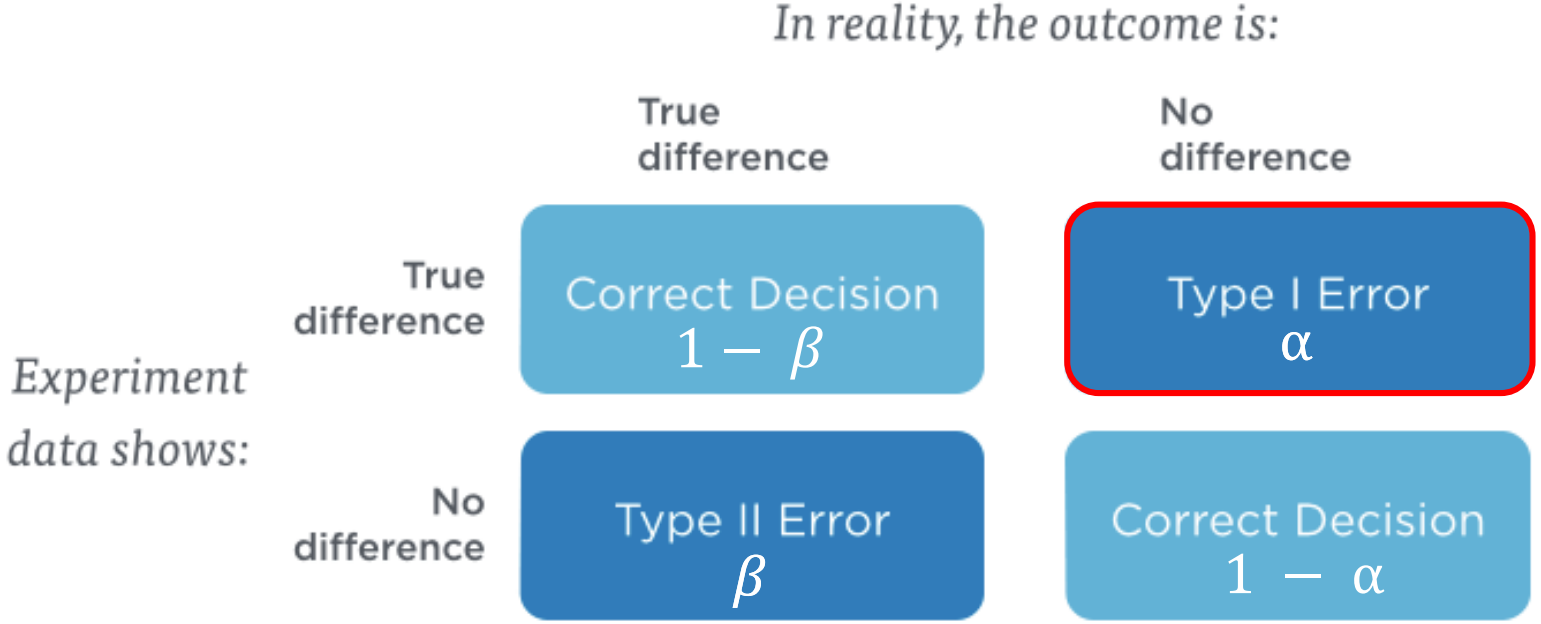


Rejecting vs Failing to Reject the Null

- You never PROVE the null
- You can reject (i.e. significant results) or fail to reject (i.e. non-significant results), but a p-value > α is NOT saying the null is the true
  - Proof of impossibility or negative proof (e.g. proving something is absent)
  - Can’t prove something as you think, but you can prove the opposite is FALSE and this is why we have the null hypothesis
  - Makes discussions about non-significant results very difficult

p-value & Type I Error

- p-value = probability of making a *type I error* when there is no true effect
- Traditionally we are willing to accept a 5% chance of results occurring if the null is actually true ($\alpha < 0.05$)



Very Low p-values

- Very low p-values are often suspect
- NEVER report a p-value = 0
  - ★There is always some probability that you could get the result by chance
- Instead report ‘p-value < 0.001’ or something

General Guidelines

A statistical results does not tell us what we should believe. It tells us how much we should change what we believe.

Non-significant: insufficient evidence to require any change

Significant ($p<0.05$): increase credence to a useful extent

Highly significant ($p<0.001$): increase credence markedly

Effect Sizes

- Effect size (general) - quantitative measure of the magnitude of a phenomenon (https://en.wikipedia.org/wiki/Effect_size)
- Standardized effect (size) – unitless measure that often incorporates error (e.g., Cohen's d, correlation coefficient)
- WARNING – statistical significance and clinically or biologically relevance are NOT the same thing

Why does effect size matter?

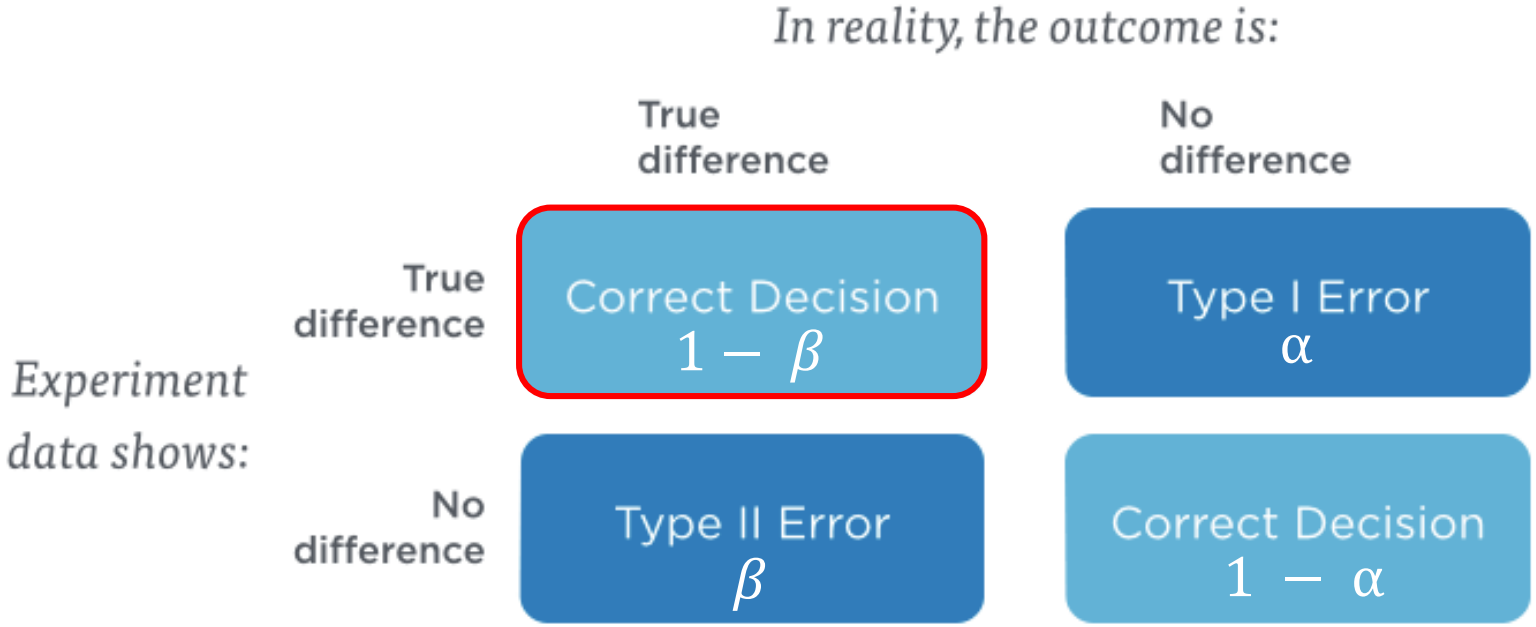
- While **statistical significance** shows that an effect exists in a study, **practical significance** shows that the effect is large enough to be meaningful in the real world
- Statistical significance alone can be misleading as it's influenced by sample size. In contrast, effect sizes are independent if sample size

Effect size example

- Question: does a new weight loss intervention perform better than an established intervention?
- Study design: control intervention & experimental intervention (n=13,000 participants in each group)
- After 6 months: mean weight loss for experimental intervention (mean = 10.6 lbs, sd = 6.7) was marginally higher than the control intervention (mean = 10.5 lbs, sd = 6.8). Resulted in a p-value = 0.01
- Does an additional weight loss difference of 0.1 pounds really matter?

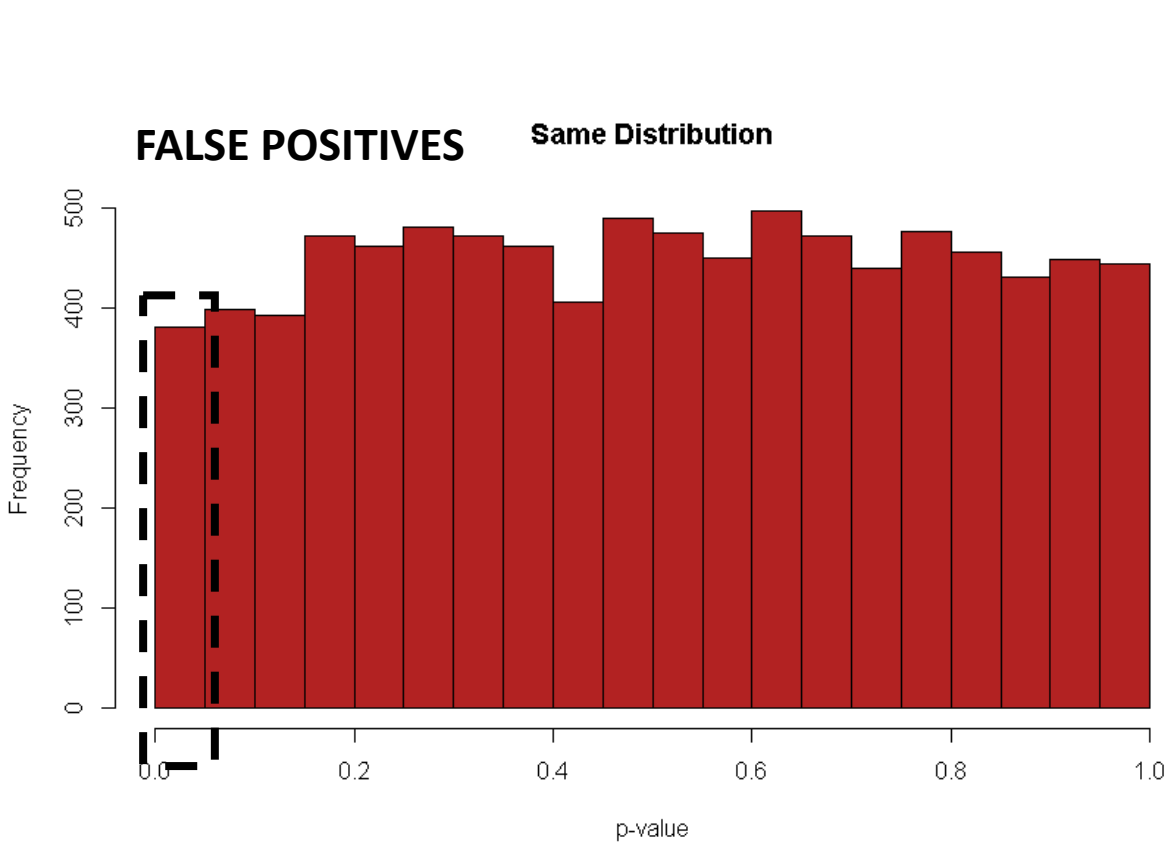
Power & Type II Error

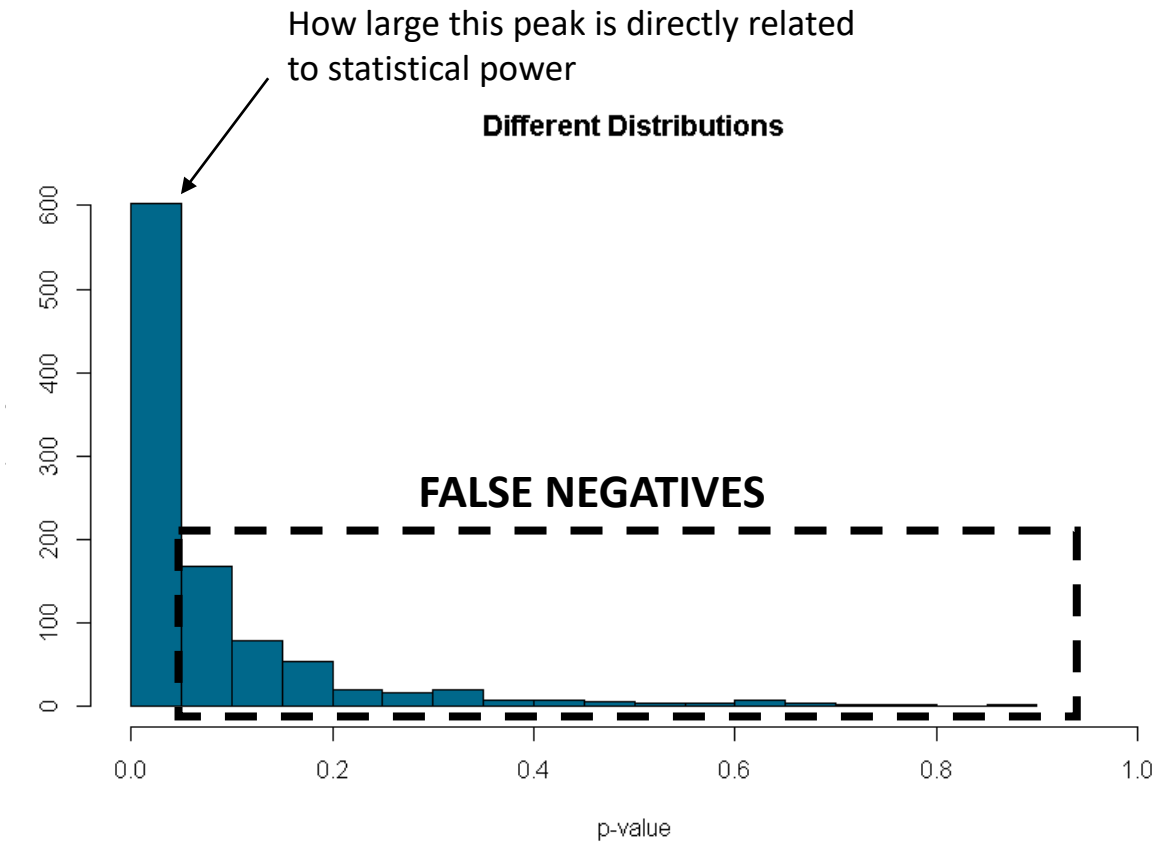
- power = probability of the detecting a true difference of a *stated size*
- power = 1 - β



Power

Histogram of p-values from multiple t-tests sampling either from the same distribution or a different distribution





Power Calculations

- 5 components – need 4 and can solve for the missing component
  1. Effect size
  2. Variance
  3. Significance Level (α)
  4. Sample size (n)
  5. Power ($1 - \beta$)
- A priori power calculations always needed for grants
- Never perform an observed or post-hoc power analysis, even if an editor requests it

Bayesian Statistics

Religions of Statistics

Two ‘religions’ in statistics - Frequentist vs. Bayesian

- **Frequentist** - relies on p-values; most ‘accepted’; taught in most introductory textbooks; easy to apply and relies solely on the data collected
- **Bayesian** - includes prior beliefs; relies on probability of being right (posterior probability) rather than probability of being wrong (p-value); requires advanced courses; harder to apply

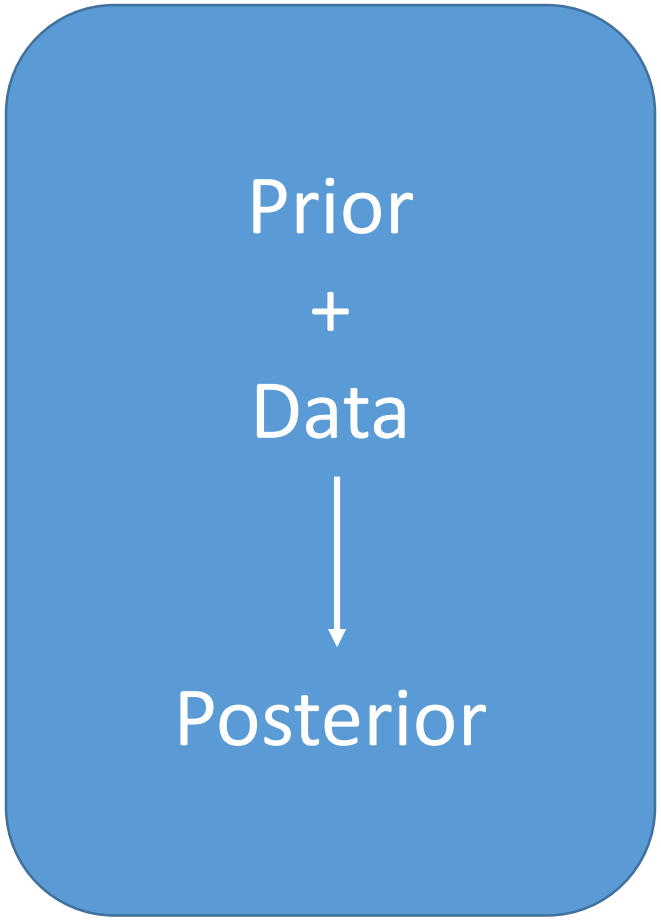
Bayesian Statistics

3 steps to a Bayesian analysis for the effectiveness of a product:

1. Determine the ‘prior likelihood’ that the product is effective.

2. Use the new evidence from the trial to calculate how much to modify our prior view

3. Use exact mathematical rules to combine the prior likelihood with the new evidence to produce a ‘posterior likelihood’



Bayesian Statistics

- Probability of a hypothesis being TRUE
- No p-values, instead get credible intervals
- Credible interval directly interprets the probability that the true value of a parameter lies within a specified range based on the data
- Example: testing treatment of drug A vs drug B. 95% probability that treatment A is 10-20% more effective than drug B
- Frequentist confidence intervals: probability that repeated samples would produce intervals containing the true value (i.e. confidence placed on interval itself instead of parameter)
- Pro: results are more intuitive
- Con: computationally intensive and need to be careful about priors

Conclusions

What Did We Learn?

- **Statistical inference** quantifies the uncertainty in our conclusions related to patterns in the data.
- **Scope of inference** forces us to think about who/what we want to/are able to extend our conclusions to.
- In general, **causal inference** in only possible in randomized experiments
- **DAGs** are used to illustrate the research question & summarize analytical strategy
- **Statistical hypotheses** clearly state comparison we would like to make including the statistical parameter of interest.
- **p-values** allow us to assign a quantitative value to how confident we are that there is no difference/effect. But this value should be combined with our prior knowledge (i.e. effect sizes)
- **p-values** do NOT address how confident we are that we ‘right’