EPI66 - Tópicos de Pesquisa I

Uso de DAGs para a identificação de confundidores na pesquisa em saúde

Ricardo de Souza Kuchenbecker
Rodrigo Citton P. dos Reis - citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Programa de Pós-Graduação em Epidemiologia

Porto Alegre, 2023

Introdução

Introdução

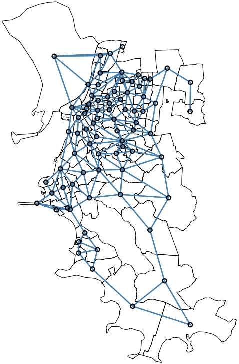
- Pesquisa epidemiológica e relações entre variáveis.
- Modelos estatísticos.
    - Propósitos dos modelos estatísticos: predição *versus* explicação (exploratório, causalidade)[1] , [2].
    - Modelos estatísticos: o papel das variáveis, e a relação de dependência entre estas.
- Na estatística, especificamos relações de dependência por meio de distribuições de probabilidades.

[1] Breiman, L. Statistical modeling: the two cultures. *Statistical Science*, 16:199-231, 2001.

[2] Shmueli, G. To explain or to predict. *Statistical Science*, 25:289-310, 2010.

Introdução

- **Grafos:** na matemática e ciência da computação, grafos são usados para modelar objetos, representados por vértices (nós), e as suas relações, representadas por arestas.



Introdução

- **DAGs: grafos acíclicos dirigidos** (*Directed acyclic graphs*)
  - Diagramas causais;
  - Modelos causais;
  - Modelos gráficos;
  - Modelos de equações estruturais não paramétricos;
  - Modelos causais estruturais.

Preliminares

Como duas variáveis podem estar associadas na população?

$$X \longrightarrow Y$$

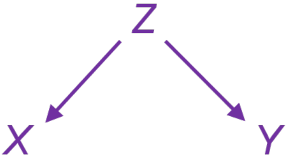
- Duas variáveis X e Y serão **associadas** na população se X causa Y.

Como duas variáveis podem estar associadas na população?

$$X \longleftarrow Y$$

- X e Y serão associadas na população se Y causa X.

Como duas variáveis podem estar associadas na população?



- Por fim, X e Y serão associadas na população se existir alguma variável Z que causa **ambas** X e Y.

Como duas variáveis podem estar associadas na população?

$$X \longrightarrow Y \qquad X \longleftarrow Y \qquad X \longleftarrow Z \longrightarrow Y$$

- X e Y não podem ser associadas na população por qualquer outra razão.
- Se X e Y são associadas na população, então **pelo menos uma** das situações acima deve ser verdade.

O que queremos dizer com associação “na população”?

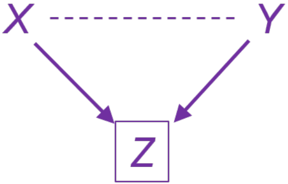
- Na terminologia estatística, X e Y são associadas “na população” significa que estas variáveis são **marginalmente associadas**.
- Se X e Y são marginalmente associadas, então, para um indivíduo em particular, saber a respeito de X nos dá alguma informação sobre o valor provável de Y, e vice-versa.
- Suponha, por simplicidade, X e Y dicotômicas. Se X e Y são marginalmente associadas, então

$$\Pr(Y = 1|X = 1) \neq \Pr(Y = 1|X = 0),$$

e

$$\Pr(X = 1|Y = 1) \neq \Pr(X = 1|Y = 0).$$

Como duas variáveis podem estar associadas em uma subpopulação?

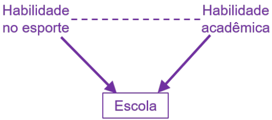


- Suponha que Z é um **efeito** tanto de X como de Y.
- Então, X e Y serão **associadas dentro do estrato** (subpopulação) de Z, mesmo se na população estas variáveis forem independentes.
- X e Y serão **condicionalmente associadas** (dado Z), mesmo que sejam marginalmente independentes (não associadas).
- A caixa ao redor de Z denota que estamos estratificando (condicionando) em Z.

Como duas variáveis podem estar associadas em uma subpopulação?

- A reta tracejada denota a **associação condicional induzida** pela estratificação/codicionamento em Z.

Como duas variáveis podem estar associadas em uma subpopulação?



- Suponha que uma escola aceita alunos ou porque são "bons" nos **esportes**, ou porque são "bons" **academicamente**; ou ainda, alunos que são "bons" nos dois.
- Suponha que a habilidade acadêmica e a habilidade no esporte sejam **independentes** na população.
  - Se selecionarmos ao acaso um aluno desta escola, e este nos informa que não possui habilidades esportivas, o que podemos dizer sobre suas habilidades acadêmicas?
- **Dentro da escola**, existirá uma associação (negativa) entre habilidade acadêmica e habilidade nos esportes.

Resumindo

- X e Y serão **associadas na população** se:
  - X causa Y.
  - Y causa X.
  - existe uma Z que é causa comum de X e Y.
- X e Y serão **associadas em subpopulações definadas por** Z se Z é um **efeito** de X e Y.

Uma avaliação empírica

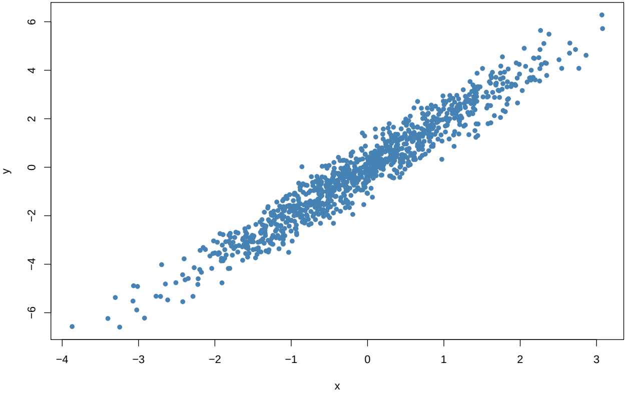
Suponha que conhecemos “por completo” a relação entre as variáveis Y, X e C, dada pelas seguintes equações

$$\begin{aligned} X &= C + \epsilon_X, \ \epsilon \sim N(0, \sigma_{\epsilon_X}), \\ Y &= 1.5X + 0.5C + \epsilon_Y, \ \epsilon \sim N(0, \sigma_{\epsilon_Y}). \end{aligned}$$

Ou seja, neste caso, temos que $X \rightarrow Y$ e $X \leftarrow C \rightarrow Y$ na população. Digamos que estamos interessados no efeito de X em Y (o qual sabemos que é 1.5).

Uma avaliação empírica

Suponha também que obtivemos uma amostra de tamanho $n = 1000$ da população.



Uma avaliação empírica

Com os dados observados, ajustamos dois modelos de regressão linear para a variável Y:

1. Modelo com apenas a variável X ("não ajustado" para confundidores);
2. Modelo com as variáveis X e C ("ajustado").

Deveríamos observar um efeito viesado de X em Y no primeiro modelo, e um efeito não-viesado no segundo modelo.

Uma avaliação empírica

Tabela 1: Ajuste por confundior

	Não ajustado			Ajustado		
Variáveis	Estimates	CI	p	Estimates	CI	p
(Intercept)	-0.01	-0.04 – 0.02	0.586	0.00	-0.03 – 0.03	0.987
x	1.89	1.86 – 1.92	$<$0.001	1.54	1.48 – 1.61	$<$0.001
c				0.44	0.36 – 0.51	$<$0.001

Uma avaliação empírica

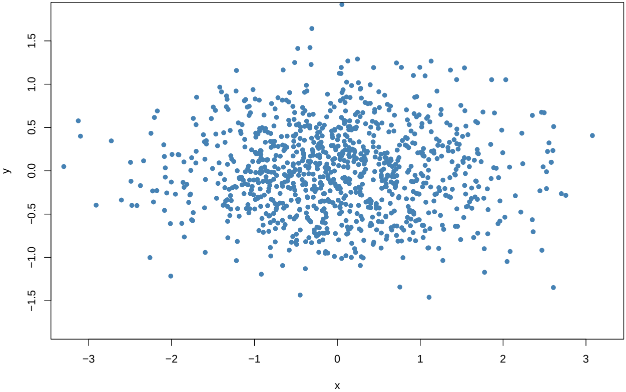
Agora suponha que a verdadeira relação entre as variáveis Y, X e Z, é dada pelas seguintes equações

$$\begin{aligned} Y &= 0X + \epsilon_Y,\ \epsilon \sim N(0, \sigma_{\epsilon_Y}), \\ Z &= X + Y + \epsilon_Z,\ \epsilon \sim N(0, \sigma_{\epsilon_Z}). \end{aligned}$$

Ou seja, neste caso, temos que X e Y são independentes e $X \rightarrow Z \leftarrow Y$ na população. Sabemos, neste caso, que o efeito de X em Y é nulo.

Uma avaliação empírica

Suponha mais uma vez que obtivemos uma amostra de tamanho $n = 1000$ da população.



Uma avaliação empírica

Com os dados observados, mais uma vez ajustamos dois modelos de regressão linear para a variável Y:

1. Modelo com apenas a variável X ("não ajustado" para confundidores);
2. Modelo com as variáveis X e Z ("ajustado").

Deveríamos observar um efeito não-viesado de X em Y no primeiro modelo, e um efeito viesado no segundo modelo (associação induzida pelo condicionamento de Z).

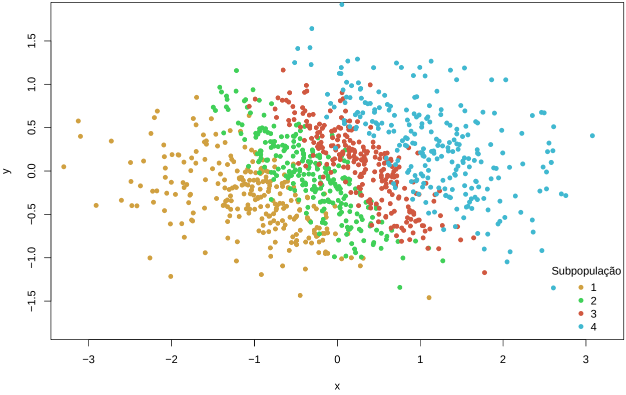
Uma avaliação empírica

Tabela 2: Ajuste por colisor

	Não ajustado			Ajustado		
Variáveis	Estimates	CI	p	Estimates	CI	p
(Intercept)	0.00	-0.03 – 0.03	0.986	-0.00	-0.02 – 0.01	0.393
x	-0.01	-0.04 – 0.02	0.639	-0.86	-0.89 – -0.84	<0.001
z				0.88	0.85 – 0.90	<0.001

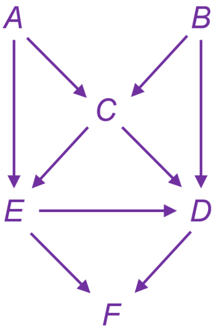
Uma avaliação empírica

- O que está acontecendo aqui?
- "Ajustar sempre" não é o correto?



DAGs: uma introdução mais formal

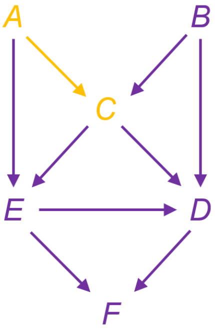
Um exemplo



Grafo acíclico dirigido

- Este é um exemplo de um **grafo acíclico dirigido** (DAG) causal (diagrama causal).
- É **dirigido**, pois cada aresta é uma seta de ponta única.
- É **causal**, pois as setas representam nossas suposições a respeito da direção da influência causal.
- É **acíclico**, pois não contém ciclos: nenhuma variável causa a si mesma.

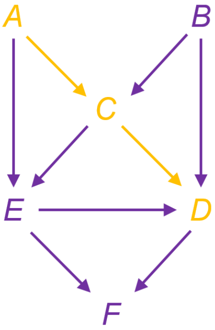
Terminologia



Pais e filhos

- A é **pai** (ou **mãe**) de C.
- C é **filho** (ou **filha**) de A.

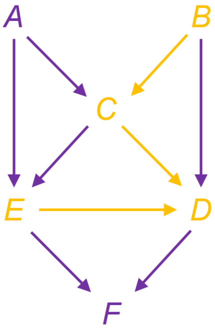
Terminologia



Ancestrais e descendentes

- A é um **ancestral** de D.
- D é **descendente** de A.
- A também é um **ancestral** de C.
- C também é um **descendente** de A.
  - Ou seja, pais são ancestrais, e filhos são descendentes.

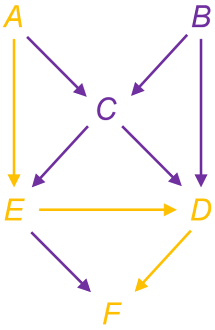
Terminologia



Caminho

- Este é um **caminho** de E para B.

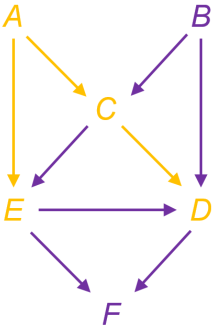
Terminologia



Caminho dirigido

- Este é um **caminho dirigido** de A para F (todas as setas no caminho apontam "para frente").

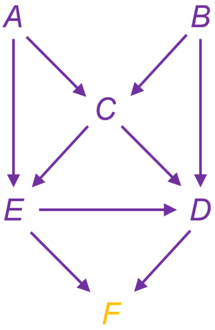
Terminologia



Caminho back-door

- Este é um **caminho porta dos fundos** de E para D (o caminho começa com uma seta **chegando em** E).

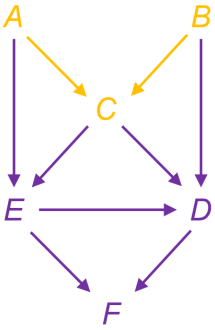
Terminologia



Collider

- F é um **colisor** desde que duas pontas de setas se encontram em F.

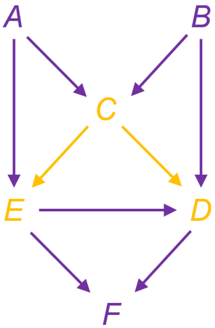
Terminologia



Nota

- Note que C é um colisor no caminho $A \rightarrow C \leftarrow B$.

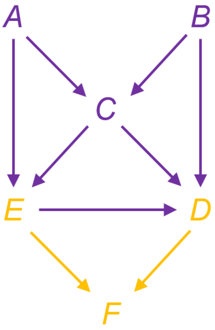
Terminologia



Nota

- No entanto, C NÃO É um colisor no caminho $E \leftarrow C \rightarrow D$.
- Assim, a definição de um colisor é em relação ao caminho que está sendo considerado.

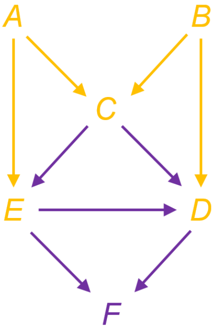
Terminologia



Caminho bloqueado

- O caminho $E \rightarrow F \leftarrow D$ é **bloqueado** desde que este contenha um colisor (F).

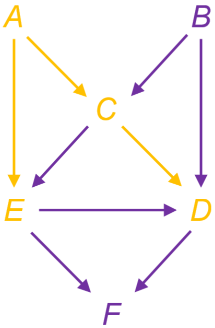
Terminologia



Caminho bloqueado

- Este caminho também é bloqueado (em C).

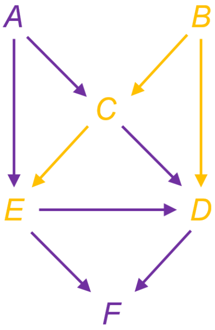
Terminologia



Caminho aberto

- Um caminho que não contém um colisor está **aberto**. Aqui temos um exemplo.

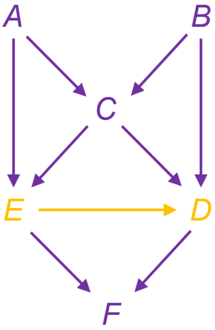
Terminologia



Caminho aberto

- E outro.

Terminologia



Caminho aberto

- E outro.

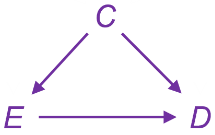
Construindo um DAG

$$E \longrightarrow D$$

Passo 1

- O primeiro passo na construção de um DAG para um problema particular é escrever a **exposição** e o **desfecho** de interesse, com uma seta da exposição para o desfecho.
- Esta seta representa o **efeito causal** que queremos estimar.

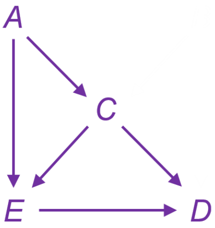
Construindo um DAG



Passo 2

- Se existir qualquer **causa comum** C de E e D, devemos colocá-lo no grafo, com setas de C para E e de C para D.
- Devemos incluir C no grafo, independentemente deste ter sido ou não mensurado em nosso estudo.

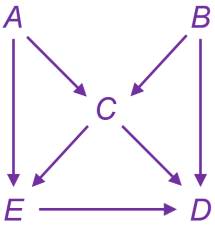
Construindo um DAG



Passo 3

- Continuamos assim, adicionando ao diagrama qualquer variável (observada ou não observada) que é uma **causa comum** de duas ou mais variáveis já existentes no diagrama.

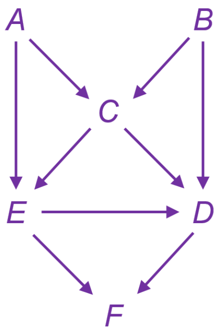
Construindo um DAG



Passo 3

- Continuamos assim, adicionando ao diagrama qualquer variável (observada ou não observada) que é uma **causa comum** de duas ou mais variáveis já existentes no diagrama.

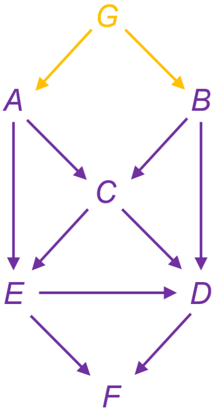
Construindo um DAG



Passo 3

- Se quisermos, podemos também incluir **outras variáveis**, mesmo que eles não sejam causas comuns de outras variáveis no diagrama.
- Por exemplo, F.
- Vamos supor que finalizamos nesse ponto. As variáveis e setas que NÃO estão em nosso grafo representam nossas **suposições causais**.

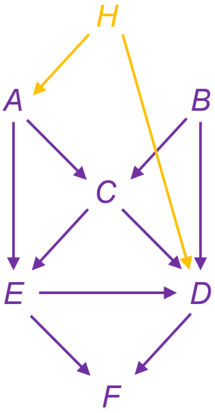
Construindo um DAG



Quais são as nossas suposições?

- Por exemplo, estamos fazendo a suposição que não há uma causa comum G de A e B.

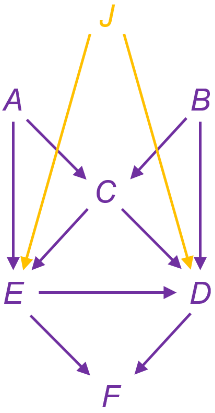
Construindo um DAG



Quais são as nossas suposições?

- E que não há uma causa comum H de A e D.

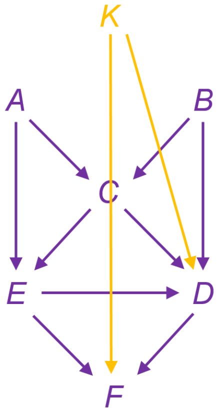
Construindo um DAG



Quais são as nossas suposições?

- E que A, B e C representam TODAS as causas comuns de E e D; não há uma causa comum adicional J.

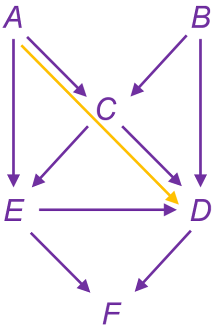
Construindo um DAG



Quais são as nossas suposições?

- E que não há uma causa comum adicional K de D e F.

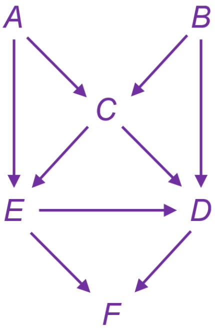
Construindo um DAG



Quais são as nossas suposições?

- Portanto, cada seta omitida também representa uma suposição.
- Por exemplo, estamos assumindo que todo o efeito de A em D atua por meio de C e E.

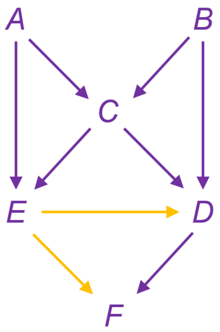
O critério back-door: existe confundimento?



Qual o próximo passo?

- **SE** acreditarmos em nosso diagrama causal, podemos proceder para determinar se a relação $E \rightarrow D$ está **confundida** ou não.
- Isto é feito utilizando o **critério porta dos fundos**.
- O critério porta dos fundos é aplicado em duas partes:
  1. a primeira parte define se existe ou não confundimento.
  2. se existir, a segunda parte determina se é possível controlar o confundimento.

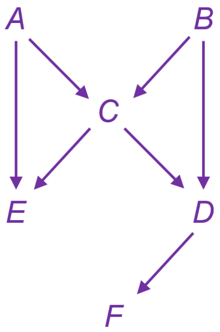
O critério back-door: existe confundimento?



Passo 1

- Primeiro removemos todas as setas **saindo da exposição**.

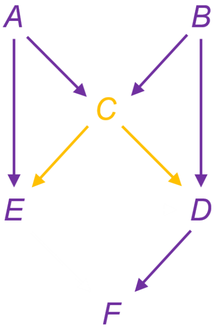
O critério back-door: existe confundimento?



Passo 2

- Em seguida, procuramos por caminhos abertos a partir da exposição até o desfecho.
- Relembrando: um caminho aberto não contém um colisor.

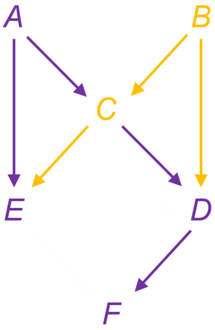
O critério back-door: existe confundimento?



Passo 2

- Este é um caminho aberto?

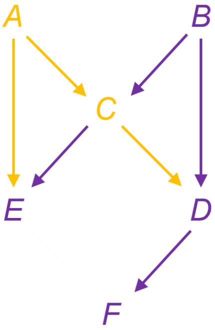
O critério back-door: existe confundimento?



Passo 2

- Este é um caminho aberto?

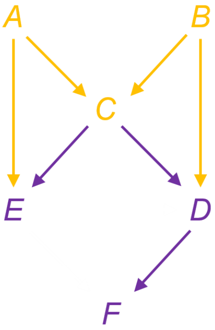
O critério back-door: existe confundimento?



Passo 2

- Este é um caminho aberto?

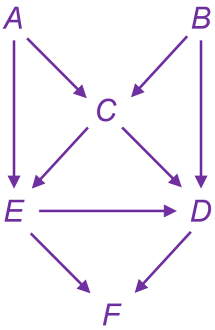
O critério back-door: existe confundimento?



Passo 2

- Este é um caminho aberto?

O critério back-door: existe confundimento?



Existe confundimento?

- Identificamos três caminhos porta dos fundos abertos de E para D. Assim, há confundimento.
- Próxima pergunta: podemos usar alguns ou todos de A, B, C, F para controlar esse confundimento?
- Existe um conjunto $\mathcal{S}$ de variáveis tal que se estratificarmos (ajustarmos) por elas, podemos concluir que o efeito causal existe no estrato?

O critério back-door

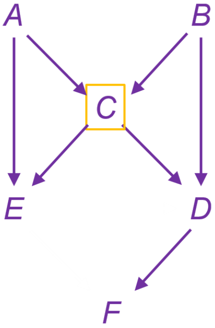
A segunda parte do critério da porta dos fundos nos permite determinar, com base em nosso diagrama causal, se um conjunto de covariáveis candidato é ou não suficiente para controlar o confundimento:

O critério back-door

(i) Primeiro, o conjunto candidato $\mathcal{S}$ não deve conter **descendentes da exposição**.
(ii) Em seguida, removemos todas as setas que saem da exposição.
(iii) Então, nós **juntamos com uma linha tracejada** quaisquer duas variáveis que compartilham um filho que esteja ela mesma em $\mathcal{S}$ ou que tenha um descendente em $\mathcal{S}$.
(iv) Existe um caminho aberto de E para D que não passa por um membro de $\mathcal{S}$?

Se NÃO, então $\mathcal{S}$ é **suficiente** para controlar para confundimento.

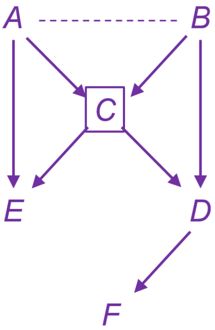
O critério back-door: podemos controlar o confundimento?



O critério back-door: passos (i) e (ii)

- C é suficiente?
- C não é um descendente de E, então o passo (i) é satisfeito.
- Todas as setas saindo da exposição já foram removidas (passo (ii)).

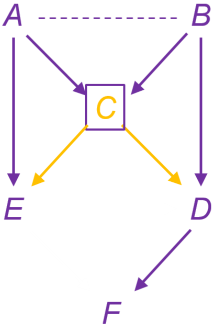
O critério back-door: podemos controlar o confundimento?



passo (iii)

- **Conectamos** A e B com uma linha tracejada, pois eles compartilham um filho (C) que está em nosso conjunto candidato (C).
- Nenhuma outra variável precisa ser conectada desta forma.

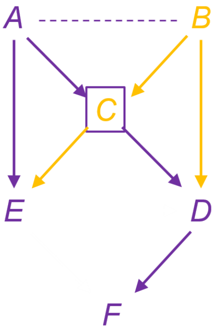
O critério back-door: podemos controlar o confundimento?



passo (iv)

- Agora procuramos por caminhos abertos de E para D e vemos se estes todos passo por C.
- Este está OK!

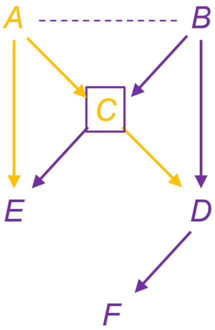
O critério back-door: podemos controlar o confundimento?



passo (iv)

- Este também!

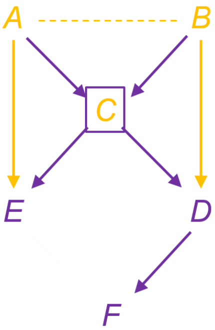
O critério back-door: podemos controlar o confundimento?



passo (iv)

- Este também!

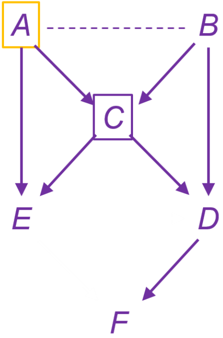
O critério back-door: podemos controlar o confundimento?



passo (iv)

- PORÉM, aqui está um caminho aberto de E para D que NÃO passa por C
- Assim, controlar apenas por C NÃO é suficiente.

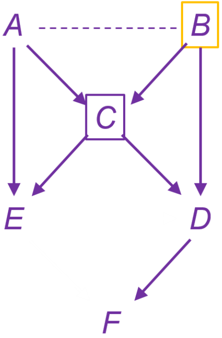
O critério back-door: podemos controlar o confundimento?



Qual é a solução?

- Devemos controlar adicionalmente para A.

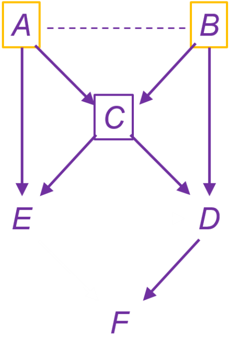
O critério back-door: podemos controlar o confundimento?



Qual é a solução?

- Ou B.

O critério back-door: podemos controlar o confundimento?



Qual é a solução?

- Ou ambos A e B para controlar para o confundimento.

Bons estudos!