

MIKUKLÁŠSKÝ CONFOUNDING

Ondřej Vencálek
Univerzita Palackého v Olomouci
`ondrej.vencalek@upol.cz`

Miku(k)láš
2025-12-04, Praha

Obsah

Korelace a kauzalita

Děti a čápi

Historie: Pearson, Galton a ten třetí (první)

Tři základní dílky stavebnice kauzálního uvažování

Confounder

Mediator

Collider

Očistná hra

Hry 1, 2 a 3

Jak na to: back-door path

Řešení her 1 až 3

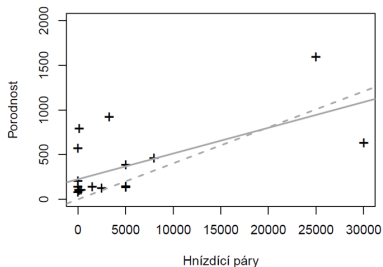
Hry 4 a 5 a jejich řešení

Motto:

Even today, if you stop statistician in the street and ask, “What does confounding mean to you?” you will probably get one of the most convoluted and confounded answers you ever heard from a scientist.

J. Dvořák: O dětech, čápech a kauzalitě (Mikukláš 2021)

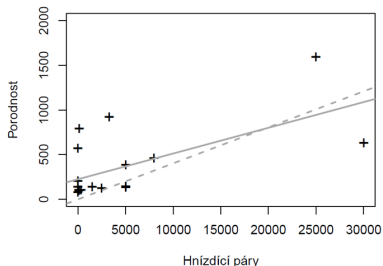
Stát	Čápi [páry]	Porodnost [tis./rok]
Albánie	100	83
Belgie	1	118
Bulharsko	5000	117
Dánsko	9	59
Francie	140	774
Itálie	5	551
Maďarsko	5000	124
Německo	3300	901
Nizozemsko	4	188
Polsko	30 000	610
Portugalsko	1500	120
Rakousko	300	87
Rumunsko	5000	367
Řecko	2500	106
Španělsko	8000	439
Švýcarsko	150	82
Turecko	25 000	1576



$\hat{\rho} \doteq 0,62$ a *statisticky významně* se liší od 0 s p -hodnotou 0,008.

J. Dvořák: O dětech, čápech a kauzalitě (Mikukláš 2021)

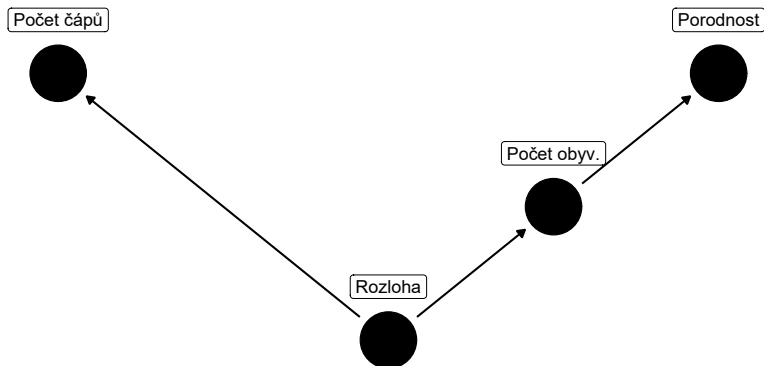
Stát	Čápi [páry]	Porodnost [tis./rok]
Albánie	100	83
Belgie	1	118
Bulharsko	5000	117
Dánsko	9	59
Francie	140	774
Itálie	5	551
Maďarsko	5000	124
Německo	3300	901
Nizozemsko	4	188
Polsko	30 000	610
Portugalsko	1500	120
Rakousko	300	87
Rumunsko	5000	367
Řecko	2500	106
Španělsko	8000	439
Švýcarsko	150	82
Turecko	25 000	1576



$\hat{\rho} \doteq 0,62$ a *statisticky významně se liší od 0 s p-hodnotou 0,008.*

poznámka (ČR od roku 2021): 112, 101, 91, 84, ?

J. Dvořák: O dětech, čápech a kauzalitě (Mikukláš 2021)



Nalezenou korelaci mezi *porodností* a *počtem hnízdících párů čápů* je možné vysvětlit vlivem dalšího faktoru, který působí na obě tyto veličiny. Tímto faktorem je *rozloha daného státu*.

Korelace – trocha historie

- ▶ **Francis Galton (1822–1911)**, výzkum dědičnosti, antropometrické údaje
Galton, F. (1888): Co-Relations and Their Measurement, Chiefly from Anthropometric Data. *Proceedings of the Royal Society of London*. 45 (273–279): 135–145.
- ▶ **Karl Pearson (1857–1936)**,
Pearson, K. (1896): Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London*. Series A, 187, 253–318, (viz str. 265).

Korelace – historie méně známá

Pearson (1896) na str. 261 uvádí:

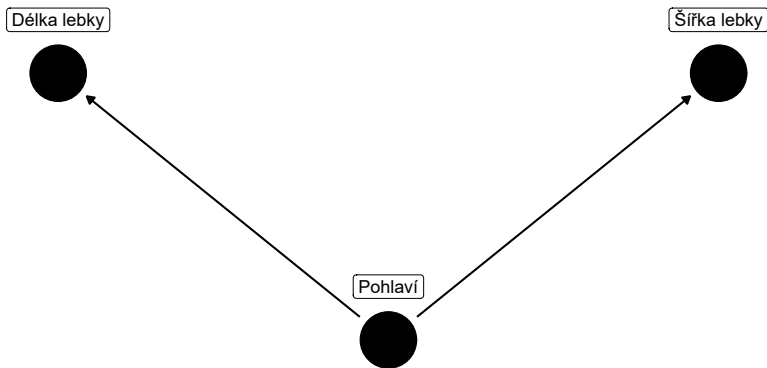
Fundamental theorems of correlation were for the first time and almost exhaustively discussed by BRAVAIS ('Analyse mathe'matique sur les probabilites des erreurs de situation d'un point.' Memoires par divers Savans, T. IX., Paris, 1846, pp. 255-332) nearly half a century ago. He deals completely with the correlation of two and three variables.

- ▶ Tuto míru asociace jako první odvodil anglický psycholog a antropolog sir [Francis Galton](#).
<https://cs.wikipedia.org/wiki/Korelace>
- ▶ [Karl Pearson](#) developed the coefficient from a similar but slightly different idea by Francis Galton.
<https://en.wikipedia.org/wiki/Correlation>
- ▶ It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by [Auguste Bravais](#) in 1844. The naming of the coefficient is thus an example of Stigler's Law. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Korelace – historie méně známá, spurious correlation

- ▶ Pearson již v roce 1899 objevil „zdánlivé/falešné korelace“ (spurious correlation)
- ▶ Vznik: dvě heterogenní populace sloučeny do jedné
- ▶ Měření 806 mužských lebek a 340 ženských lebek z pařížských katakomb
- ▶ Pearson spočítal korelaci mezi délkou lebky a její šířkou.
- ▶ Když výpočet provedl jen pro muže nebo jen pro ženy, byly korelace zanedbatelné – mezi délkou a šířkou lebky neexistovala významná asociace.
- ▶ Ale když byly obě skupiny sloučeny, korelace byla 0,197.

Korelace – historie méně známá, spurious correlation



Pohlaví je **confounderem**.

Poučení

- ▶ Zajímají nás kauzální vztahy (příčina–následek).
- ▶ Korelace není totéž co kauzalita!

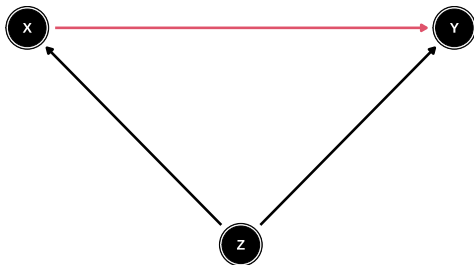
Tři základní dílky stavebnice kauzálního uvažování

Vyšetřování vztahu mezi veličinami X a Y komplikuje třetí veličina Z , která může mít tři různé role:

- ▶  Confounder (společná příčina)
- ▶  Mediator (zprostředkovatel)
- ▶  Collider (společný důsledek)

1. Confounder $X \leftarrow Z \rightarrow Y$

Společná příčina



Příklady:

- ▶ pohlaví (X = délka lebky, Y = šířka lebky)
- ▶ rozsah požáru (X = počet hasičů, Y = výše škody)
- ▶ věk (X = velikost boty, Y = čtenářské schopnosti), viz dále

1. Confounder $X \leftarrow Z \rightarrow Y$

K. Zvára: Regrese (2008):

- ▶ Jiná situace se popisuje anglickým slovem *confounding*. K takovému *matení* dochází tehdy, když vedle nezávisle proměnné x a závisle proměnné y existuje jiná (matoucí) veličina z , která ovlivňuje y nezávisle na hodnotě x , přičemž sama z také souvisí s x . Neexistuje však příčinný řetězec $x \rightarrow z \rightarrow y$.
- ▶ Situace připomíná *Simpsonův paradox* známý z kontingenčních tabulek. Ten je založen na konfrontaci závěrů o závislosti Y na X založených buď na rozdělení podmíněném danou hodnotou třetí veličiny Z nebo na rozdělení marginálním, po vyloučení veličiny Z .
- ▶ Příkladem může být výskyt *rakoviny jícnu* y (měřený například počtem onemocnění na 100 000 obyvatel), který je ovlivňován *podílem x kuřáků v populaci* a současně *spotřebou alkoholu* z . Tyto dvě doprovodné veličiny spolu nepochybně také souvisí.

1. Confounder $X \leftarrow Z \rightarrow Y$

X = velká bota (0/1) Y = úspěch u zkoušky (0/1), Z = starší věk (0/1)

$$P(Y = 1|Z = 1) = 0,8$$

$$P(Y = 1|Z = 0) = 0,3$$

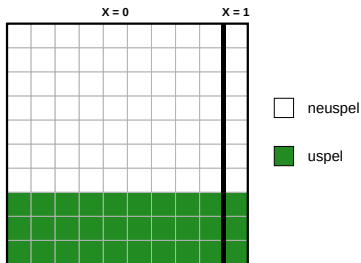
A dále

$$P(X = 1|Z = 1) = 0,9$$

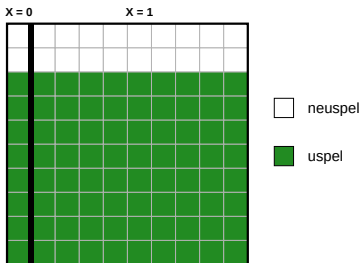
$$P(X = 1|Z = 0) = 0,1$$

1. Confounder $X \leftarrow Z \rightarrow Y$

$Z = 0$

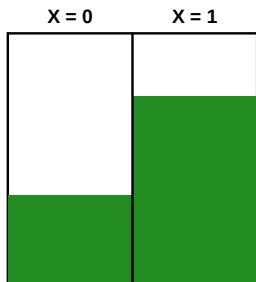


$Z = 1$



1. Confounder $X \leftarrow Z \rightarrow Y$

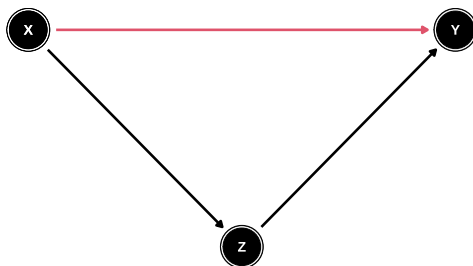
Při ignorování Z ... smícháme případy $Z = 0$ a $Z = 1$ (např. 1:1)



Nezohledněním confounderu vzniká falešná korelace mezi X a Y (vztah mezi nimi není kauzální), tedy **nezohlednění confounderu je chyba!**

2. Mediator $X \rightarrow Z \rightarrow Y$

Prostředník / zprostředkovatel / mediátor

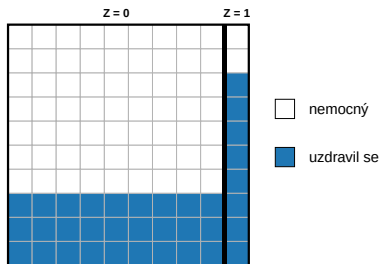


Příklady:

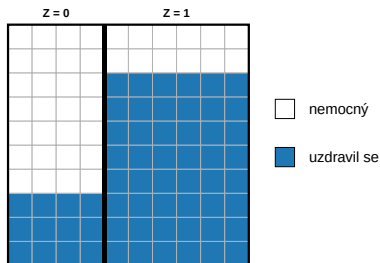
- ▶ kouř (0/1) ($X = \text{oheň (0/1)}$, $Y = \text{alarm (0/1)}$)
- ▶ zenový stav mysli (0/1) ($X = \text{lék (0/1)}$, $Y = \text{uzdravení (0/1)}$), viz dále

2. Mediator $X \rightarrow Z \rightarrow Y$

$X = 0$



$X = 1$



2.  Mediator $X \rightarrow Z \rightarrow Y$

$$P(Y = 1|Z = 1) = 0,8$$

$$P(Y = 1|Z = 0) = 0,3$$

2. Mediator $X \rightarrow Z \rightarrow Y$

$$P(Y = 1|Z = 1) = 0,8 = P(Y = 1|Z = 1, X = 1) = P(Y = 1|Z = 1, X = 0)$$

$$P(Y = 1|Z = 0) = 0,3 = P(Y = 1|Z = 0, X = 1) = P(Y = 1|Z = 0, X = 0)$$

2. Mediator $X \rightarrow Z \rightarrow Y$

$$P(Y = 1|Z = 1) = 0,8 = P(Y = 1|Z = 1, X = 1) = P(Y = 1|Z = 1, X = 0)$$

$$P(Y = 1|Z = 0) = 0,3 = P(Y = 1|Z = 0, X = 1) = P(Y = 1|Z = 0, X = 0)$$

Viděli jsme, že pro $X = 1$ bylo přitom mnohem více zdravených, neboť

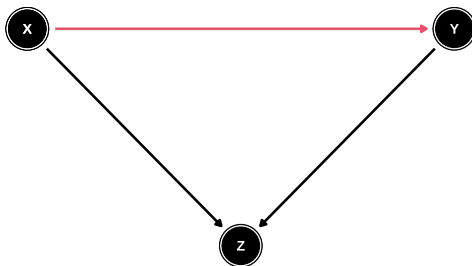
$$P(Z = 1|X = 0) = 0,1$$

$$P(Z = 1|X = 1) = 0,6$$

Adjustací na Z uzavřeme kauzální cestu z X do Y .

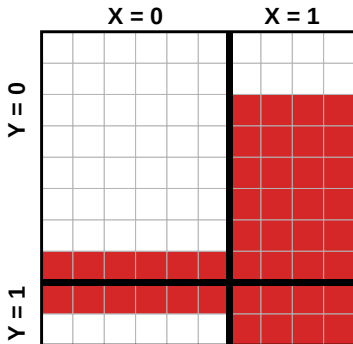
3. Collider $X \rightarrow Z \leftarrow Y$

Společný důsledek

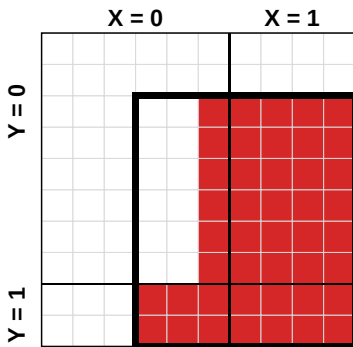


Příklady:

- ▶ úspěch (0/1) ($X = \text{talent (0/1)}$, $Y = \text{krása (0/1)}$)
- ▶ hospitalizace (0/1) ($X = \text{diabetes (0/1)}$, $Y = \text{zlomená noha (0/1)}$), Berksonův paradox

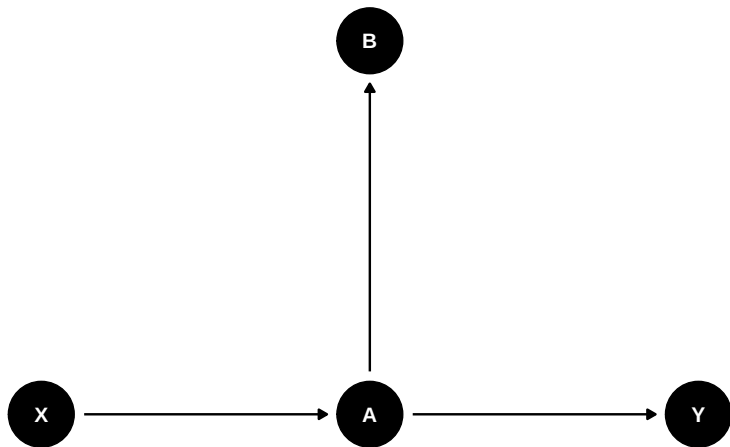


3. Collider $X \rightarrow Z \leftarrow Y$



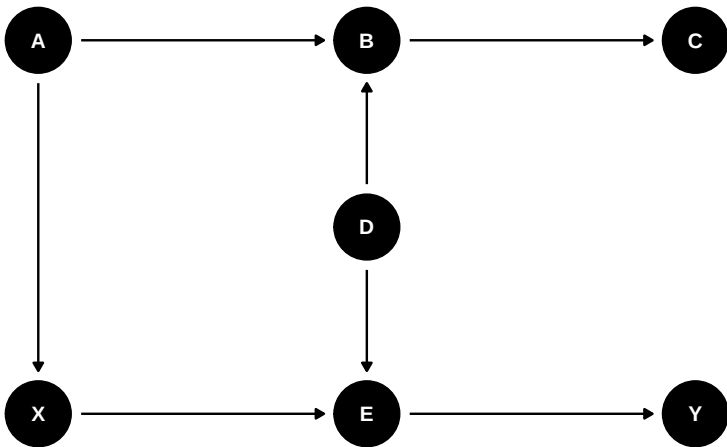
Podmíněním $Z = 1$ vzniká negativní korelace mezi X a Y .

Hra 1

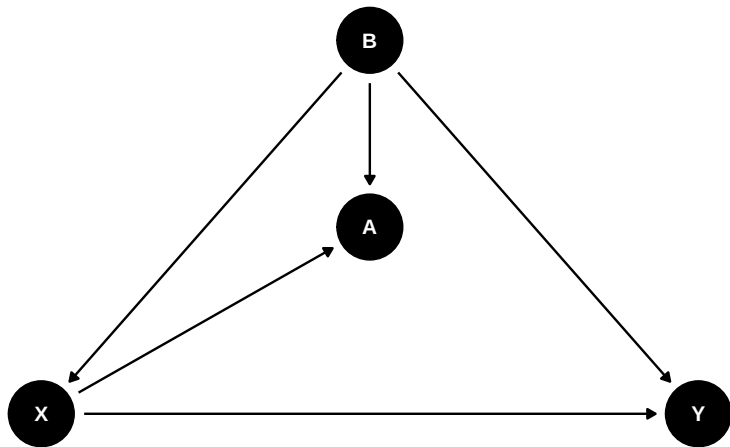


Je B confounderem? (Máme adjustovat?)

Hra 2



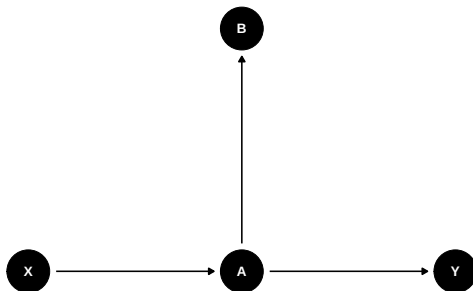
Hra 3



Jak na to: back-door path

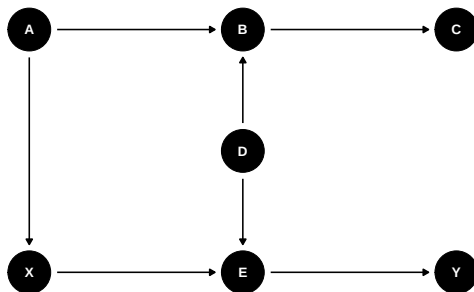
- ▶ **Back-door path** je libovolná cesta z X do Y začínající šipkou směřující do X .
- ▶ Deconfounding = zablokování všech back-door cest z X do Y .
- ▶ Toho dosáhneme kontrolováním některých proměnných Z .
- ▶ Žádná z proměnných v množině Z nesmí být potomkem X na kauzální cestě z X do Y .

Hra 1 – řešení



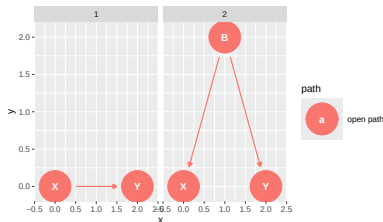
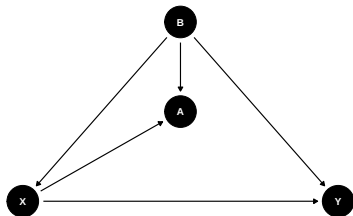
```
1 > paths(hra1, from = "X", to = "Y")
2 $paths
3 [1] "X -> A -> Y"
4 $open
5 [1] TRUE
6 > adjustmentSets(hra2)
7 {}
```

Hra 2 – řešení



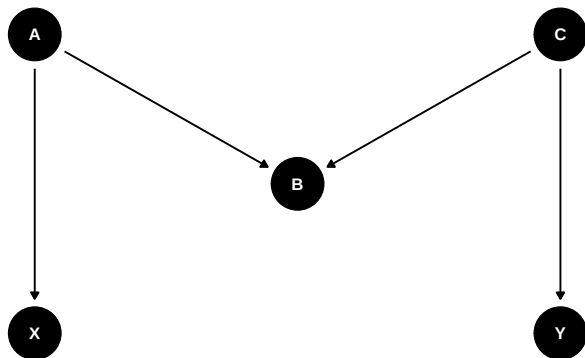
```
1 > paths(hra2, from = "X", to = "Y")
2 $paths
3 [1] "X -> E -> Y" "X <- A -> B <- D -> E -> Y"
4 $open
5 [1] TRUE FALSE
6 > adjustmentSets(hra2)
7 {}
```

Hra 3 – řešení



```
1 > paths(hra3, from = "X", to = "Y")
2 $paths
3 [1] "X -> A <- B -> Y" "X -> Y" "X <- B -> Y"
4 $open
5 [1] FALSE TRUE TRUE
6 > adjustmentSets(hra3)
7 { B }
8 > hra3 %>% ggdag_paths(from = "X", to = "Y")
```


Hra 4

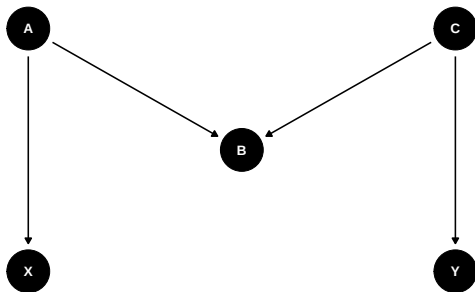


Je B confounderem?

X kouření, Y onemocnění plic,

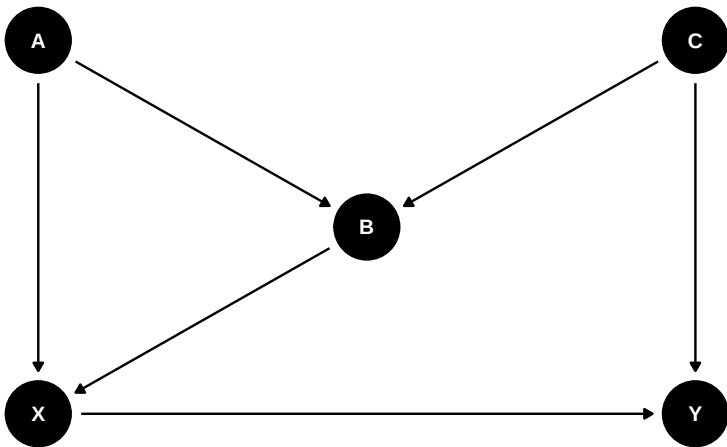
B bezpečnost. pásy, A přístup ke spol. normám, C přístup ke zdraví

Hra 4 – řešení

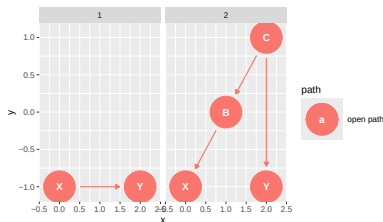
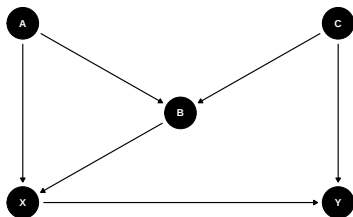


```
1 > paths(hra4, from = "X", to = "Y")
2 $paths
3 [1] "X <- A -> B <- C -> Y"
4 $open
5 [1] FALSE
6 > adjustmentSets(hra4)
7 {}
```

Hra 5



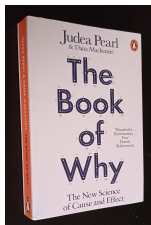
Hra 5 – řešení



```
1 > paths(hra5, from = "X", to = "Y")
2 $paths
3 [1] "X -> Y" "X <- A -> B <- C -> Y" "X <- B <- C -> Y"
4 $open
5 [1] TRUE FALSE TRUE
6 > adjustmentSets(hra5)
7 { C }
8 { A, B }
9 > hra5 %>% ggdag_paths(from = "X", to = "Y")
```

Děkuji za pozornost.

Snad jsem vás dokonale zmátl (confoundoval)



Děkuji za pozornost.

Snad jsem vás dokonale NEzmátl (NEconfoundoval)

