Leve de Wiskunde:
Causaliteit en statistiek

Joris Mooij
j.m.mooij@uva.nl

University of Amsterdam

19 april, 2024

Causale vragen in de wetenschap

Veel wetenschappelijke vragen zijn *causaal* van aard:

- Krijg je longkanker van vapen?

Causale vragen in de wetenschap

Veel wetenschappelijke vragen zijn *causaal* van aard:

- Krijg je longkanker van vapen?
- Word je slimmer van het eten van chocola?

Causale vragen in de wetenschap

Veel wetenschappelijke vragen zijn *causaal* van aard:

- Krijg je longkanker van vapen?
- Word je slimmer van het eten van chocola?
- Hoe goed beschermde het COVID-19 vaccin tegen ziekenhuisopname?

Causale vragen in de wetenschap

Veel wetenschappelijke vragen zijn *causaal* van aard:

- Krijg je longkanker van vapen?
- Word je slimmer van het eten van chocola?
- Hoe goed beschermde het COVID-19 vaccin tegen ziekenhuisopname?

Algemene vorm: als actie X wordt uitgevoerd, wat gebeurt er met Y?

Causale vragen in de maatschappij

Veel maatschappelijke vraagstukken zijn ook causaal van aard:

- Hoeveel uur extra moet je studeren om voor je examen 1 punt hoger te halen?

Algemene vorm: als actie X wordt uitgevoerd, wat gebeurt er met Y?

Causale vragen in de maatschappij

Veel maatschappelijke vraagstukken zijn ook causaal van aard:

- Hoeveel uur extra moet je studeren om voor je examen 1 punt hoger te halen?
- Hoeveel neemt de inflatie in NL af als de ECB de rente met 1 procent punt verhoogt?

Algemene vorm: als actie X wordt uitgevoerd, wat gebeurt er met Y?

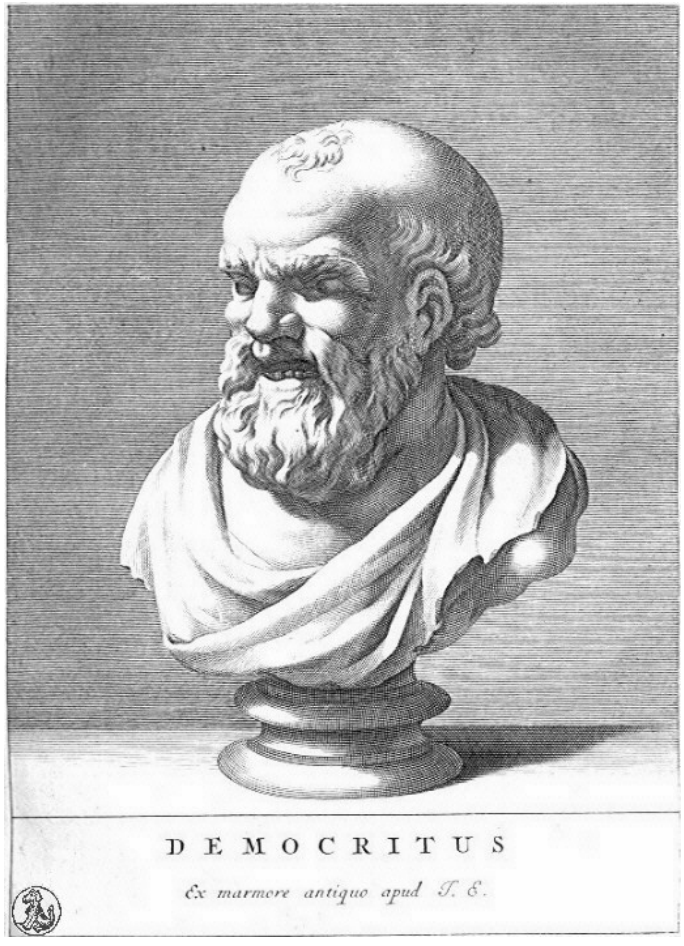
Causale vragen in de maatschappij

Veel maatschappelijke vraagstukken zijn ook causaal van aard:

- Hoeveel uur extra moet je studeren om voor je examen 1 punt hoger te halen?
- Hoeveel neemt de inflatie in NL af als de ECB de rente met 1 procent punt verhoogt?
- Zouden vrouwelijke promovendi die een geslachtsverandering ondergaan meer kans maken om *cum laude* te promoveren?

Algemene vorm: als actie X wordt uitgevoerd, wat gebeurt er met Y?

Geschiedenis van causaliteit in het kort: Democritus



"Ik zou liever een ware oorzaak ontdekken dan het koninkrijk van Perzië bezitten."

Democritus (ca. 460–370 b.C.)

Geschiedenis van causaliteit in het kort: Hume

"Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and infer the existence of the one from that of the other."

David Hume (1711–1776), *Treatise of Human Nature*

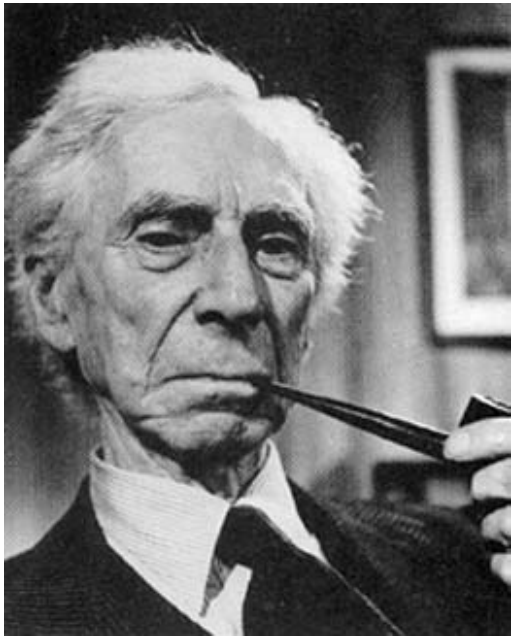
Geschiedenis van causaliteit in het kort: Hume

"Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and infer the existence of the one from that of the other."

David Hume (1711–1776), *Treatise of Human Nature*

(Maar hoe zit het dan met het kraaien van de haan vlak voor zonsopkomst?)

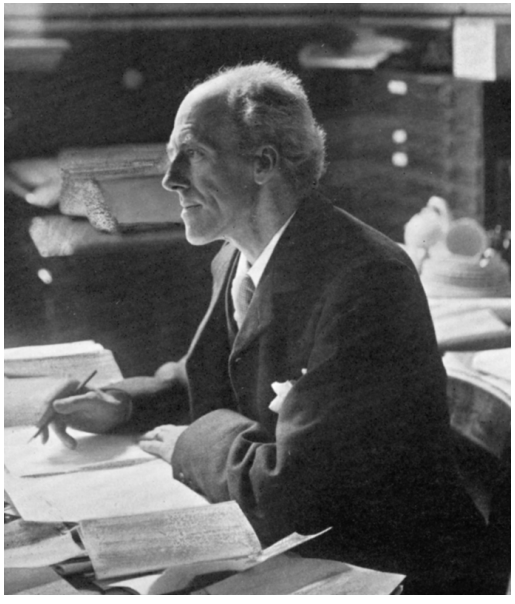
Geschiedenis van causaliteit in het kort: Russell



"All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs. **The law of causality**, I believe, like much that passes muster among philosophers, **is a relic of a bygone age, surviving**, like the monarchy, **only because it is erroneously supposed to do no harm**."

Bertrand Russell (1872–1970), *On The Notion Of Cause*

Geschiedenis van causaliteit in het kort: Pearson



"Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect."

Karl Pearson (1857–1936), *The Grammar of Science*

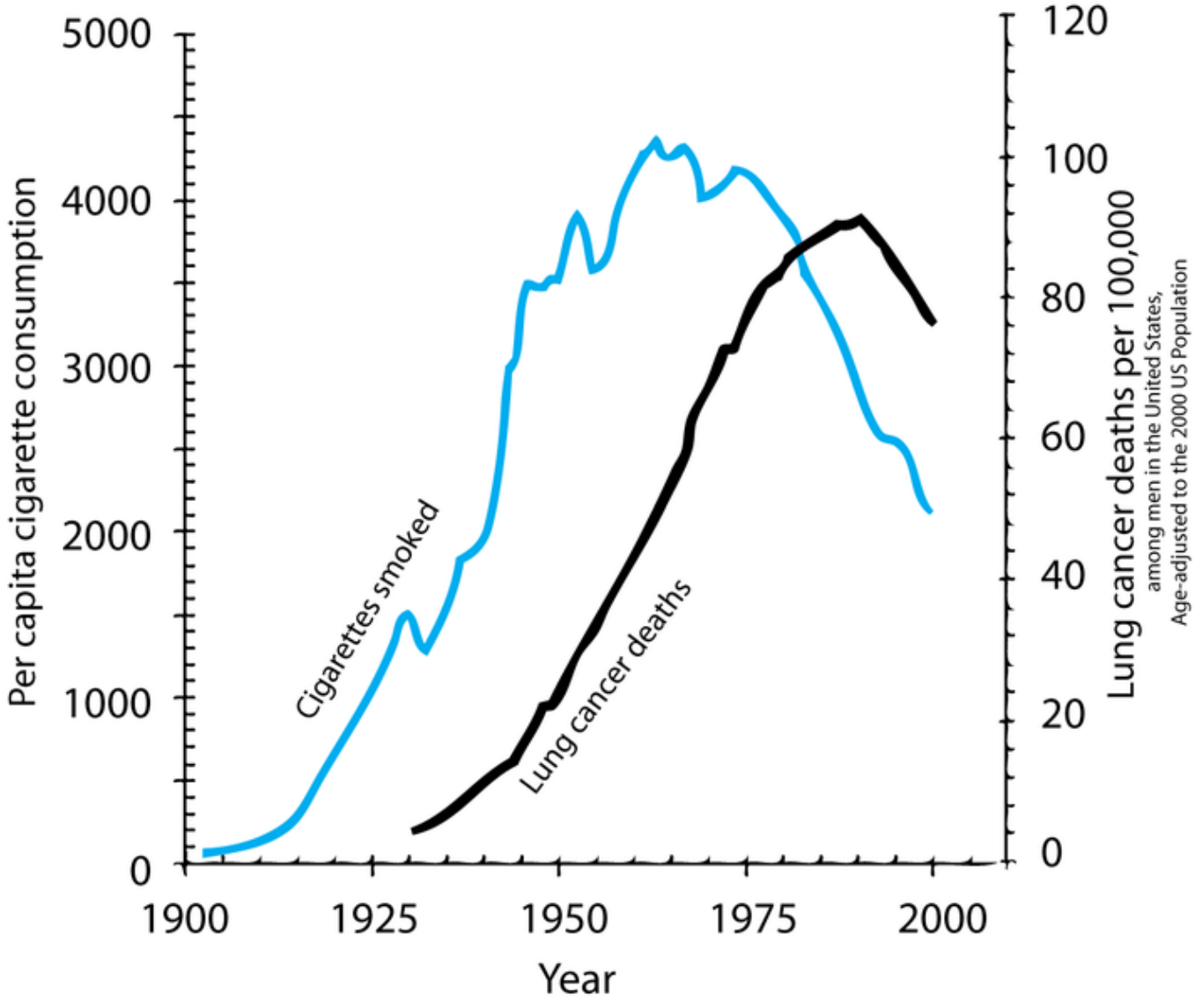
Vandaag de dag. . .

. . . is causaliteit een bloeiend multi-disciplinair vakgebied:

- epidemiologie
- econometrie
- genetica
- machine learning
- wiskunde, statistiek
- kunstmatige intelligentie
- informatica
- . . .

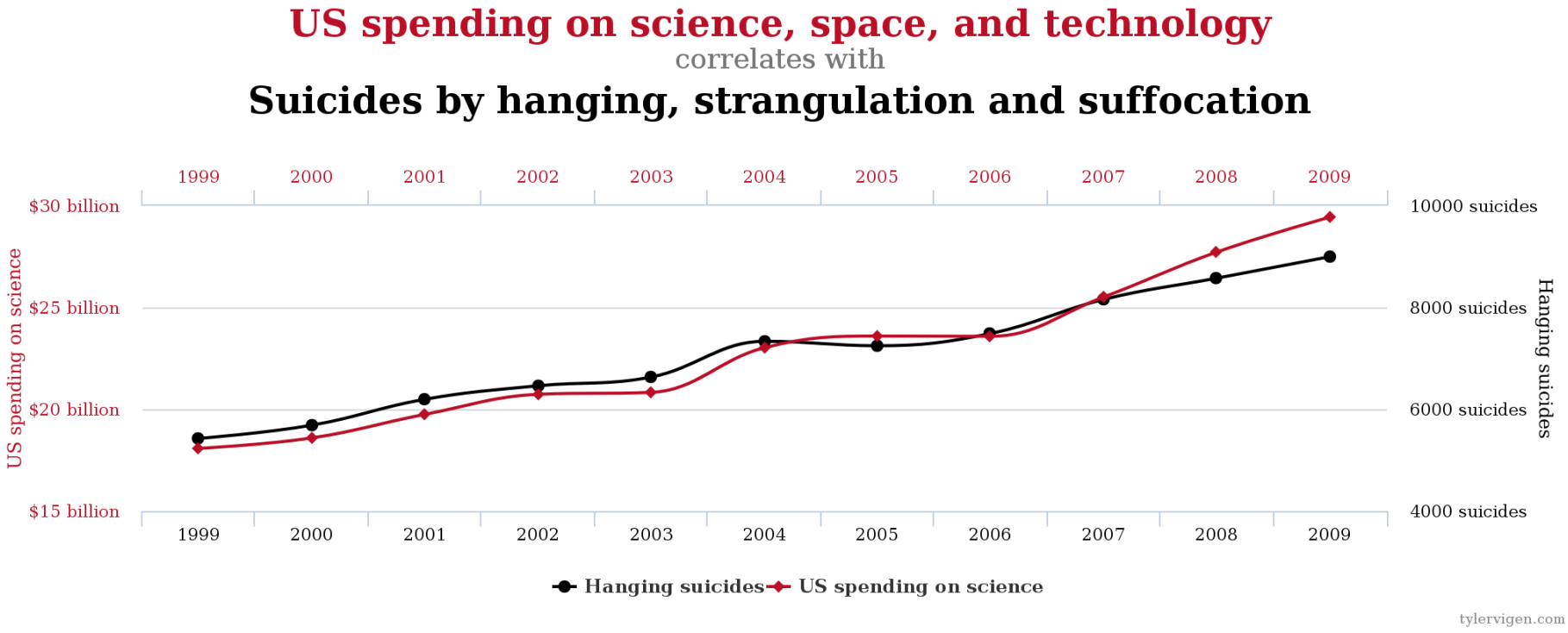
Causale KI is onlangs opgenomen in Gartner's "Hype Cycle for Emerging Technologies" als 1 van de 25 innovatieve technologieën met een verwachte grote impact in de komende 5–10 jaar.

Veroorzaakt roken longkanker?



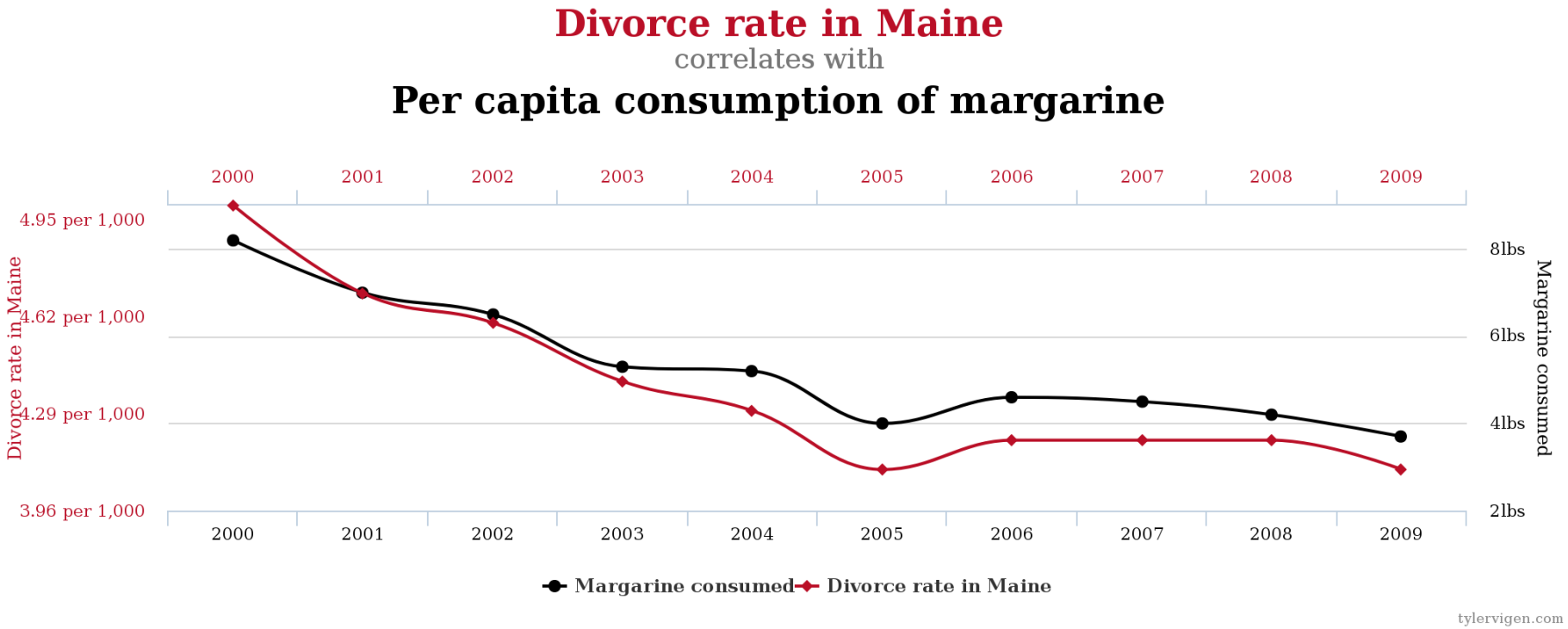
Source: `https://scpoecon.github.io/ScPoEconometrics/causality.html`

Correlatie of oorzakelijk verband?



Source: `http://tylervigen.com/spurious-correlations`

Correlatie of oorzakelijk verband?



Source: `http://tylervigen.com/spurious-correlations`

Statistische significantie

- Is de correlatie wel significant?
- Of is het gewoon 'toeval'?
- Statistische valkuil: **multiple testing**
  (als je maar lang genoeg zoekt naar een patroon in de data, dan vind je altijd wel iets).

Op zoek naar een definitie

Definition (deterministische systemen (schets))

Als een actie X de *mogelijke waarden* van Y verandert, zeggen we: X **veroorzaakt** Y.

Op zoek naar een definitie

Definition (deterministische systemen (schets))

Als een actie X de *mogelijke waarden* van Y verandert, zeggen we: X **veroorzaakt** Y.

Op zoek naar een definitie

Definition (stochastische systemen (schets))

Als een actie X de *kansverdeling* van Y verandert, zeggen we: X **veroorzaakt** Y.

Op zoek naar een definitie

Definition (stochastische systemen (schets))

Als een actie X de *kansverdeling* van Y verandert, zeggen we: X **veroorzaakt** Y.

Nobel prijzen en chocolade consumptie

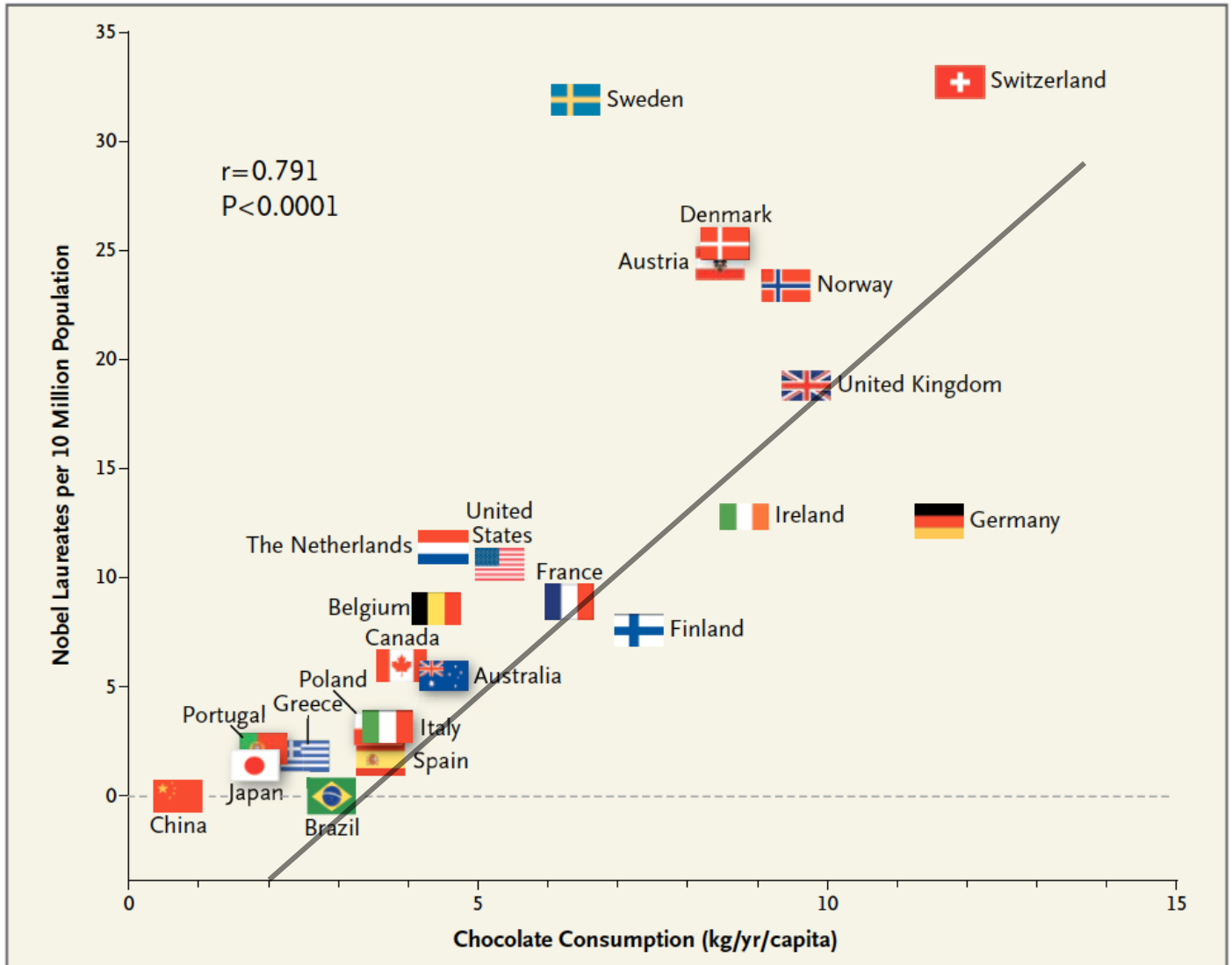
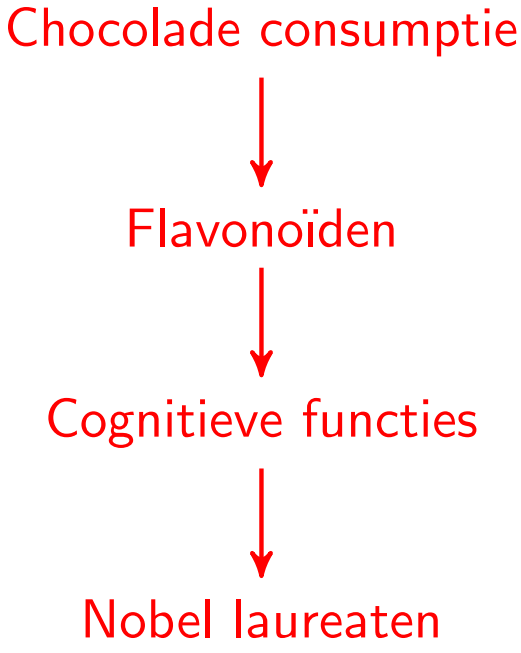


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Messerli, *New England Journal of Medicine*, 2012

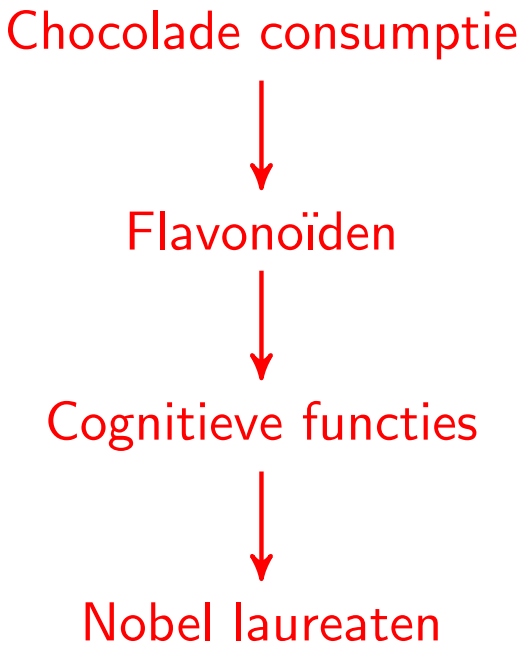
Causale hypotheses

Hypothese H1:

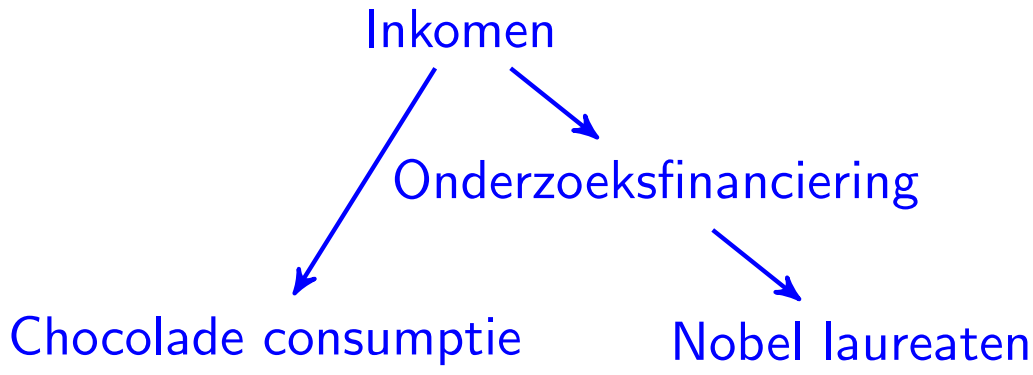


Causale hypotheses

Hypothese H1:

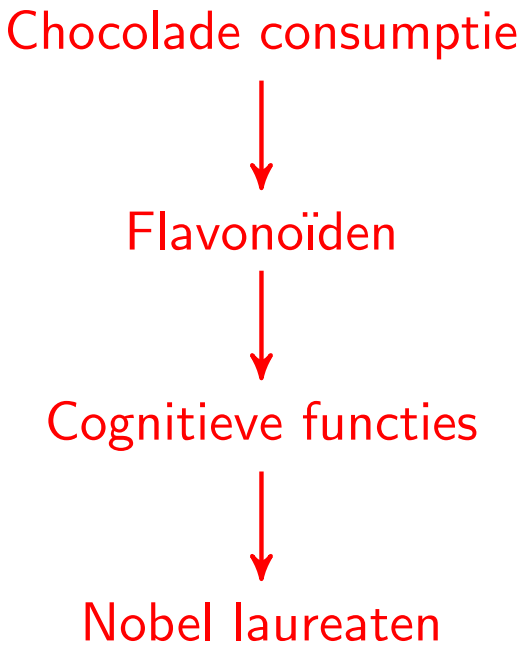


Hypothese H2:

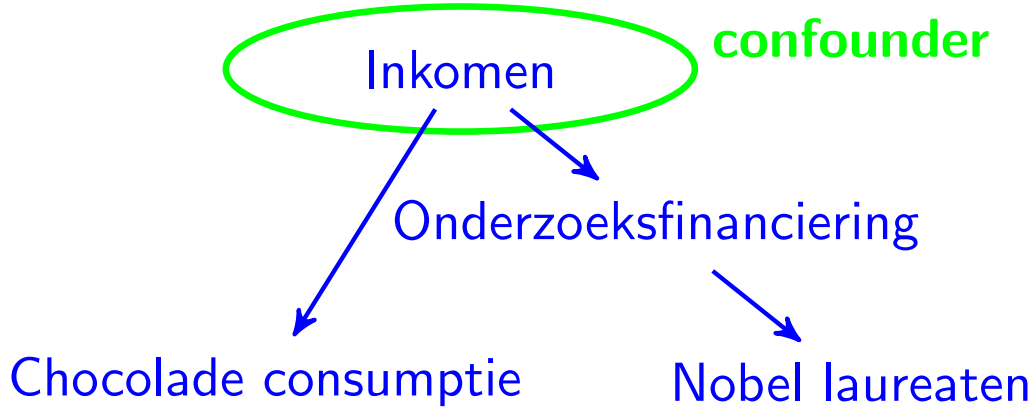


Causale hypotheses

Hypothese H1:



Hypothese H2:



Stel elke Nederlander zou dubbel zoveel chocolade eten?

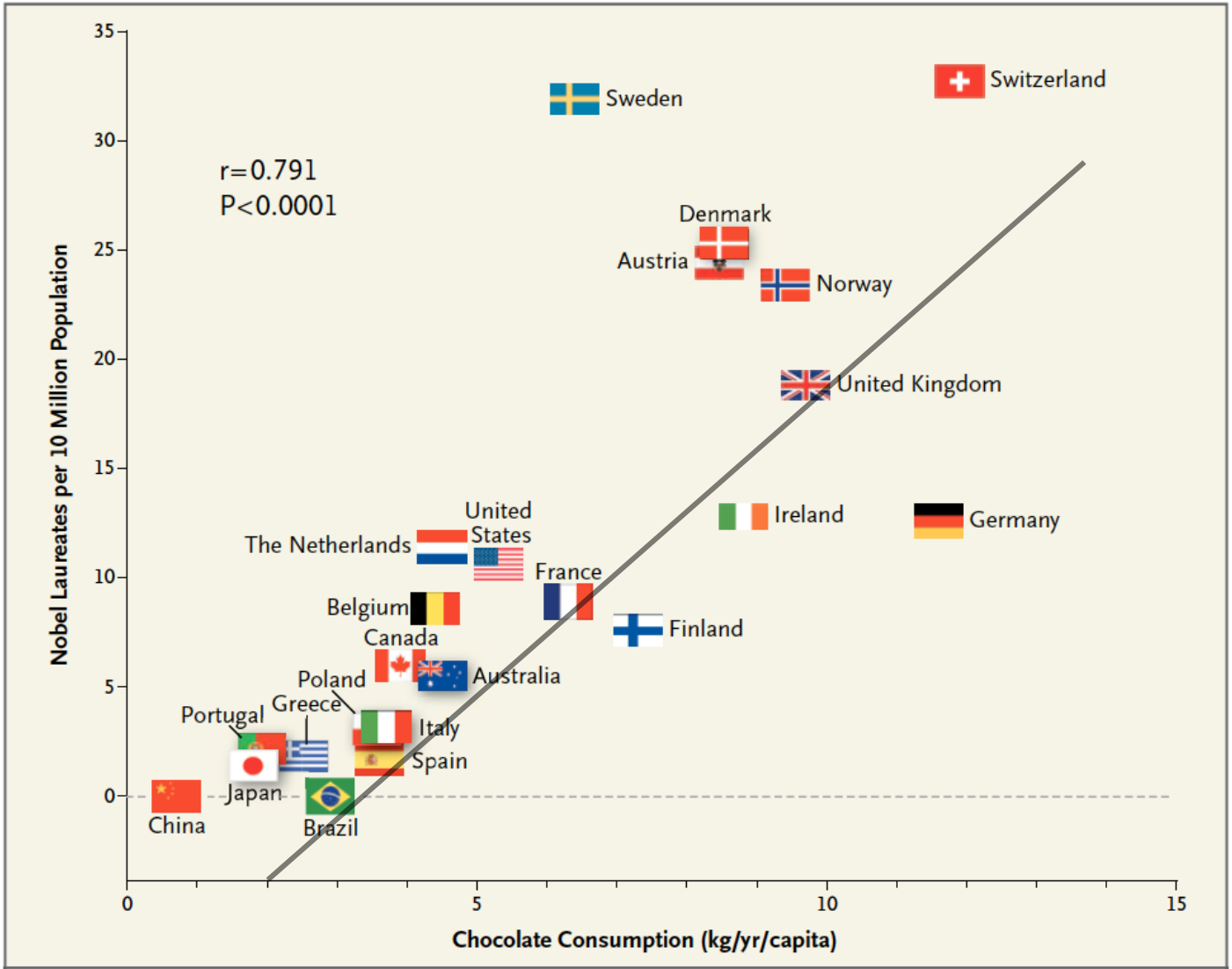


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Stel elke Nederlander zou dubbel zoveel chocolade eten?

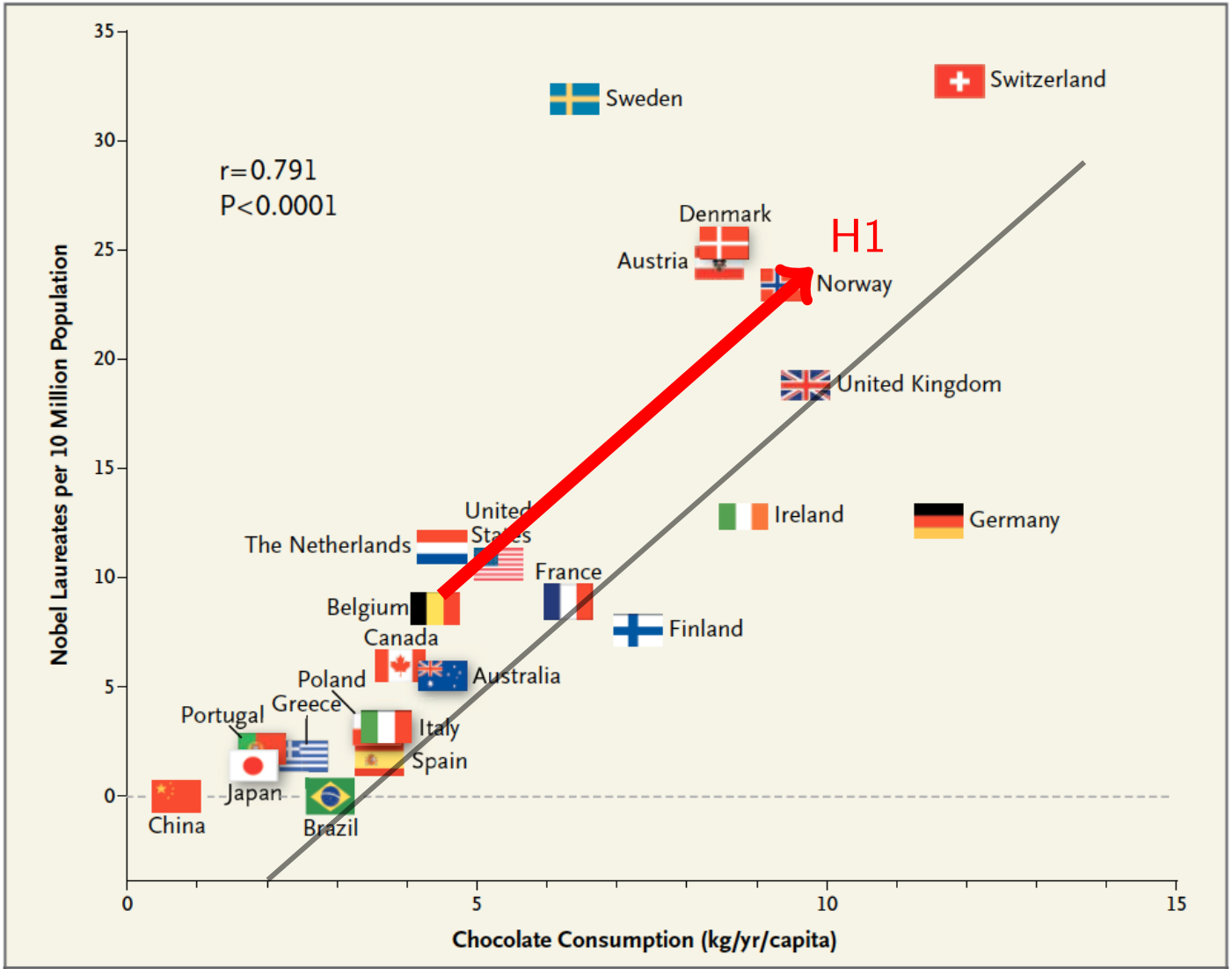


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Stel elke Nederlander zou dubbel zoveel chocolade eten?

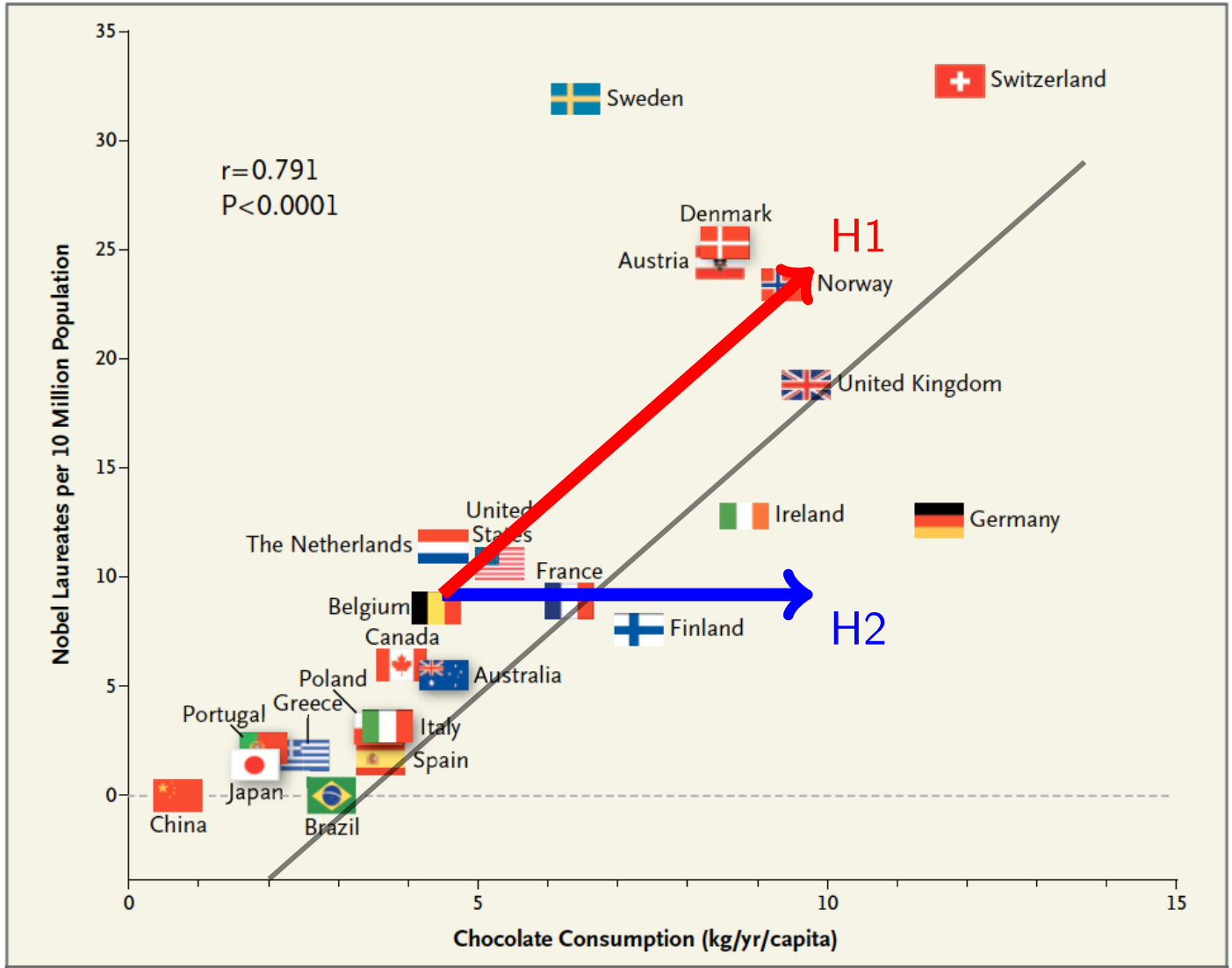
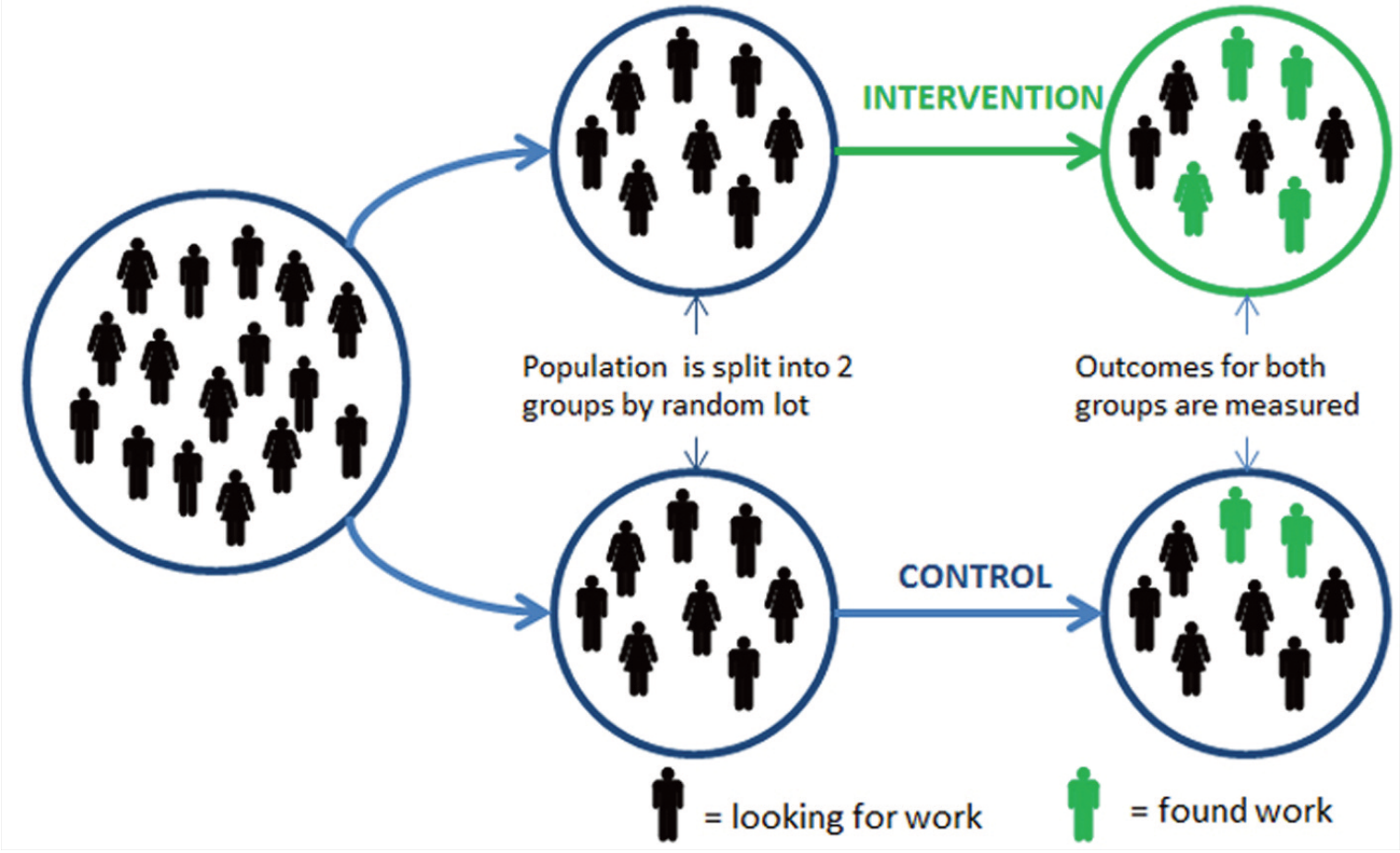


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Gerandomiseerde studies met controlegroep

Gouden standaard voor het schatten van effecten



Jan Baptista van Helmont, 1579–1644

Beperkingen van RCTs

Beperkingen van gerandomiseerde studies met controlegroep:

- Logistieke aspecten
- Ethische aspecten
- Inclusie criteria

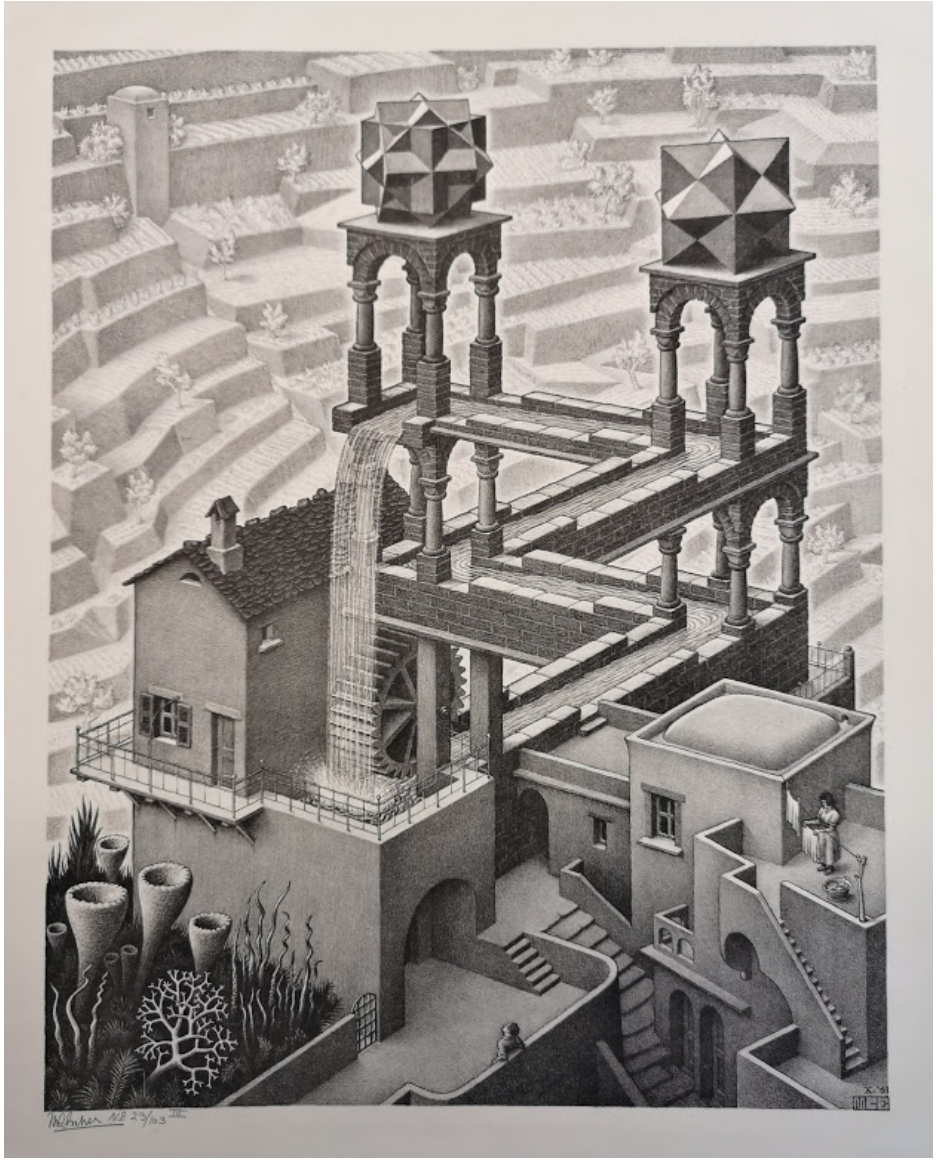
Beperkingen van RCTs

Beperkingen van gerandomiseerde studies met controlegroep:

- Logistieke aspecten
- Ethische aspecten
- Inclusie criteria

Alternatief:
Kunnen we gebruik maken van *observationele* data?

Simpson's paradox: a statistische illusie



Maurits C. Escher (1898–1972)

Simpson's paradox in EPR data

Variabele	Betekenis	Waarden	Voorbeeld
X	Actie	A or B	type COVID-19 vaccin
Y	Uitkomst	$+$ or $-$	ziekenhuisopname na COVID-19

Simpson's paradox in EPR data

Variabele	Betekenis	Waarden	Voorbeeld
X	Actie	A or B	type COVID-19 vaccin
Y	Uitkomst	$+$ or $-$	ziekenhuisopname na COVID-19

EPR data (10,000 individuen):

	♂+♀	
	$+$	$-$
A	2500	2500
B	3000	2000

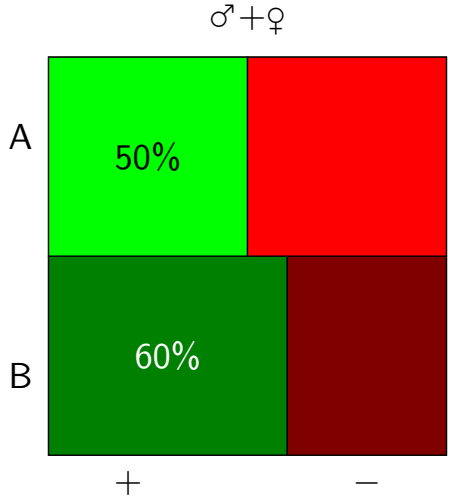
Simpson's paradox in EPR data

Variabele	Betekenis	Waarden	Voorbeeld
X	Actie	A or B	type COVID-19 vaccin
Y	Uitkomst	$+$ or $-$	ziekenhuisopname na COVID-19

EPR data (10,000 individuen):

	♂+♀	
	$+$	$-$
A	2500	2500
B	3000	2000

Fracties van positieve uitkomsten gevisualiseerd:



Simpson's paradox in EPR data

Variabele	Betekenis	Waarden	Voorbeeld
X	Actie	A or B	type COVID-19 vaccin
Y	Uitkomst	$+$ or $-$	ziekenhuisopname na COVID-19
Z	Geslacht	♂ / ♀	man / vrouw

EPR data (10,000 individuen):

♂+♀

	+	−
A	2500	2500
B	3000	2000

=

♂

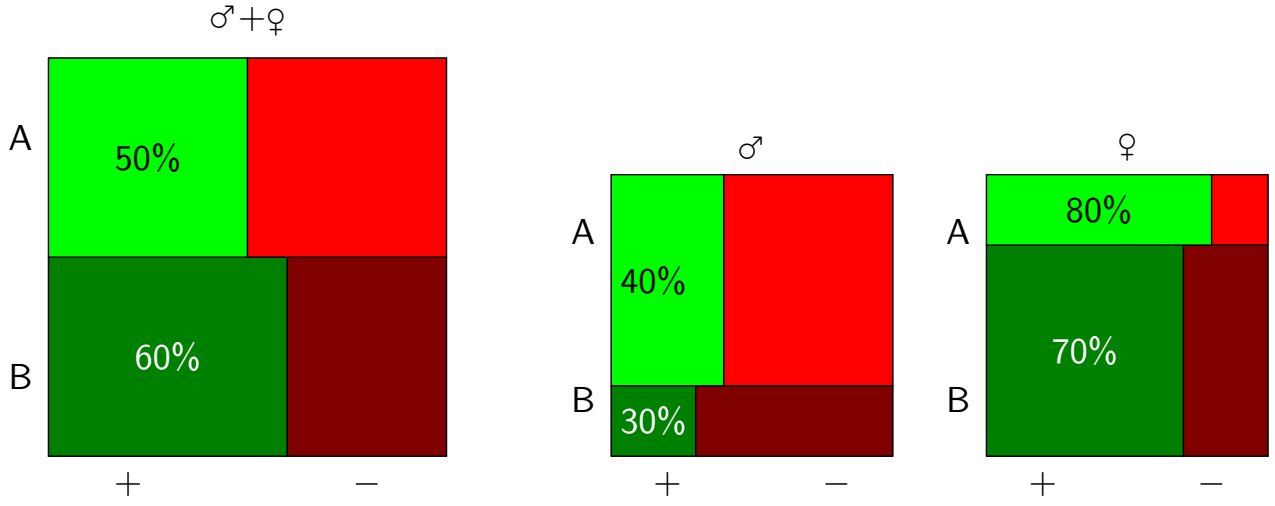
	+	−
A	1500	2250
B	375	875

+

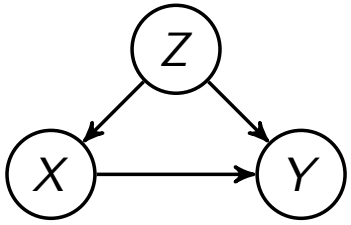
♀

	+	−
A	1000	250
B	2625	1125

Fracties van positieve uitkomsten gevisualiseerd:



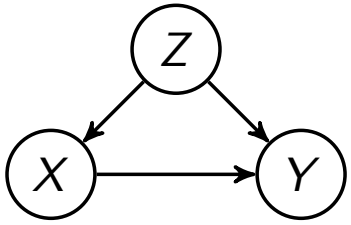
Causale hypotheses voor EPR data



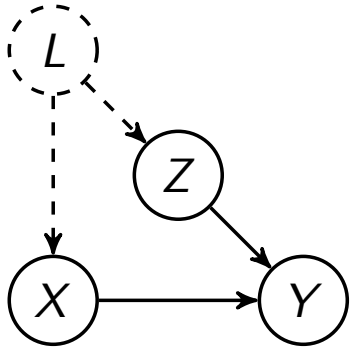
Prefereer actie A

X actie
Z geslacht
Y uitkomst
L verborgen variabele

Causale hypotheses voor EPR data



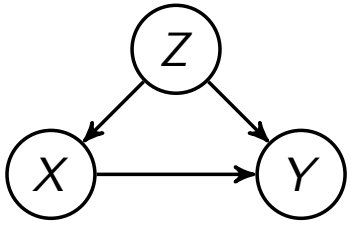
Prefereer actie A



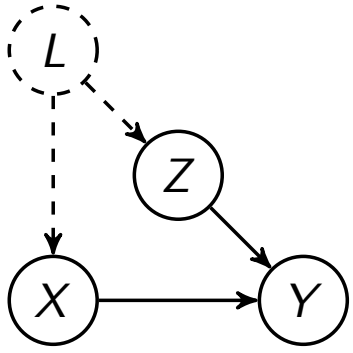
Prefereer actie A

X	actie
Z	geslacht
Y	uitkomst
L	verborgen variabele

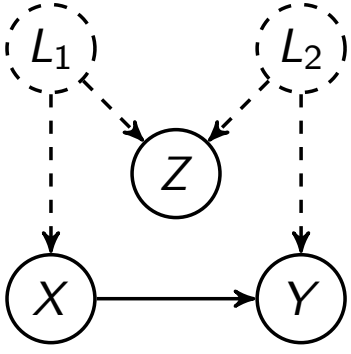
Causale hypotheses voor EPR data



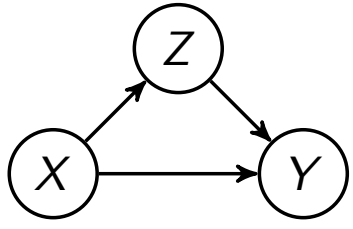
Prefereer actie A



Prefereer actie A



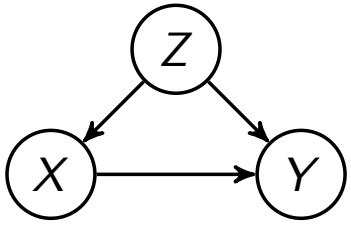
Prefereer actie B



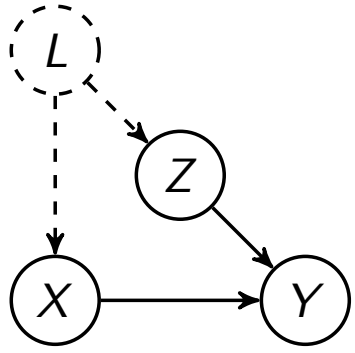
Prefereer actie B

X	actie
Z	geslacht
Y	uitkomst
L	verborgen variabele

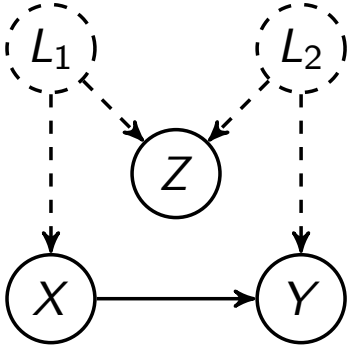
Causale hypotheses voor EPR data



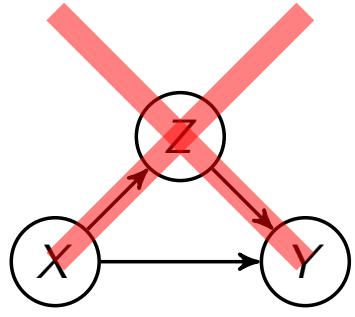
Prefereer actie A



Prefereer actie A



Prefereer actie B



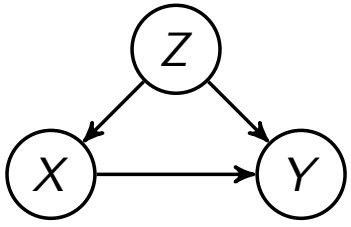
Prefereer actie B

X	actie
Z	geslacht
Y	uitkomst
L	verborgen variabele

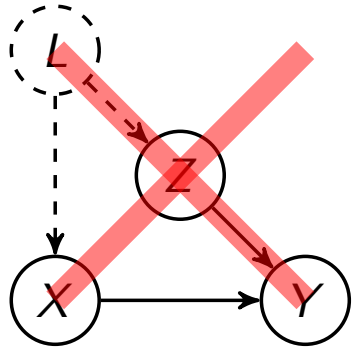
Aannames:

- actie X heeft geen invloed op geslacht Z

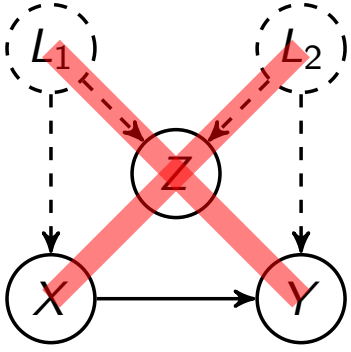
Causale hypotheses voor EPR data



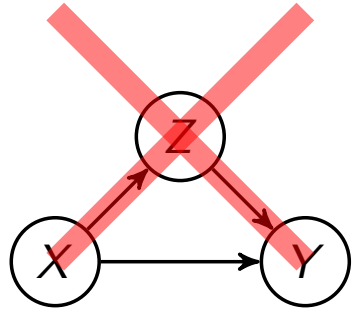
Prefereer actie A



Prefereer actie A



Prefereer actie B



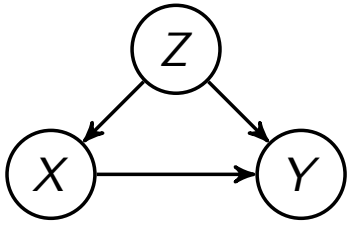
Prefereer actie B

X	actie
Z	geslacht
Y	uitkomst
L	verborgen variabele

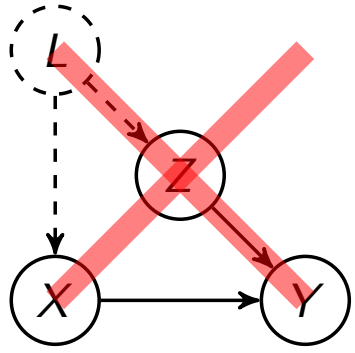
Aannames:

- actie X heeft geen invloed op geslacht Z
- geen verborgen gezamenlijke oorzaak (confounder) van Z en ..

Causale hypotheses voor EPR data

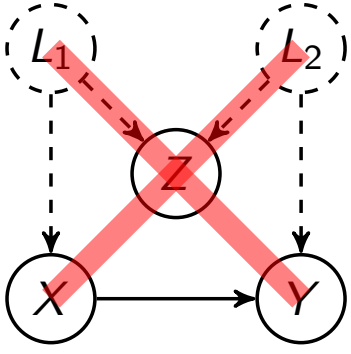


Prefereer actie A

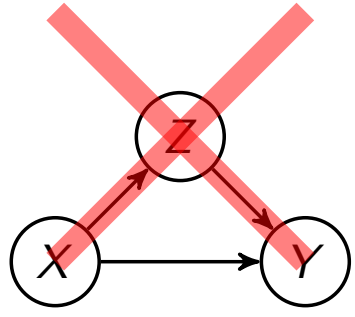


Prefereer actie A

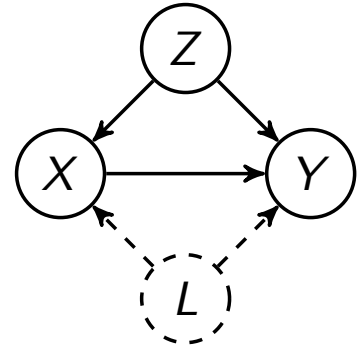
X actie
Z geslacht
Y uitkomst
L verborgen variabele



Prefereer actie B



Prefereer actie B



Onduidelijk!

Aannames:

- actie X heeft geen invloed op geslacht Z
- geen verborgen gezamenlijke oorzaak (confounder) van Z en ..

Natuurlijke grenzen

Theorem (Manski & Nagin, *Sociological Methodology*, 1998)

Voor discrete variabelen X, Y, zelfs met verborgen gezamenlijke oorzaken,

$$\begin{aligned} p(X = x, Y = y) &\leq p(Y(x) = y) \\ &\leq p(X = x, Y = y) + 1 - p(X = x). \end{aligned}$$

Proof.

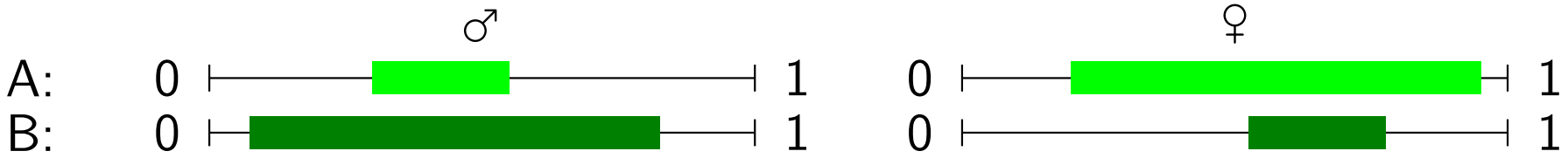
Gebruik makend van consistentie ($Y = Y(X)$) en elementaire kansrekening:

$$\begin{aligned} p(X = x, Y = y) &= p(X = x, Y(x) = y) \\ &\leq p(Y(x) = y) \\ &= p(X = x, Y(x) = y) + p(X \neq x, Y(x) = y) \\ &\leq p(X = x, Y = y) + p(X \neq x) \\ &= p(X = x, Y = y) + 1 - p(X = x) \end{aligned}$$

Natuurlijke grens

We kunnen de natuurlijke grens toepassen op de EPR data.

De kans op een positieve uitkomst onder actie X,
$p(Y = + \,|\, \mathrm{do}(X), Z)$, moet in de volgende intervallen liggen:

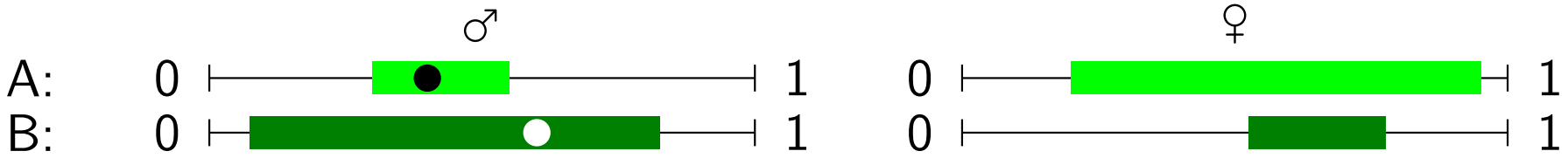


- Uit alleen de observationele data kunnen we **niet concluderen** welke actie te prefereren is.

Natuurlijke grens

We kunnen de natuurlijke grens toepassen op de EPR data.

De kans op een positieve uitkomst onder actie X,
$p(Y = + \,|\, \mathrm{do}(X), Z)$, moet in de volgende intervallen liggen:

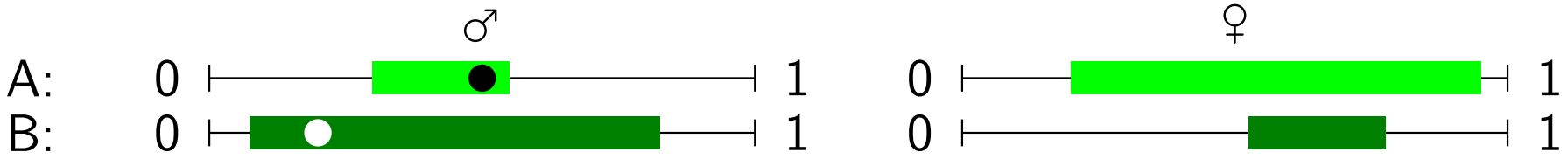


- Uit alleen de observationele data kunnen we **niet concluderen** welke actie te prefereren is.

Natuurlijke grens

We kunnen de natuurlijke grens toepassen op de EPR data.

De kans op een positieve uitkomst onder actie X,
$p(Y = + \,|\, \mathrm{do}(X), Z)$, moet in de volgende intervallen liggen:

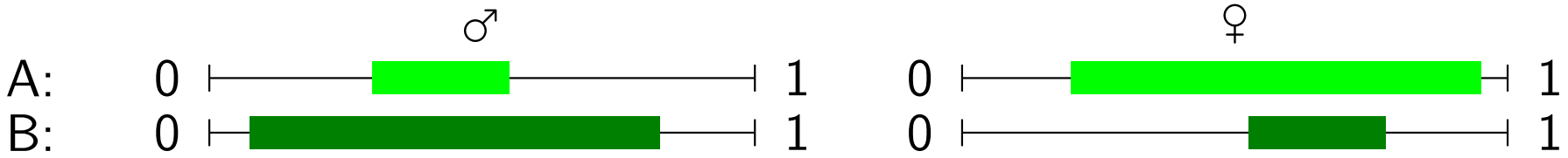


- Uit alleen de observationele data kunnen we **niet concluderen** welke actie te prefereren is.

Natuurlijke grens

We kunnen de natuurlijke grens toepassen op de EPR data.

De kans op een positieve uitkomst onder actie X,
$p(Y = + \,|\, \mathrm{do}(X), Z)$, moet in de volgende intervallen liggen:

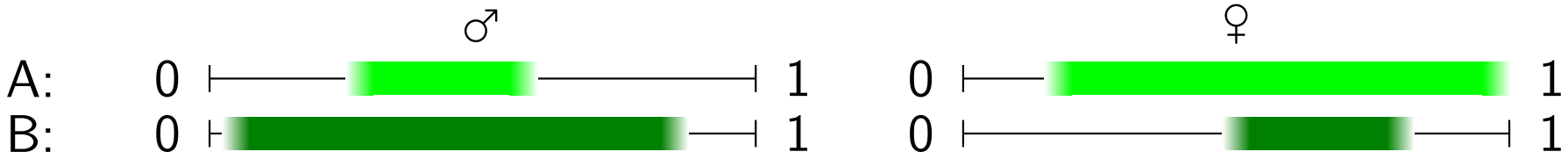


- Uit alleen de observationele data kunnen we **niet concluderen** welke actie te prefereren is.

Natuurlijke grens

We kunnen de natuurlijke grens toepassen op de EPR data.

De kans op een positieve uitkomst onder actie X,
$p(Y = + \,|\, \mathrm{do}(X), Z)$, moet in de volgende intervallen liggen:

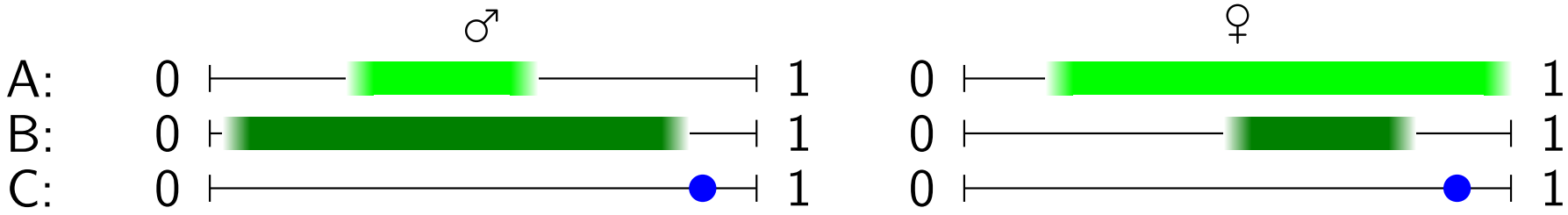


- Uit alleen de observationele data kunnen we **niet concluderen** welke actie te prefereren is.
- De grootste onzekerheid is “causaal”; daarbovenop komt “statistische” onzekerheid omdat we extrapoleren vanuit een (mogelijk kleine) steeproef.

Natuurlijke grens

We kunnen de natuurlijke grens toepassen op de EPR data.

De kans op een positieve uitkomst onder actie X,
$p(Y = + \,|\, \mathrm{do}(X), Z)$, moet in de volgende intervallen liggen:



- Uit alleen de observationele data kunnen we **niet concluderen** welke actie te prefereren is.
- De grootste onzekerheid is "causaal"; daarbovenop komt "statistische" onzekerheid omdat we extrapoleren vanuit een (mogelijk kleine) steeproef.
- Maar: de data **is informatief**. Bijvoorbeeld, als een derde actie C wordt geevalueerd in een RCT, dan hebben we minder proefpersonen nodig om de beste actie te vinden.

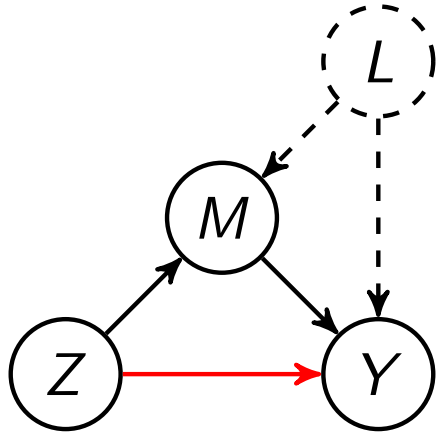
Simpson's paradox: echt voorbeeld

- Toelating van master studenten bij University College Berkeley
- 45% van alle mannelijke studenten werd toegelaten, maar slechts 30% van alle vrouwelijke studenten
- Op afdelingsniveau verdwijnt het verschil bijna geheel
- Verklaring: vrouwelijke studenten schreven zich vaker in bij een afdeling met lagere toelatingskans

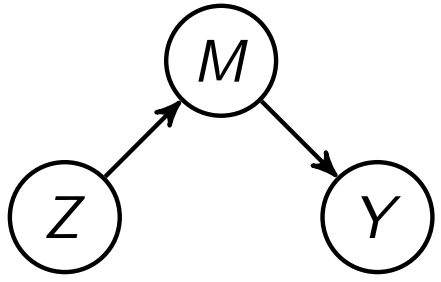
Dus, is hier sprake van discriminatie naar geslacht?

Causale hypotheses toetsen

Z	geslacht
Y	toelating
M	afdeling
L	verborgen variabele (wiskundige vaardigheden)



$Z \not\perp\!\!\!\perp Y \mid M$



$Z \perp\!\!\!\perp Y \mid M$

Aantal studenten: 4257. Statistische onafhankelijkheidstoetsen:

$Z \perp\!\!\!\perp Y$:	$p = 4 \times 10^{-22}$	(sterke associatie)
$Z \perp\!\!\!\perp Y \mid M$:	$p = 0.00135$	(zwakke voorwaardelijke associatie)

Dit suggereert dat het selectieproces (nauwelijks) discriminerend is naar geslacht.

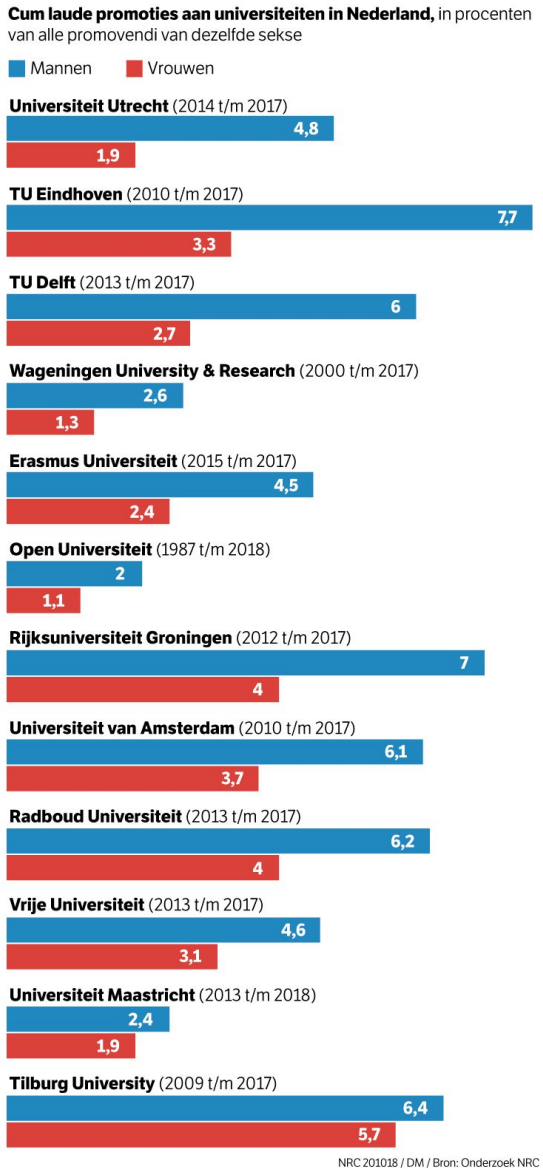
Discriminatie en causaliteit

Menu | nrc

Helft van de promovendi is vrouw, maar cum laude krijgen ze zelden

Wetenschappelijk bedrijf Aan alle Nederlandse universiteiten hadden mannen de afgelopen jaren meer kans om cum laude te promoveren dan vrouwen. De criteria voor cum laude promoveren zijn niet objectief gedefinieerd, dus er is volop ruimte voor genderbias.

Ellen de Bruin 19 oktober 2018 Leestijd 7 minuten

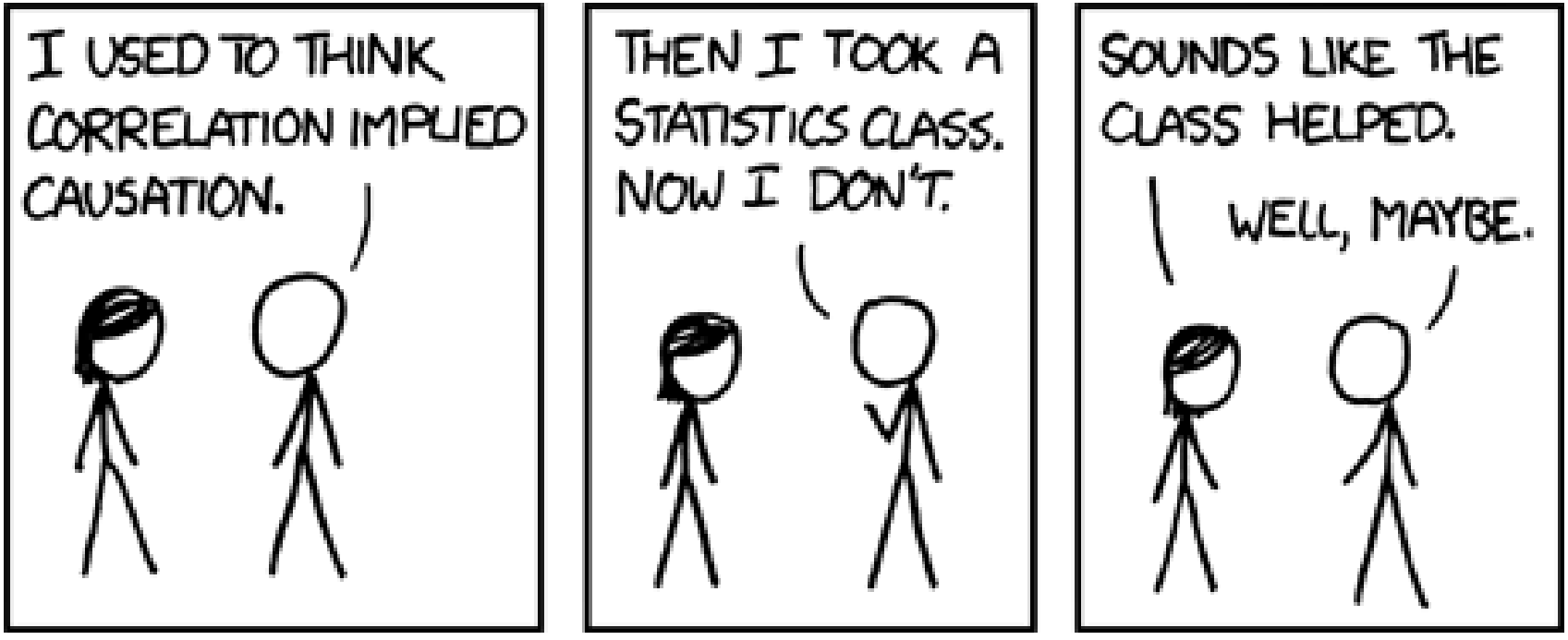


Cum laude promoties aan universiteiten in Nederland, in procenten van alle promovendi van dezelfde sekse
Mannen Vrouwen

Bron: NRC

Conclusie

Take-home: Wiskundige modellen stellen ons in staat om causaliteit beter te begrijpen en betrouwbaardere conclusies te trekken uit data.



xkcd.org