

Curs 04 – Modelare și eșantionare

Corelația

Erori de modelare

Erori de eșantionare

10/16/2022

Structura cursului

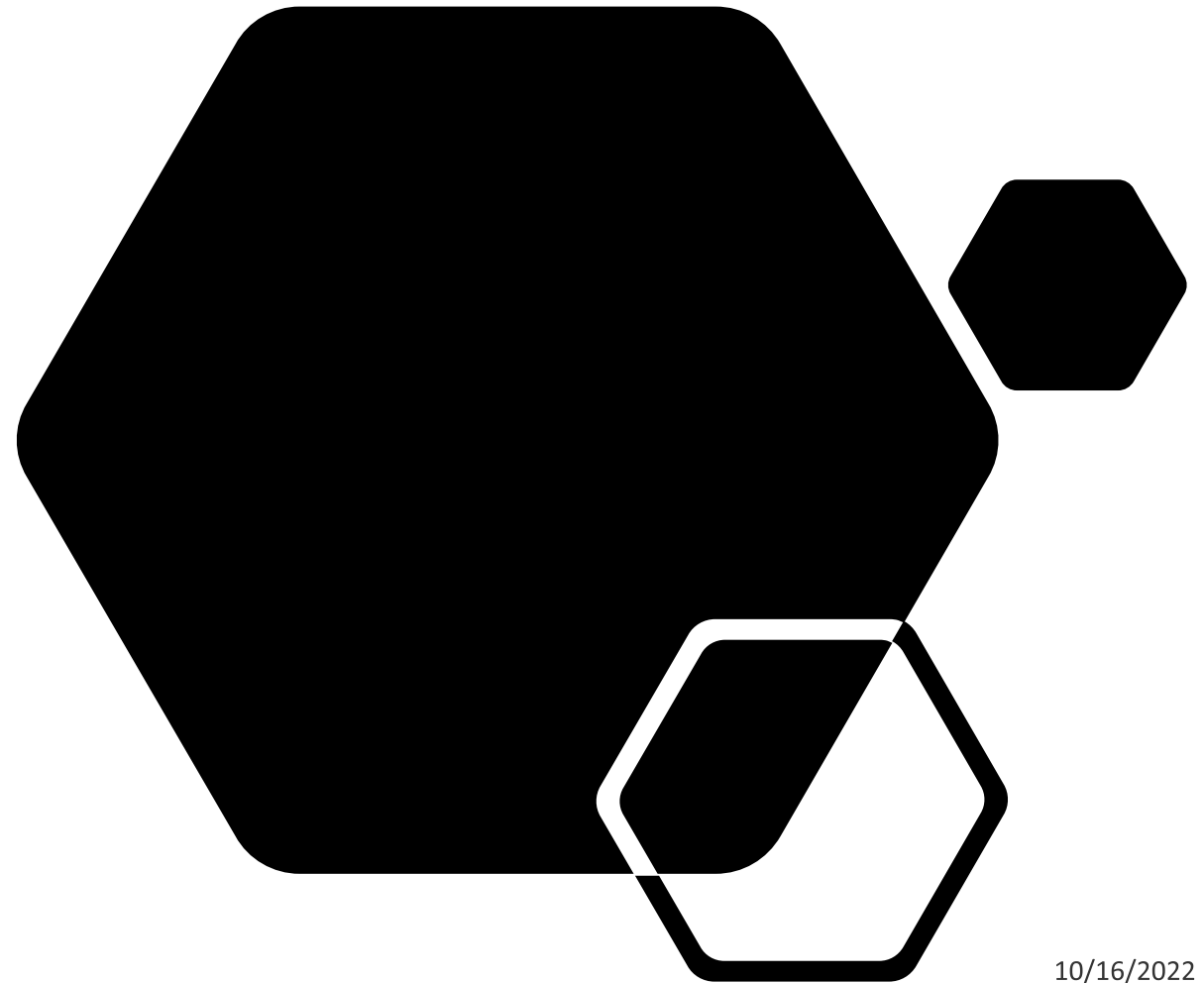
1. Why?
2. Cauzalitate
3. Măsurare
4. Modelare și eșantionare
 - Corelația
 - Erori de modelare
 - Erori de eșantionare
5. Tehnici de analiză
6. Predicție
7. Programare și ML
8. ML și Deep Learning
9. Producția
10. Why Privacy?
11. Privacy Preserving Algorithms
12. Privacy Architectures and Federated Learning

Corelația

Corelația este necesară, dar insuficientă pentru cauzalitate

„Corelația nu înseamnă cauzalitate”

Corelația poate deriva din alte relații de cauzalitate sau din co-incidente repetate



10/16/2022

Covariația variabilelor

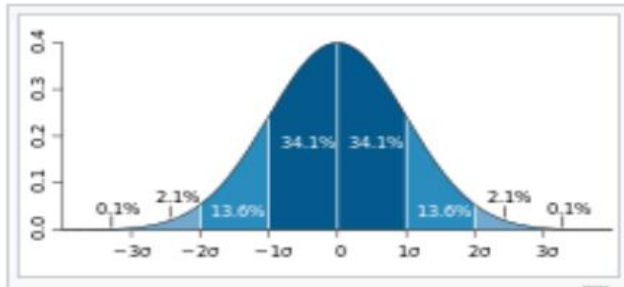
- Dacă estimăm mai bine valorile lui A știind valorile lui B, decât la întâmplare, atunci spunem că:
 - A și B sunt asociate (vezi clarificări [aici](#))
 - A și B covariază (adică variază împreună)
 - A și B sunt corelate (mai ales pentru variabilele numerice)
- Și variabilele [ordinale](#) sau [nominale](#) pot fi asociate
 - Județul și etnia
 - Țara și limba maternă
 - Genul și prenumele

Covarianță și corelație

- **Covarianța** măsoară **sensul** relației dintre 2 variabile
- Dacă variabilele evoluează...
 - În același sens: cov. pozitivă
 - În sensuri diferite: cov. negativă
 - Fără relație: cov. nulă
- Mărimea covarianței depinde de scala variabilelor
 - **Nu** măsoară intensitatea asocierii!
- **Corelația** măsoară **sensul și intensitatea** relației dintre 2 variabile
- Corelația este covarianța variabilelor standardizate
 - Unitatea de măsură naturală este înlocuită de abaterea standard
- Mărimea corelației este standardizată
 - **Corelația măsoară intensitatea asocierii**
 - Rang: [-1, 1]

Relația statistică

Varianță și abaterea standard a unei variabile



$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

- σ = population standard deviation
- N = the size of the population
- x_i = each value from the population
- μ = the population mean

Covarianță vs. corelație pt. 2 variabile

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

$$\text{Correlation} = \frac{\text{Cov}(x,y)}{\sigma_x * \sigma_y}$$

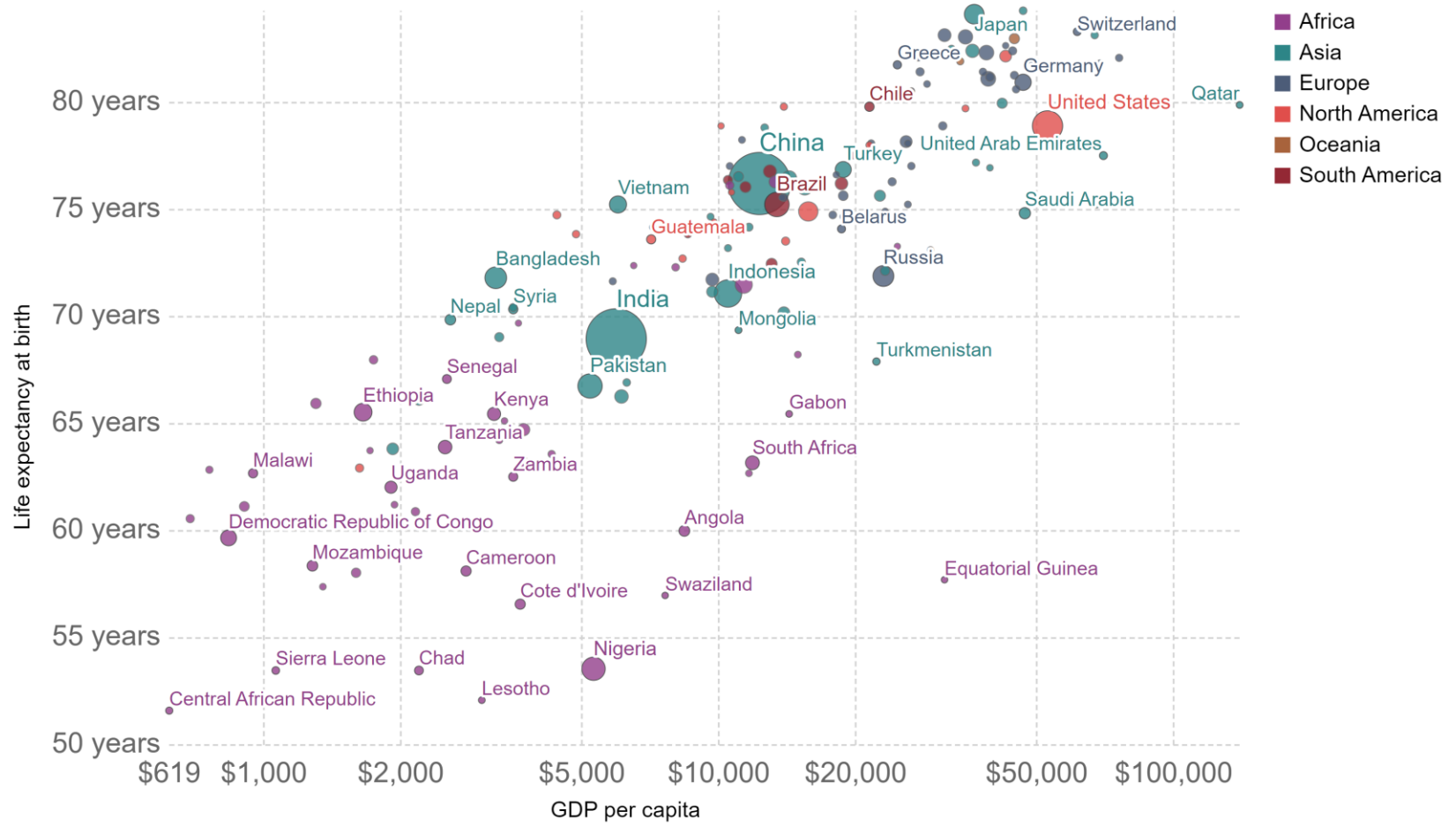
where:

- cov is the covariance
- σ_x is the standard deviation of X
- σ_y is the standard deviation of Y

Corelație pozitivă

Life expectancy vs. GDP per capita, 2016

GDP per capita is measured in 2011 international dollars, which corrects for inflation and cross-country price differences.

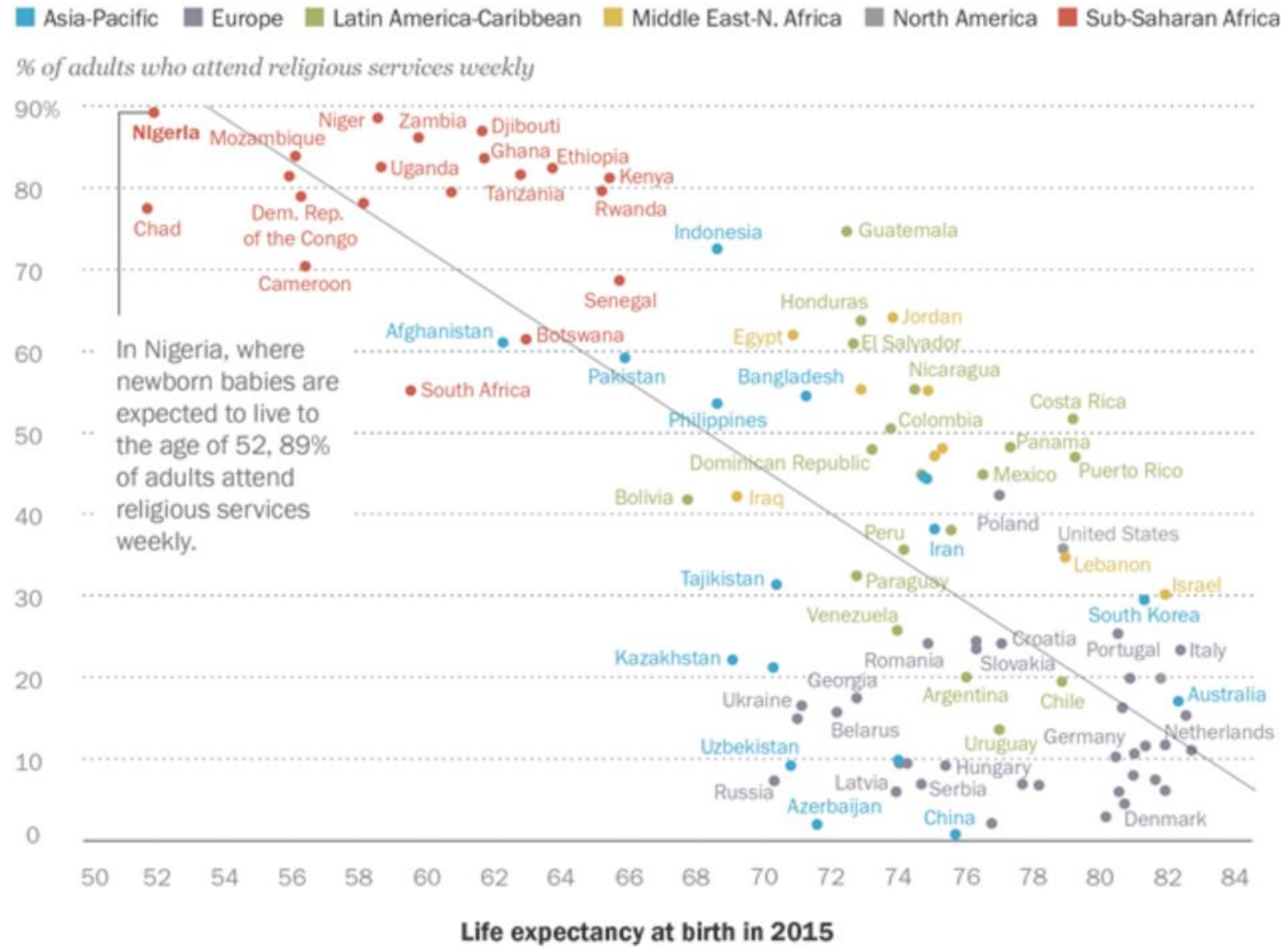


Source: Clio-Infra & UN Population Division ; Maddison Project Database (2018)

OurWorldInData.org/life-expectancy • CC BY

Corelație negativă

Weekly worship attendance is most common where life is shortest



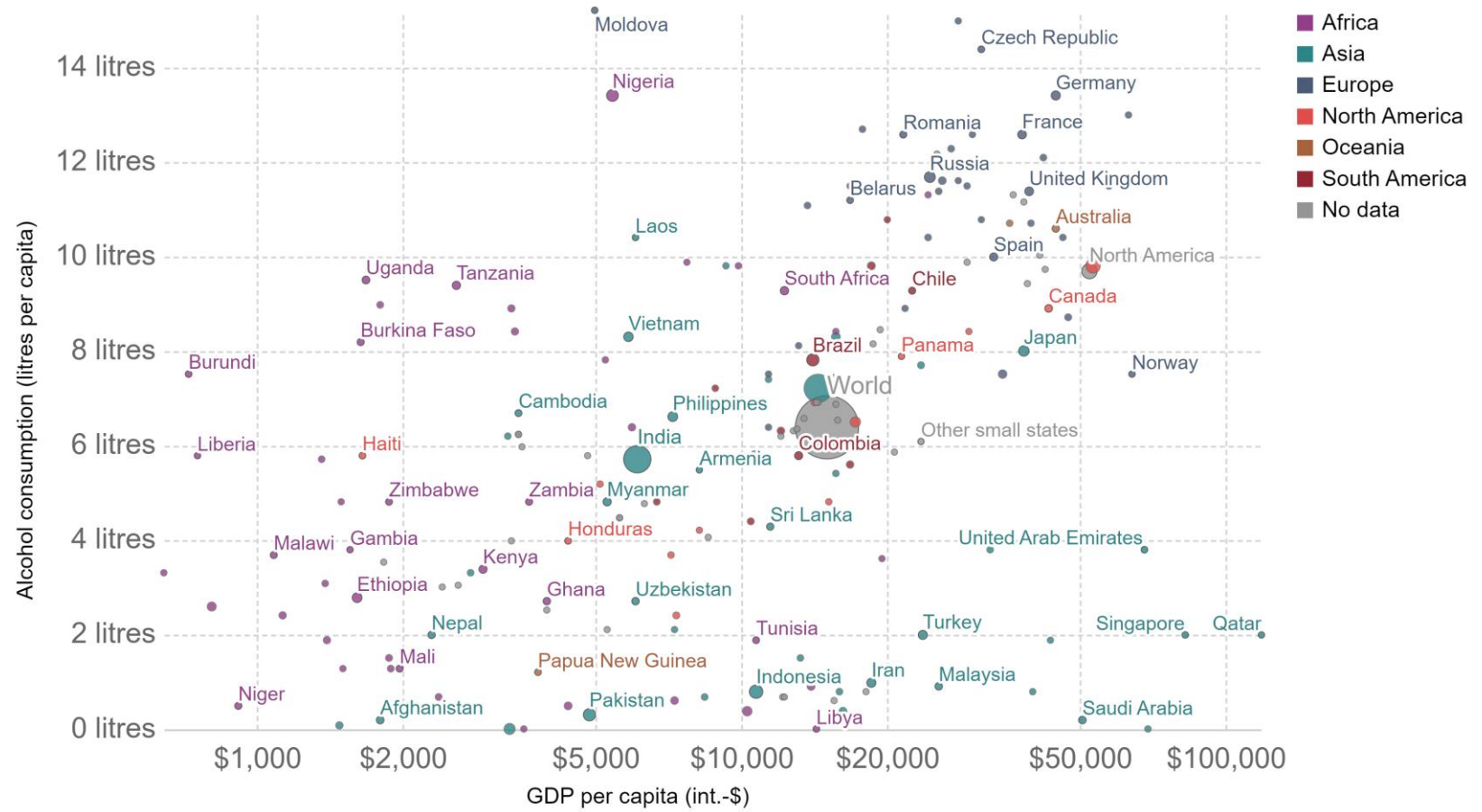
Note: Graphic shows 101 of 102 countries and territories surveyed; United Nations does not report life expectancy at birth for Kosovo.
 Source: Pew Research Center surveys, 2008-2017. Life expectancy data from United Nations World Population Prospects (2017).
 "The Age Gap in Religion Around the World"

Corelație nulă/slabă

Alcohol consumption vs. GDP per capita, 2016

Average annual alcohol consumption measured in liters of pure alcohol per person (15 years of age or older).
Gross domestic product (GDP) per capita measured is adjusted for differences in price levels between countries (measured in international-\$).

Our World
in Data



Source: World Bank, Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data

CC BY

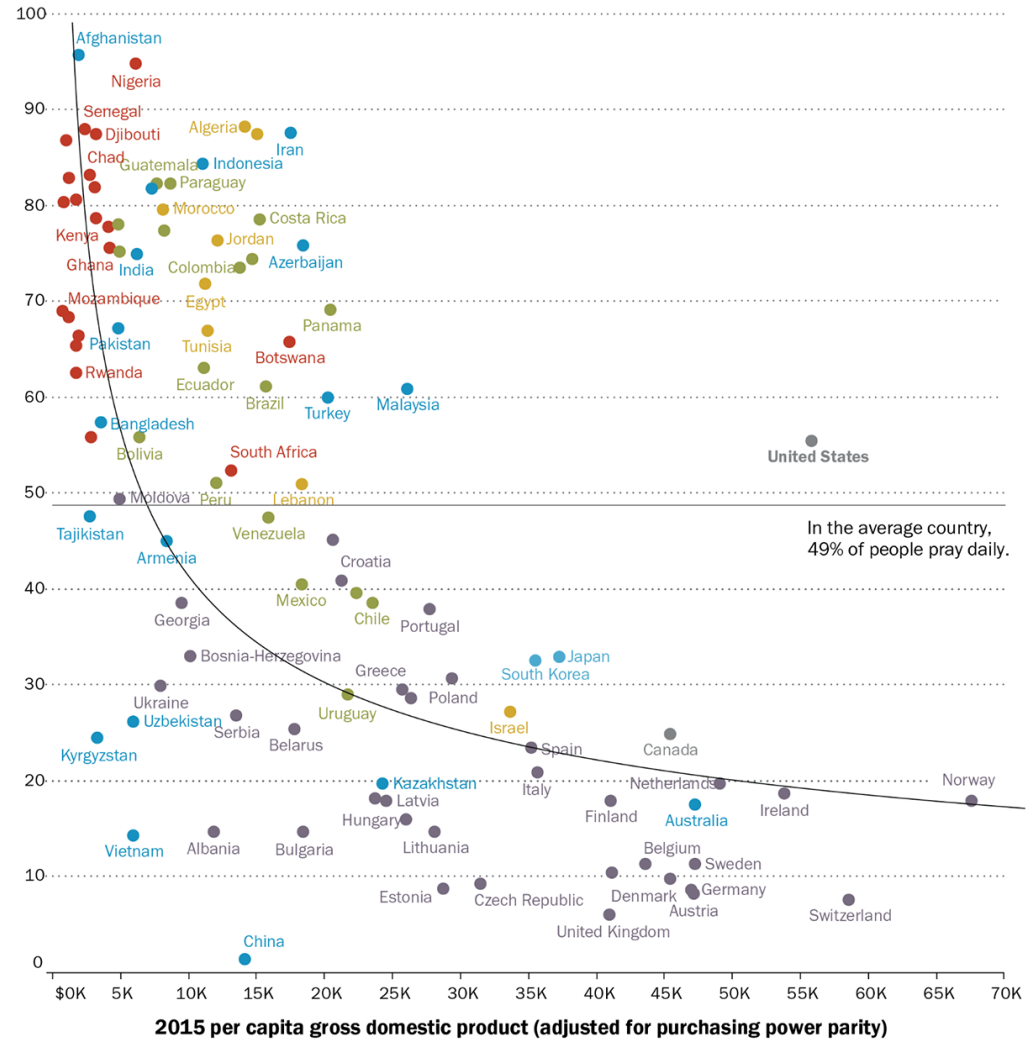
Sursă

Corelație non-lineară

Daily prayer is more common in the U.S. than in many other wealthy countries

% who say they pray daily

■ Asia-Pacific ■ Europe ■ Latin America-Caribbean ■ Middle East-N. Africa ■ North America ■ Sub-Saharan Africa



Note: Graphic shows 102 of 105 countries and territories surveyed; the International Monetary Fund does not report gross domestic product (GDP) figures for Kosovo, Palestinian territories or Puerto Rico. The measure of daily prayer may be less reflective of the extent of religious life in countries where non-Abrahamic religions are dominant (such as China, where Buddhism ranks among the most popular religions) than in countries where Abrahamic religions are dominant.

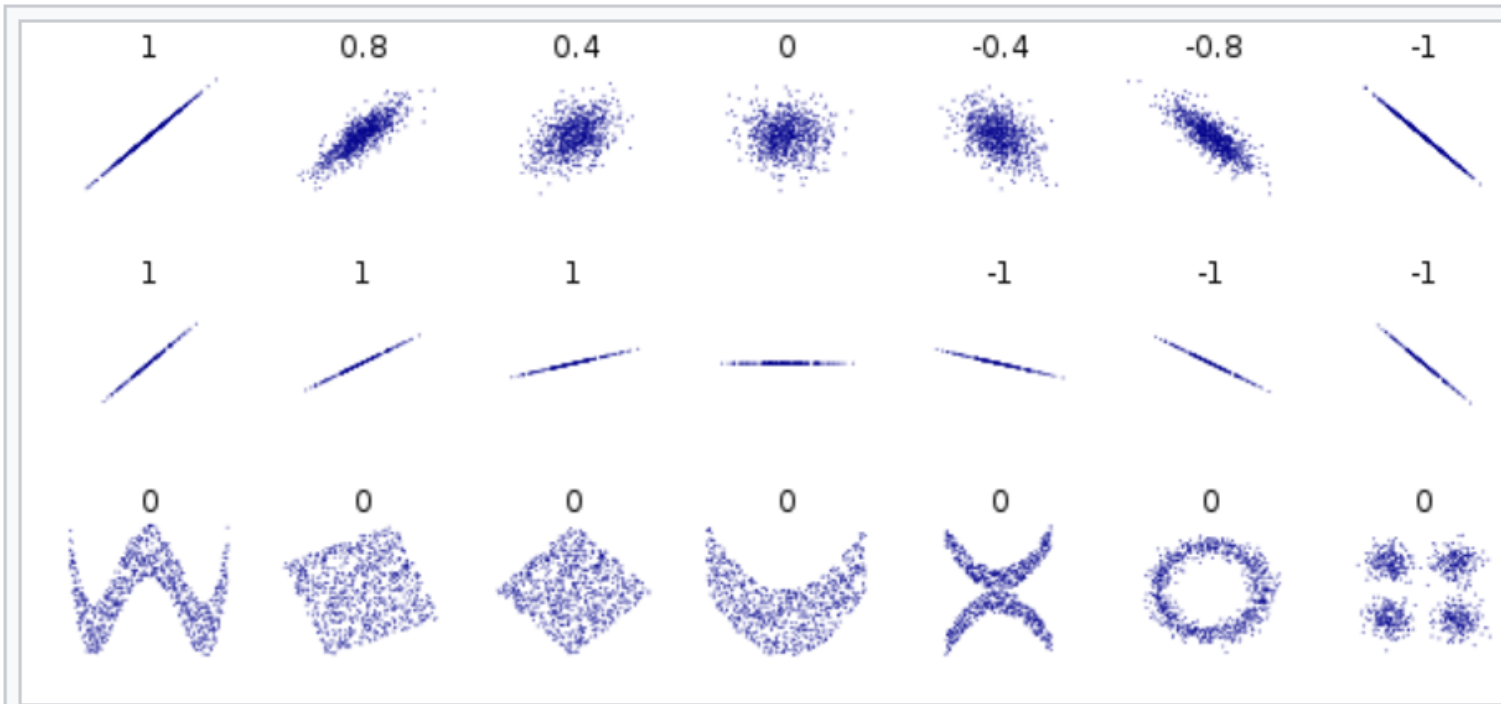
Source: Pew Research Center surveys, 2008-2017. GDP data from the International Monetary Fund World Economic Outlook Database, October 2015. Per capita GDP figures are adjusted for purchasing power parity.

"The Age Gap in Religion Around the World"

PEW RESEARCH CENTER

Coeficientul de corelație R

- Dacă două variabile numerice co-variază liniar putem calcula coeficientul de corelație **Bravais-Pearson: R**
 - R măsoară intensitatea relației liniare dintre două variabile numerice
 - R variază între -1 și 1
 - Dacă $R = 0$ atunci variabilele nu sunt asociate
- Există și alte măsuri ale asocierii
 - Covarianța
 - Alți coeficienți speciali pentru variabile ordinale sau nominale

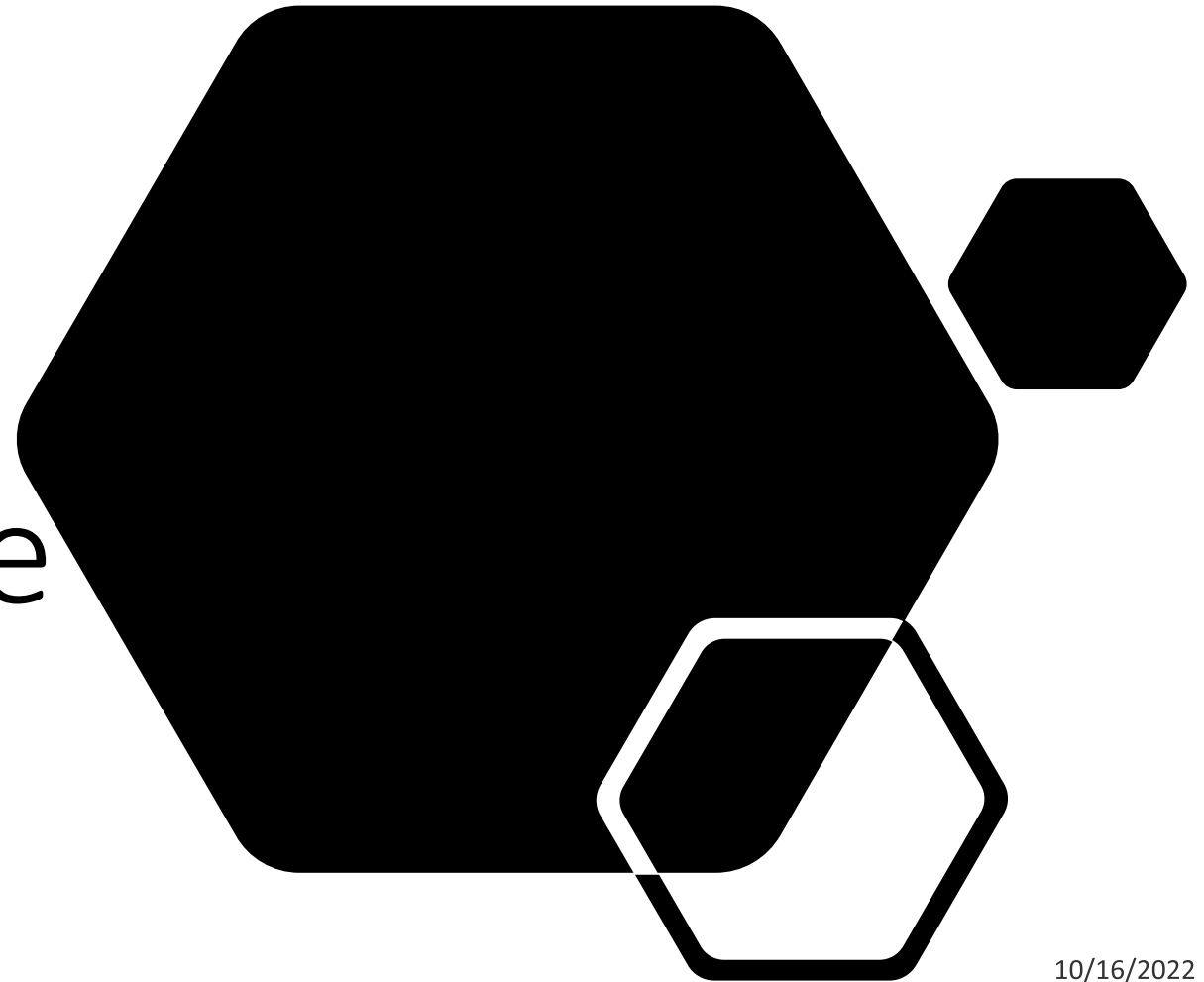


Several sets of (x, y) points, with the Pearson correlation coefficient of x and y for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

Corelații

- Care e coeficientul de corelație dintre...
 - Vârsta și anul nașterii unei persoane?
 - Numărul de căsătorii și numărul de divorțuri ale unei persoane?
 - Lungimea părului și mărimea pantofului?
 - Numărul de animale de companie avute și codul poștal din București?

Erori de modelare

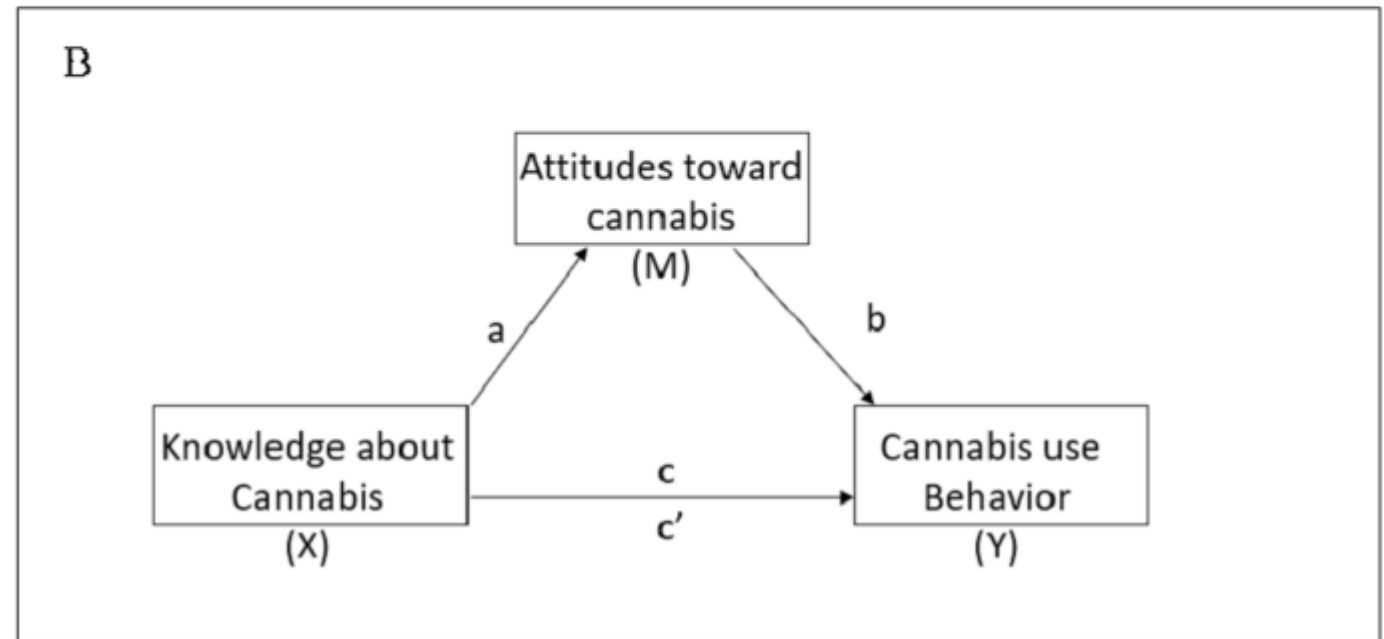
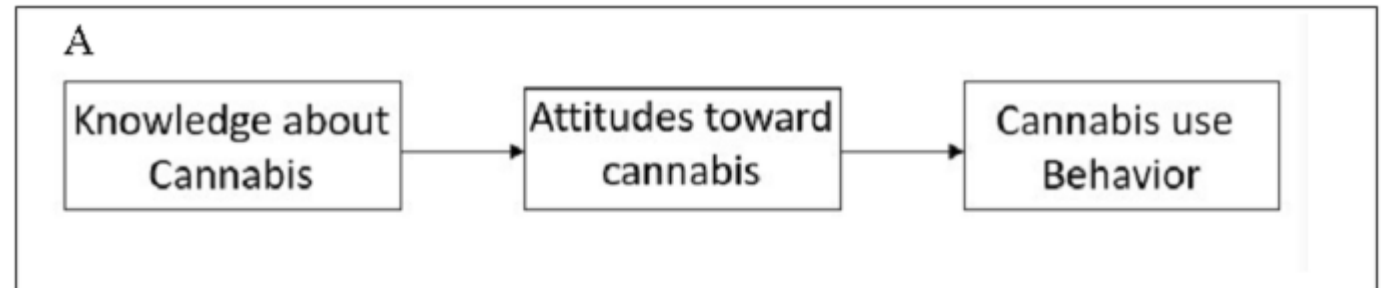


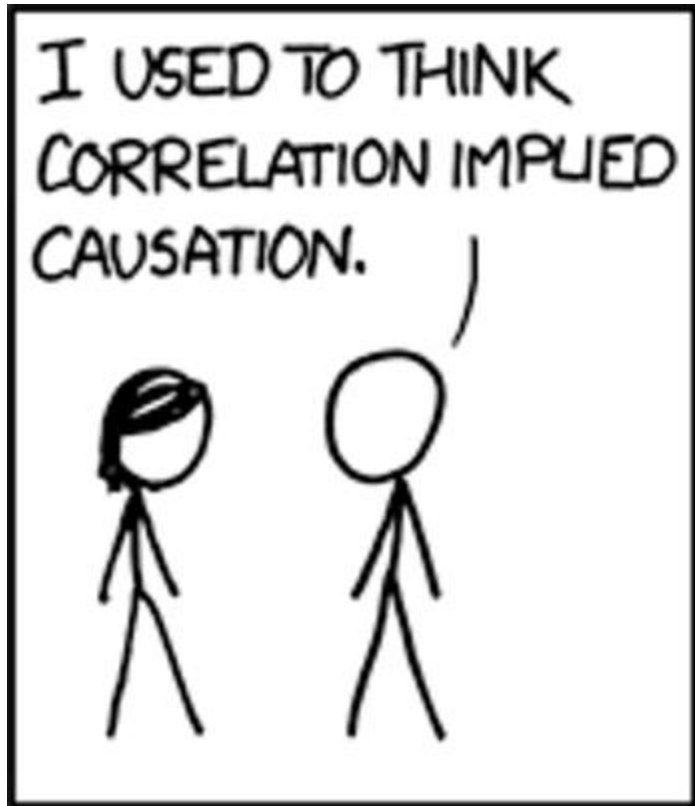
10/16/2022



Modelarea

- Modelul face legătura dintre corelații și cauzalitate
- Aceleași corelații pot fi modelate diferit
 - Și greșit 😊
- Model = diagnoză

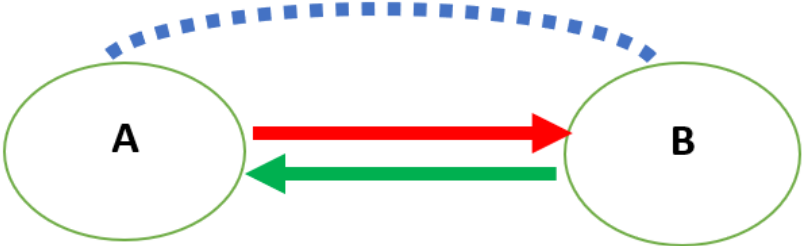




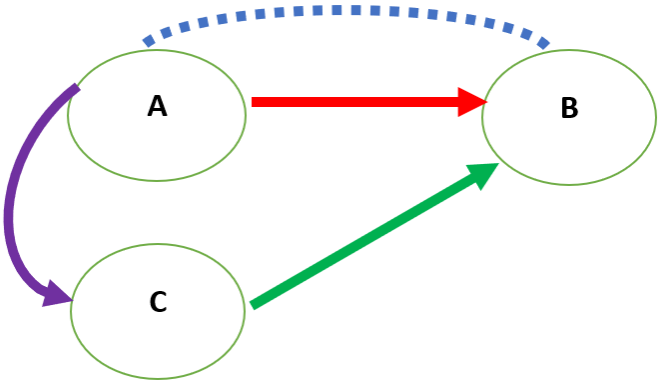
Schemele grafice ale erorilor de modelare



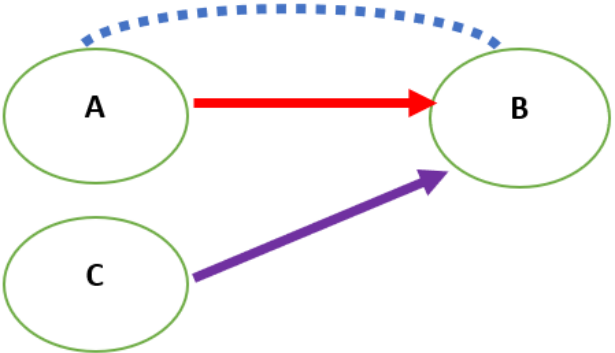
1. Corelații co-incidentale



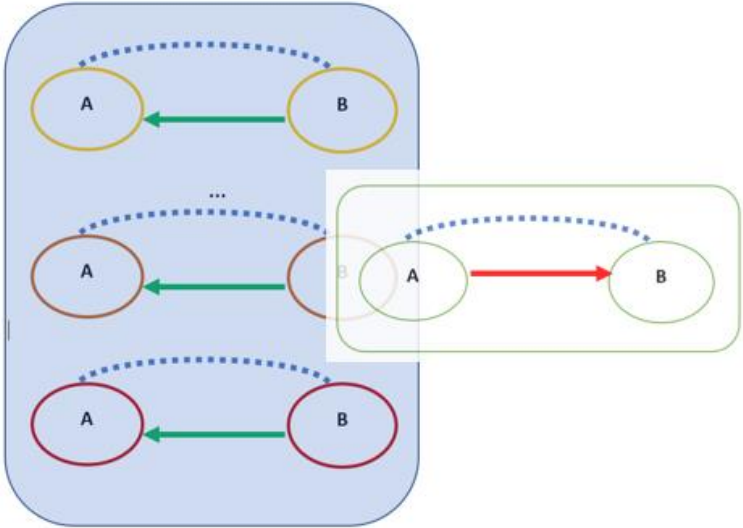
2. Sensul cauzării



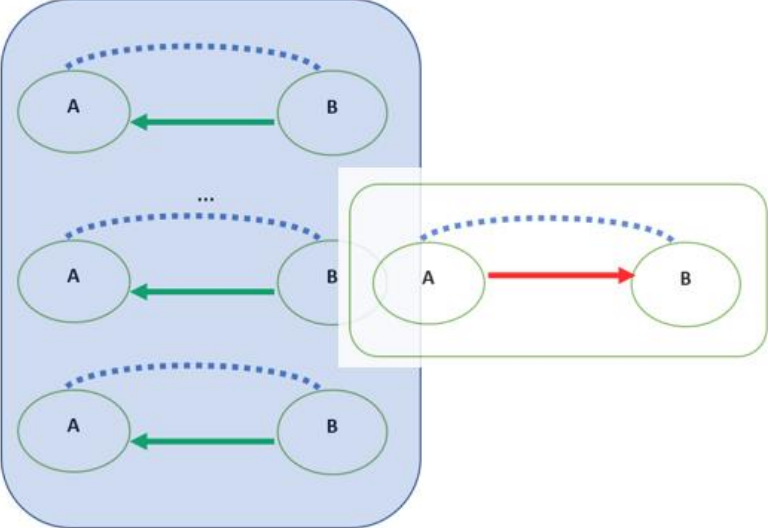
3. Corect din rațiuni false



6. Eroarea sistematică de măsurare



4. Paradoxul lui Simpson



5. Eroarea ecologică

Erorile de modelare

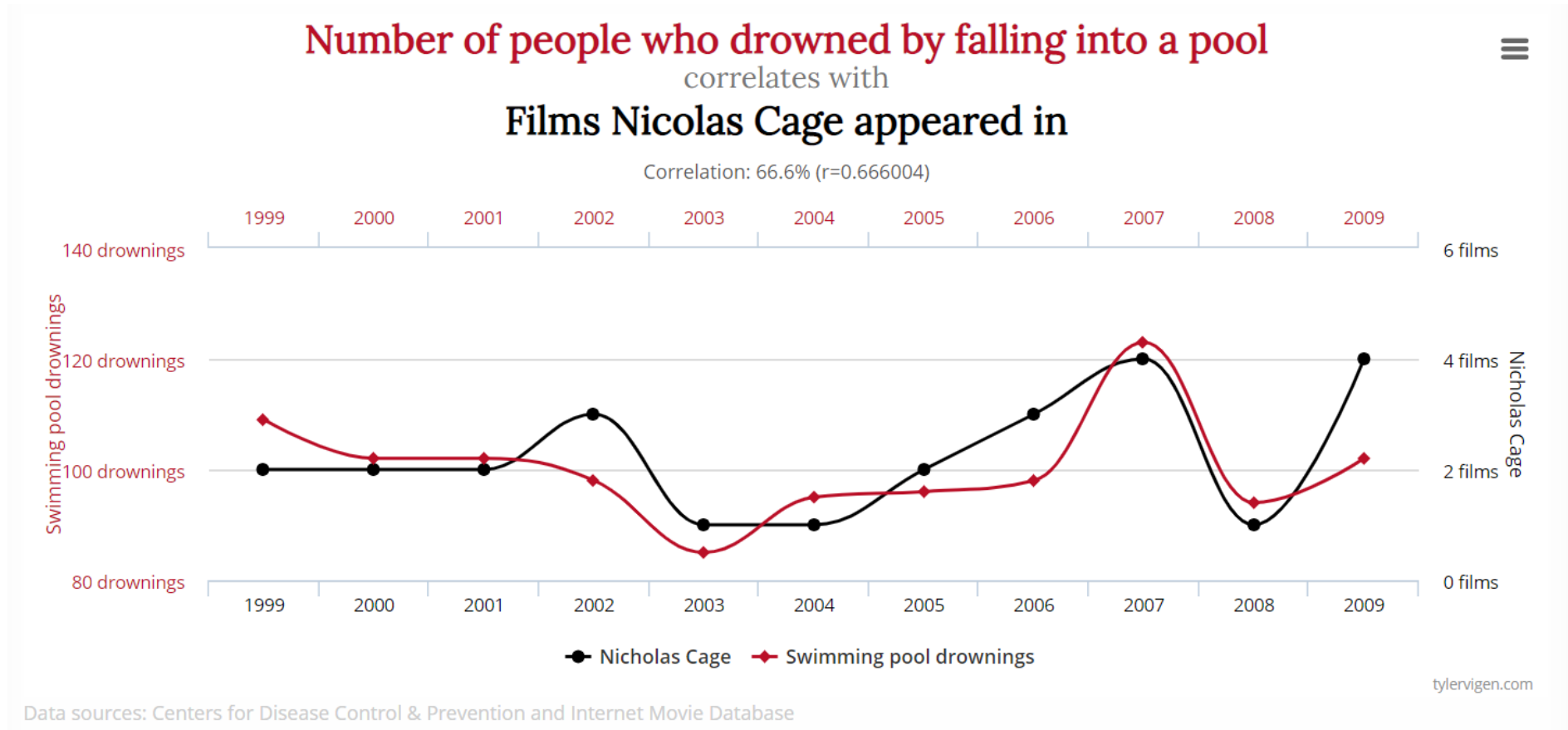
- **Corelații co-incidentale:** nu există nicio cauzalitate, doar co-incidente repetate apărute aleatoriu într-un volum mare de date
- **Sensul cauzării:** presupusul efect este de fapt cauza, sau relația este circulară
- **A fi corect din rațiuni false:** alte variabile corelate din model sunt de fapt cauza
- **Paradoxul lui Simpson:** o corelație per ansamblu pozitivă în populație e agregată din relații cauzale negative în interiorul subpopulațiilor
- **Eroarea ecologică:** prin observarea corelației la nivelul colectivităților de indivizi, sensul relației cauzale la nivel de individ se inversează
- **Eroarea sistematică de măsurare:** lipsește din modelul de măsurare variabila ce cauzează bias-ul

1. Corelații coincidentale

- Corelația nu înseamnă cauzalitate
- Uneori corelația nu înseamnă nimic
- Cu suficiente date putem găsi patternuri aleatorii ce par sistematice



Corelații iluzorii - coincidentale



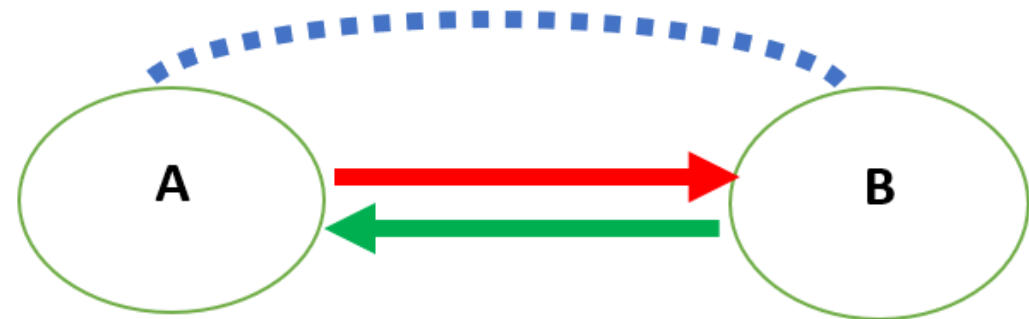
Tyler Vygen, Spurious correlations

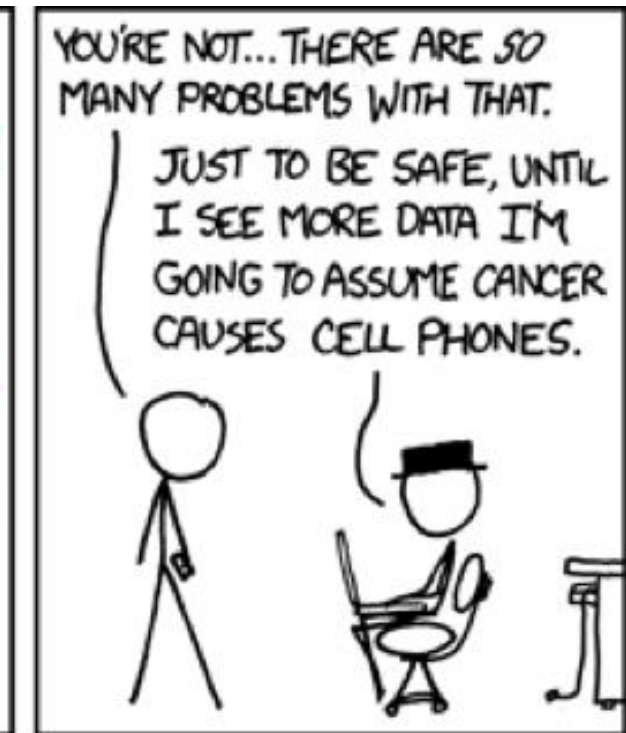
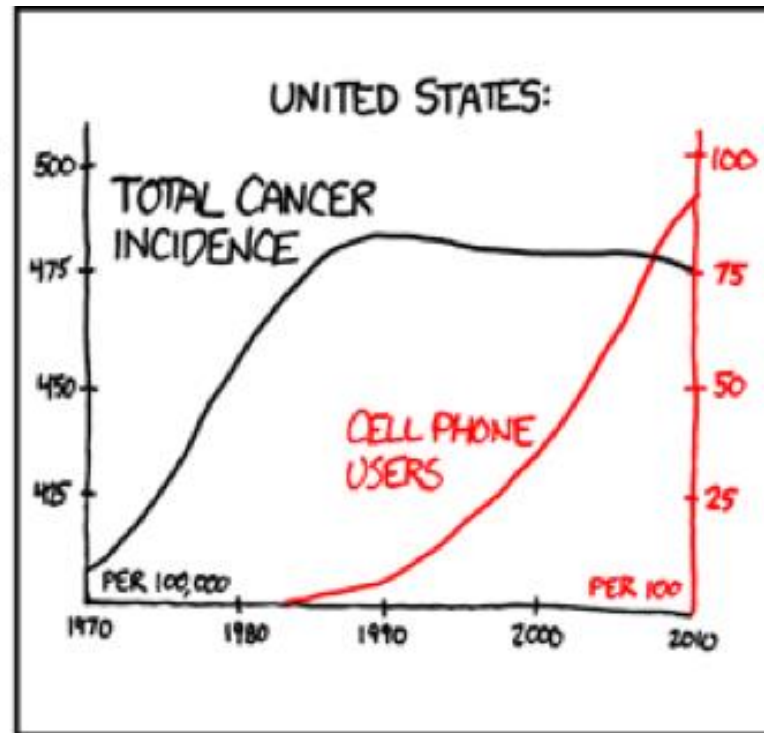
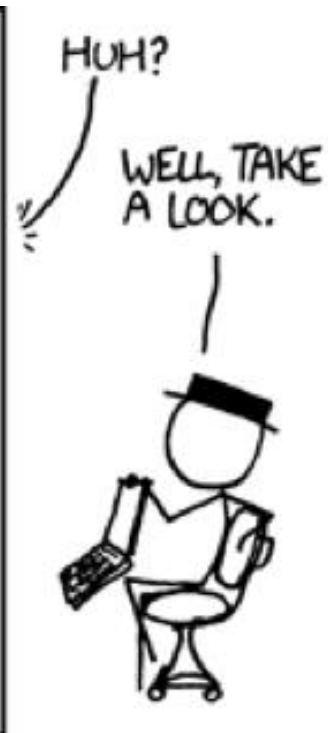


© marketoonist.com

2. Sensul cauzării

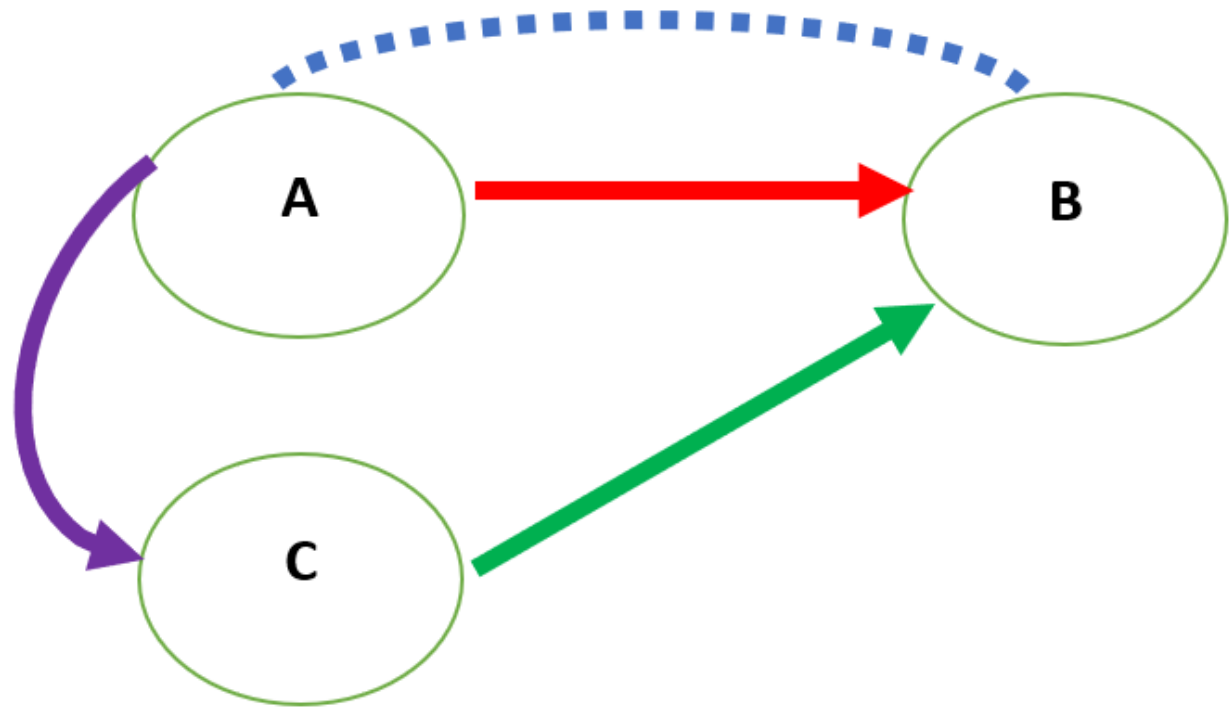
- Care este relația dintre competență și săracie?
- Este sărăcia produsă de incompetența personală?
- Este sărăcia cauza incompetenței personale?





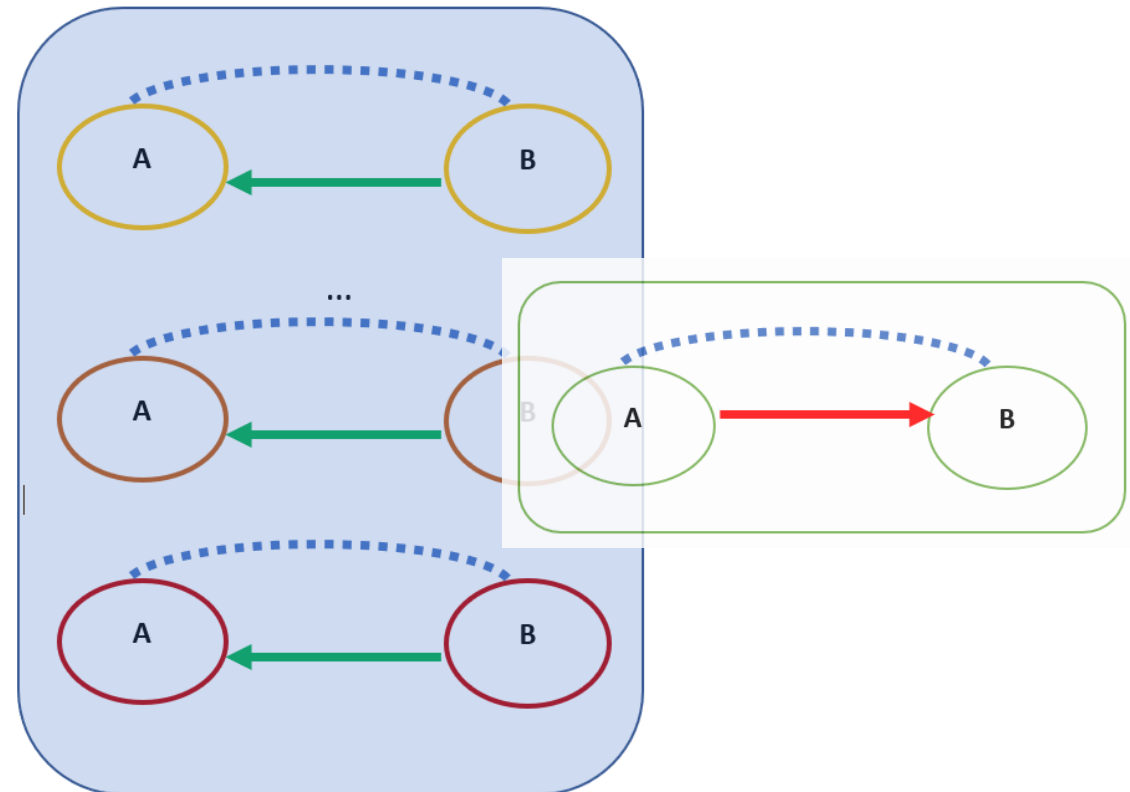
3. Corect din rațiuni false

- Sporește usturoiul purtat la gât imunitatea?
- Sunt femeile mai puțin capabile de muncă intelectuală?
- Relația dintre A și C schimbă semnificația cauzalității dintre A și B



4. Paradoxul lui Simpson sau al variabilei omise

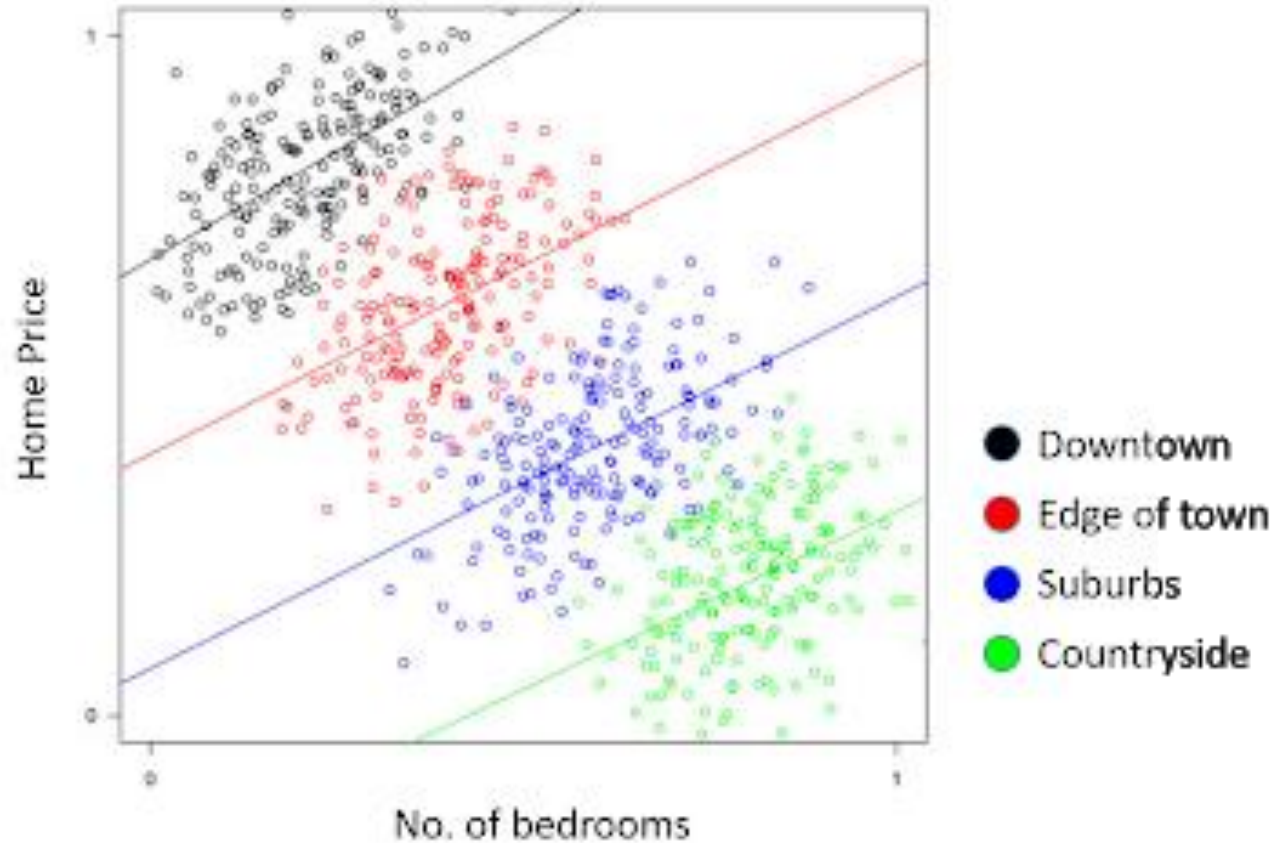
- Populația nu este omogenă
- Subpopulații definite de o terță variabilă, omisă din analiză
- Cauzalitatea din subpopulații este inversată sau ascunsă prin agregare



Paradoxul lui Simpson

Modificarea unei relații
atunci când ținem cont
de o terță variabilă
(introducerea unei
variabile de control)

Eg: subpopulații definite
de poziția în oraș (centru
/ periferie)

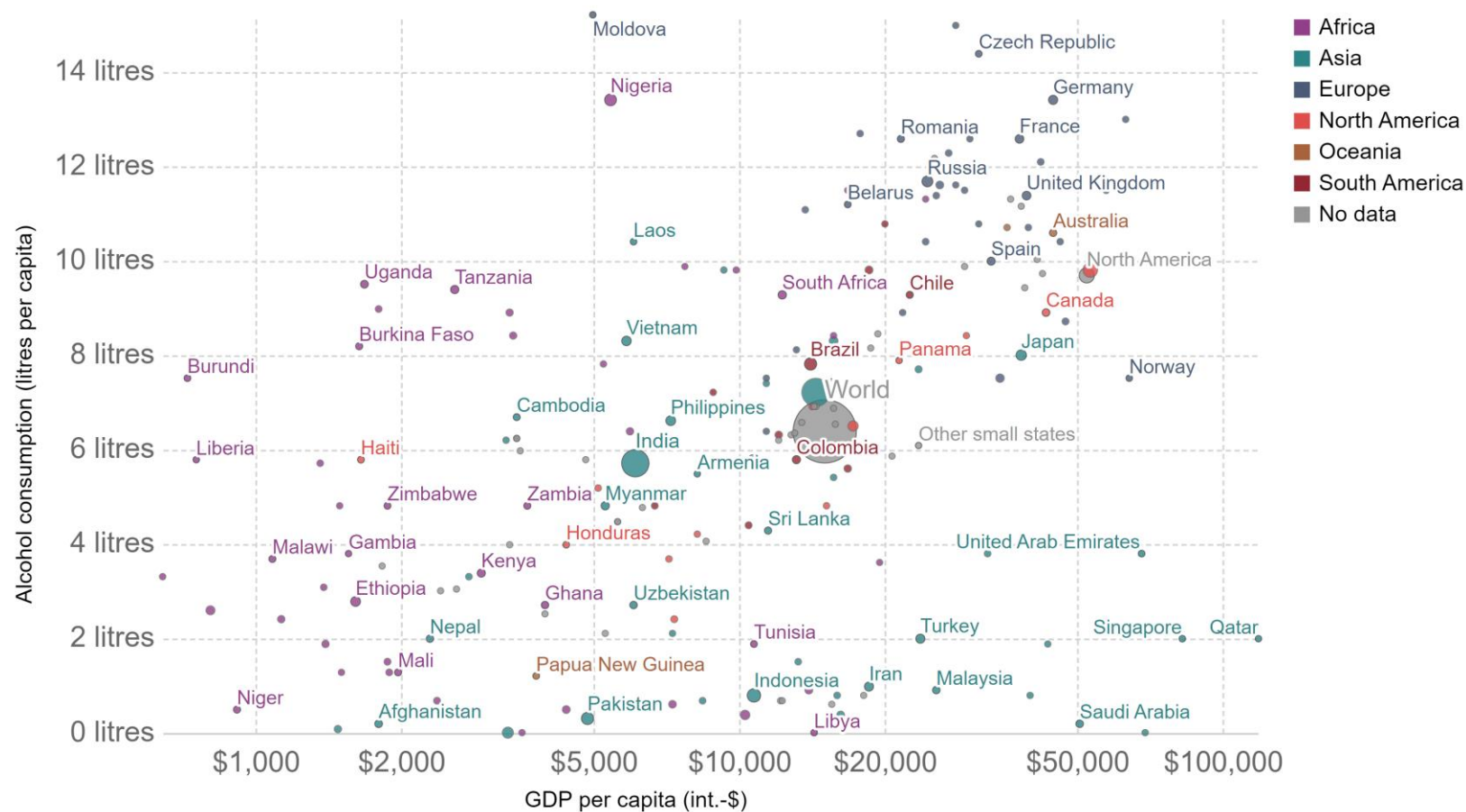


Corelație slabă vs. pozitivă + nulă: controlul variabilei „religie”

Alcohol consumption vs. GDP per capita, 2016

Average annual alcohol consumption measured in liters of pure alcohol per person (15 years of age or older). Gross domestic product (GDP) per capita measured is adjusted for differences in price levels between countries (measured in international-\$).

Our World in Data



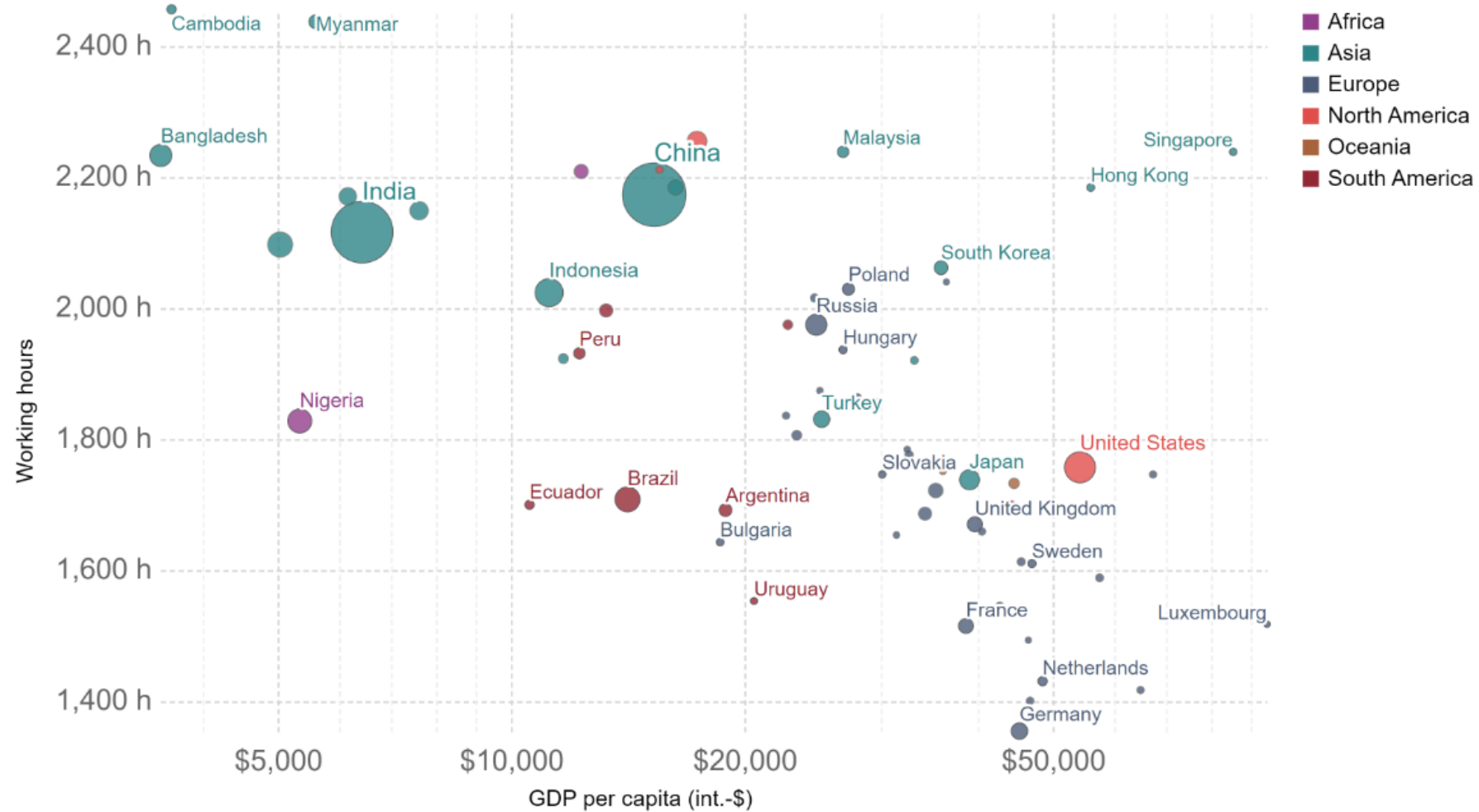
Source: World Bank, Population (Gapminder, HYDE(2016) & UN (2019)), Our World In Data

CC BY

Sursă

Average annual working hours vs. GDP per capita, 2017

Working hours are the annual average per employed person. GDP per capita is measured in constant international-\$. This means it is adjusted for price differences between countries and adjusted for inflation to allow comparisons between countries and over time.

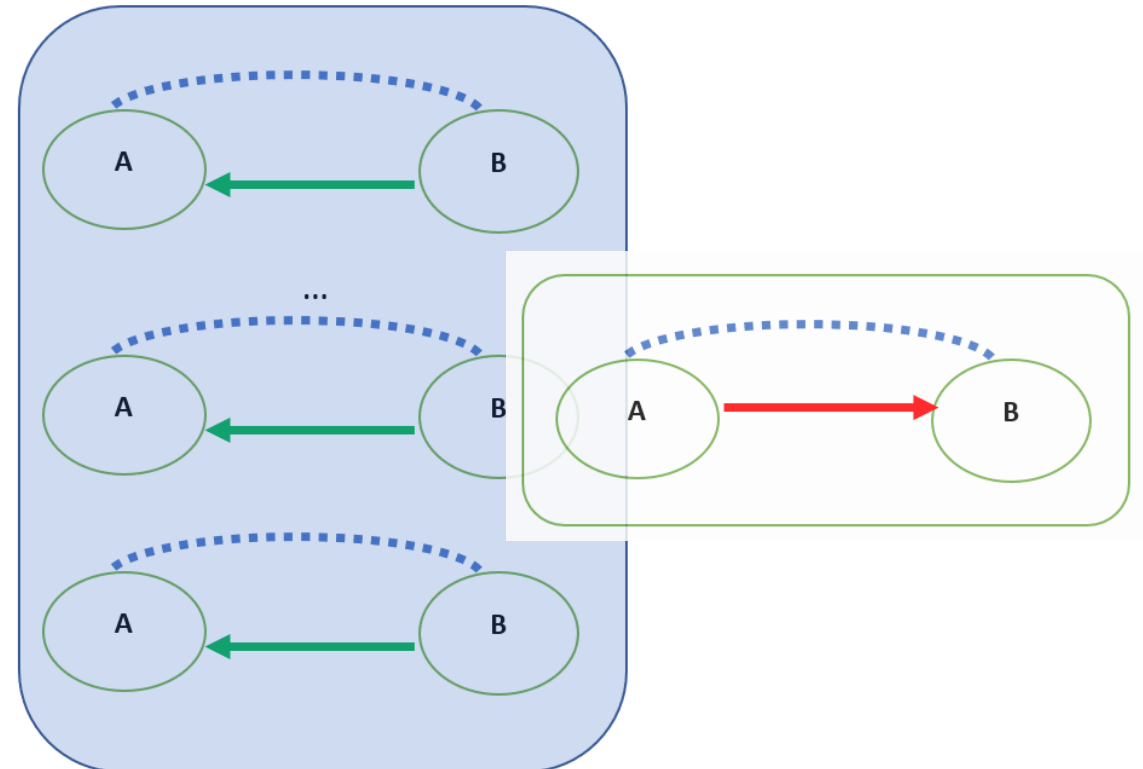


Source: Feenstra et al. (2015) Penn World Tables 9.1, World Bank

OurWorldInData.org/working-hours/ • CC BY

5. Eroarea ecologică (de agregare)

- O populație compusă din colectivități de indivizi (omogene)
- Relația causală se petrece la nivelul indivizilor
- Sensul relației cauzale la nivel individual este inversat prin agregare



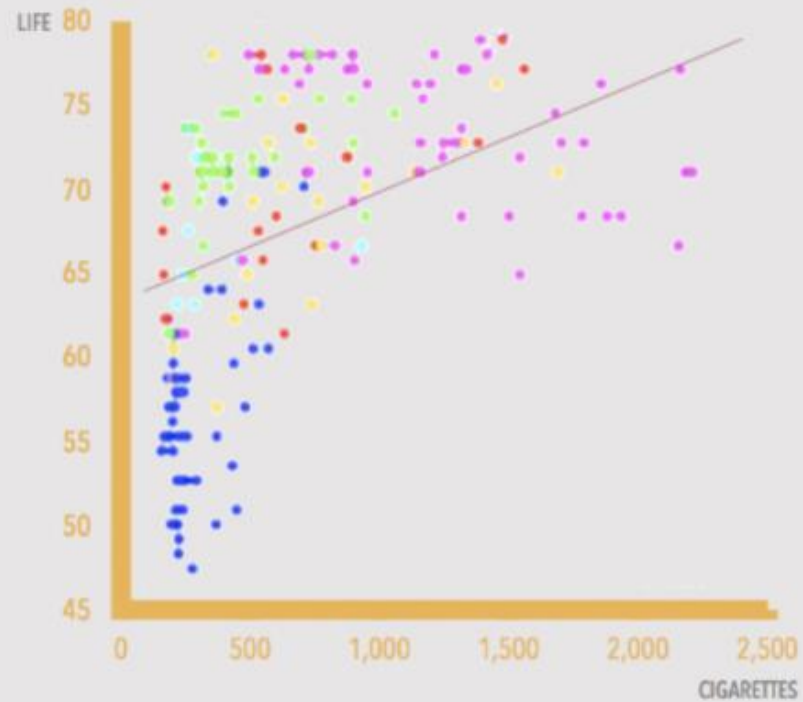
Eroarea ecologică

- Modificarea unei relații atunci când trecem de la unități individuale la unități agregate – sau invers
- Dacă în țările în care se muncește mult sărăcia e mai mare decât în celelalte, rezultă oare că în fiecare țară oamenii care muncesc mult sunt mai săraci decât ceilalți?
 - Relație la nivel agregat: unitatea = țara
 - Relație la nivel individual, în interiorul țărilor: unitatea = individul



Eroarea ecologică / de agregare

Is smoking cigarettes good for your health?



Statistical Summary

Coefficients	Estimate	Std. Error	DF	t-value	p-value
(Intercept)	65.075840	0.855974	183	76.025515	<.00001
Cigarettes	0.006915	0.000855	183	8.090493	<.00001

This relationship is definitely **statistically significant**.

One extra cigarette per year adds 0.006915 years to your life.

So **4 cigarettes a day will add 10 years to your life.**

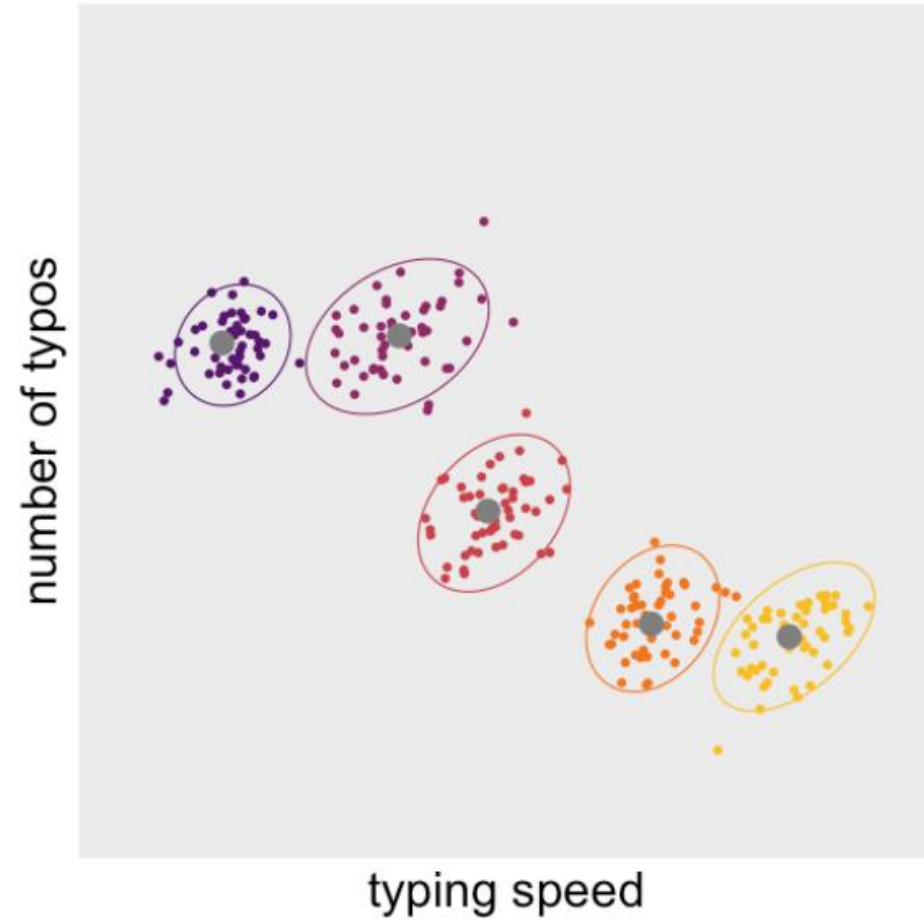
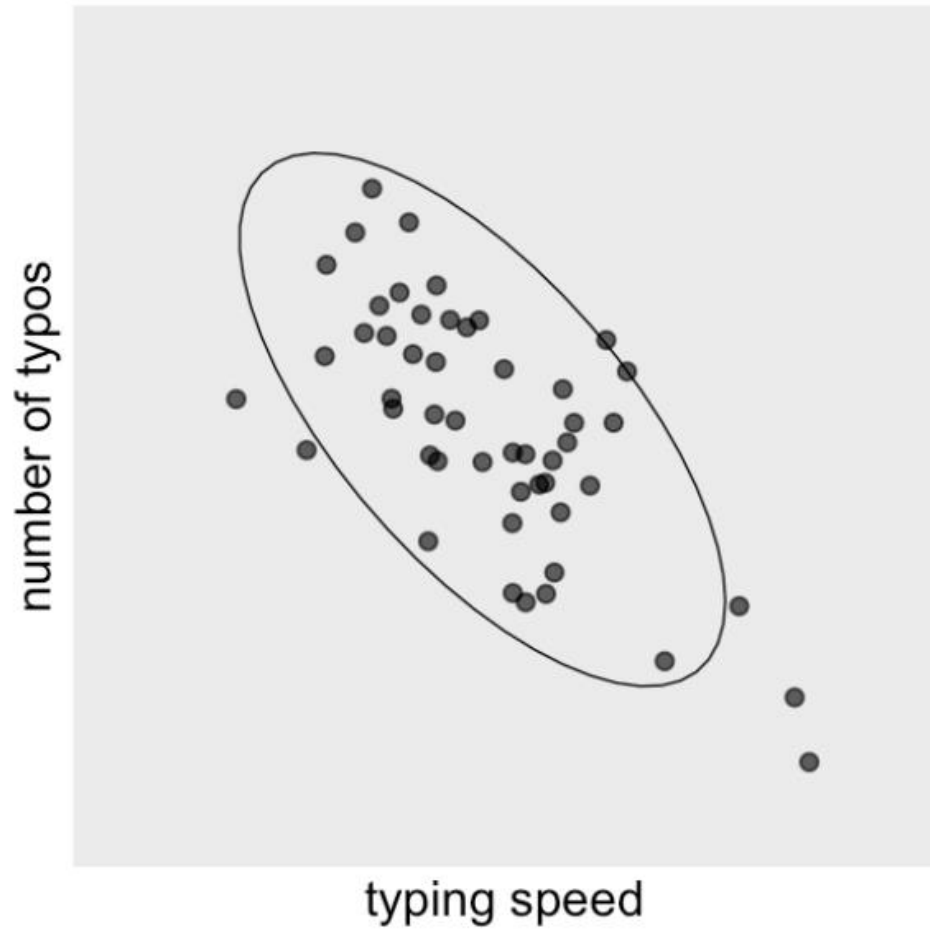
Eroarea ecologică / de agregare

Is smoking cigarettes good for your health?

If we look at individual level data, we can see that the individual life expectancy decreases as smoking increases.



Eroarea ecologică / de agregare

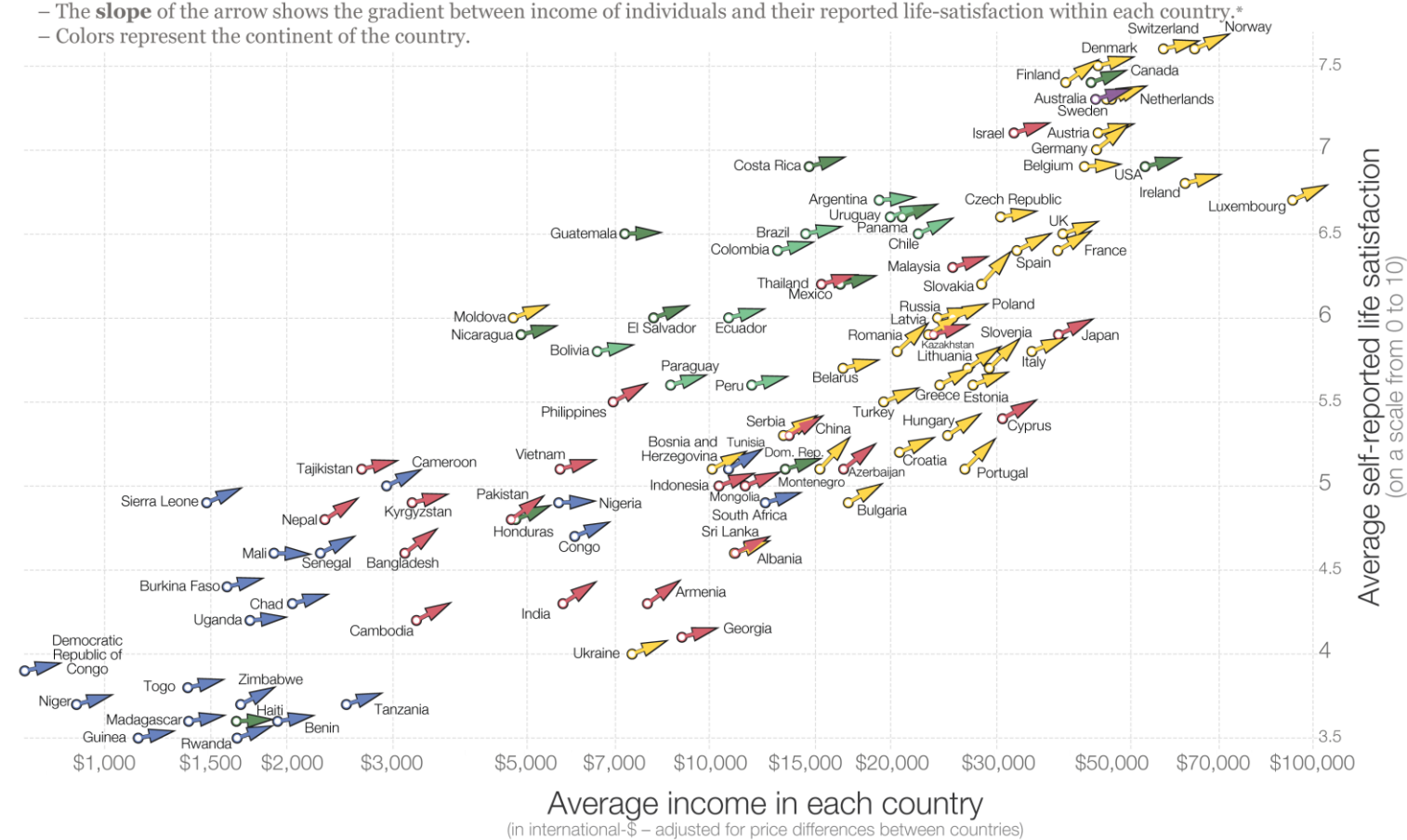


Grafic care arată
simultan relația
la nivel agregat
și la nivel
individual

People in richer countries tend to be happier and within all countries richer people tend to be happier



- The **position** of the arrow shows the average life satisfaction reported by the population of a country (vertical axis) and the average income of that country (horizontal axis).
- The **slope** of the arrow shows the gradient between income of individuals and their reported life-satisfaction within each country.*
- Colors represent the continent of the country.

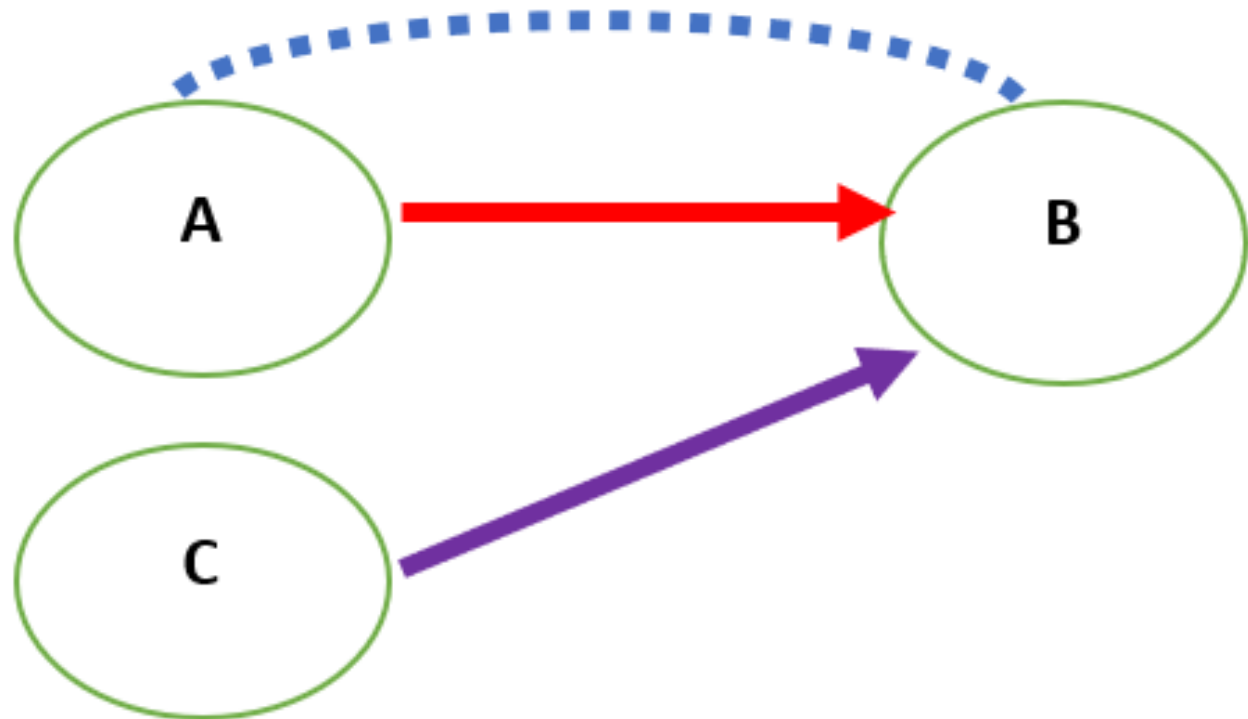


Sursă

* The gradients correspond, country by country, to the regression coefficients between income quintiles and the related average life satisfaction reported by people within each income quintile.
Data sources: World Bank for data on incomes by quintile (based on income shares by quintile and GDP per capita as the mean income); Gallup World Poll for life satisfaction by income quintile. The visualization is available at OurWorldinData.org. There you find the research and more visualizations on life satisfaction. Licensed under CC-BY-SA by the author Max Roser.

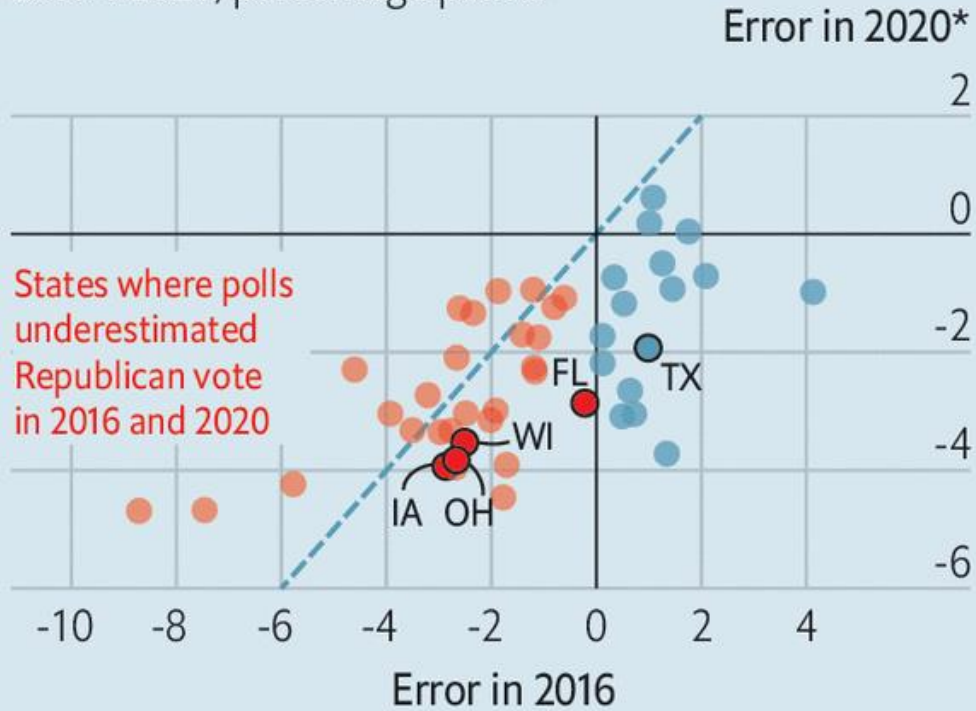
6. Eroarea sistematică de măsurare

- Eroarea sistematică de măsurare a lui A prin B e o eroare de modelare
- A = construct, B=indicator
- Lipsește din model variabila C, care explică bias-ul



Double fault

United States, presidential elections, polls' error in predicting two-party Democratic vote share 2016 v 2020, percentage points



Sources: *New York Times*; *The Economist*

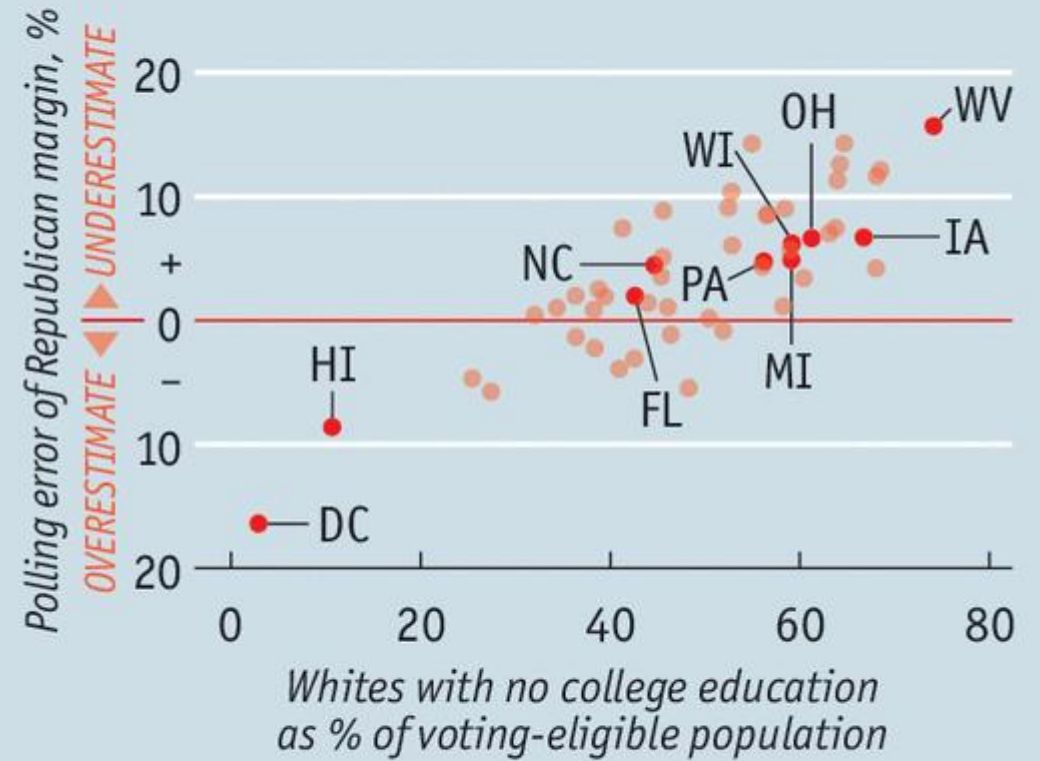
*States with over 98% of votes counted

The Economist

[Sursa](#): The Economist

Where the votes were

US state polling errors and demography, 2016



Sources: *Atlas of US Presidential Elections*; *Census Bureau*; *FiveThirtyEight*; *The Economist*

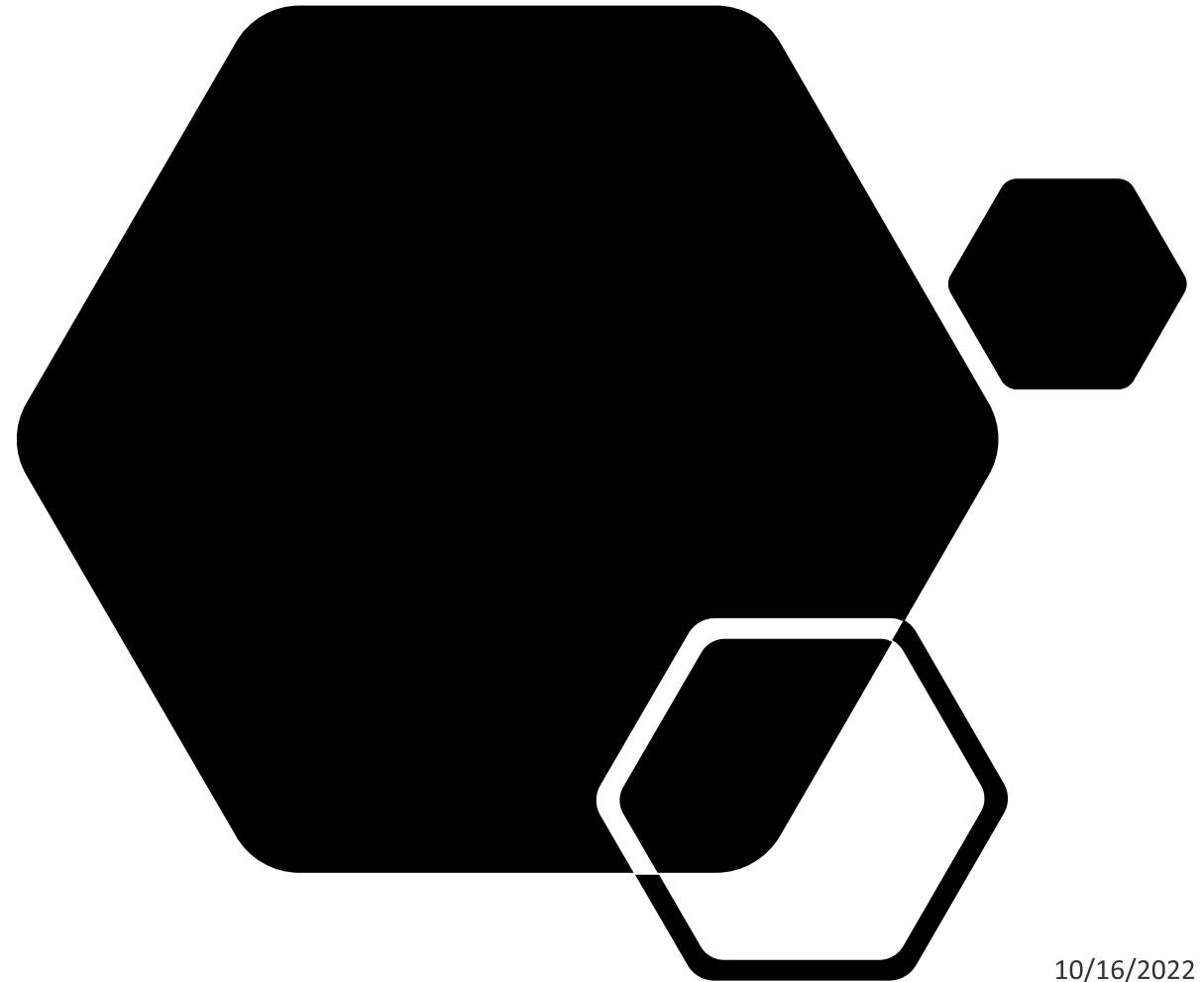
Economist.com

[Sursa](#): The Economist

Erori de eșantionare

Eșantionare și generalizare

Probabilitatea de eroare (p)



10/16/2022

Sondaje de opinie și erori de eșantionare

În cazul sondajelor apar și erori de eșantionare

Eg sondaje **CATI**

- Computer-assisted telephone interviewing

Cei care răspund sunt...

- mai agreabili (OCEAN)
- mai încrezători
- mai liberali ([Vox](#))

One pollster's explanation for why the polls got it wrong

The kind of people who answer polls are really weird, and it's ruining polling.

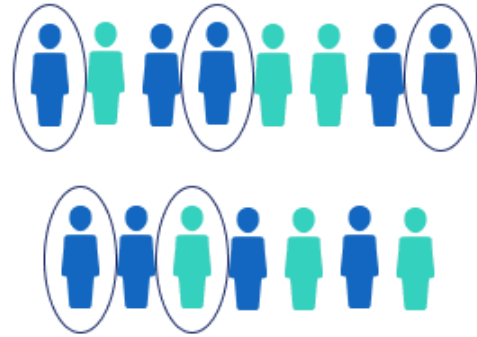
By Dylan Matthews | dylan@vox.com | Nov 10, 2020, 9:20am EST



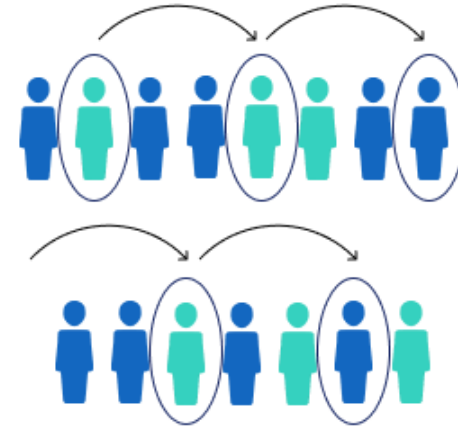
Eșantion și populație

- Nu putem studia toți membrii populației,
 - Studiem un eșantion
 - Generalizare (inducție) = tragem concluzii despre întreaga populație
- **Eșantionul probabilistic** permite **generalizarea**
 - Estimările obținute din eșantion sunt valabile, cu o marjă de eroare, pentru populație
- **Eșantionul non-probabilistic** nu permite generalizarea statistică
 - Eșantion de disponibilitate: comod, dar ne-reprezentativ
 - Eșantion teoretic: permite explorarea sistematică a variației
 - Eșantioanele non-probabilistice permit crearea de tipologii orientative, dar nu de generalizări sau tendințe la nivelul populației

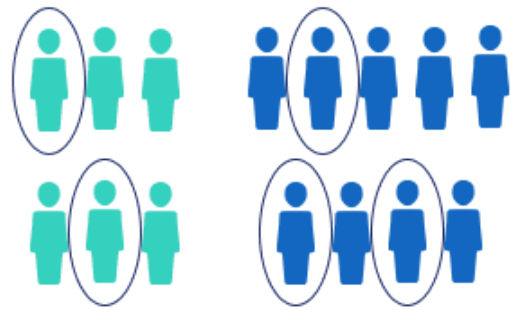
Simple random sample



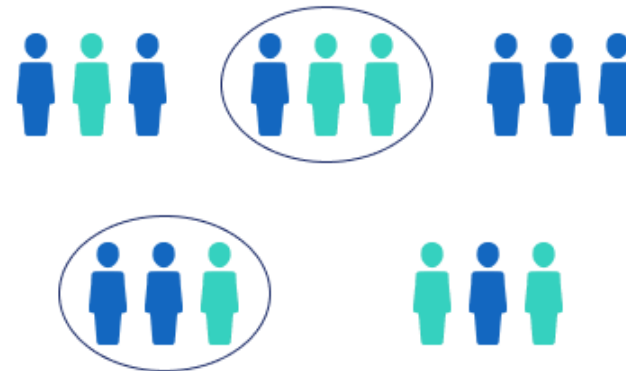
Systematic sample



Stratified sample

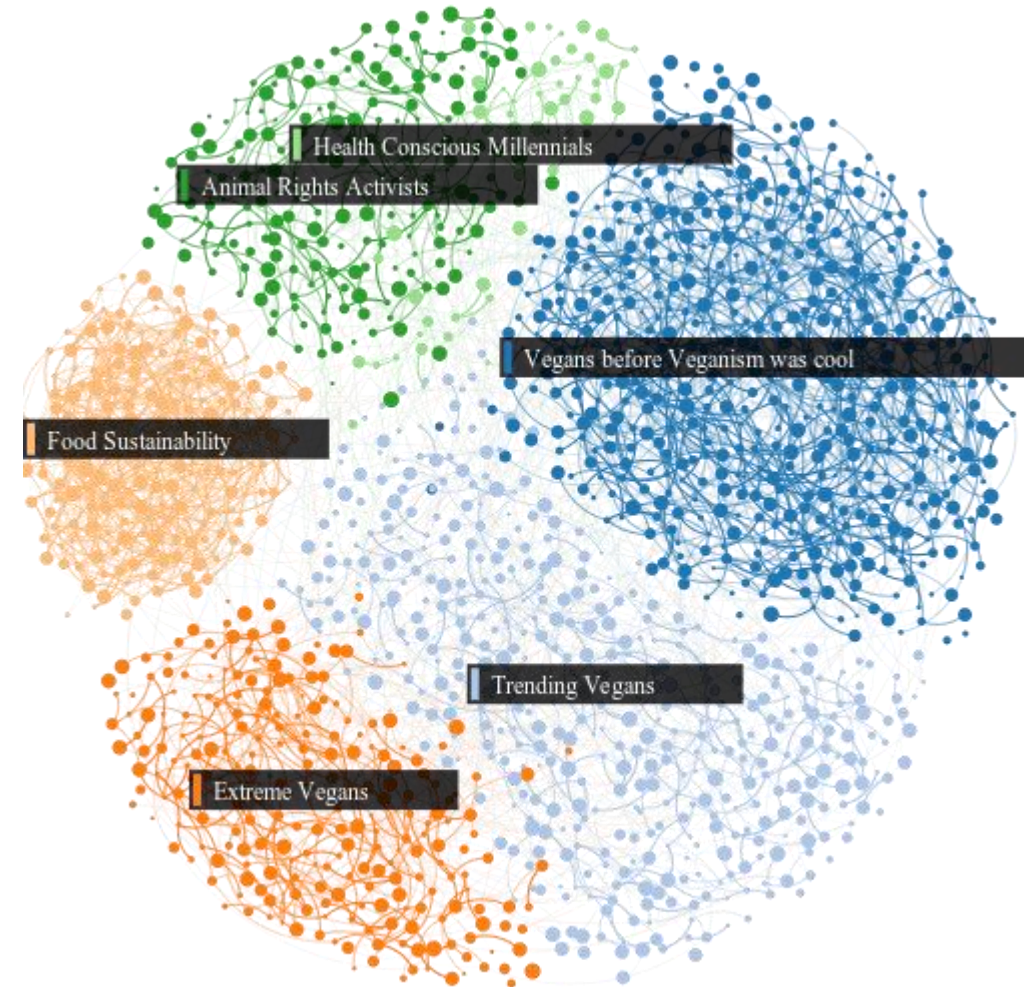


Cluster sample



Exemplu

- Sunt utilizatorii Twitter reprezentativi pentru populație?
 - „The Affinio algorithm ingests massive amounts of Twitter audience data – or social signals – and uncovers naturally-occurring actionable insights derived from a significant, unbiased, representative audience.” ([Affinio](#))



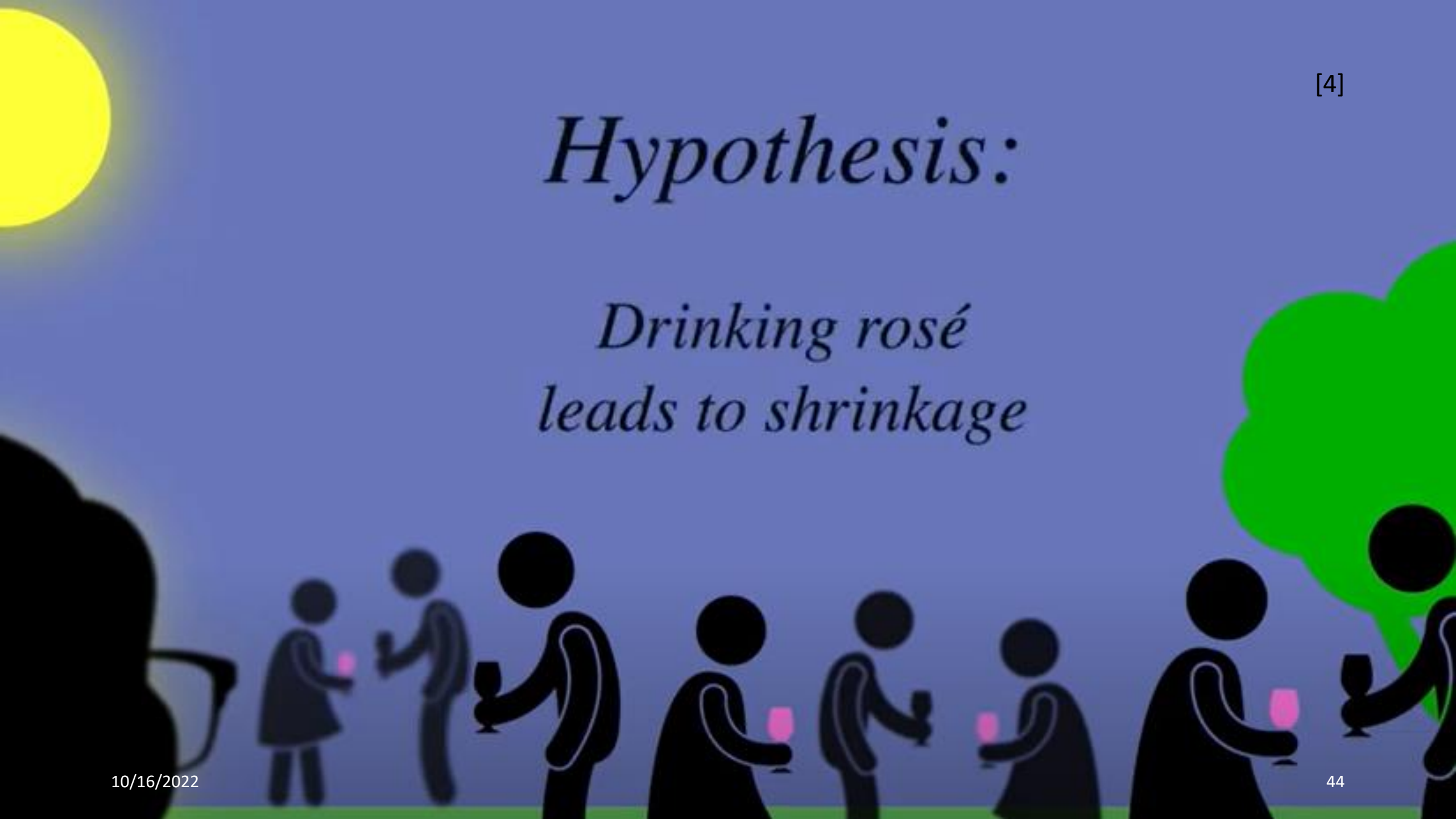
([Davies 2019](#))

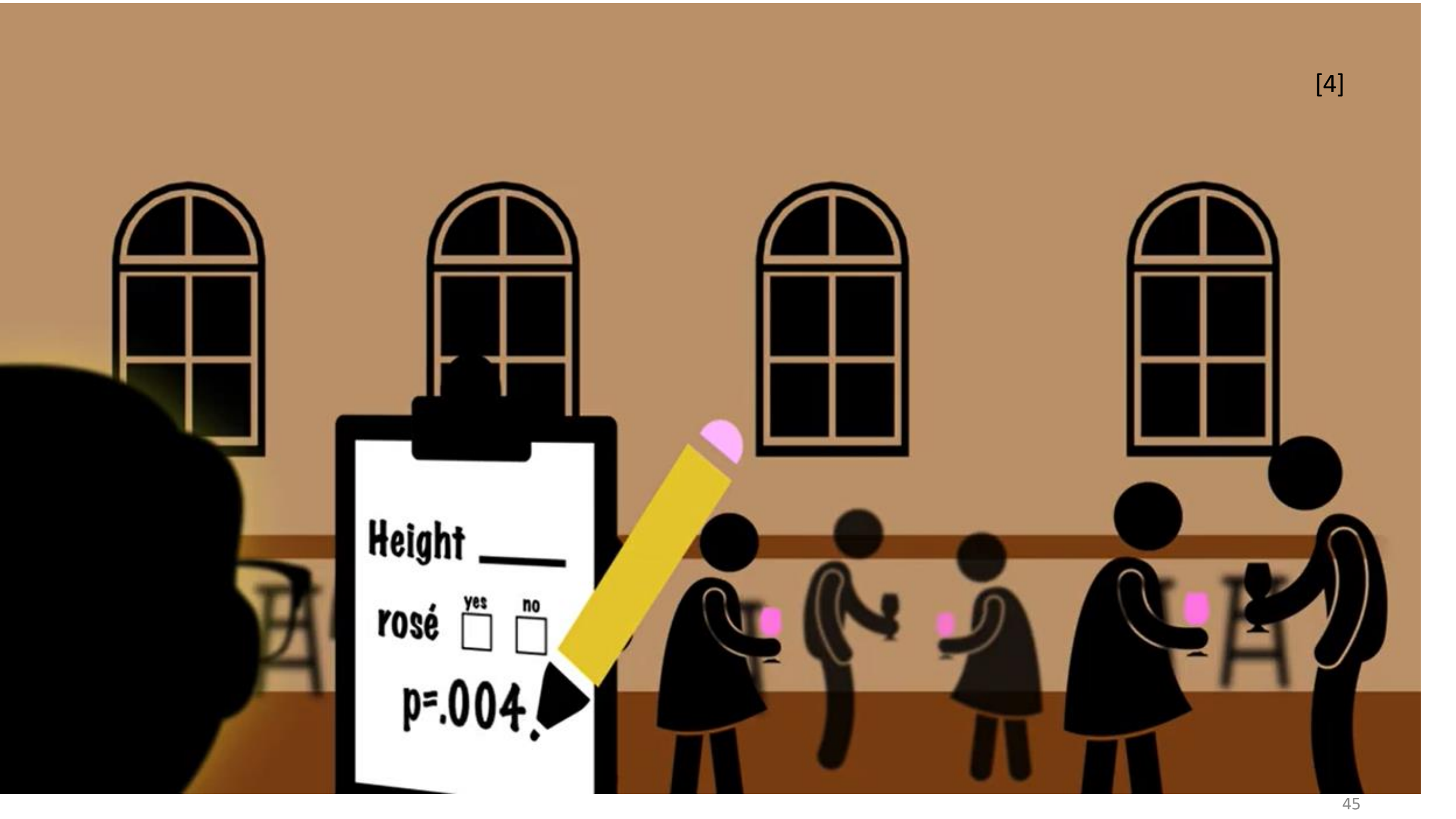
Semnificația statistică (p sau sig.)

- Dacă avem **estimări ale asocierii variabilelor, într-un eșantion**, putem calcula p sau sig:
 - Sig. = „semnificația statistică” (statistical significance)
 - p = „probabilitatea de eroare”
- p = probabilitatea ca valoarea obținută în eșantion să fie datorată șansei și nu unui proces real din populație
- De regulă dacă **$p < 5\%$** considerăm că acea valoare este reală
- **Valoarea p nu exclude erorile de modelare!**
- **Dacă eșantionul nu este aleator, nu putem generaliza la populație!**

Hypothesis:

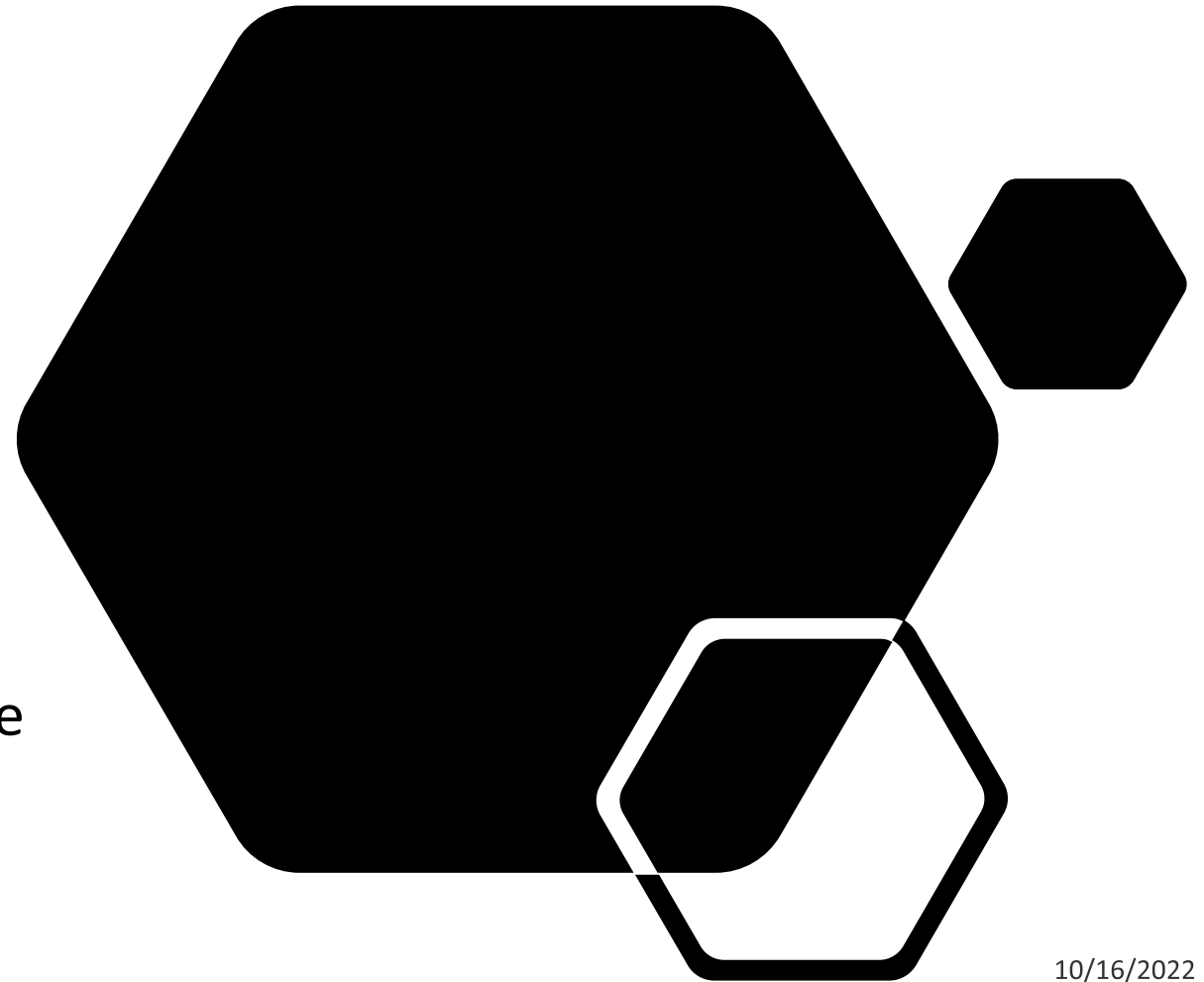
*Drinking rosé
leads to shrinkage*





Erori de modelare – studiu de caz

F. Messerli, Consumul de ciocolată și premiile
Nobel



10/16/2022

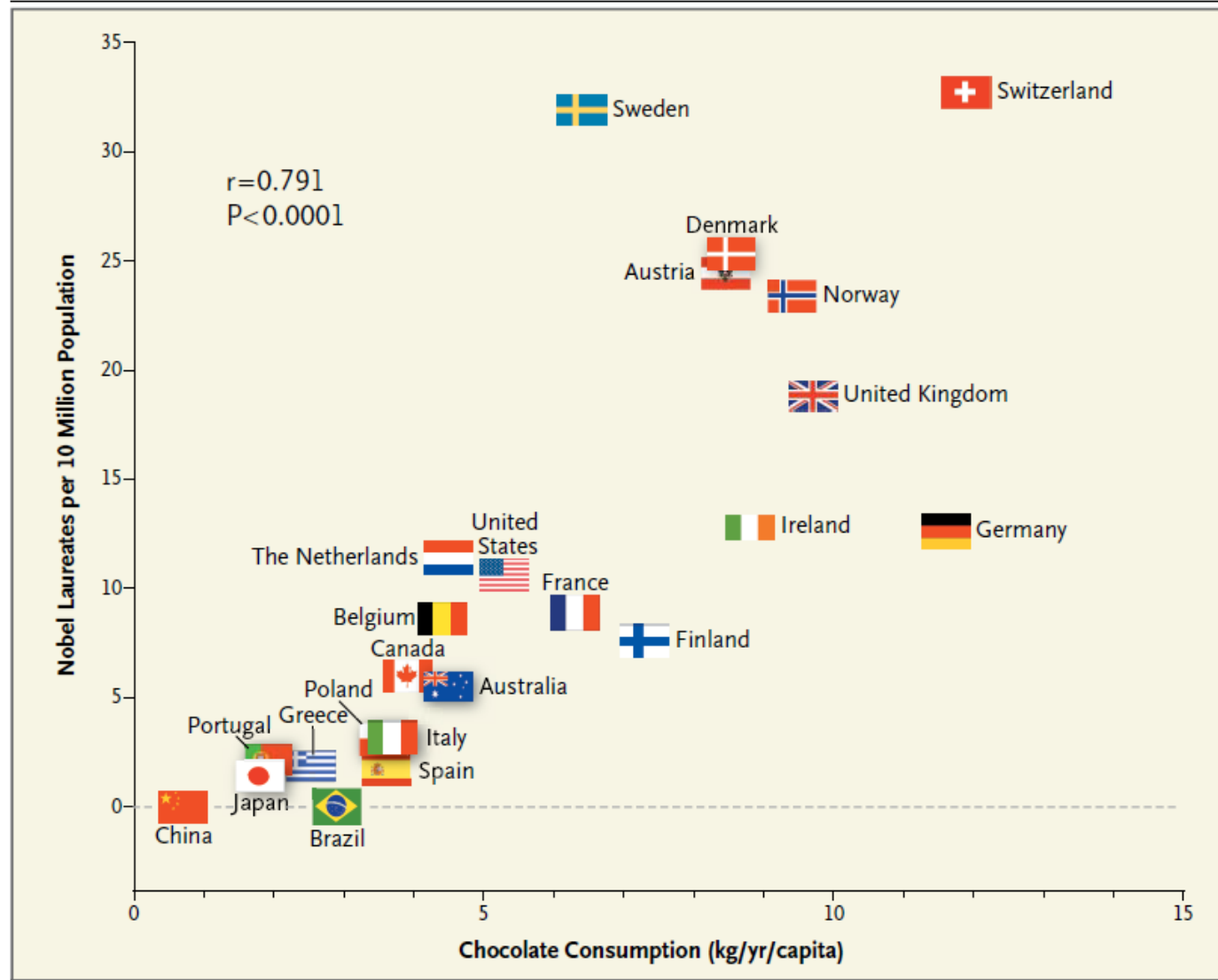
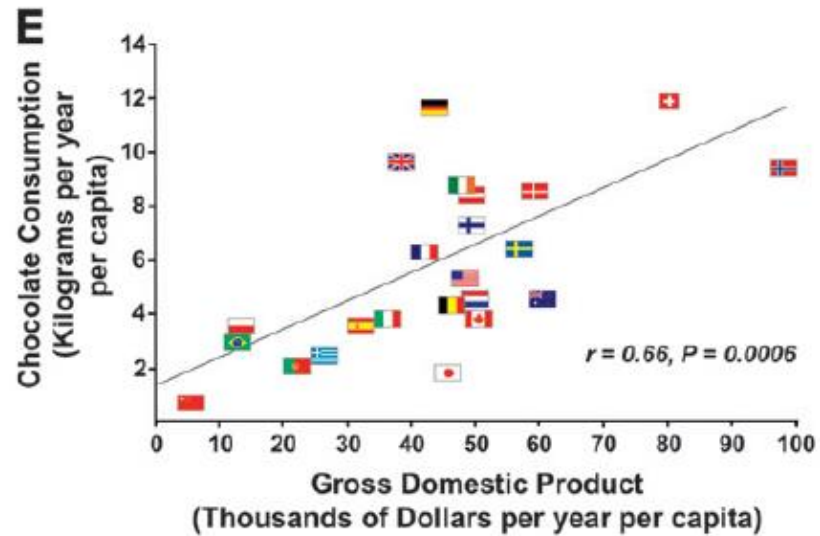
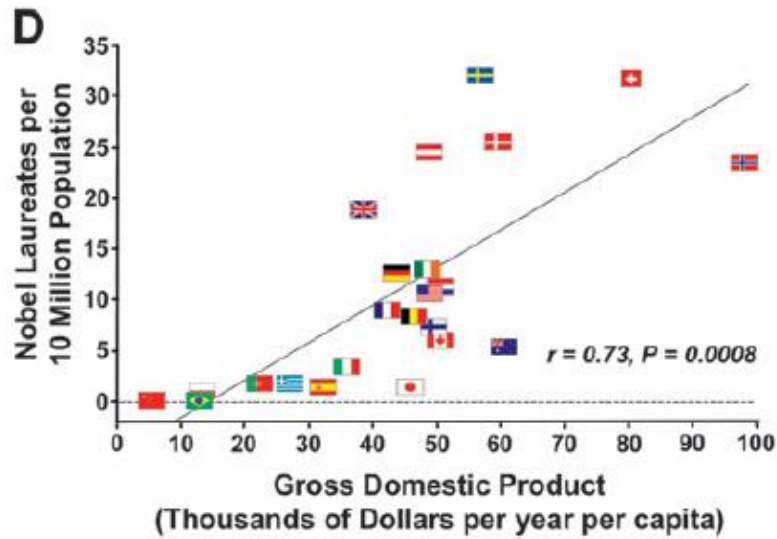
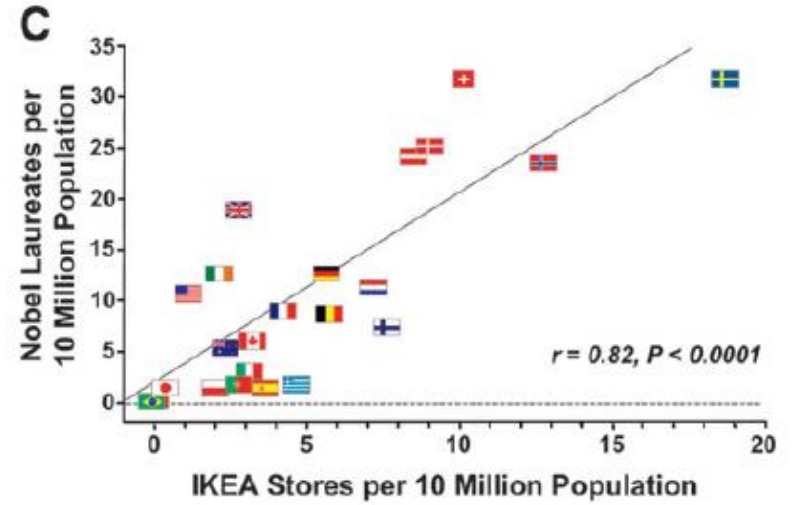
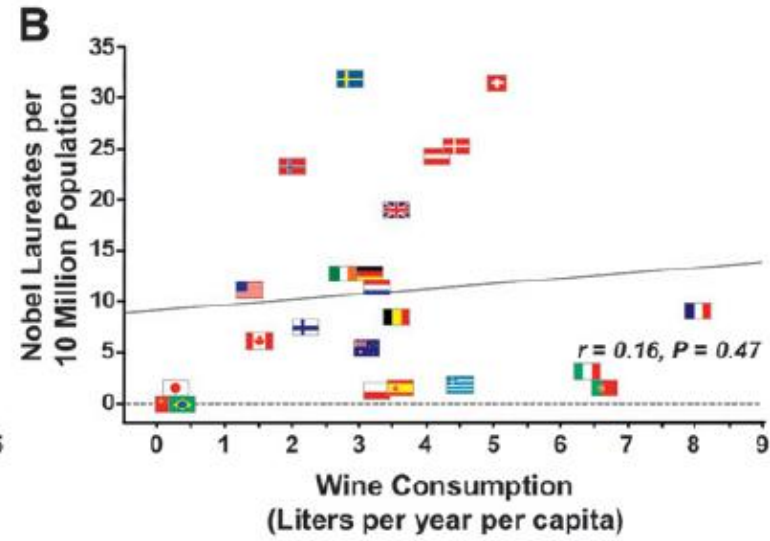
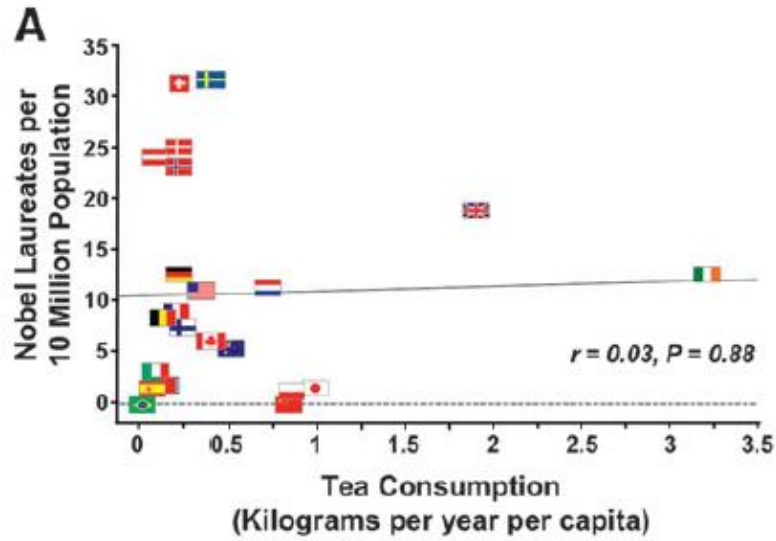
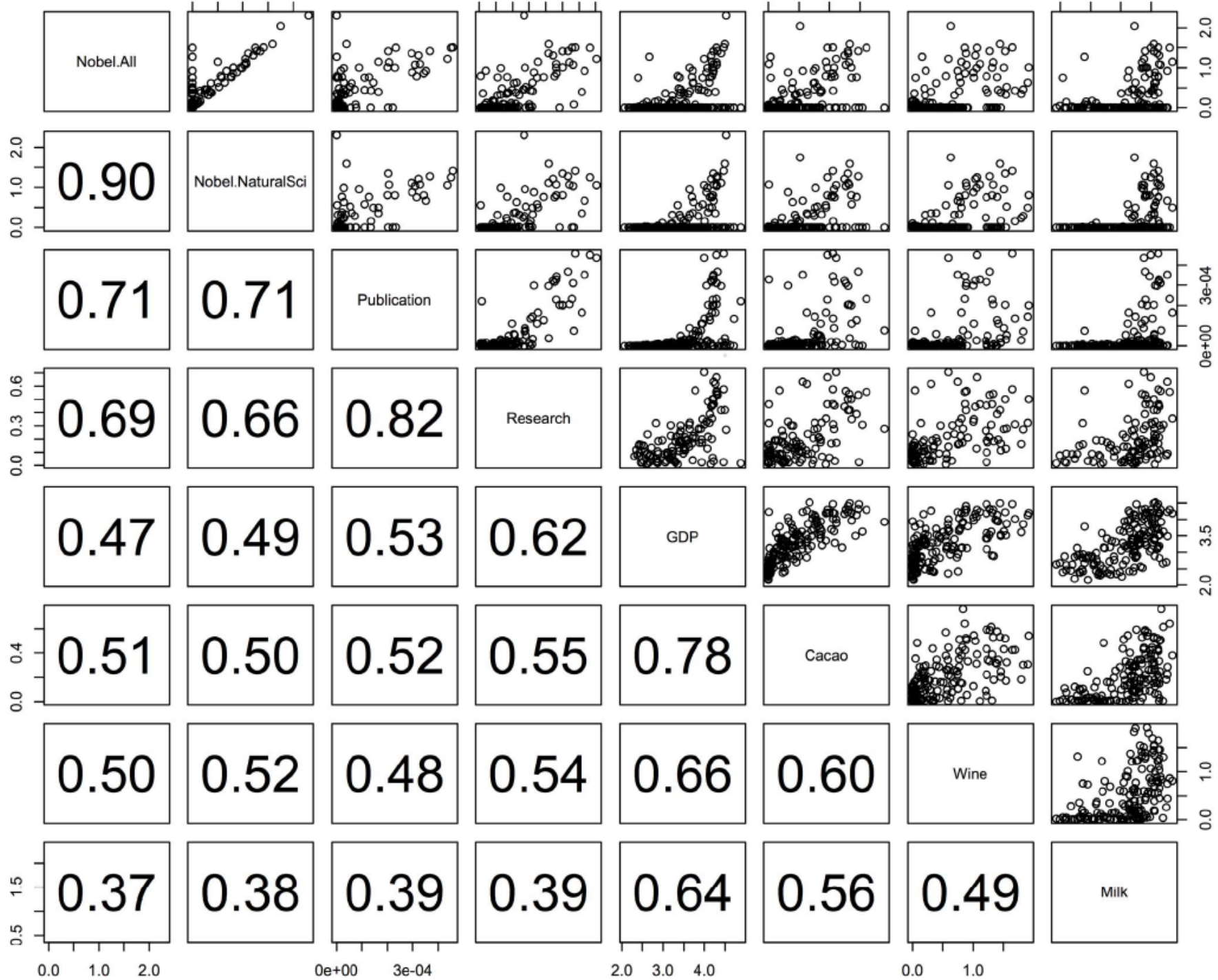


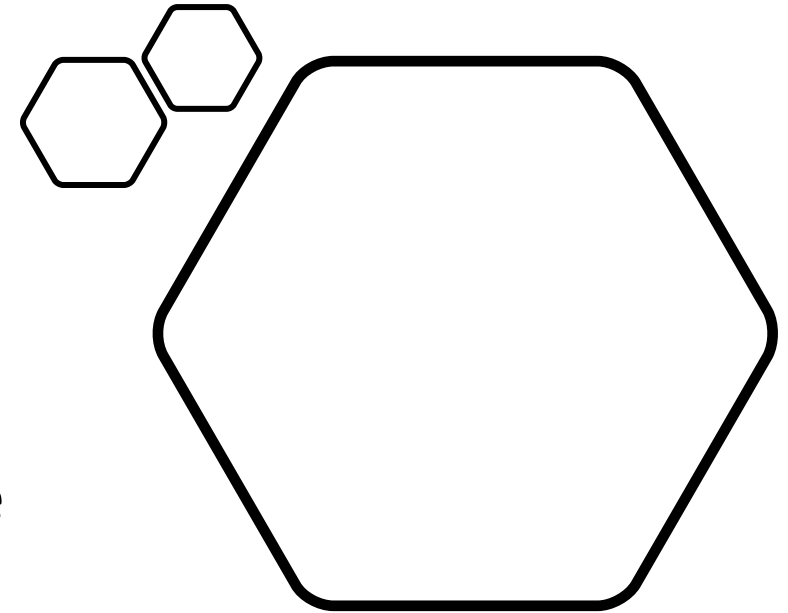
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.





Concluzii

- Corelația este un posibil simptom al cauzalității
 - Corelația este observabilă în norul de puncte
 - Coeficienții de asociere măsoară intensitatea relației
- Modelul face legătura dintre corelație și explicație
- Erori de modelare
 - Corelații coincidentale
 - Sensul cauzării
 - A fi corect din rațiuni false
 - Paradoxul lui Simpson
 - Eroarea ecologică
 - Erorile sistematice de măsurare
- Erorile pot proveni nu doar din model, ci și din eșantionare

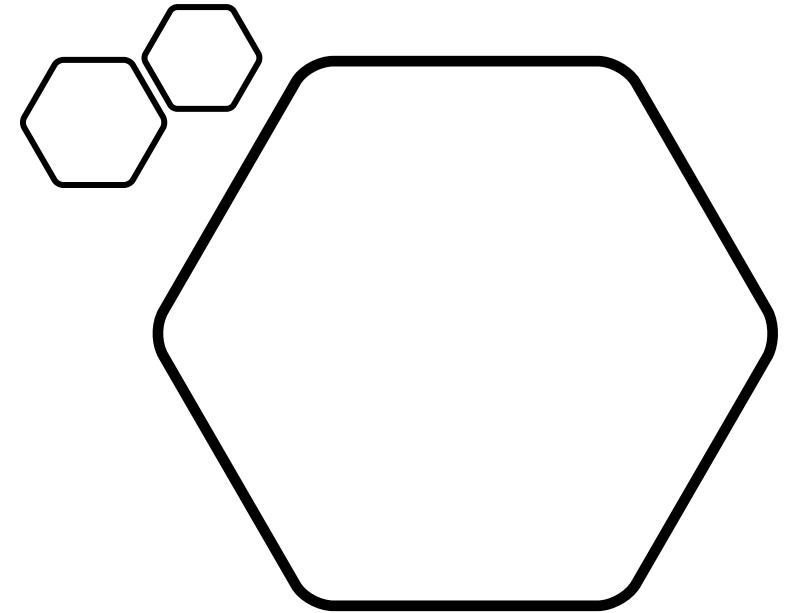


Bibliografie

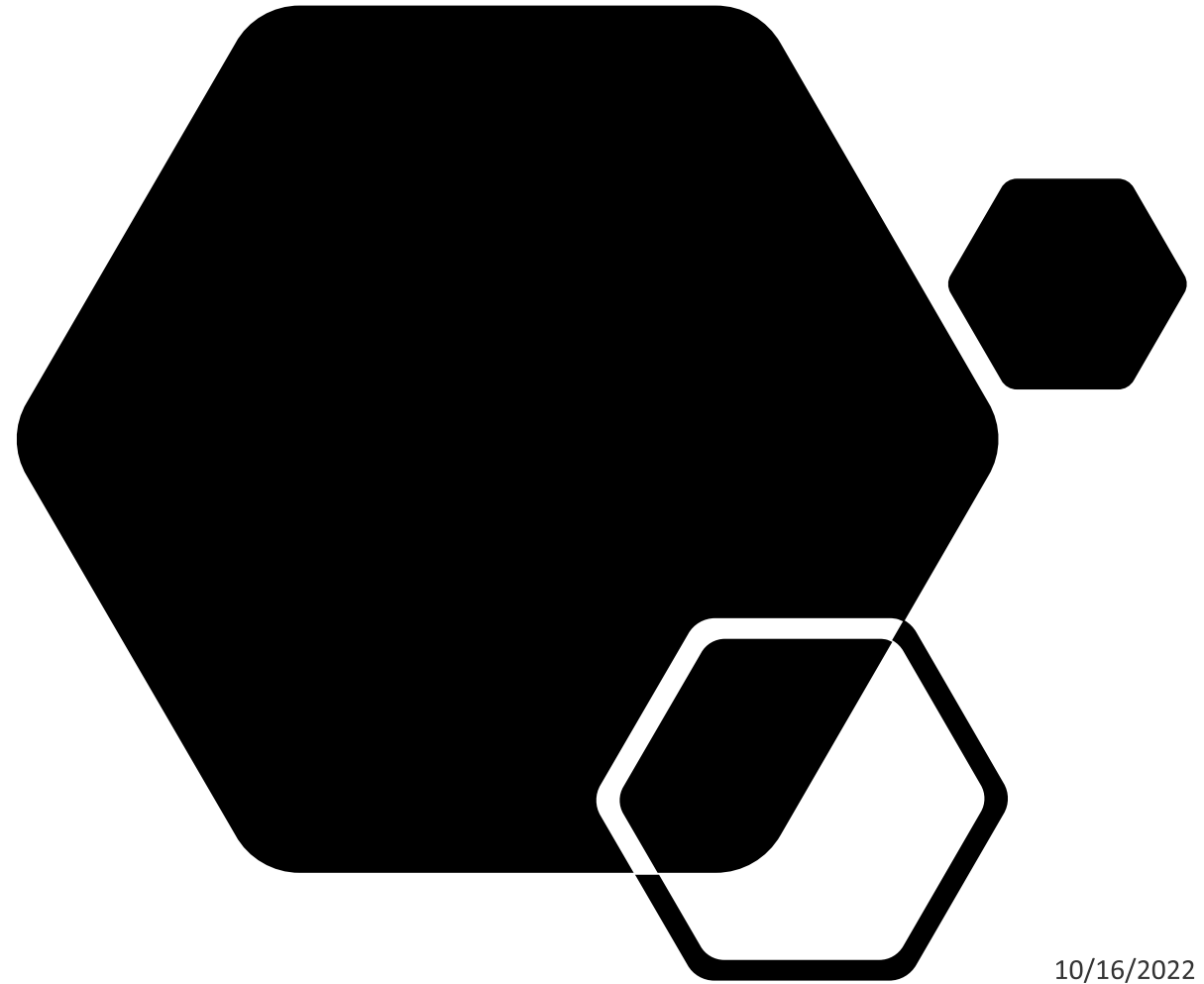
- [1] Vigen, T. (2015). *Spurious correlations*. Hachette books. <http://www.tylervigen.com/spurious-correlations>
- [2] Steve Borgatti, Simpson's Paradox.
<https://sites.google.com/site/ba762researchmethods/materials/handouts/simpsonsparadox>
- [3] Kievit, R., Frankenhuis, W. E., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology*, 4, 513.
- [4] F. Perry Wilson, Confounding: What Did the Researcher Miss?,
<https://www.youtube.com/watch?v=HMTs5AZtjpE>
- [5] Messerli, F.H. Chocolate Consumption, Cognitive Function, and Nobel Laureates. [NEJM](#). (2012)
- [6] Maurage, P., Heeren, A., & Pesenti, M. (2013). Does chocolate consumption really boost Nobel award chances? The peril of over-interpreting correlations in health studies. *The Journal of Nutrition*, 143(6), 931-933.
- [7] Doi, H., Heeren, A., & Maurage, P. (2014). Scientific activity is a better predictor of Nobel award chances than dietary habits and economic factors. *PLoS One*, 9(3), e92612.
- [8] Shona McCombes, An introduction to sampling methods, <https://www.scribbr.com/methodology/sampling-methods/>
- [9]. Wikipedia contributors. Correlation and dependence.
https://en.wikipedia.org/wiki/Correlation_and_dependence
- [10] Zeiger, J. S., Silvers, W. S., Fleegler, E. M., & Zeiger, R. S. (2020). Attitudes about cannabis mediate the relationship between cannabis knowledge and use in active adult athletes. *Journal of Cannabis Research*, 2, 1-13.

Concluzii

- Fenomenele sunt măsurate prin indicatori
- Procesul măsurat vs. procesul măsurării
- Două modele de măsurare
 - Modelul reflectiv: constructul cauzează indicatorii
 - Modelul formativ: indicatorii compun constructul
- Tendință centrală și variabilitate
- Erori de măsurare
 - Sistemice (validitate) vs. aleatorii (fidelitate)
 - Stilul de răspuns
 - Tipul de scală



Extra time



10/16/2022

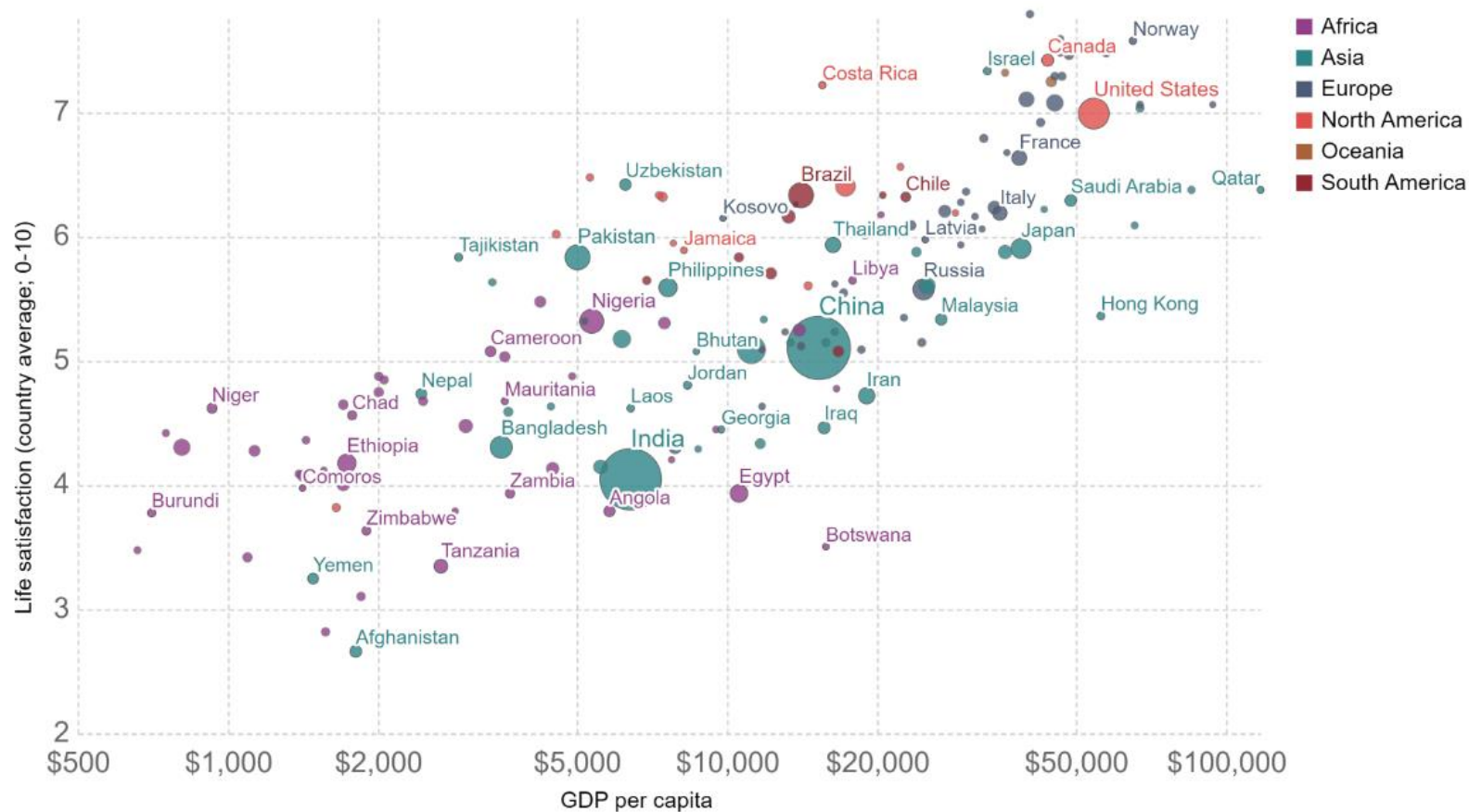
Corelație pozitivă

Self-reported Life Satisfaction vs GDP per capita, 2017

Our World in Data

The vertical axis shows the national average of the self-reported life satisfaction on a scale ranging from 0-10, where 10 is the highest possible life satisfaction.

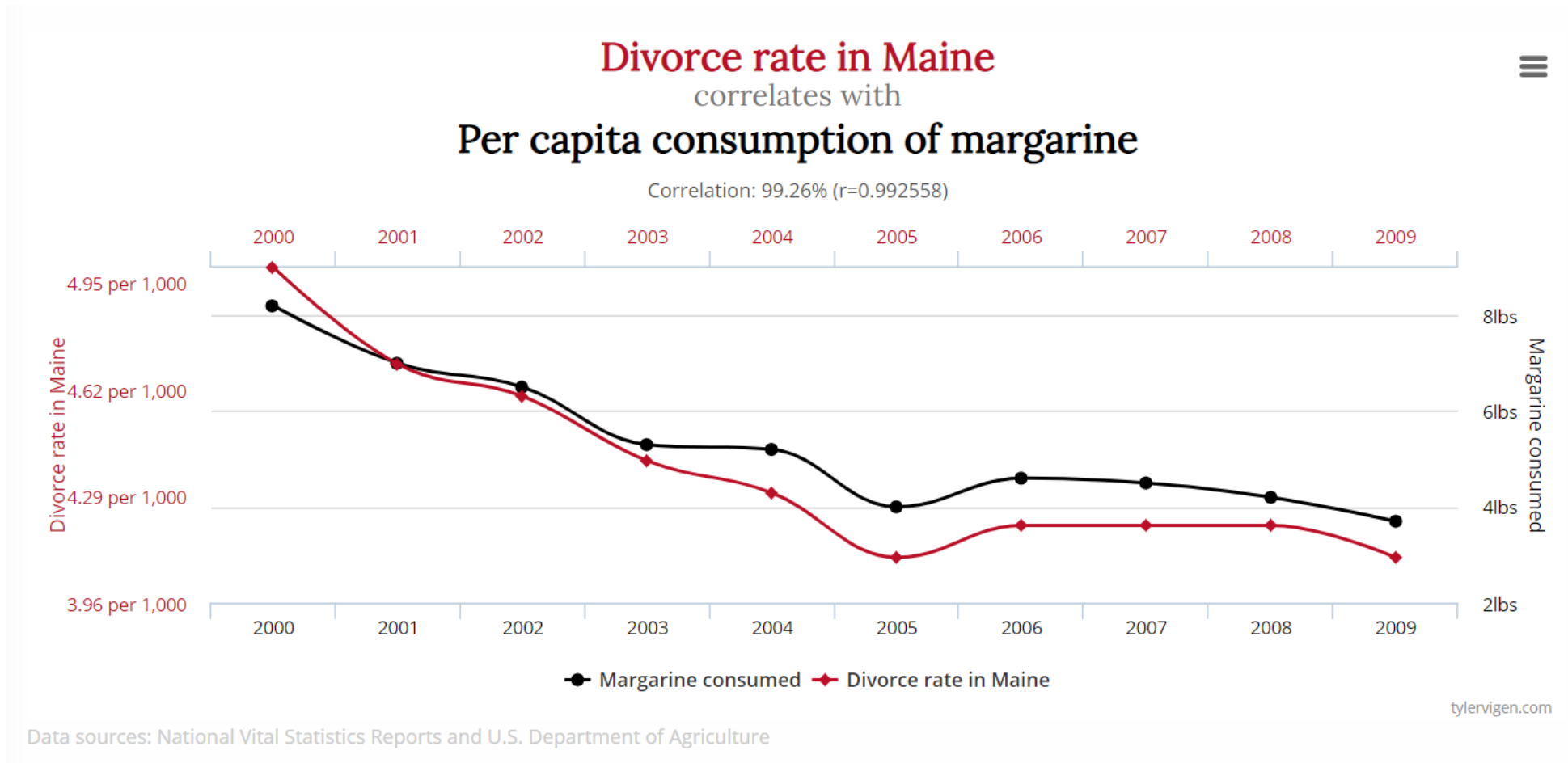
The horizontal axis shows GDP per capita adjusted for inflation and cross-country price differences.



Source: World Happiness Report (2019), World Bank

OurWorldInData.org/happiness-and-life-satisfaction/ • CC BY

Sursă

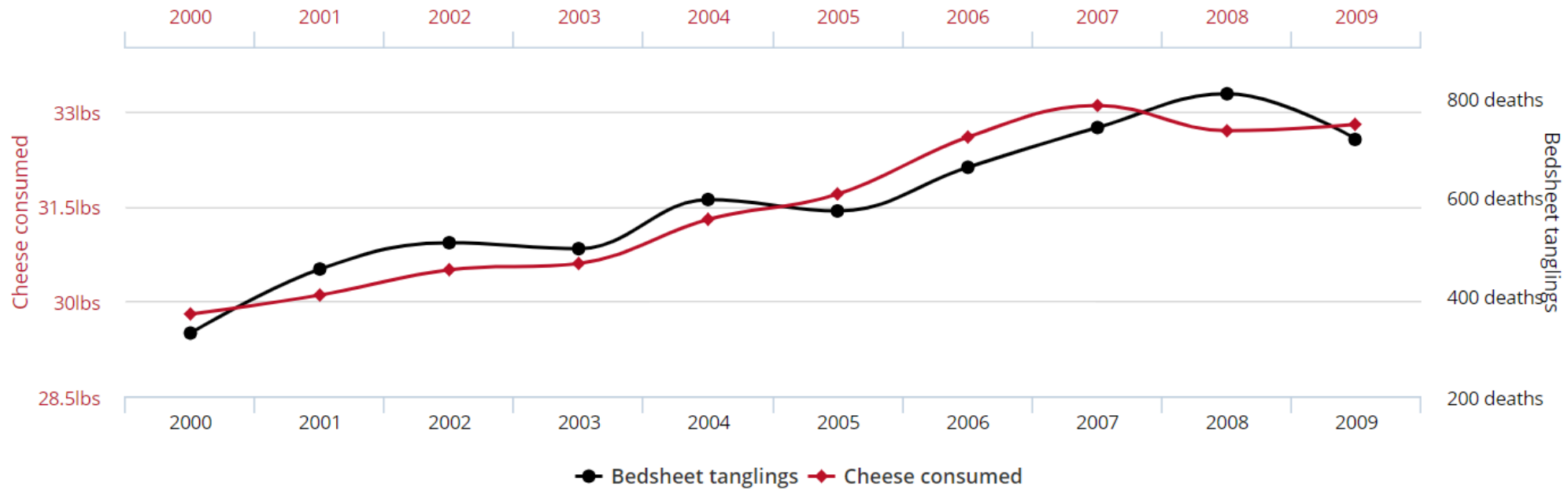


Tyler Vygen, Spurious correlations

Per capita cheese consumption correlates with Number of people who died by becoming tangled in their bedsheets



Correlation: 94.71% (r=0.947091)



tylervigen.com

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Tyler Vigen, Spurious correlations