# Analiza datelor. Modelare și eșantionare

Proiectul învățământului superior din Moldova – DATASEA

Răzvan Rughiniș

razvan.rughinis@upb.ro

# Concepte principale

Corelația

Erori de modelare

Erori de eșantionare

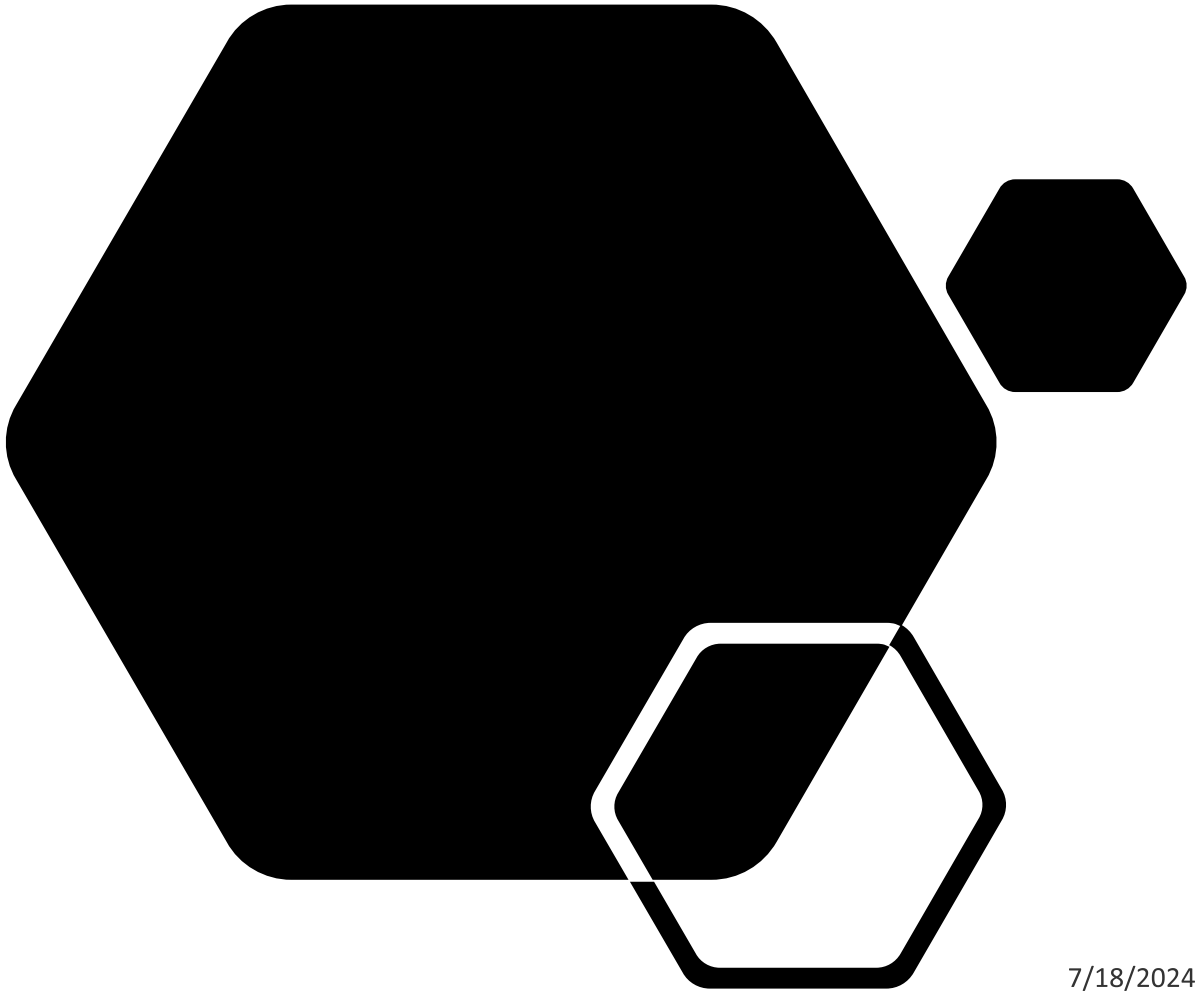
Semnificația statistică

# Corelația

Covariația

Asocierea

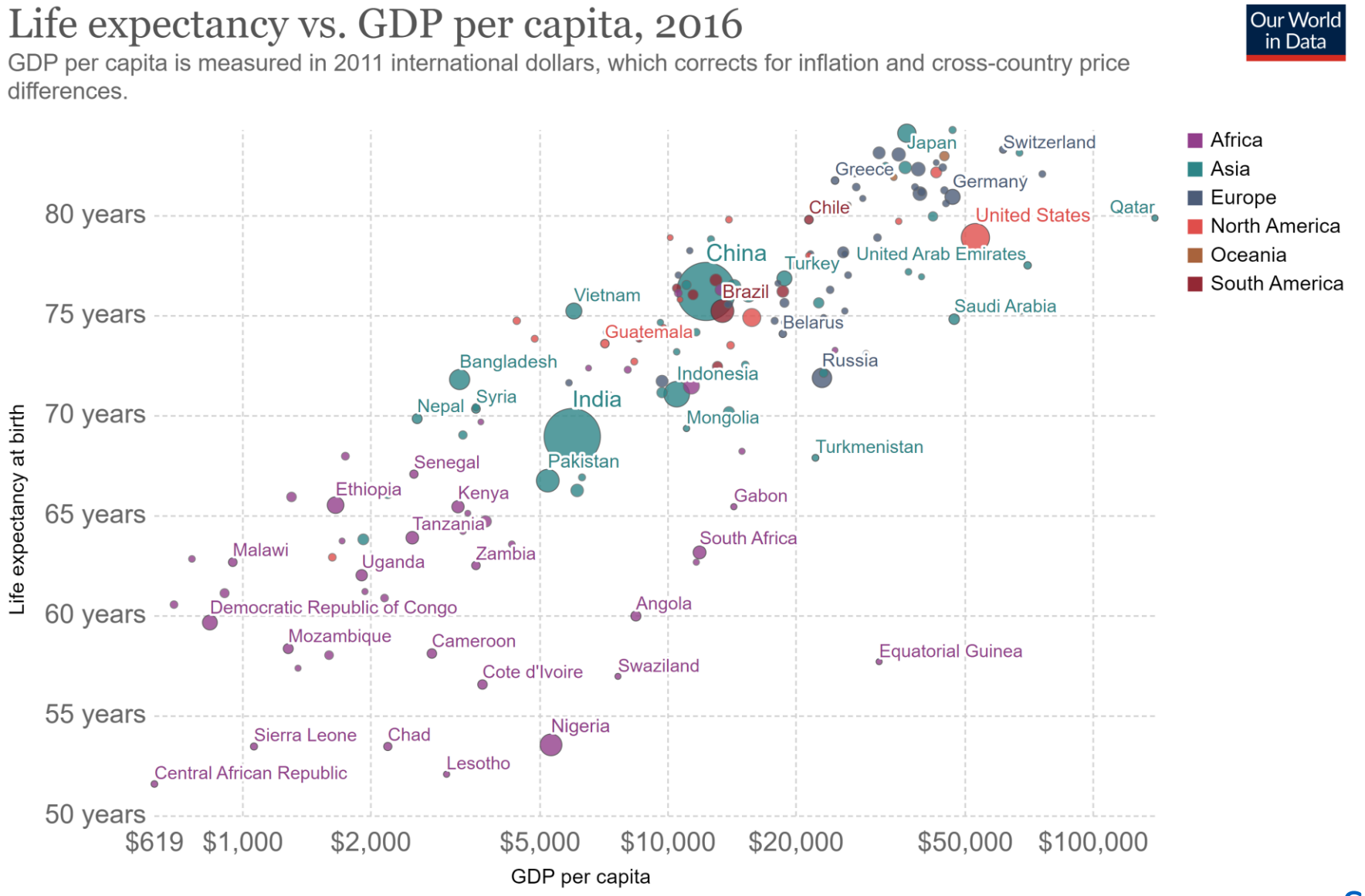
Corelația



# Covariația sau asocierea variabilelor

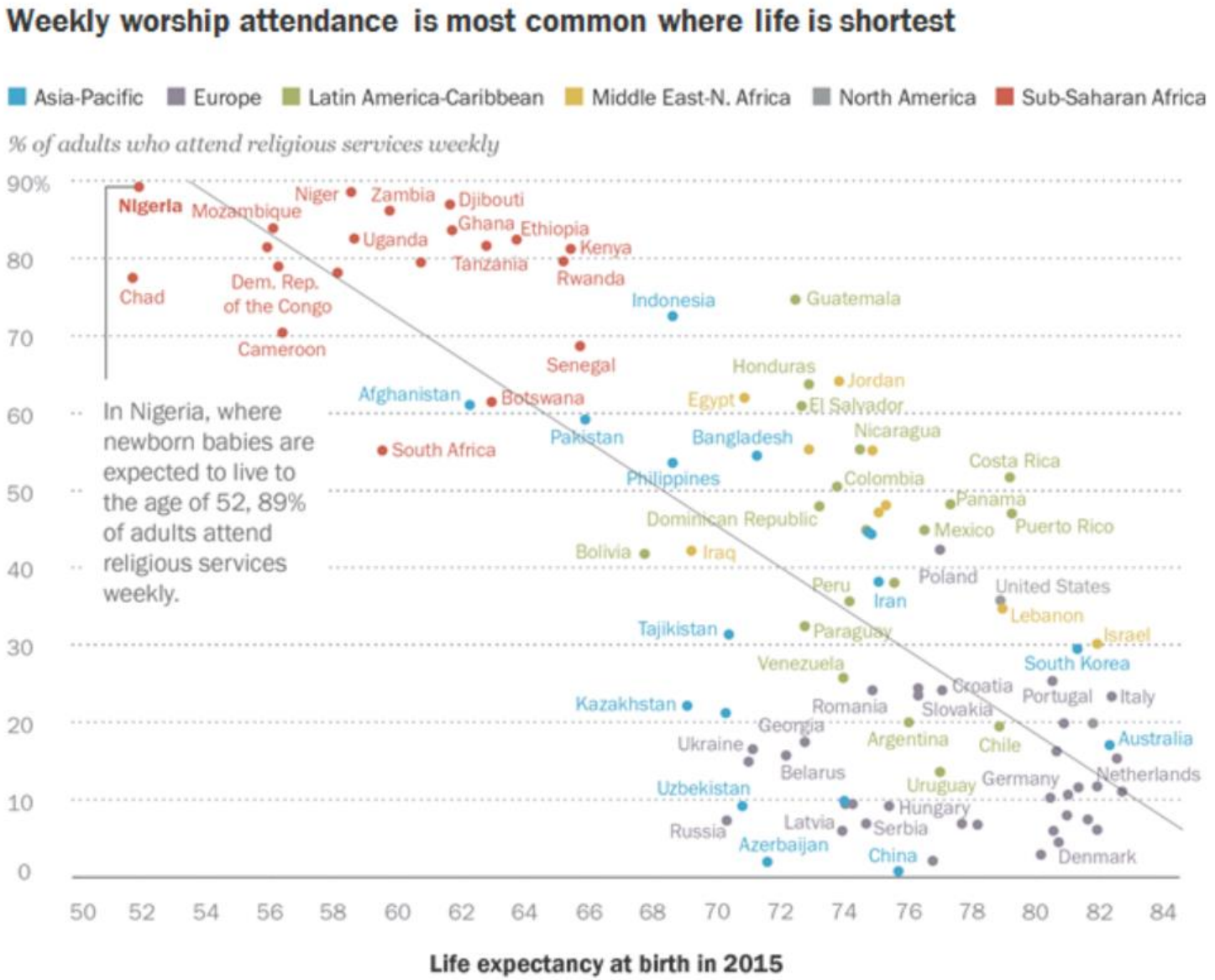
- Dacă estimăm mai bine valorile lui A știind valorile lui B, decât la întâmplare, atunci spunem că:
  - A și B sunt asociate (vezi clarificări aici)
  - A și B covariază (adică variază împreună)
  - A și B sunt corelate (mai ales pentru variabilele numerice)
- Și variabilele ordinale sau nominale pot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

# Corelație pozitivă



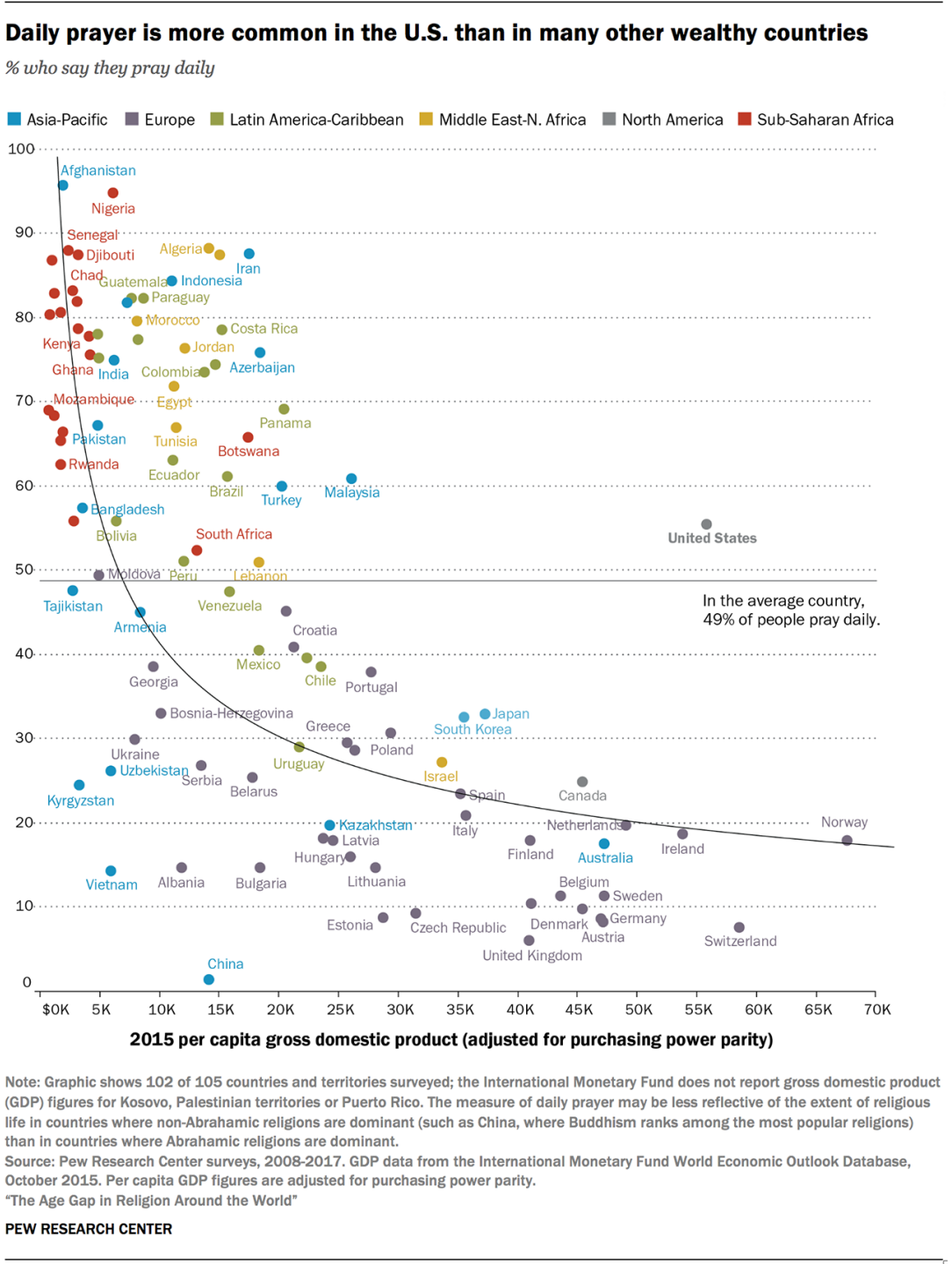

Sursă

# Corelație negativă



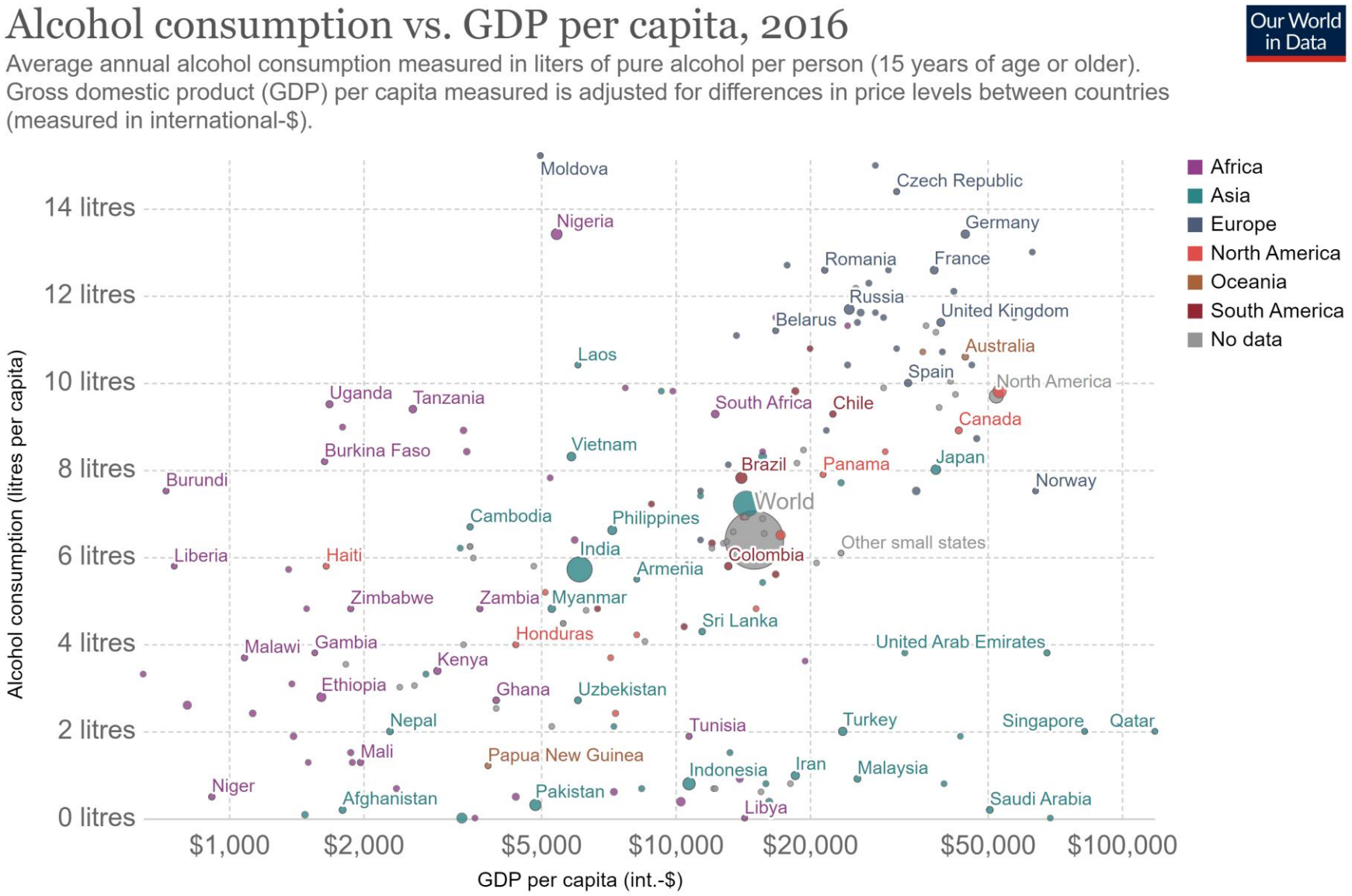

Sursă

# Asociere non-liniară




Sursă

# Corelație nulă/slabă




Sursă

# Covariață și corelație

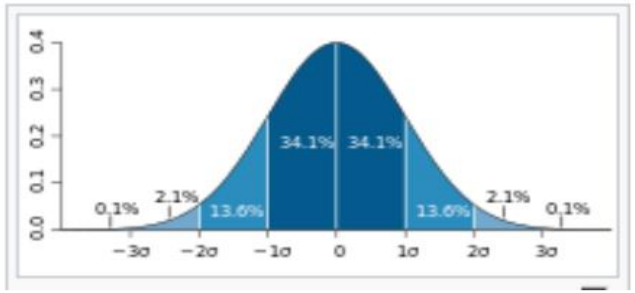
- **Covariația** măsoară **sensul** relației dintre 2 variabile

- 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 covariația variabilelor normalizate
  - Unitatea de măsură naturală este înlocuită de abaterea standard
- Mărimea corelației nu are unitate de măsură
  - **Corelația măsoară intensitatea asocierii**
  - **Rang: [-1, 1]**

# Relația statistică

## Varianță și abaterea standard a unei variabile




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

$\sigma$ = population standard deviation

$N$ = the size of the population

$x_i$ = each value from the population

$\mu$ = the population mean

## Covarianță vs. coef. corelație R pt. 2 variabile

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

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

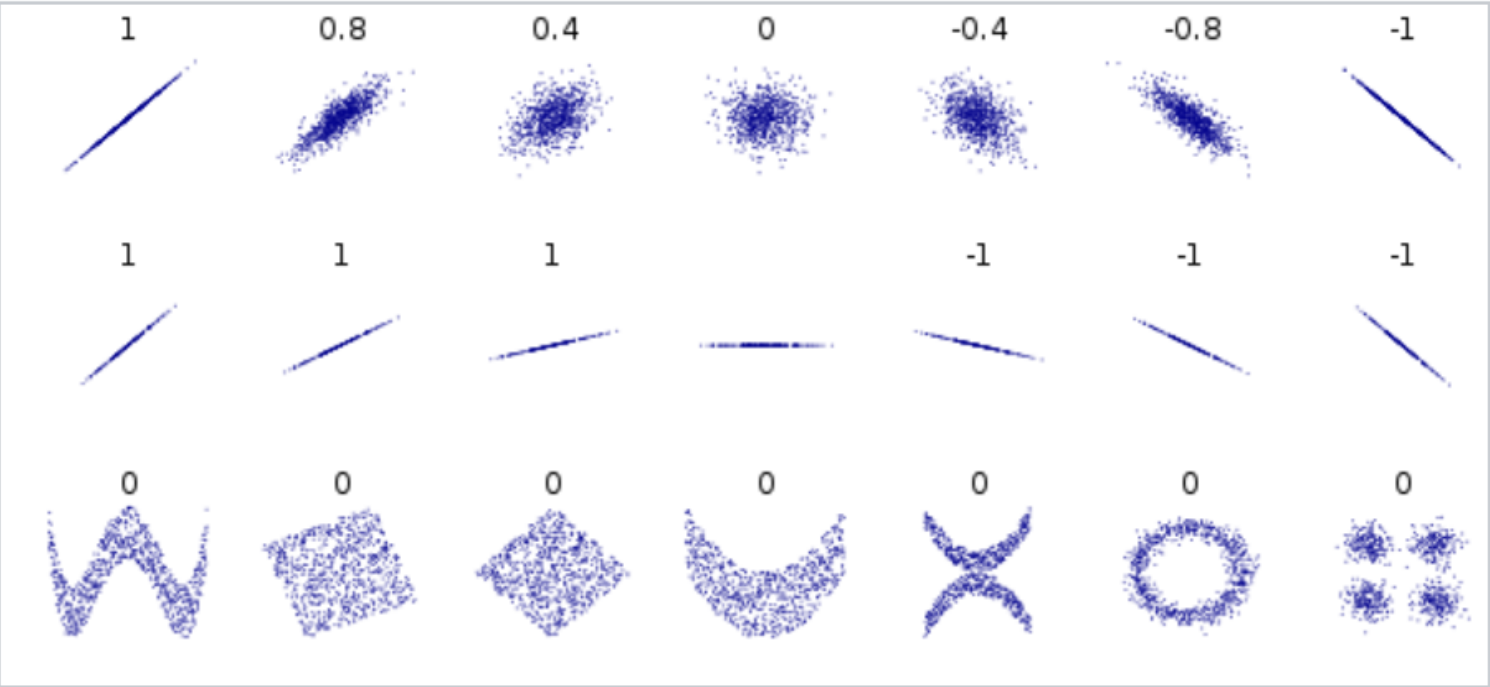
where:

- cov is the covariance
- $\sigma_X$ is the standard deviation of $X$
- $\sigma_Y$ is the standard deviation of $Y$

Sursă

# 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 Pearson 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?

**C1_8. De obicei, cat de des mergeti la biserica? * I1. Sexul Crosstabulation**

% within I1. Sexul

| | | I1. Sexul | | Total |
|---|---|---|---|---|
| | | 0 Masculin | 1 Feminin | |
| C1_8. De obicei, cat de des mergeti la biserica? | 1 Deloc | 20.6% | 13.7% | 17.0% |
| | 2 Mai rar | 40.4% | 32.2% | 36.2% |
| | 3 De cateva ori pe luna | 25.9% | 38.7% | 32.6% |
| | 4 De cateva ori pe saptamana | 11.0% | 14.2% | 12.7% |
| | 5 Zilnic | 2.0% | 1.1% | 1.6% |
| Total | | 100.0% | 100.0% | 100.0% |

Barometrul de consum cultural 2012

**C1_8. De obicei, cat de des mergeti la biserica? * Mediu de rezidenta Crosstabulation**

% within Mediu de rezidenta

| | | Mediu de rezidenta | | Total |
|---|---|---|---|---|
| | | 0 Rural | 1 Urban | |
| C1_8. De obicei, cat de des mergeti la biserica? | 1 Deloc | 16.4% | 17.4% | 17.0% |
| | 2 Mai rar | 35.9% | 36.5% | 36.2% |
| | 3 De cateva ori pe luna | 36.0% | 30.1% | 32.6% |
| | 4 De cateva ori pe saptamana | 10.1% | 14.6% | 12.7% |
| | 5 Zilnic | 1.6% | 1.5% | 1.5% |
| Total | | 100.0% | 100.0% | 100.0% |

Barometrul de consum cultural 2012

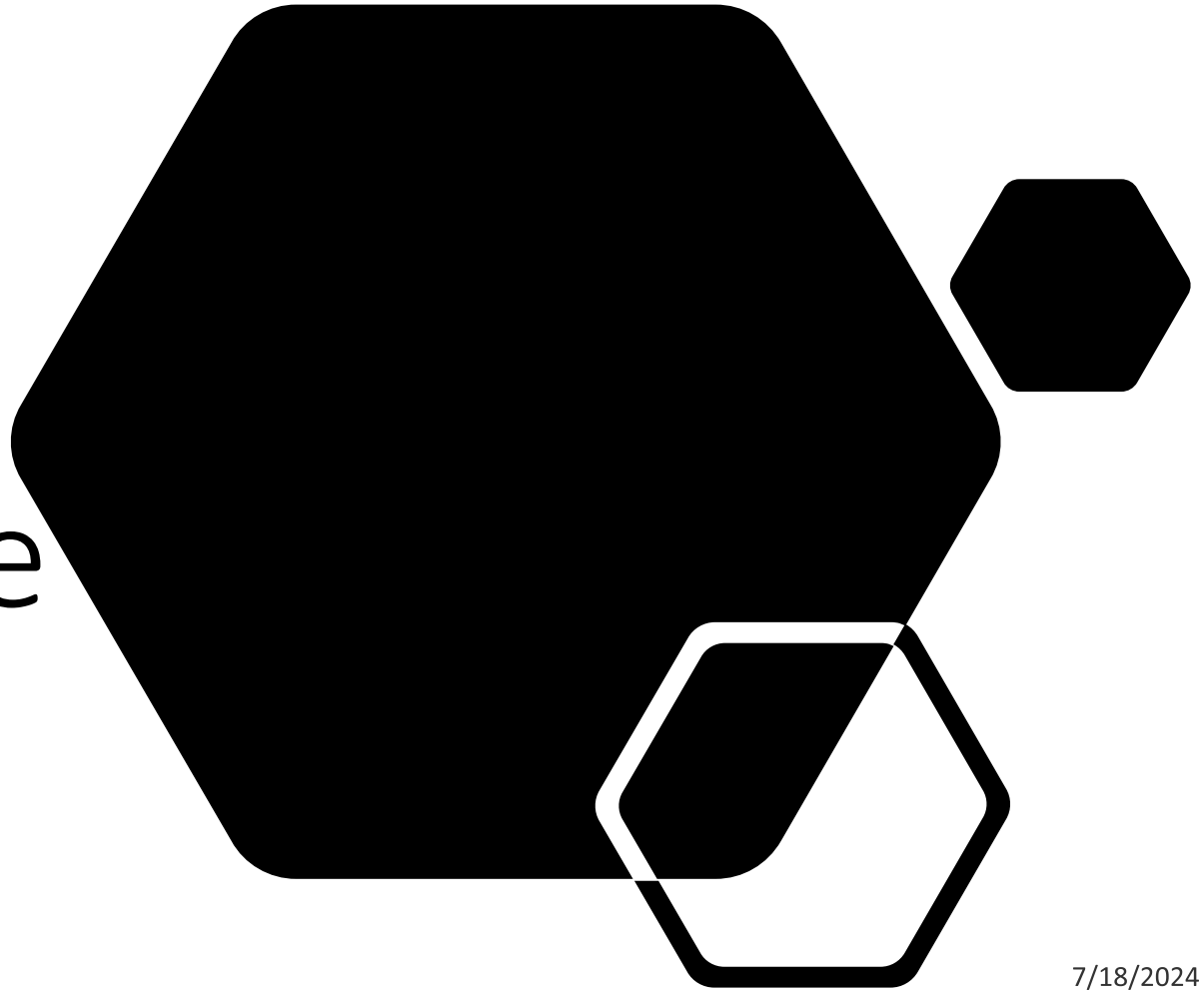
**Correlations**

| | | C1_8. De obicei, cat de des mergeti la biserica? | I1. Sexul | Mediu de rezidenta | I2. Varsta |
|---|---|---|---|---|---|
| C1_8. De obicei, cat de des mergeti la biserica? | Pearson Correlation | 1 | .120** | .009 | .081** |
| | Sig. (2-tailed) | | .000 | .676 | .000 |
| | N | 2358 | 2358 | 2358 | 2358 |
| I1. Sexul | Pearson Correlation | .120** | 1 | -.014 | .078** |
| | Sig. (2-tailed) | .000 | | .485 | .000 |
| | N | 2358 | 2371 | 2371 | 2371 |
| Mediu de rezidenta | Pearson Correlation | .009 | -.014 | 1 | -.076** |
| | Sig. (2-tailed) | .676 | .485 | | .000 |
| | N | 2358 | 2371 | 2371 | 2371 |
| I2. Varsta | Pearson Correlation | .081** | .078** | -.076** | 1 |
| | Sig. (2-tailed) | .000 | .000 | .000 | |
| | N | 2358 | 2371 | 2371 | 2371 |

**. Correlation is significant at the 0.01 level (2-tailed).

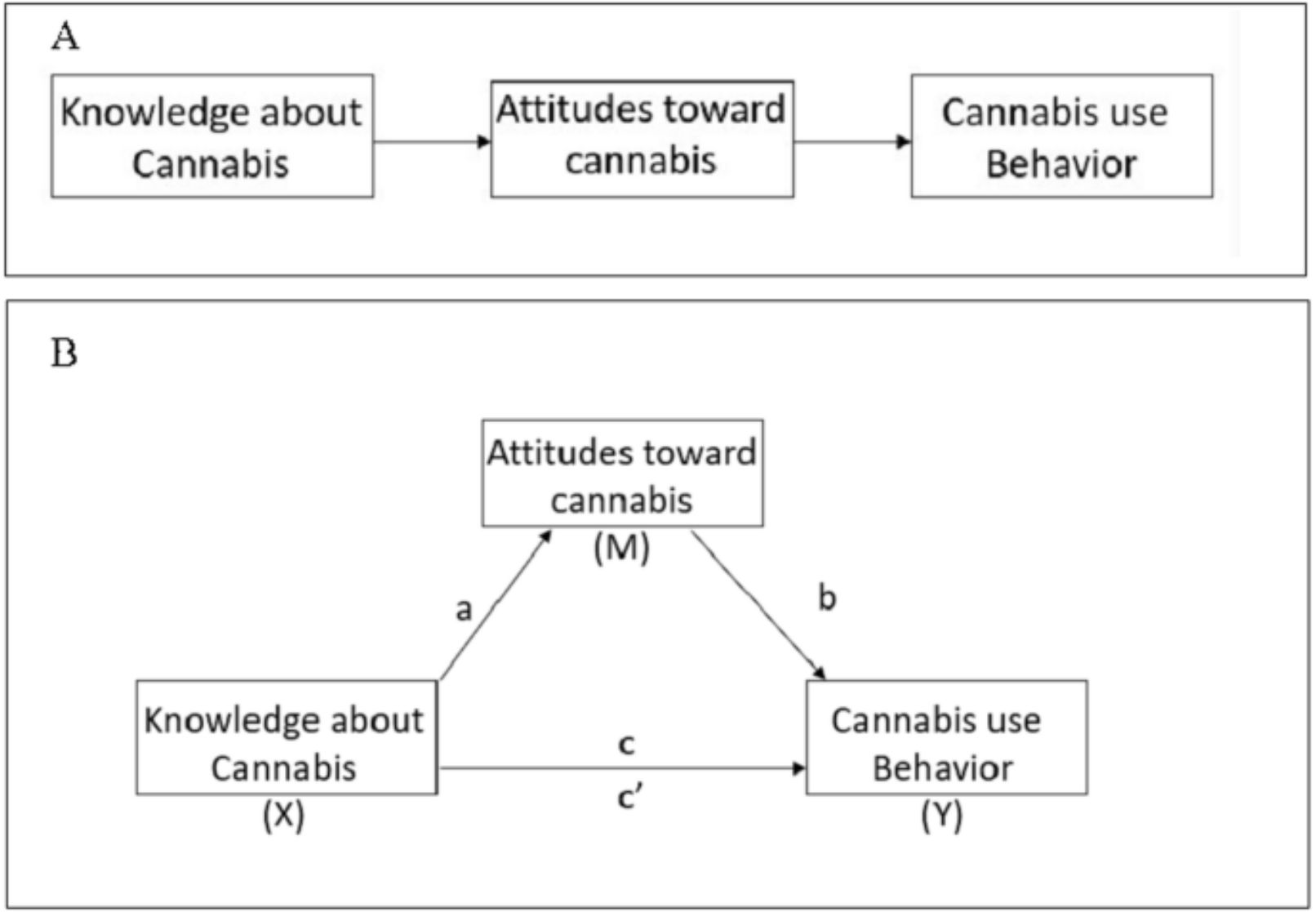
Barometrul de consum cultural 2012

# Erori de modelare

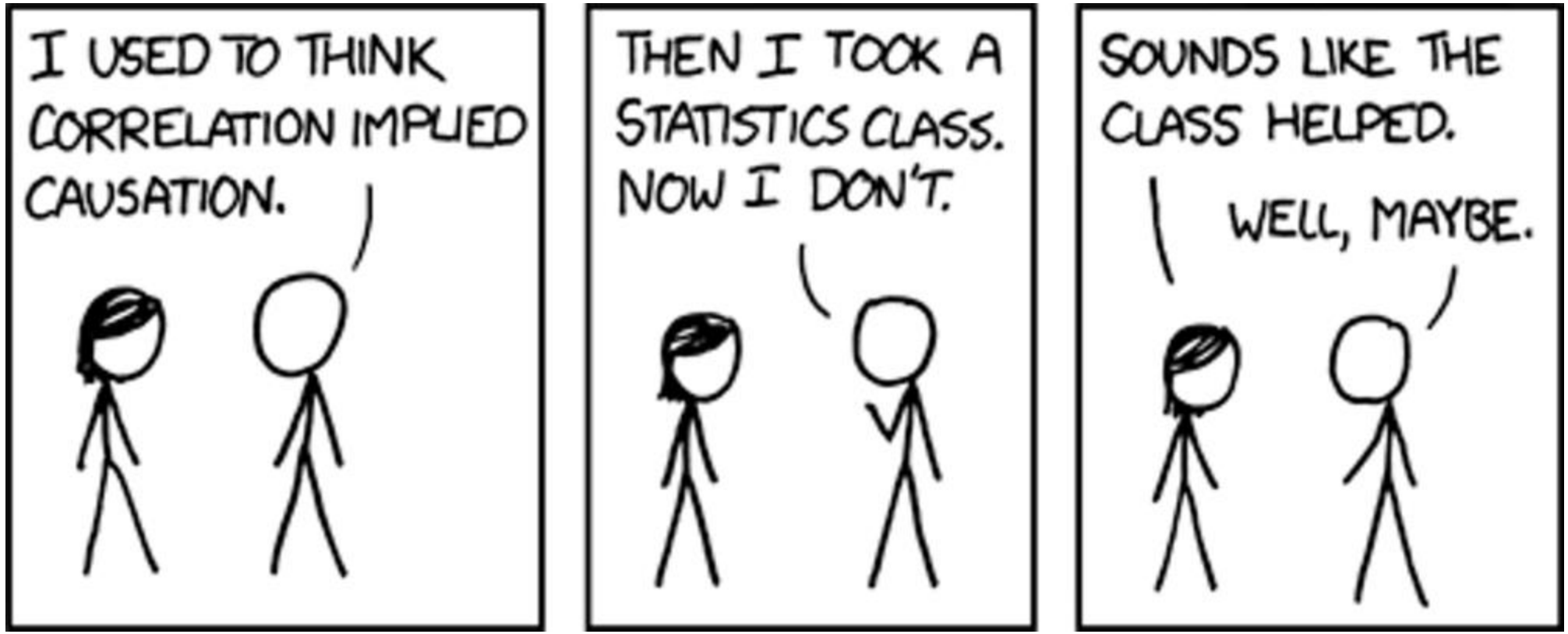


# Modelarea

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

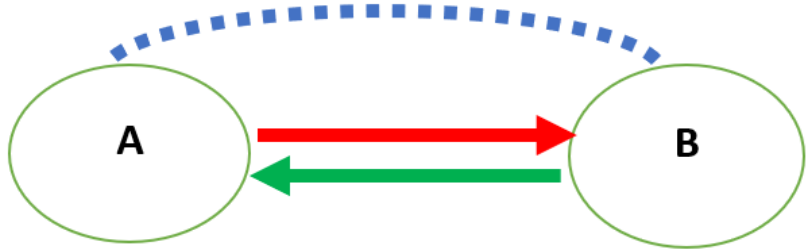



Zeiger et al.

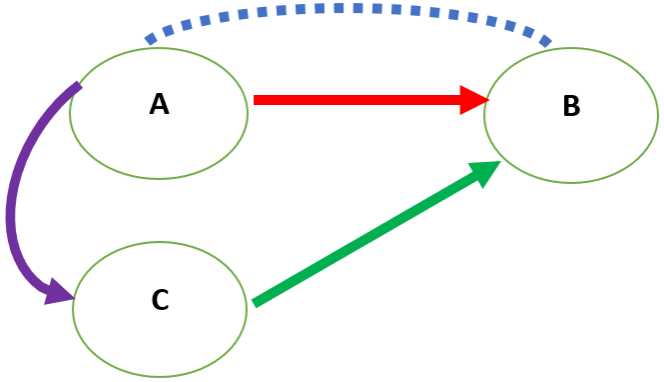

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

# Schemele grafice ale erorilor de modelare

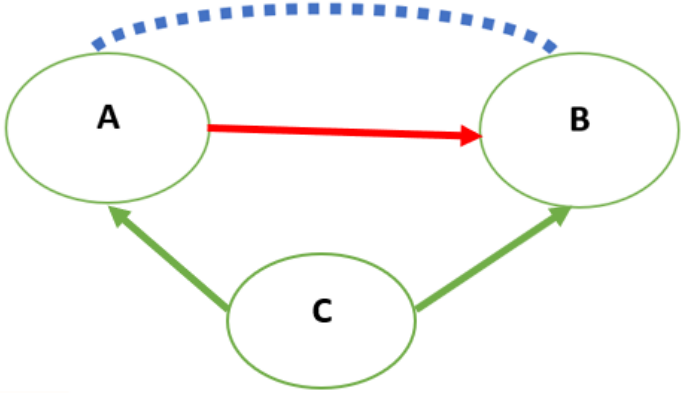

**1. Corelații co-incidentale**



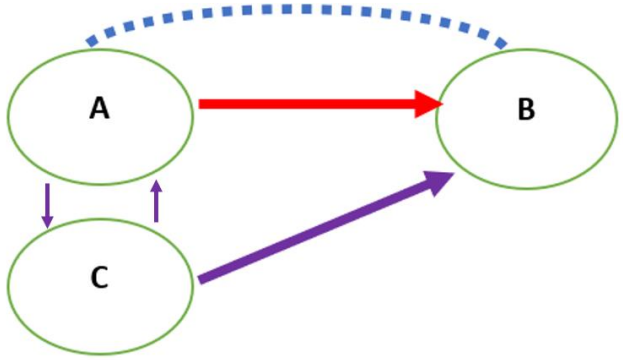

**2. Sensul cauzării**



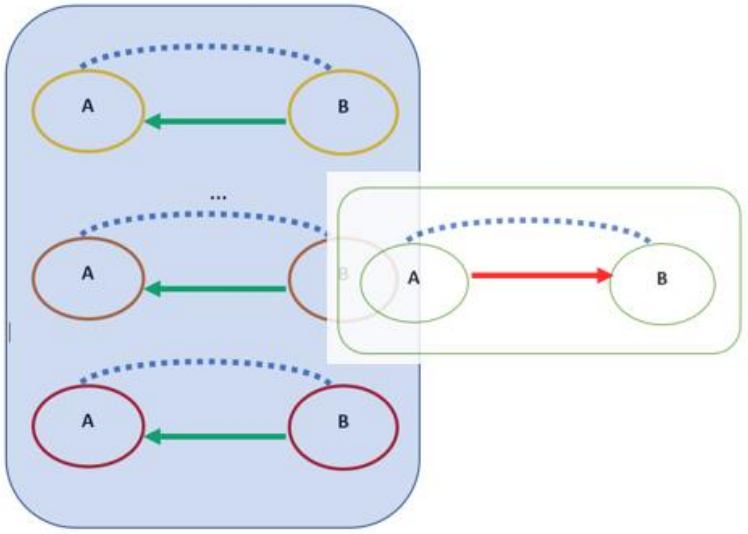

**3. Corect din rațiuni false**



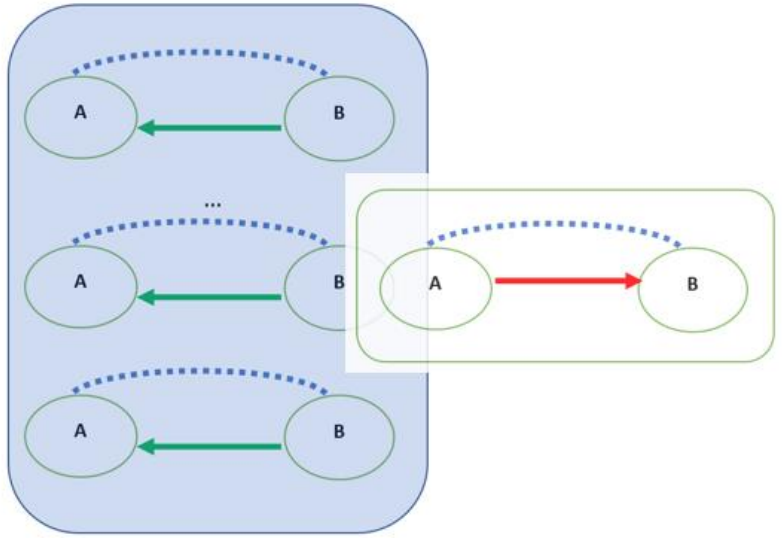

**4. Cauzalitate iluzorie**




**5. Eroarea sistematică de măsurare**




**6. Paradoxul lui Simpson**




**7. Eroarea ecologică**

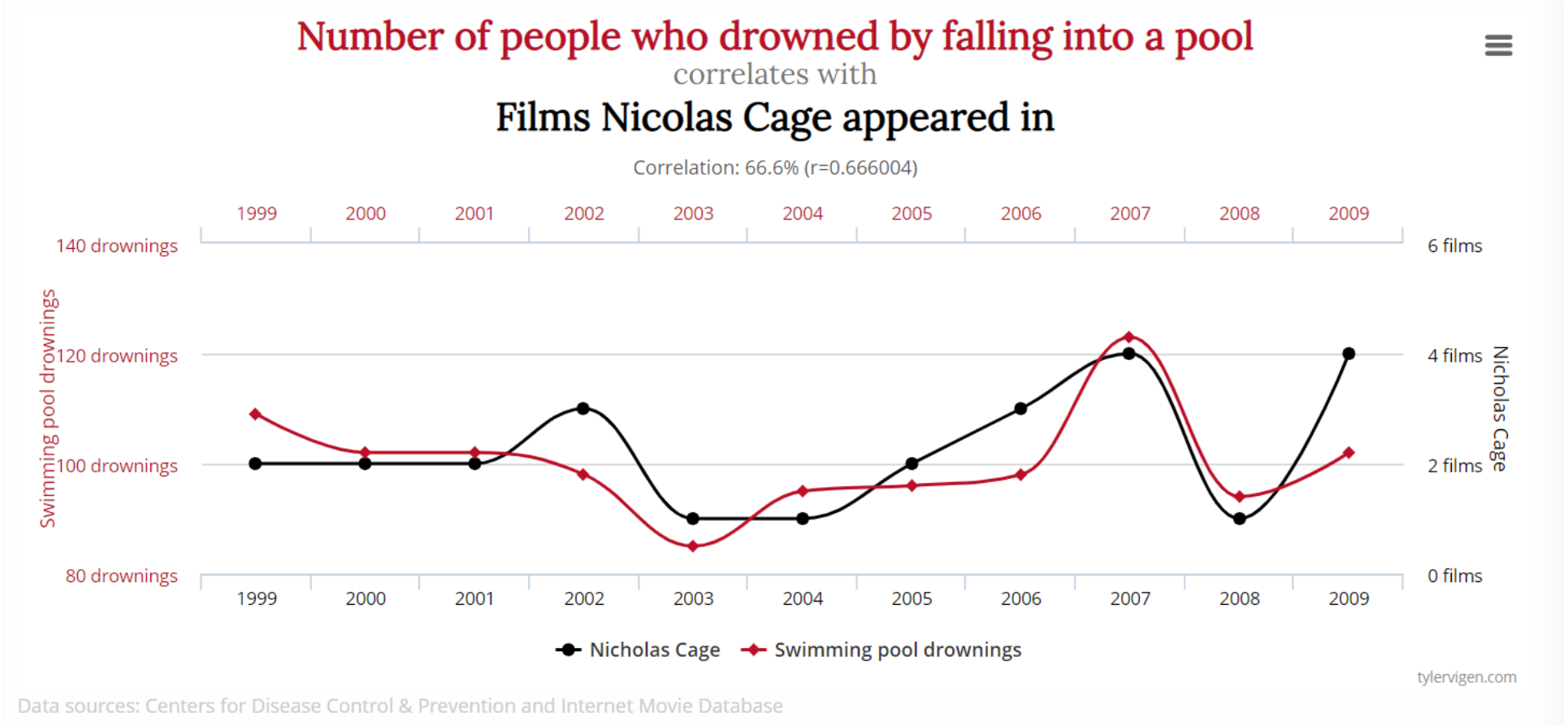
# Erorile de modelare

1. **Corelații co-incidentale**: nu există nicio cauzalitate, doar co-incidențe repetate apărute aleatoriu într-un volum mare de date
2. **Sensul cauzării**: presupusul efect este de fapt cauza, sau relația este circulară
3. **A fi corect din rațiuni false**: alte variabile intermediare din model sunt de fapt cauza, schimbând interpretarea fenomenului
4. **Cauzalitate iluzorie** (en. *spurious correlation*): variabilele sunt corelate, fiind ambele efecte ale unei terțe cauze
5. **Eroarea sistematică de măsurare**: lipsește din modelul explicativ variabila ce cauzează bias-ul în măsurare
6. **Paradoxul lui Simpson**: o corelație per ansamblu pozitivă în populație e agregată din relații cauzale negative în interiorul subpopulațiilor
7. **Eroarea ecologică**: prin observarea corelației la nivelul colectivităților de indivizi, sensul relației cauzale reale, la nivel de individ, se inversează

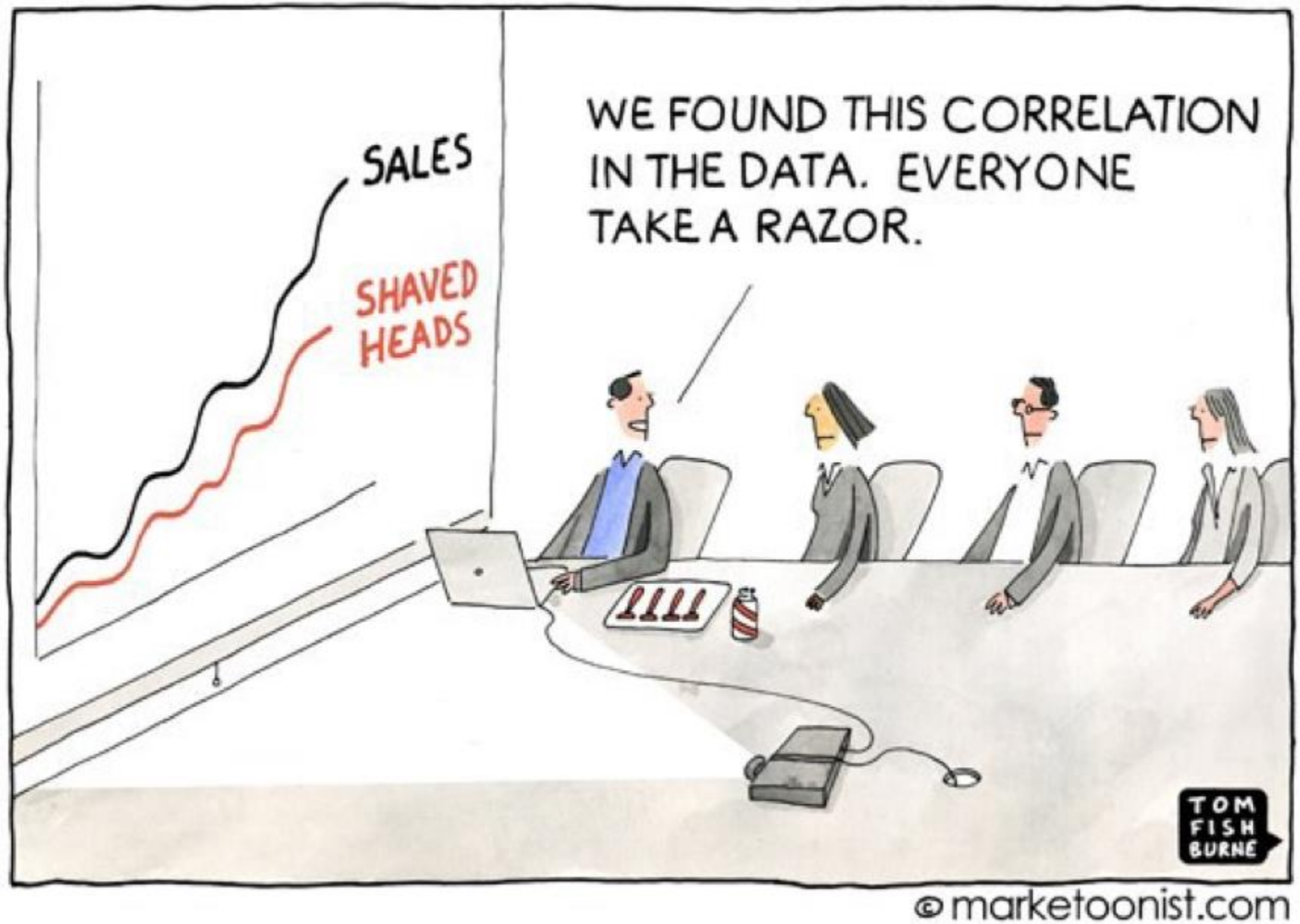
# 1. Corelații coincidentale

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

# Corelații iluzorii - coincidentale

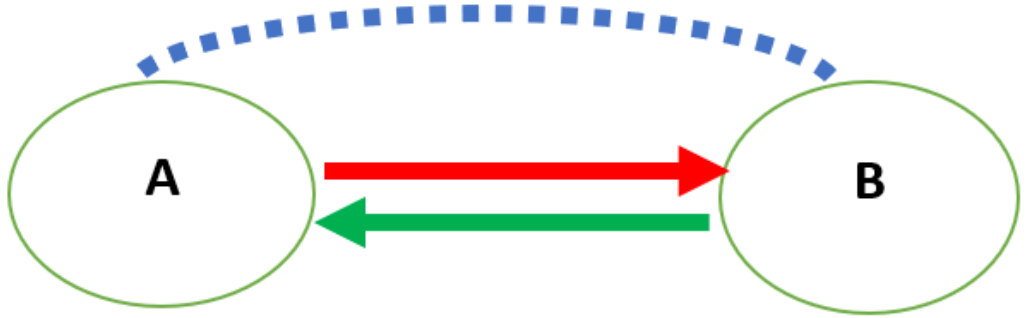



Tyler Vygen, Spurious correlations


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
©marketoonist.com

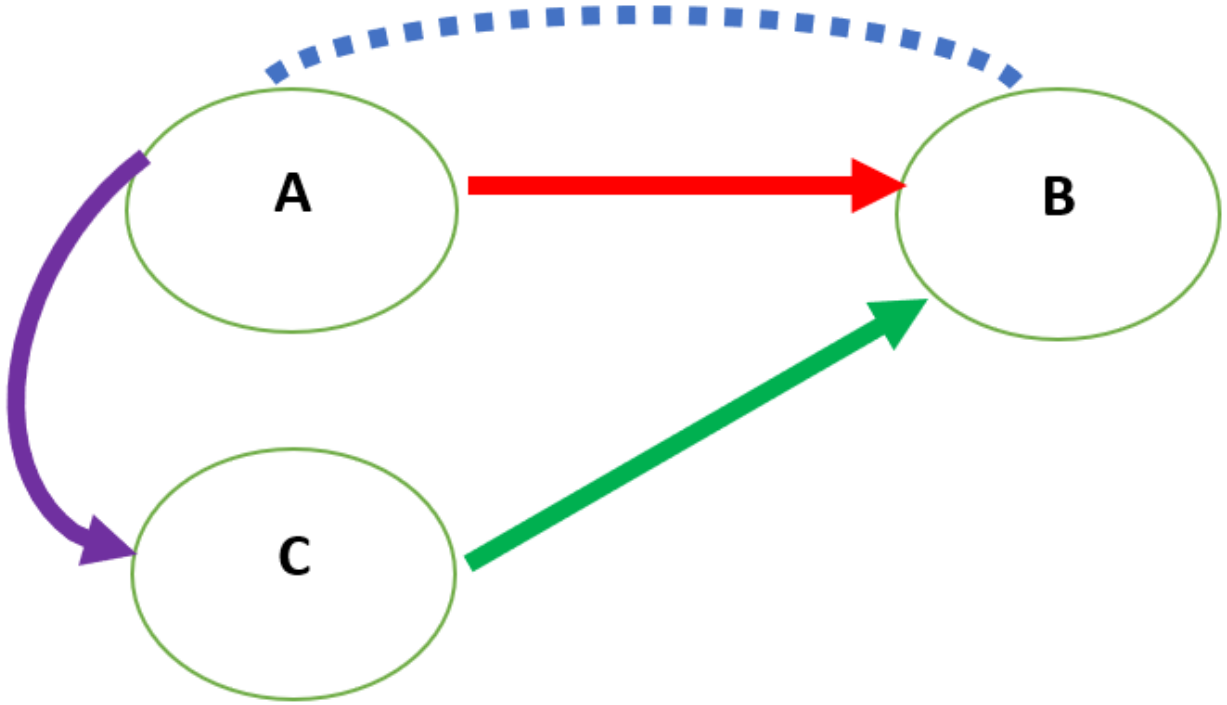
# 2. Sensul cauzării

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



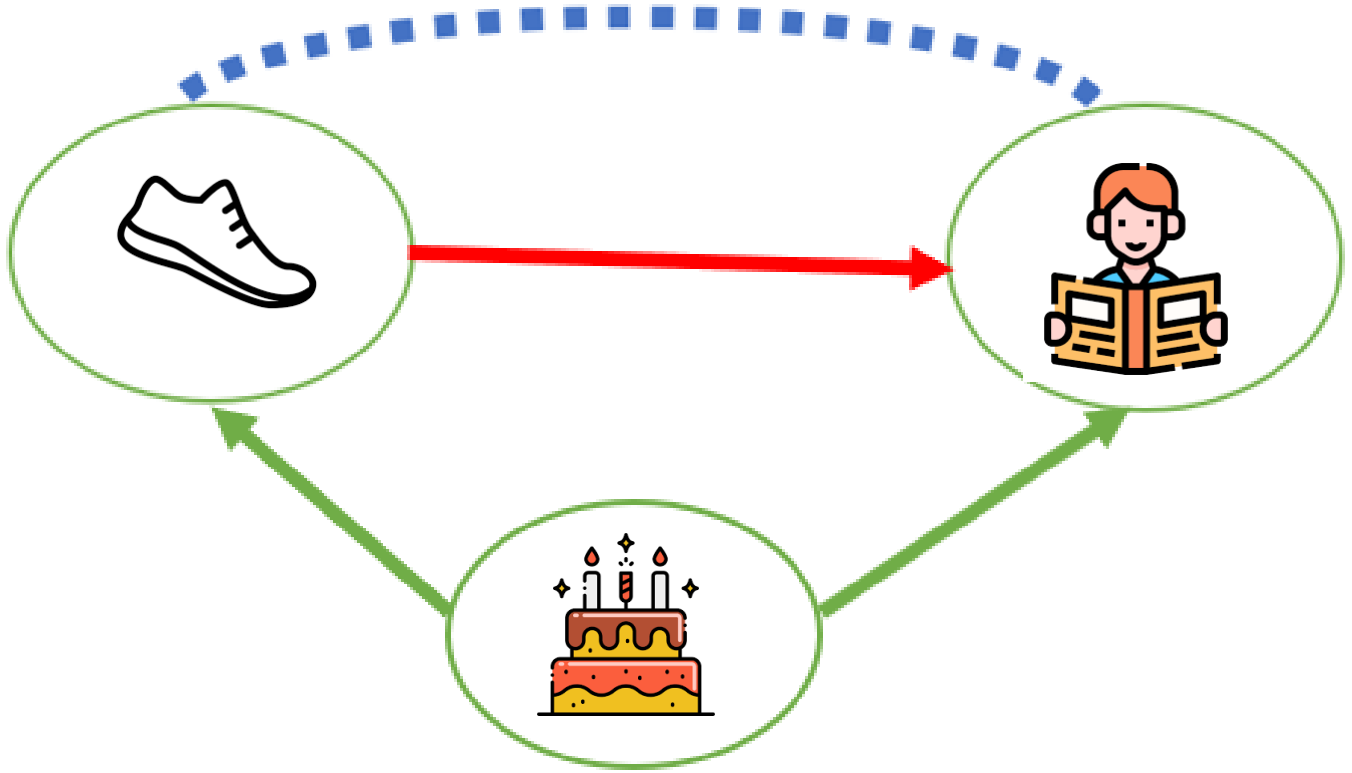
# 3. Corect din rațiuni false

- Sporește usturoiul purtat la gât imunitatea?
- Sunt femeile mai puțin capabile de muncă intelectuală?
- Înțelegerea procesului **A – C – B** schimbă interpretarea corelației dintre **A și B**



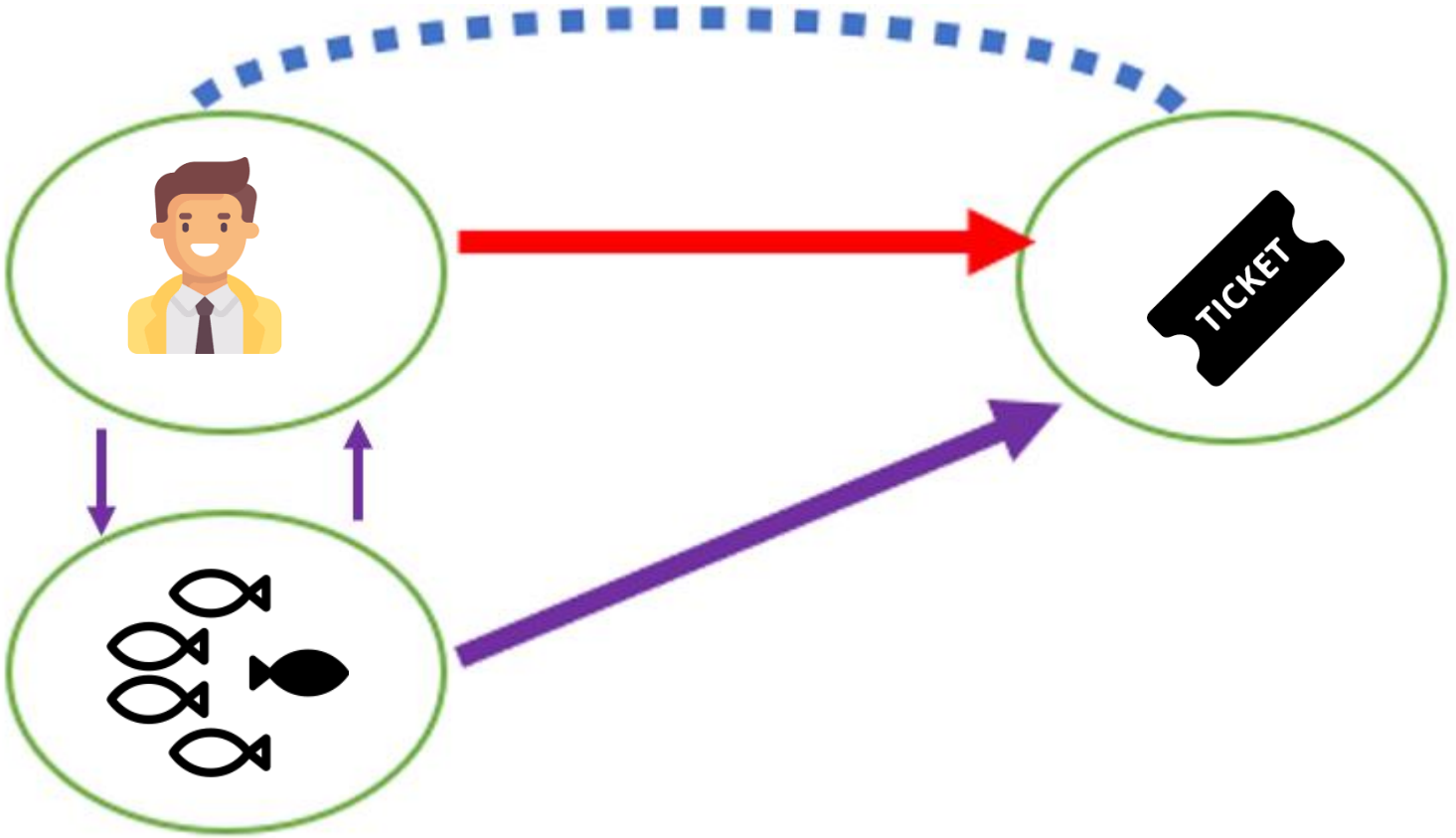
# 4. Cauzalitate iluzorie

- Pentru copii, mărimea pantofului corelează cu abilitatea de a citi
- În timp, vânzările de înghețată corelează cu rata criminalității violente
- În timp, rata pirateriei navale corelează cu încălzirea globală



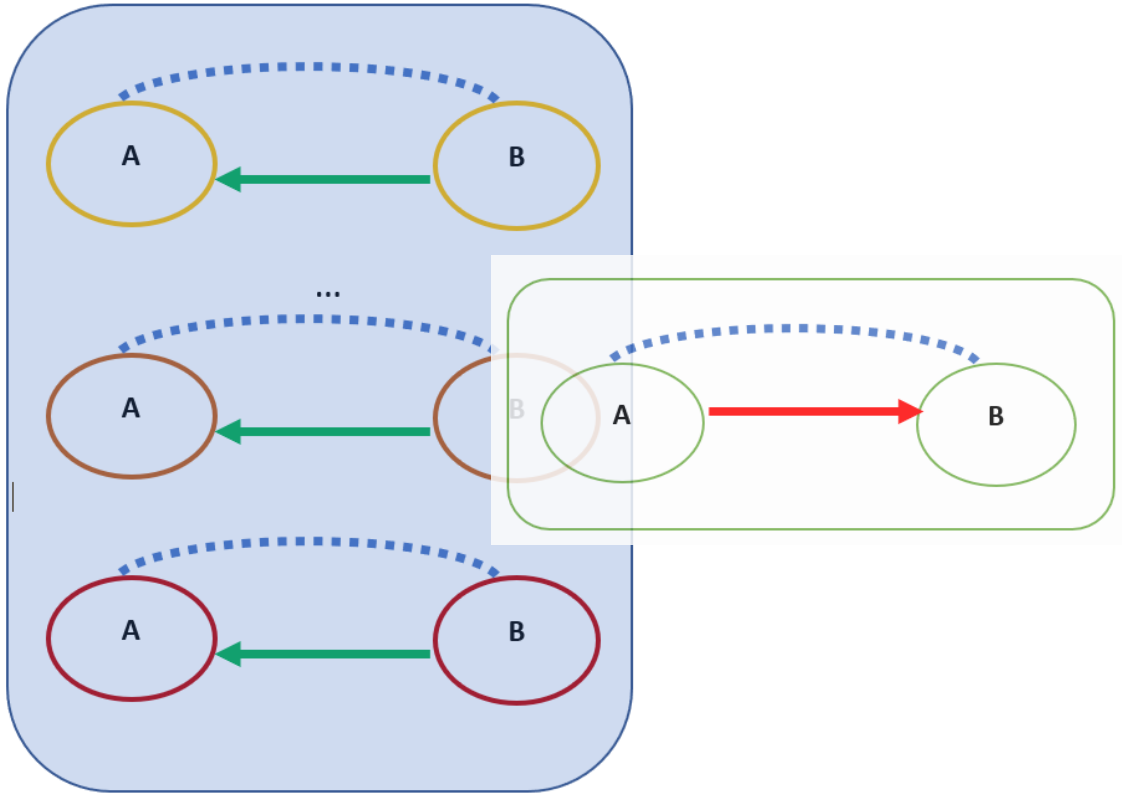
# 5. Eroarea sistematică de măsurare

- 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



# 6. Paradoxul lui Simpson sau al variabilei omise

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

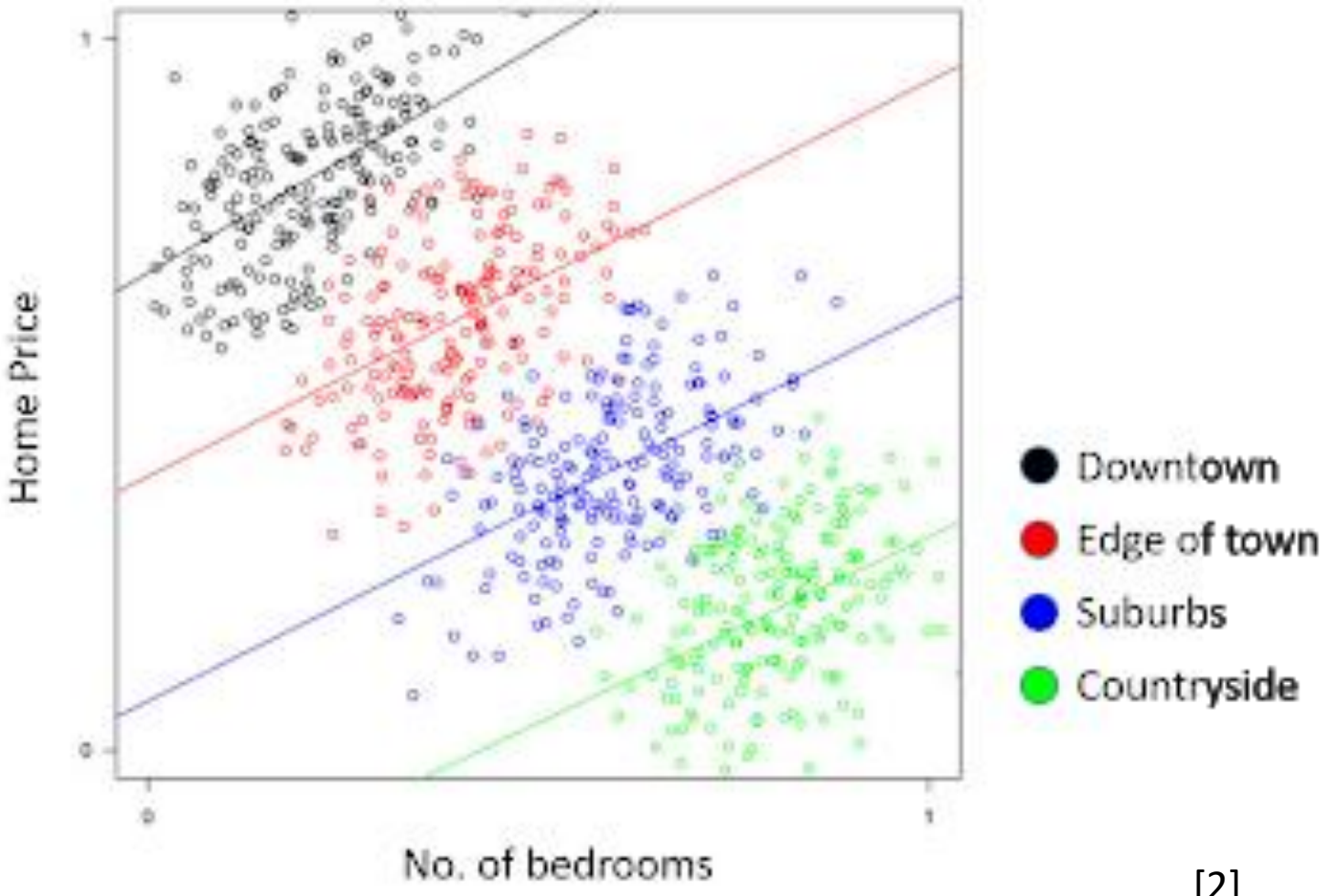


# Paradoxul lui Simpson

Modificarea unei relații atunci când ținem cont de o terță variabilă

(introducerea unei variabile de control)

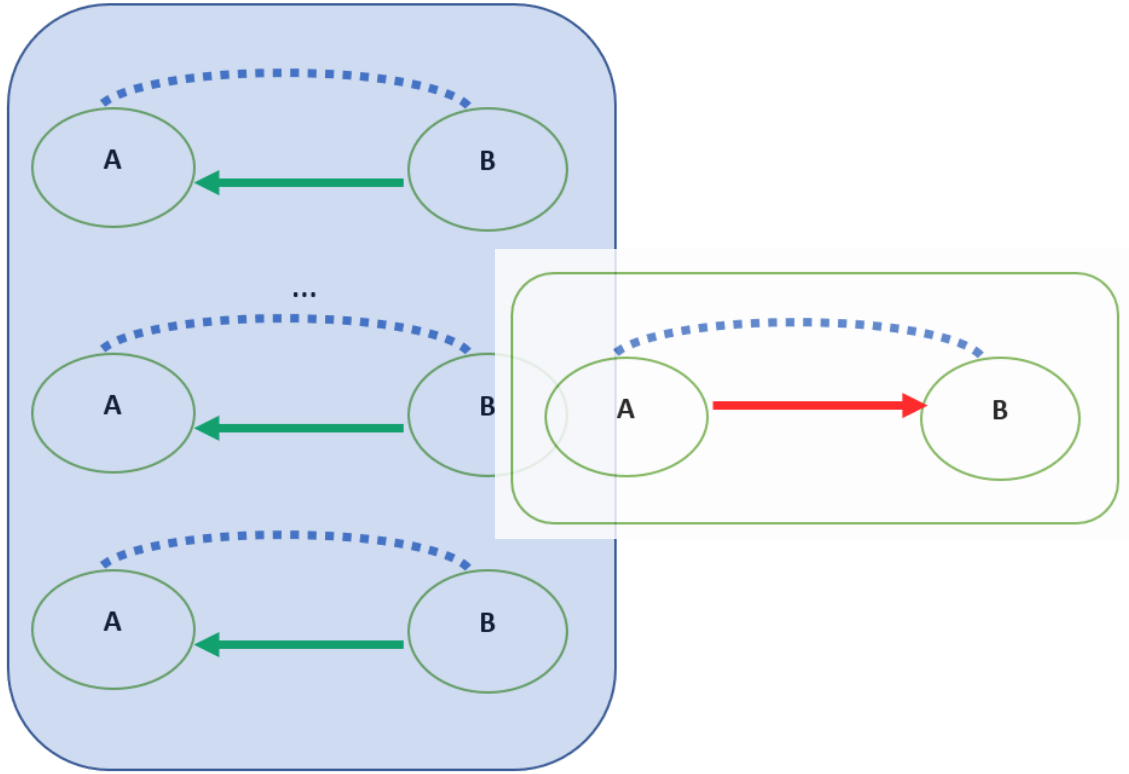
Eg: subpopulații definite de poziția în oraș (centru / periferie)




[2]

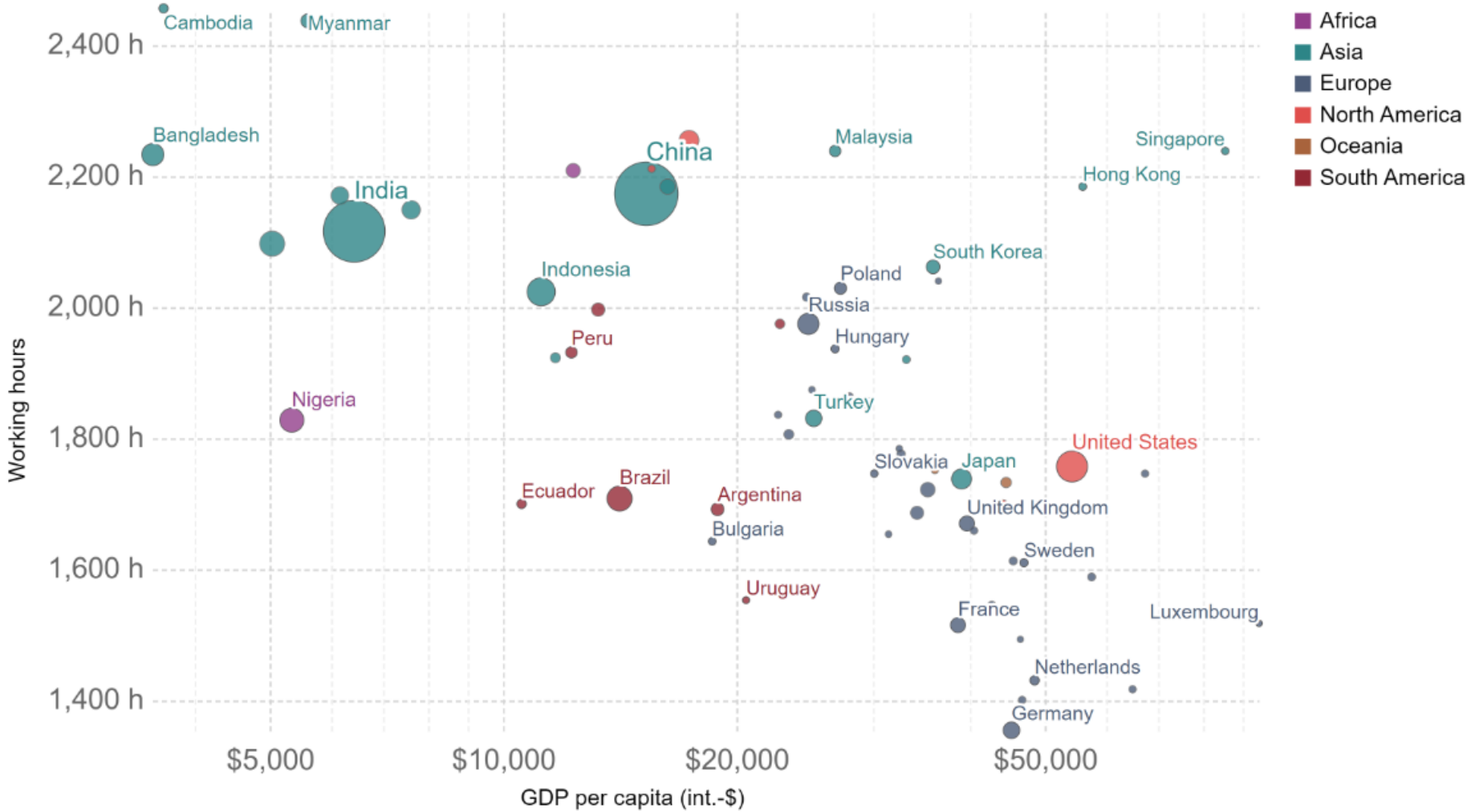
# 7. Eroarea ecologică (de agregare)

- O populație compusă din colectivități de indivizi (omogene)
- Relația cauzală se petrece la nivelul indivizilor
- Sensul relației cauzale la nivel individual este inversat prin agregare, care maschează o altă variabilă cauzală



# 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

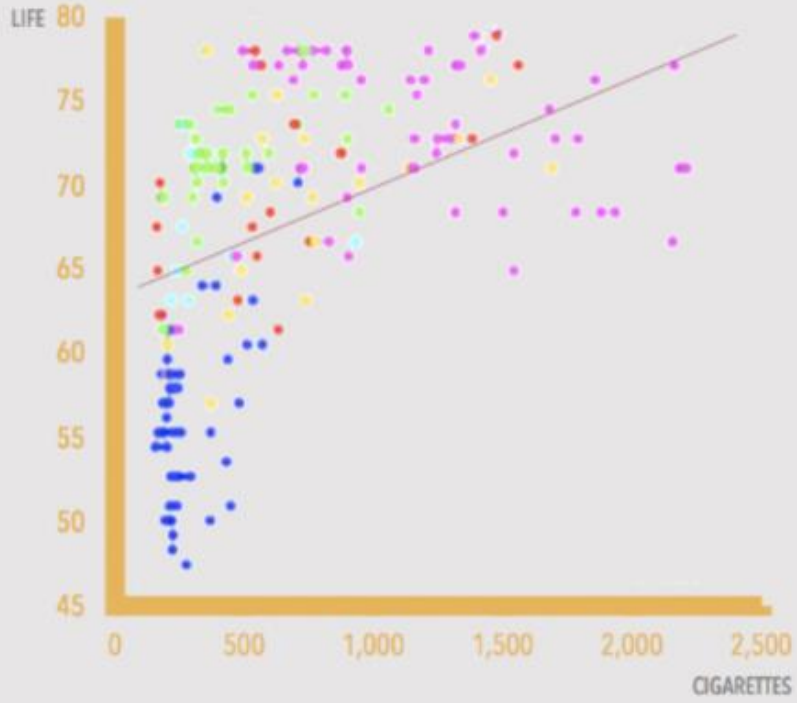
Sursă

# 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.

14

Krause, 2018

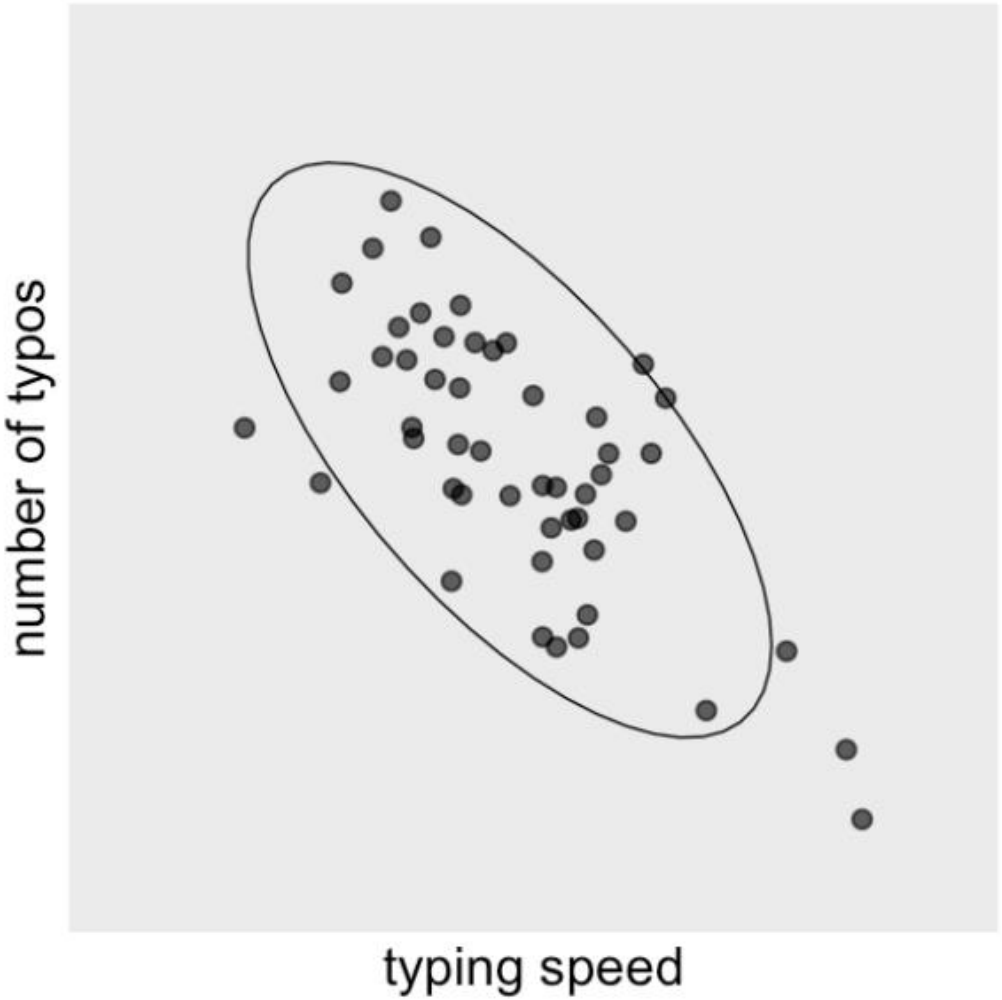
# Eroarea ecologică / de agregare

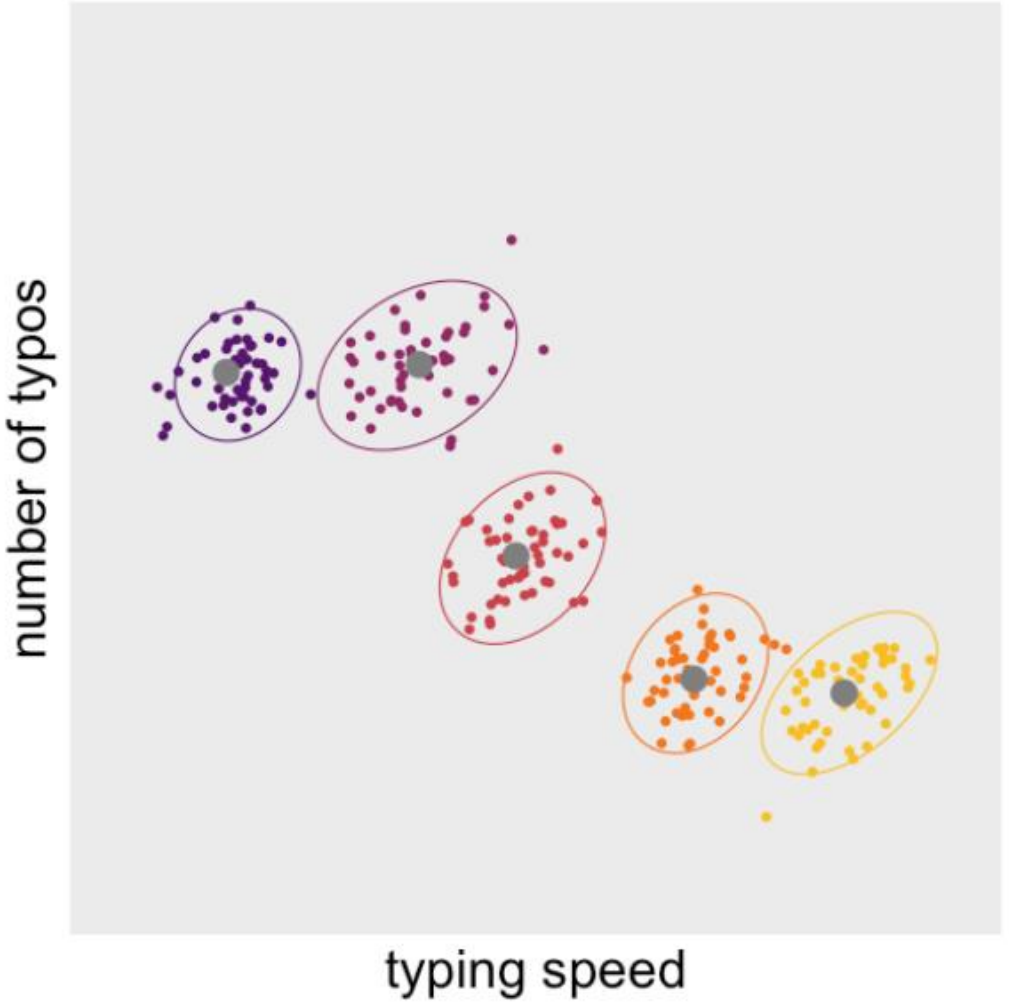
## Is smoking cigarettes good for your health?

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

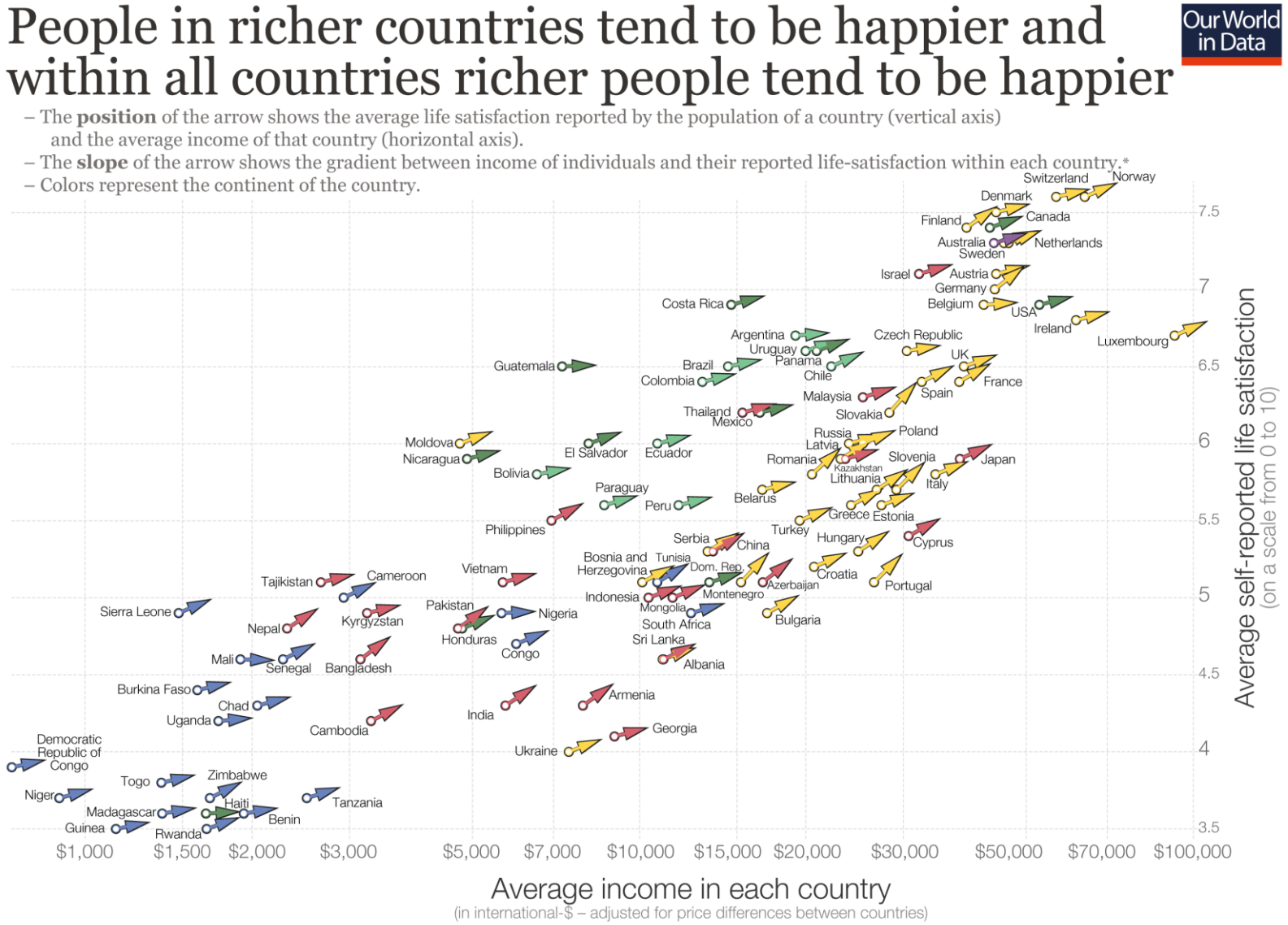

18

# Eroarea ecologică / de agregare



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

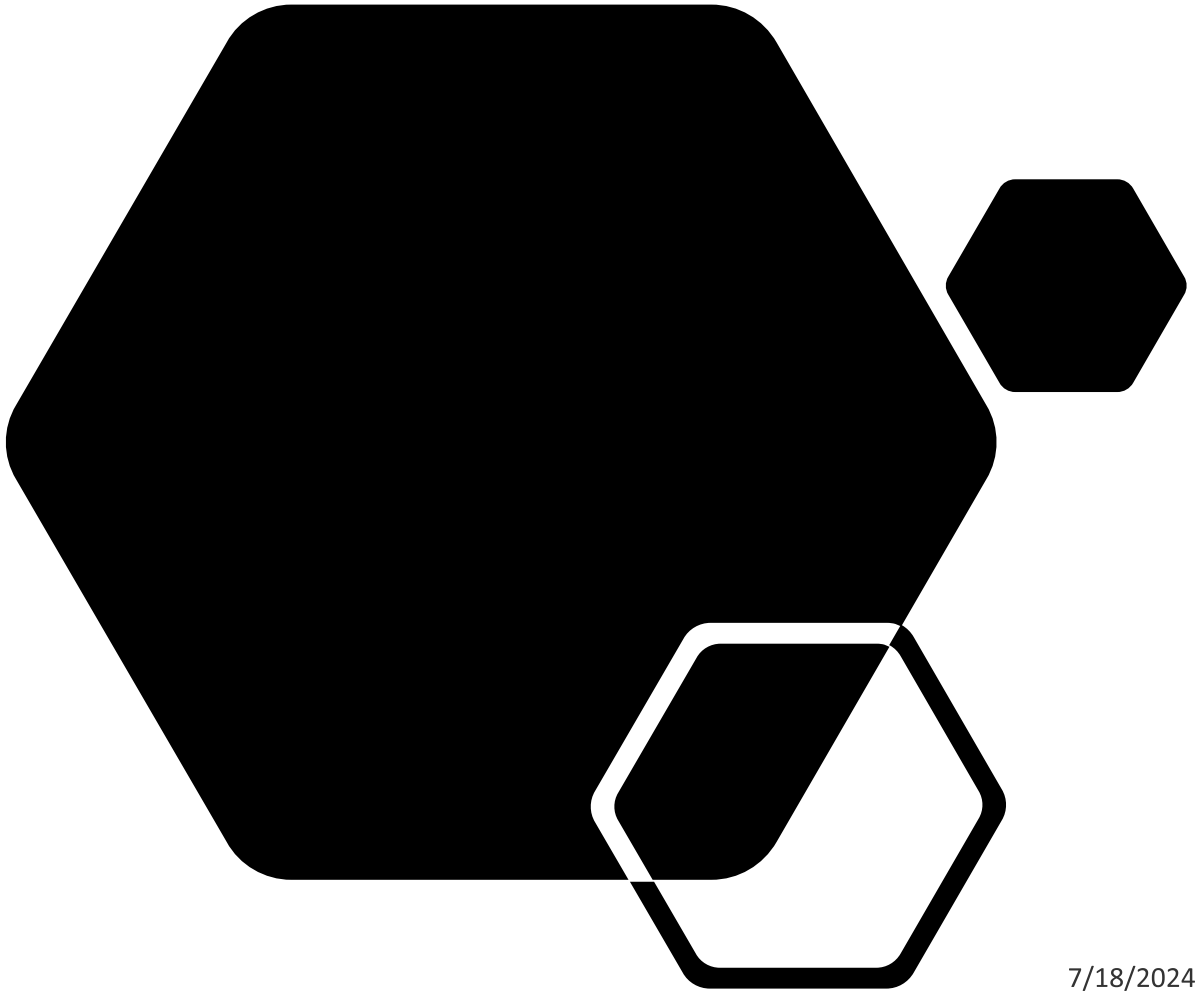



Sursă

# Erori de eșantionare

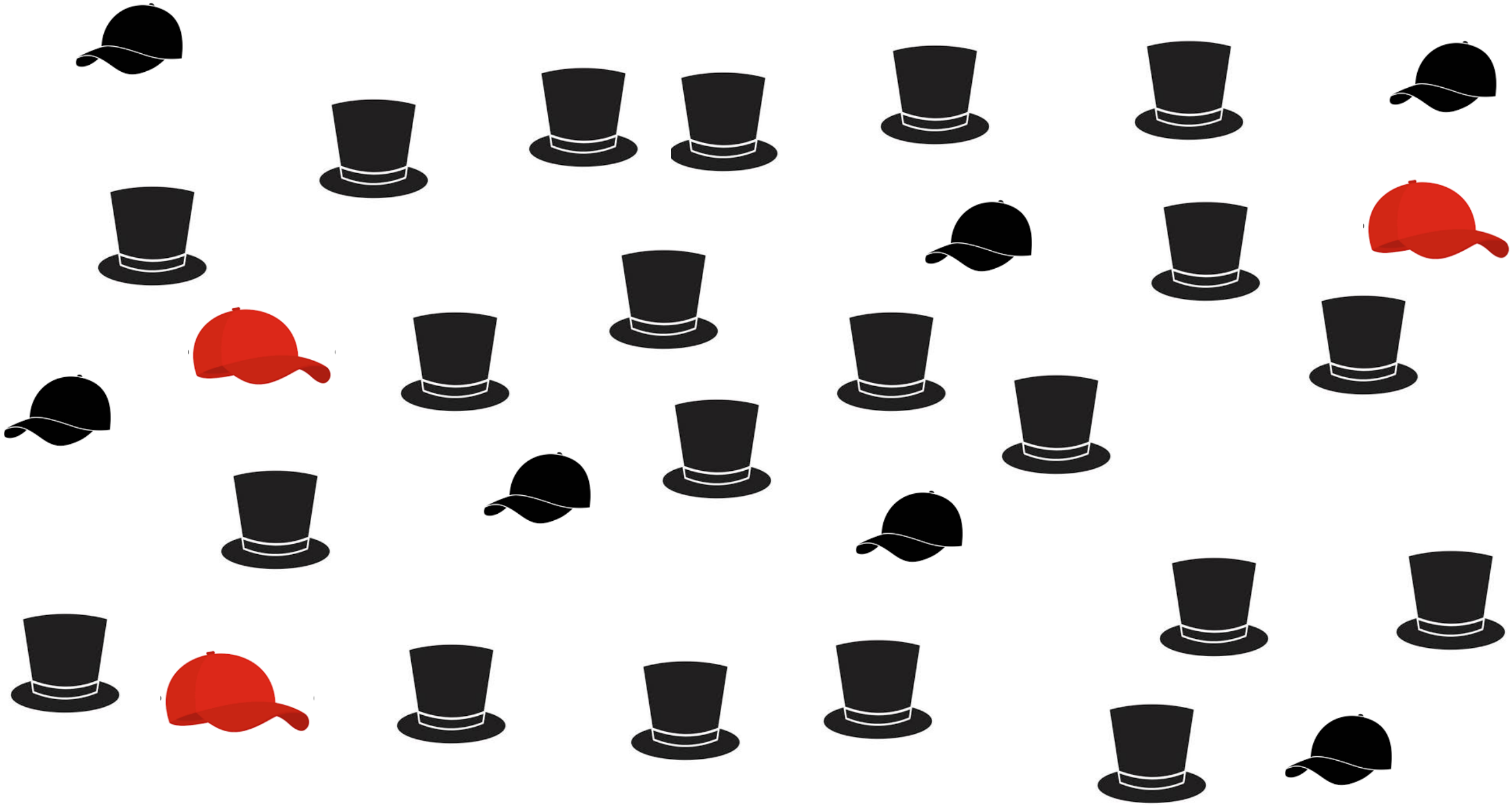
Eșantionare și generalizare

Probabilitatea de eroare (p)

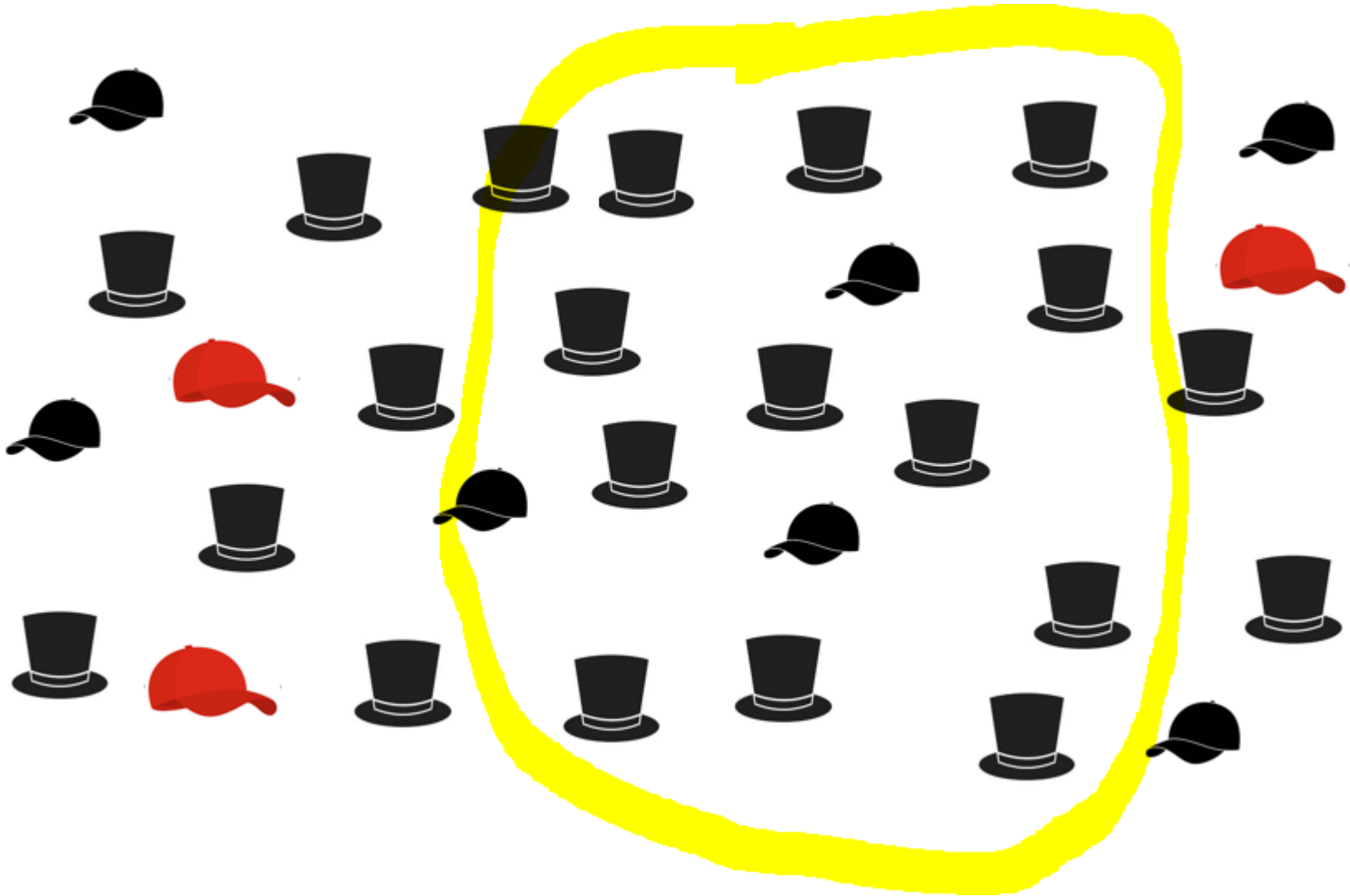


# Eșantion și populație

- Nu putem studia toți membrii populației
  - Studiem un eșantion
  - Generalizare (inducție) = tragem concluzii despre întreaga populație
- **Eșantionul probabilistic** permite **generalizarea proporțiilor, intensităților**
  - Estimările obținute din eșantion sunt valabile, cu o marjă de eroare, pentru populație
- **Eșantionul non-probabilistic** nu permite generalizarea estimărilor numerice
  - **Eșantion de disponibilitate**: comod, dar ne-reprezentativ
    - **Eșantion bulgăre de zăpadă** – pentru populații dificile
  - **Eșantion teoretic**: permite explorarea sistematică a variației, sau studiul în profunzime (eg. Eșantion de experți)
  - Eșantioanele non-probabilistice permit crearea de tipologii orientative, dar nu permit generalizări sau identificarea unor tendințe la nivelul populației



# Missing (NOT) at random



# 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)

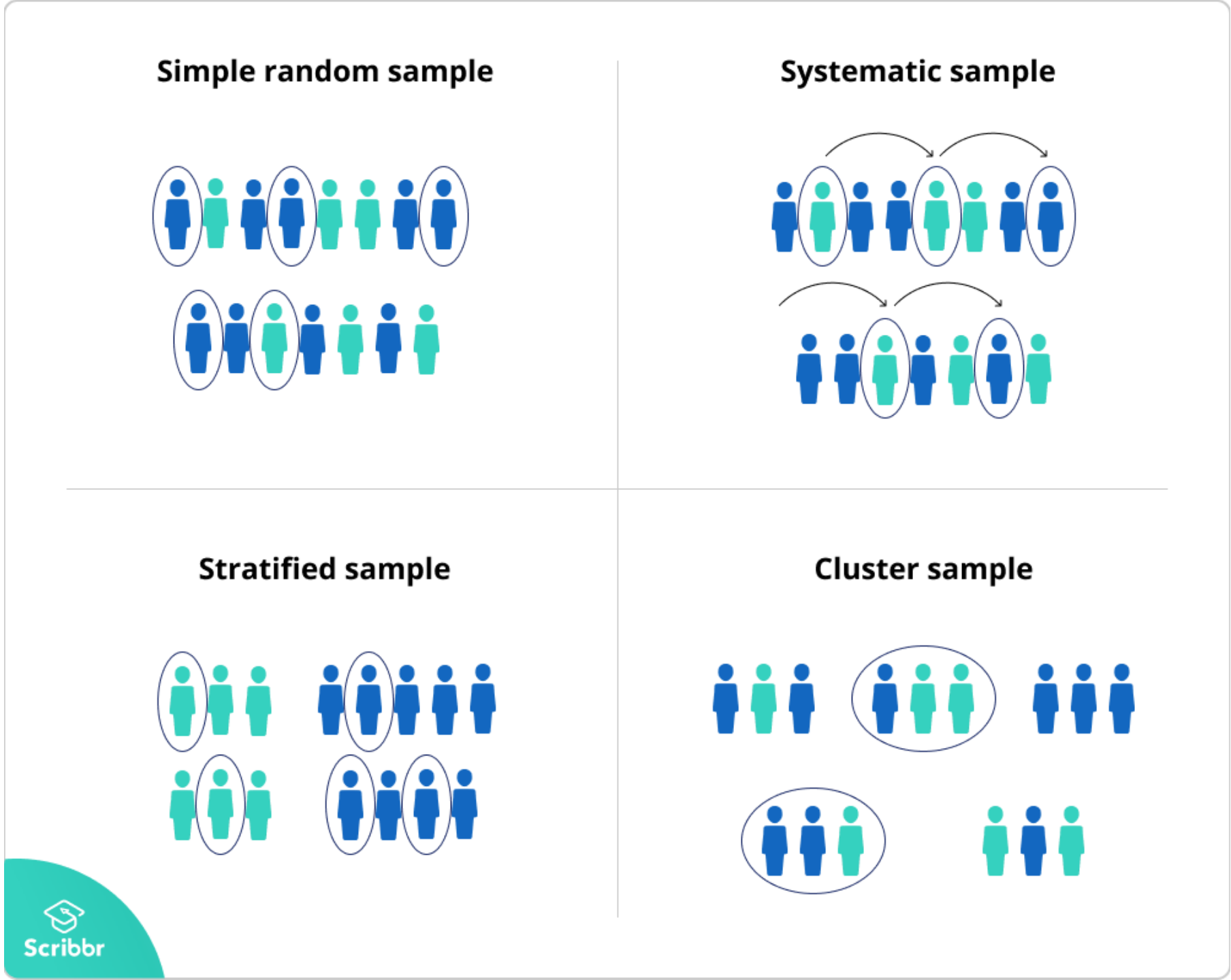
**One pollster's explanation for why the polls got it wrong**

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

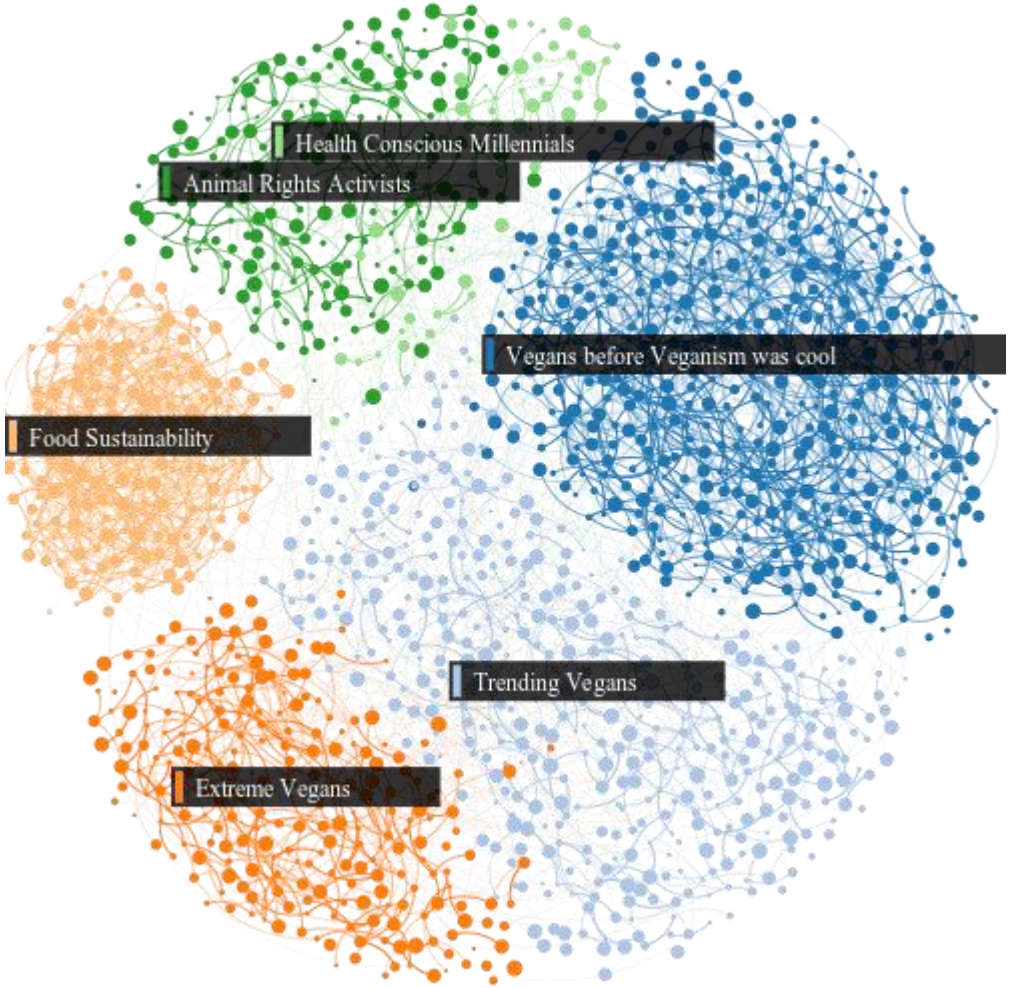
SHARE

# Eșantioane probabiliste



# 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!**
  - **Sau dacă experimentul nu a folosit randomizarea**

**C1_2. De obicei, cat de des mergeti la cumparaturi in hipermarketuri? * Mediu de rezidenta Crosstabulation**

% within Mediu de rezidenta

| | | Mediu de rezidenta | | Total |
|---|---|---|---|---|
| | | 0 Rural | 1 Urban | |
| C1_2. De obicei, cat de des mergeti la cumparaturi in hipermarketuri? | 1 Deloc | 38.7% | 13.4% | 24.2% |
| | 2 Mai rar | 22.5% | 17.6% | 19.7% |
| | 3 De cateva ori pe luna | 28.2% | 34.0% | 31.6% |
| | 4 De cateva ori pe saptamana | 10.2% | 30.0% | 21.5% |
| | 5 Zilnic | 0.4% | 4.9% | 3.0% |
| Total | | 100.0% | 100.0% | 100.0% |

Barometrul de consum cultural 2012

**C1_2. De obicei, cat de des mergeti la cumparaturi in hipermarketuri? * Singur sau cuplu Crosstabulation**

% within Singur sau cuplu

| | | Singur sau cuplu | | |
|---|---|---|---|---|
| | | .00 Stare civila - singur (necasatorit, divortat, separat, vaduv) | 1.00 Stare civila - in cuplu (casatorit, concubinaj) | Total |
| C1_2. De obicei, cat de des mergeti la cumparaturi in hipermarketuri? | 1 Deloc | 27.1% | 22.7% | 24.2% |
| | 2 Mai rar | 19.5% | 19.9% | 19.8% |
| | 3 De cateva ori pe luna | 32.0% | 31.2% | 31.5% |
| | 4 De cateva ori pe saptamana | 18.3% | 23.3% | 21.6% |
| | 5 Zilnic | 3.1% | 2.9% | 3.0% |
| Total | | 100.0% | 100.0% | 100.0% |

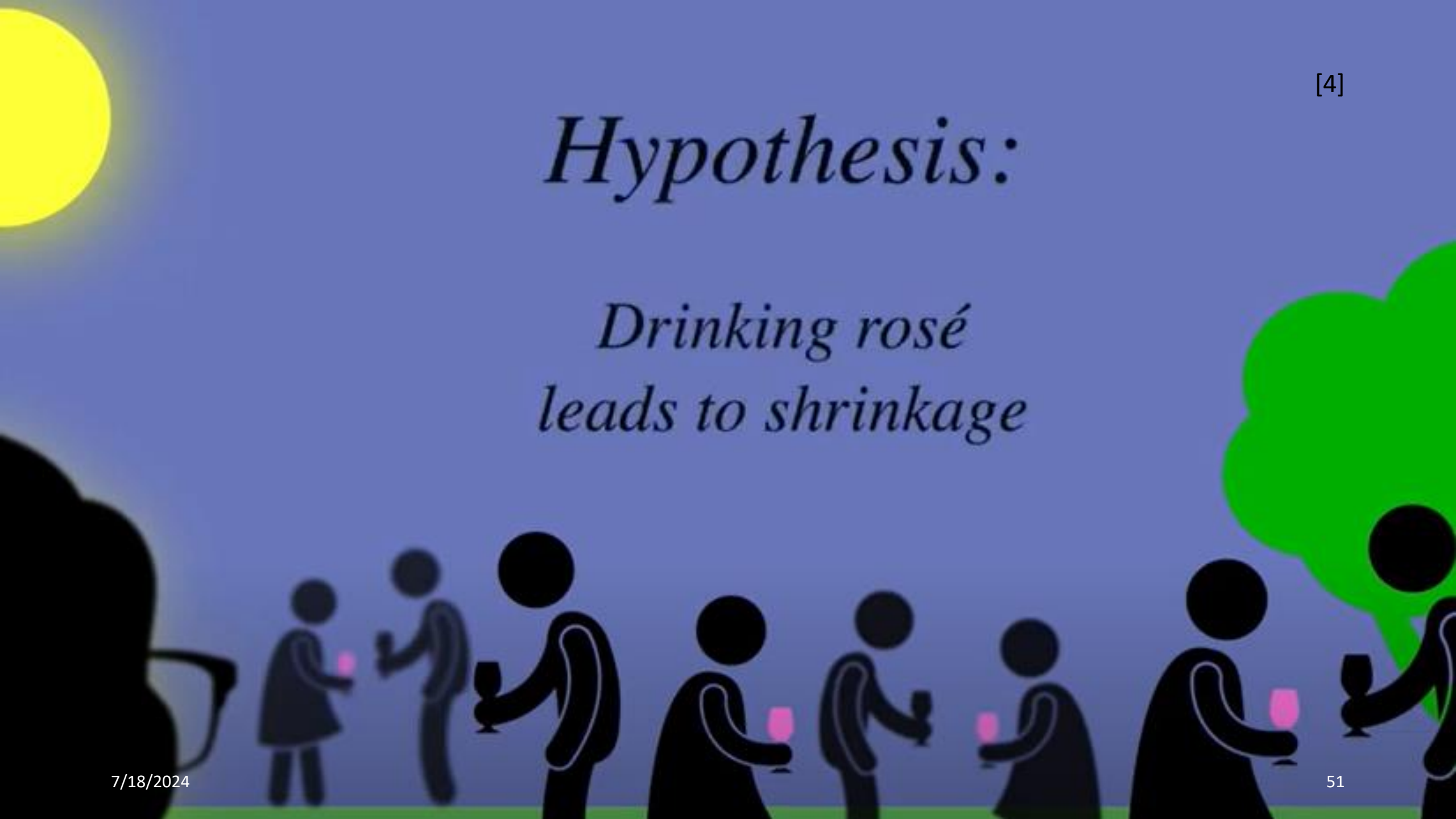
Barometrul de consum cultural 2012

**Correlations**

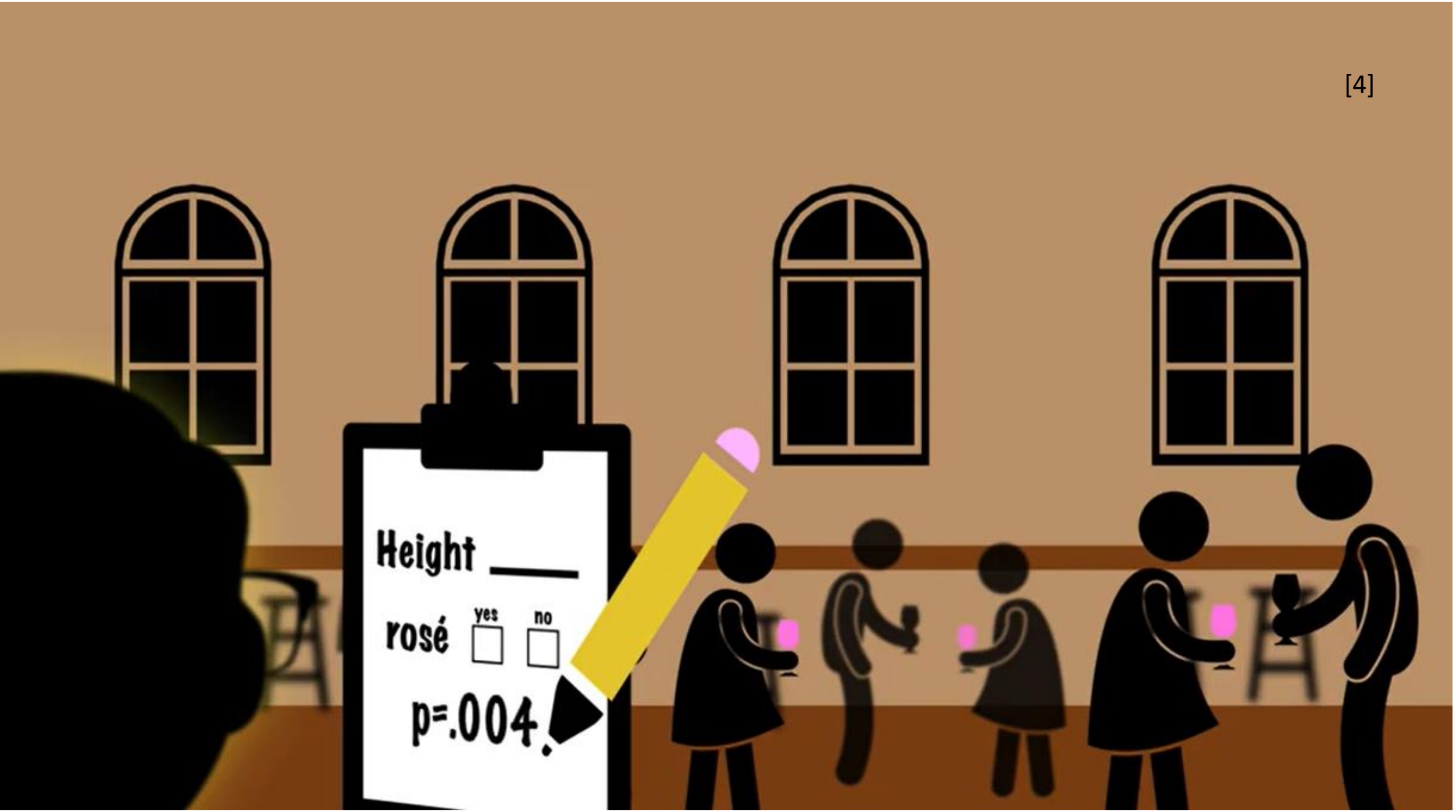
| | | C1_2. De obicei, cat de des mergeti la cumparaturi in hipermarketuri? | Mediu de rezidenta | Singur sau cuplu | I2. Varsta |
|---|---|---|---|---|---|
| C1_2. De obicei, cat de des mergeti la cumparaturi in hipermarketuri? | Pearson Correlation | 1 | .360** | .053** | -.298** |
| | Sig. (2-tailed) | | .000 | .010 | .000 |
| | N | 2363 | 2363 | 2357 | 2363 |
| Mediu de rezidenta | Pearson Correlation | .360** | 1 | -.036 | -.076** |
| | Sig. (2-tailed) | .000 | | .082 | .000 |
| | N | 2363 | 2371 | 2365 | 2371 |
| Singur sau cuplu | Pearson Correlation | .053** | -.036 | 1 | .007 |
| | Sig. (2-tailed) | .010 | .082 | | .719 |
| | N | 2357 | 2365 | 2365 | 2365 |
| I2. Varsta | Pearson Correlation | -.298** | -.076** | .007 | 1 |
| | Sig. (2-tailed) | .000 | .000 | .719 | |
| | N | 2363 | 2371 | 2365 | 2371 |

**. Correlation is significant at the 0.01 level (2-tailed).

[4]

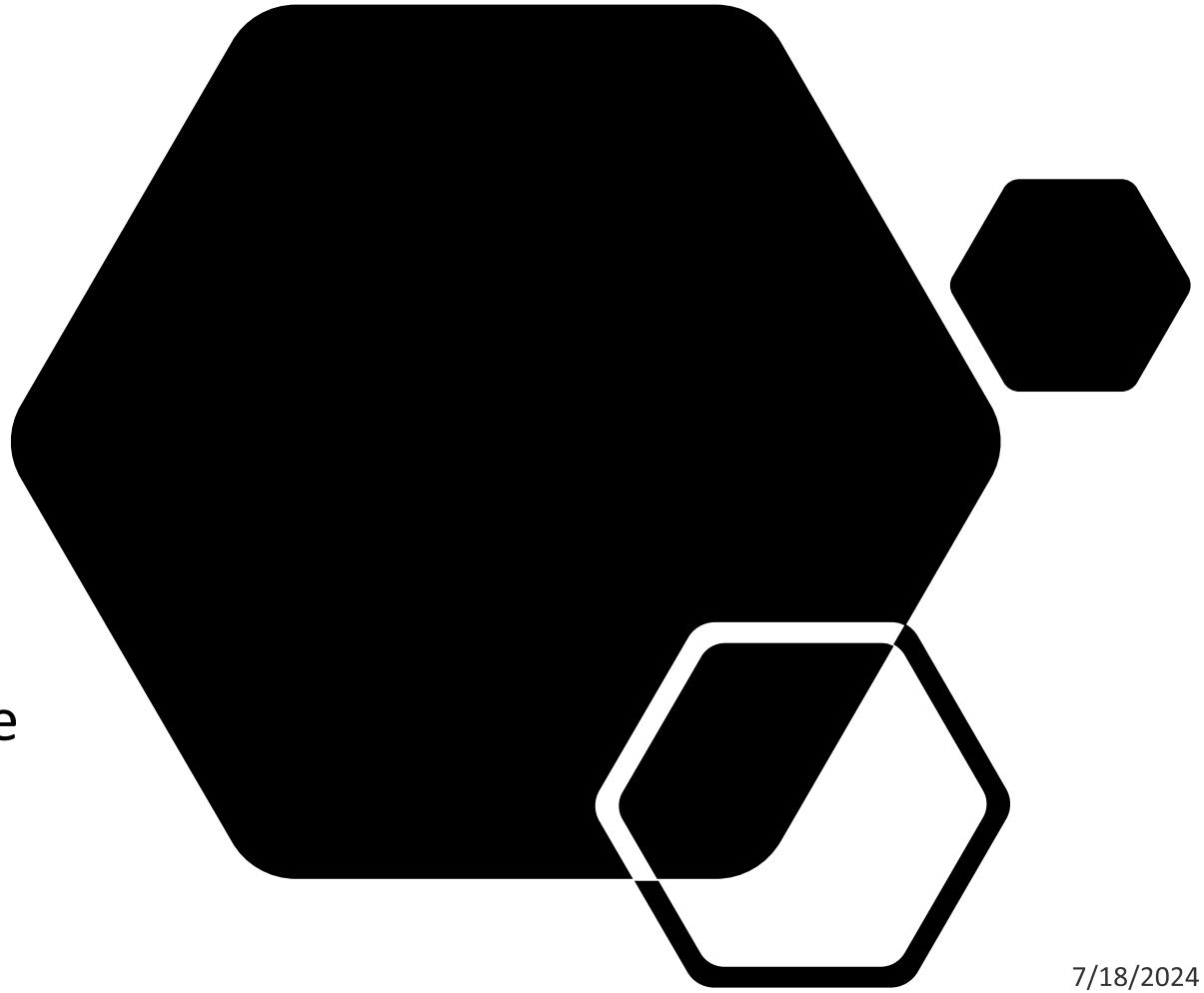


[4]



# Erori de modelare – studiu de caz

F. Messerli, Consumul de ciocolată și premiile Nobel



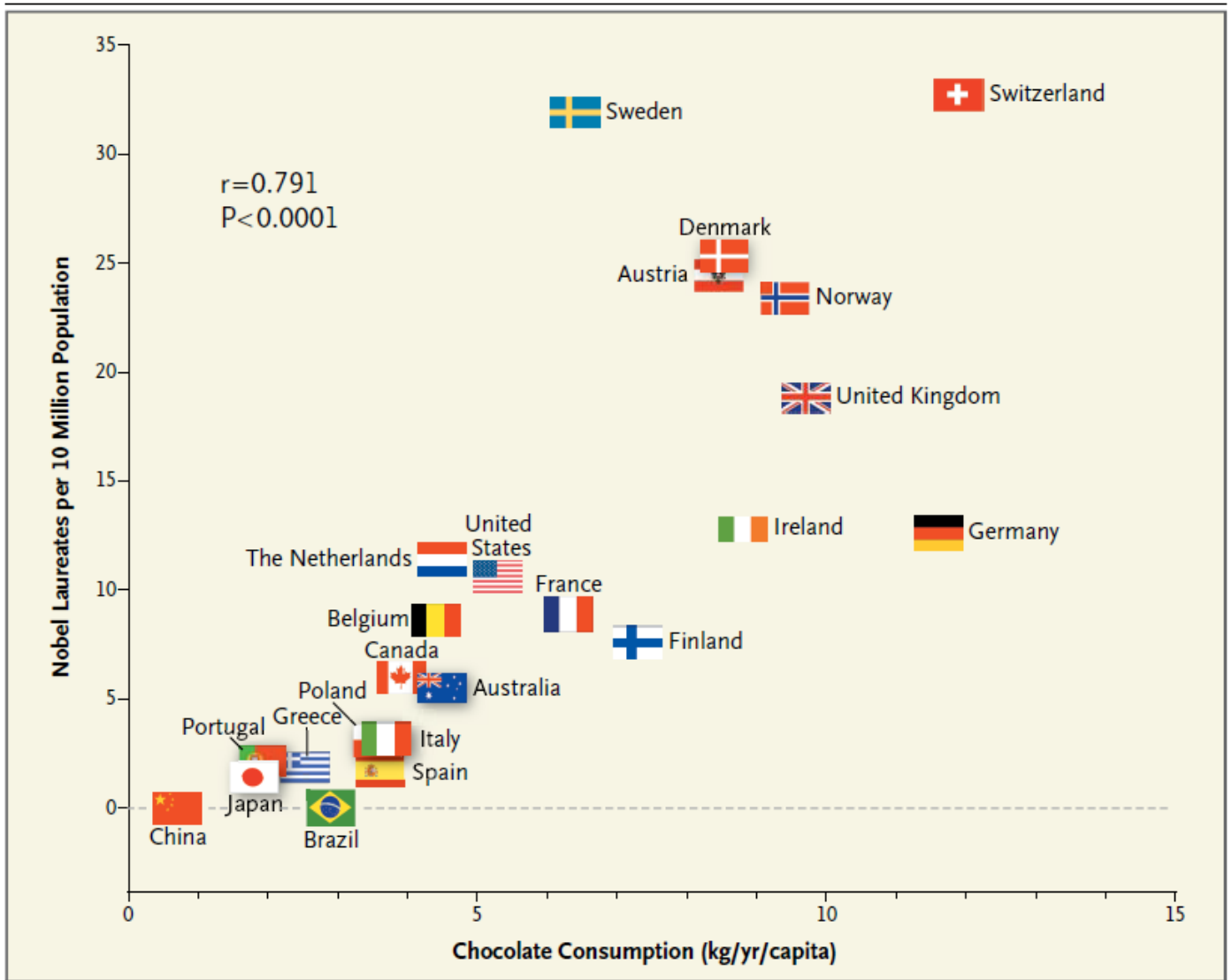
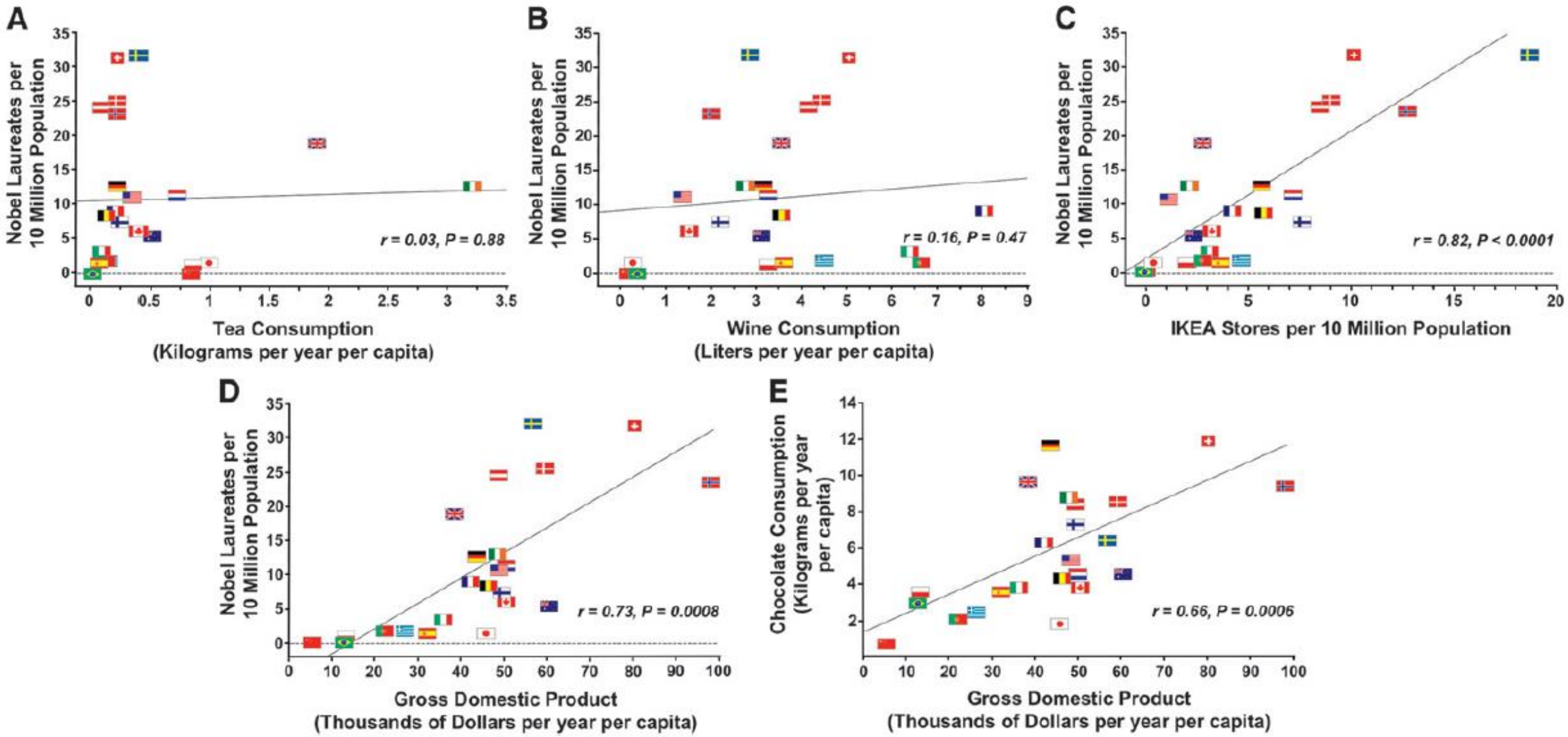
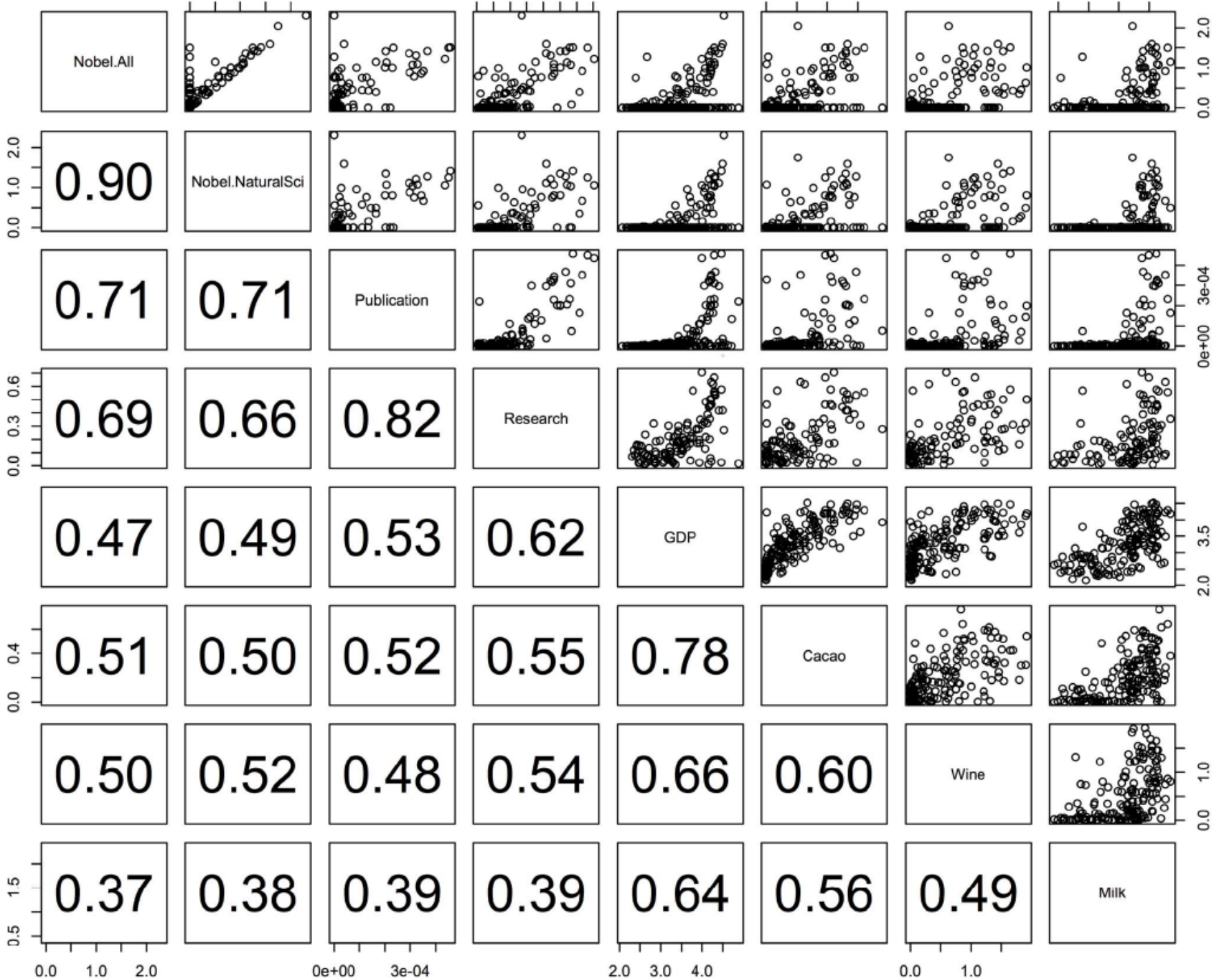
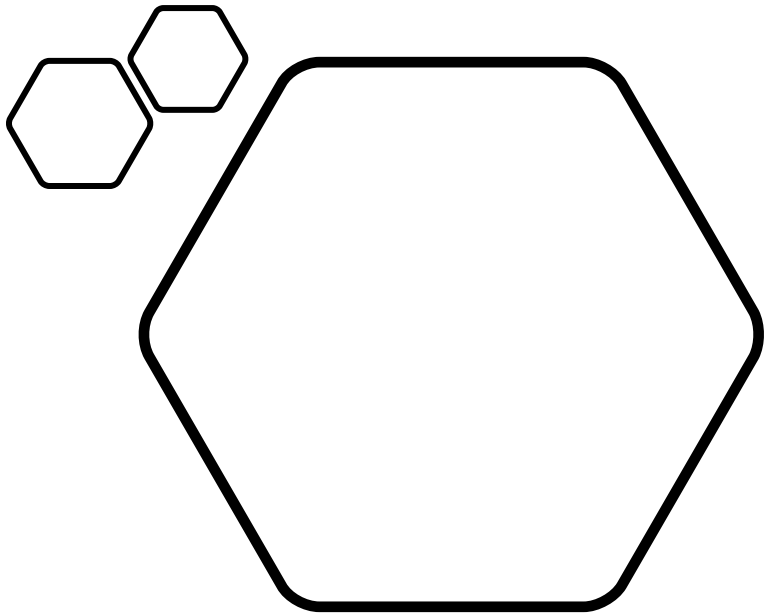



**Figure 1.** **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**


A
Nobel Laureates per 10 Million Population
Tea Consumption (Kilograms per year per capita)
r = 0.03, P = 0.88
B
Nobel Laureates per 10 Million Population
Wine Consumption (Liters per year per capita)
r = 0.16, P = 0.47
C
Nobel Laureates per 10 Million Population
IKEA Stores per 10 Million Population
r = 0.82, P < 0.0001
D
Nobel Laureates per 10 Million Population
Gross Domestic Product (Thousands of Dollars per year per capita)
r = 0.73, P = 0.0008
E
Chocolate Consumption (Kilograms per year per capita)
Gross Domestic Product (Thousands of Dollars per year per capita)
r = 0.66, P = 0.0006


Nobel.All
Nobel.NaturalSci
0.90
Publication
0.71
0.71
Research
0.69
0.66
0.82
GDP
0.47
0.49
0.53
0.62
Cacao
0.51
0.50
0.52
0.55
0.78
Wine
0.50
0.52
0.48
0.54
0.66
0.60
Milk
0.37
0.38
0.39
0.39
0.64
0.56
0.49

# Concluzii



- Corelația este un **posibil simptom** al cauzalității – nu o garanție
  - 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
  - Cauzalitate iluzorie
  - Erorile sistematice de măsurare
  - Paradoxul lui Simpson
  - Eroarea ecologică
- Erorile pot proveni nu doar din model, ci și din **eșantionare**
- **Semnificația statistică** – estimarea generalizabilității pentru eșantioane probabilistice

# 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.