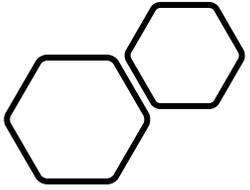


Analiza datelor. Măsurare și erori

Proiectul învățământului
superior din Moldova –
DATASEA

Răzvan Rughiniș
razvan.rughinis@upb.ro





Concepte principale

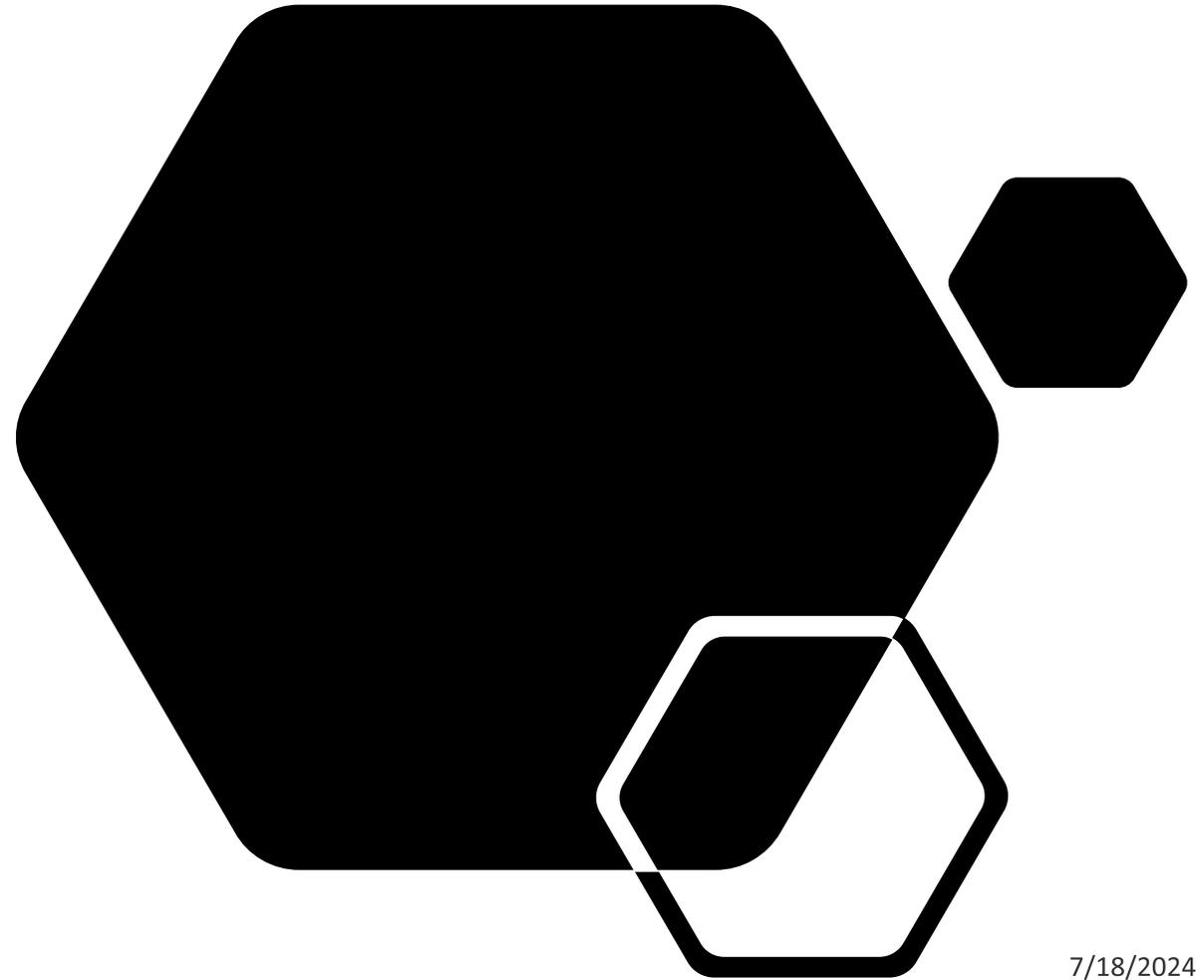
Măsurare și colectare

Distribuții

Modele de măsurare

Erori de măsurare

Măsurarea și colectarea datelor



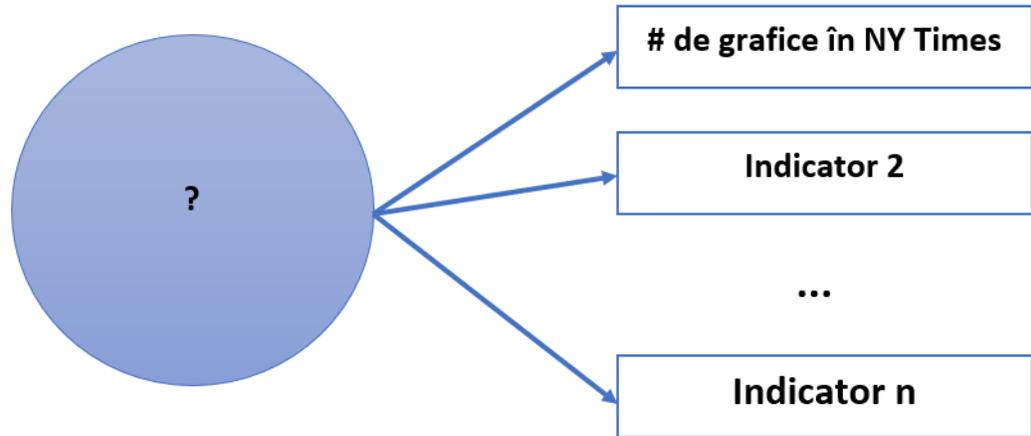
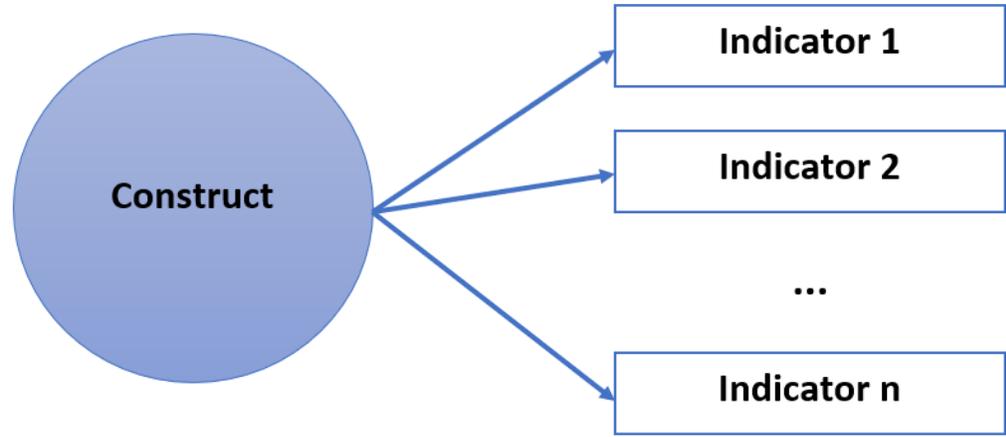
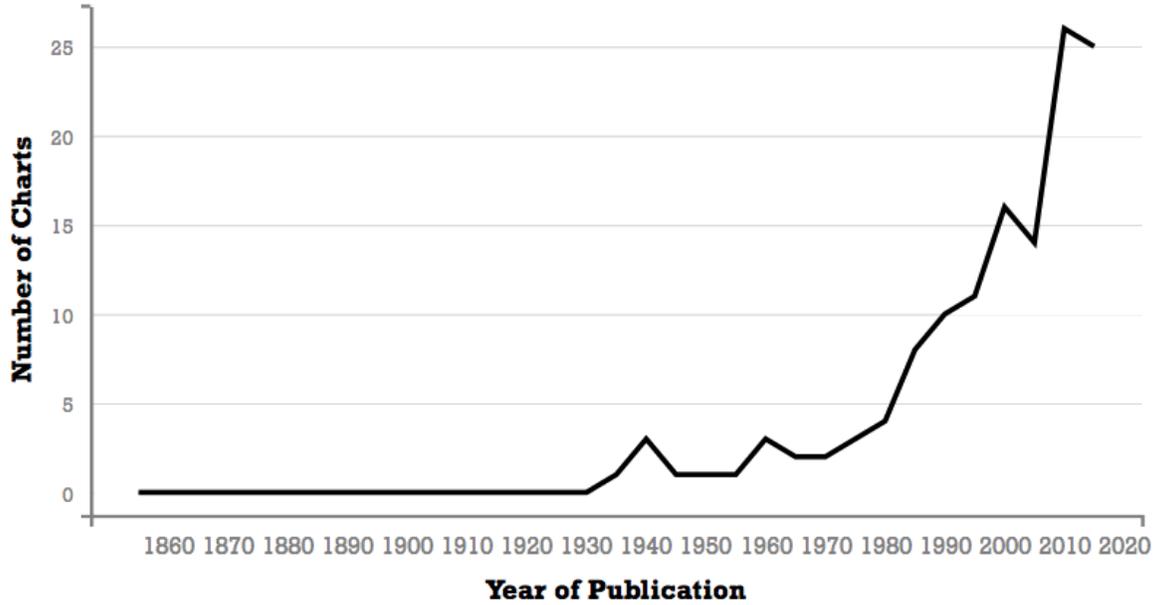
7/18/2024

Măsurare vs. colectarea datelor

- Măsurarea: transformarea **fenomenelor** în **variabile**
 - Înălțimea → un număr de cm
 - Greutatea → un număr de kg
 - Temperatura → un număr de grade
- Colectarea datelor: organizarea & stocarea datelor din:
 - Măsurători proprii
 - Măsurători din surse diferite
- Heterogenitatea datelor: fiecare sursă poate avea propria metodologie
 - Armonizarea datelor din surse diferite

The Rise of the Chart in the Newspaper

The # of Charts in a September Issue of the New York Times (Checked Every Five Years)

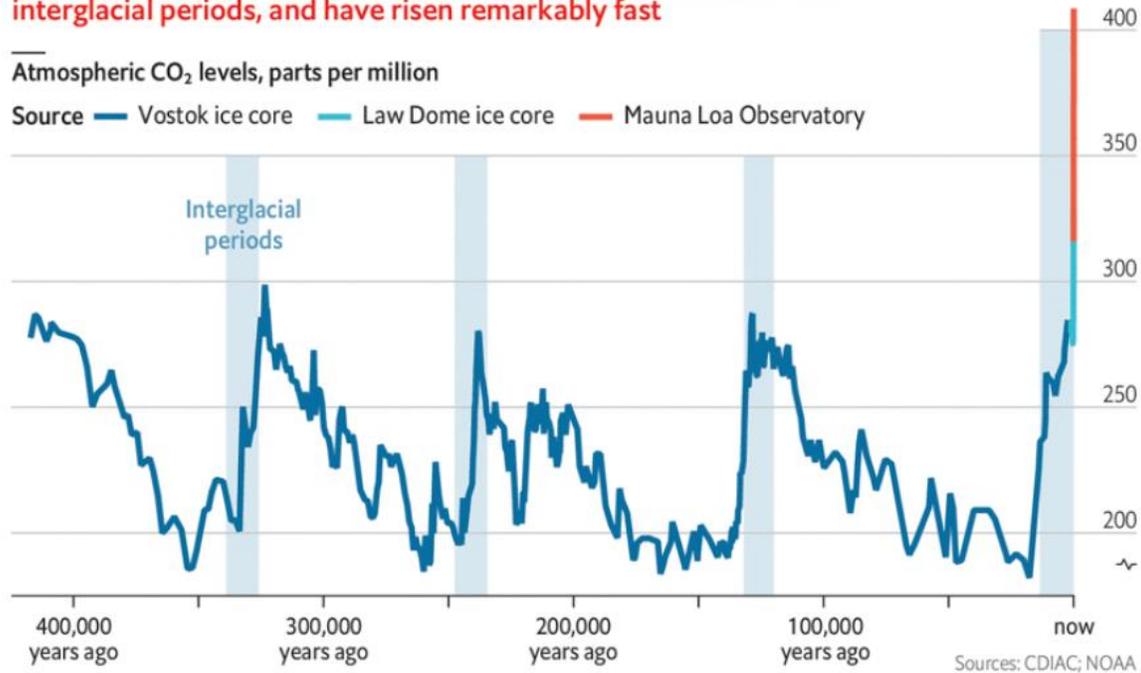


Măsurarea trecutului prin prezent

Like an ice age, in reverse; CO₂ levels are far higher than previous interglacial periods, and have risen remarkably fast

Atmospheric CO₂ levels, parts per million

Source — Vostok ice core — Law Dome ice core — Mauna Loa Observatory

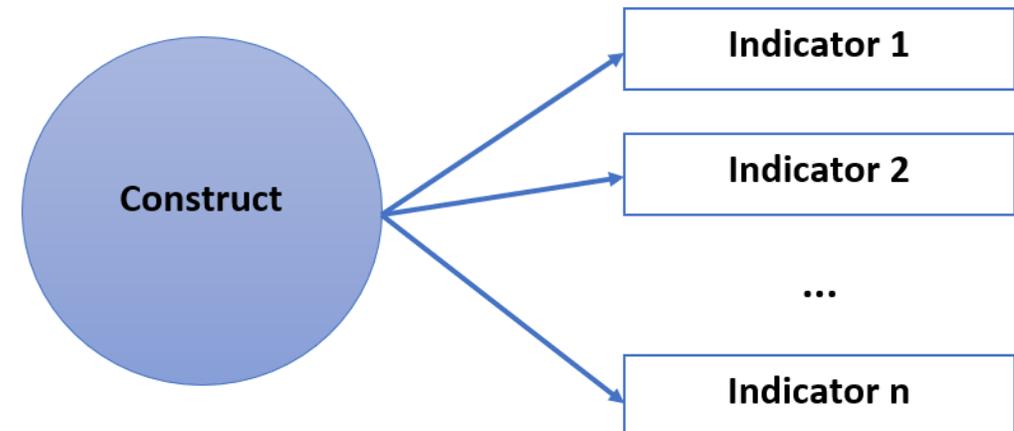


Source:
The Economist

Măsurarea CO₂ cu mii de ani în urmă prin CO₂ actual în ghețari

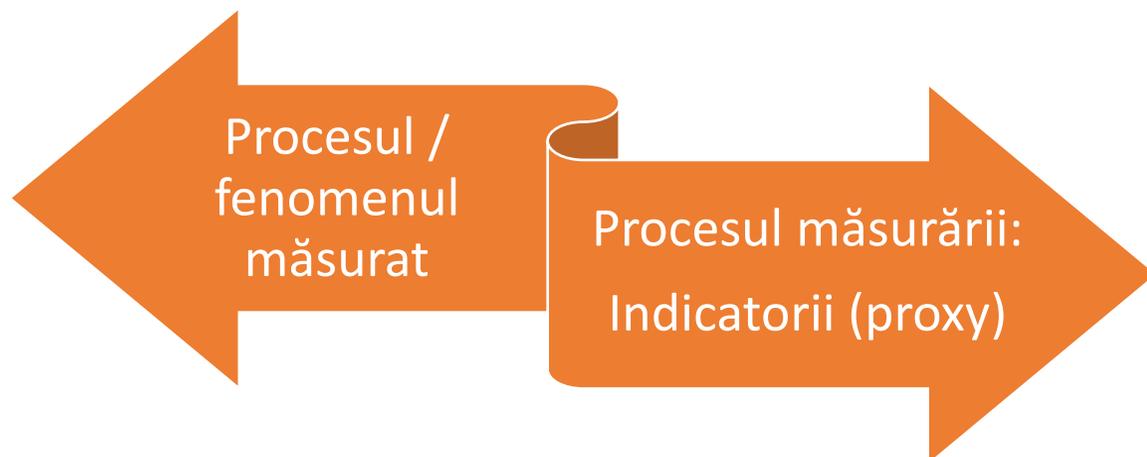
- CO₂ în trecut = constructul
- CO₂ în prezent, în ghețari = indicatorul

+ Armonizarea și agregarea datelor din surse diferite



Două competențe centrale în DS

Să facem diferența între:

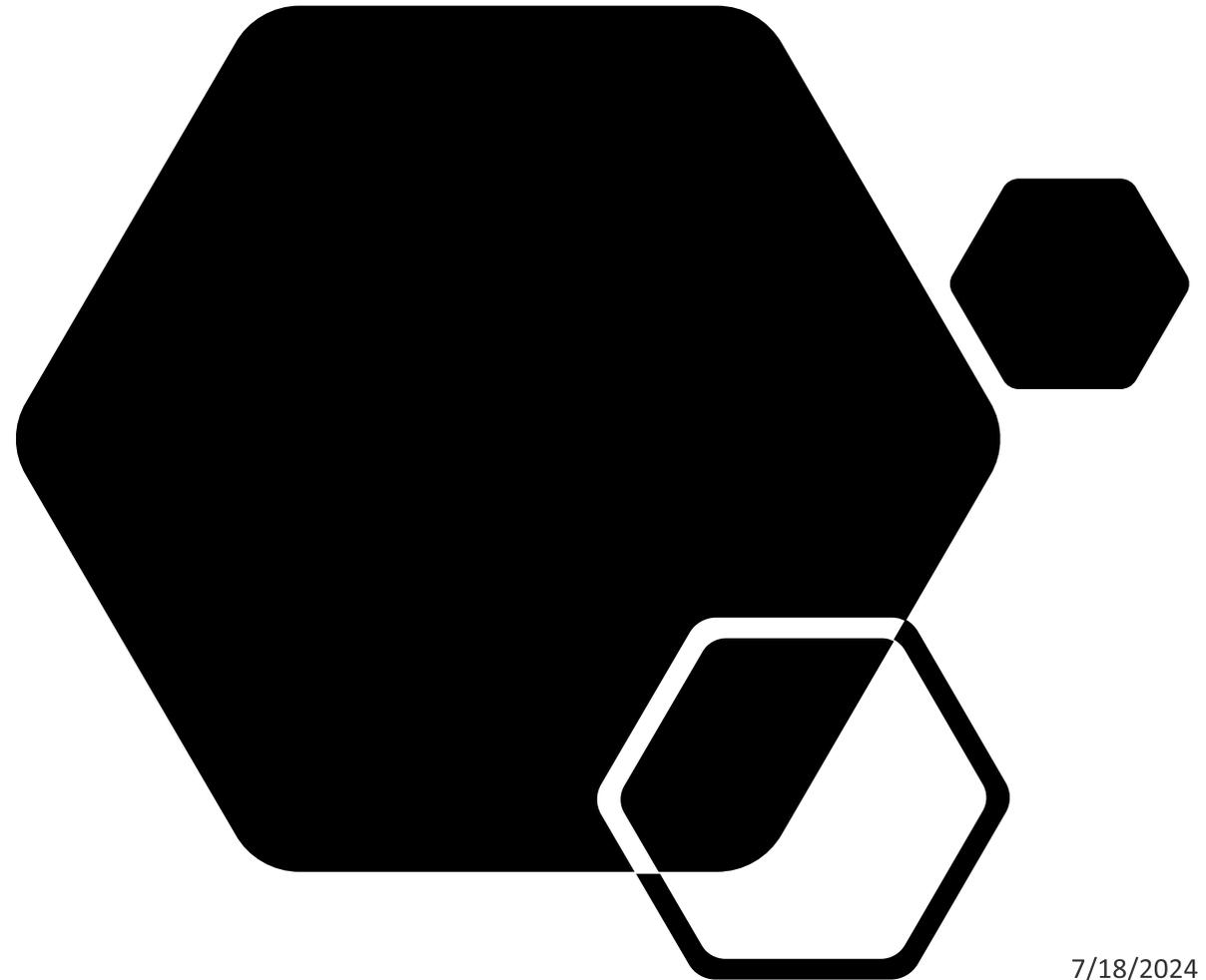


Să ținem cont mereu de:



Distribuții

Niveluri de măsurare

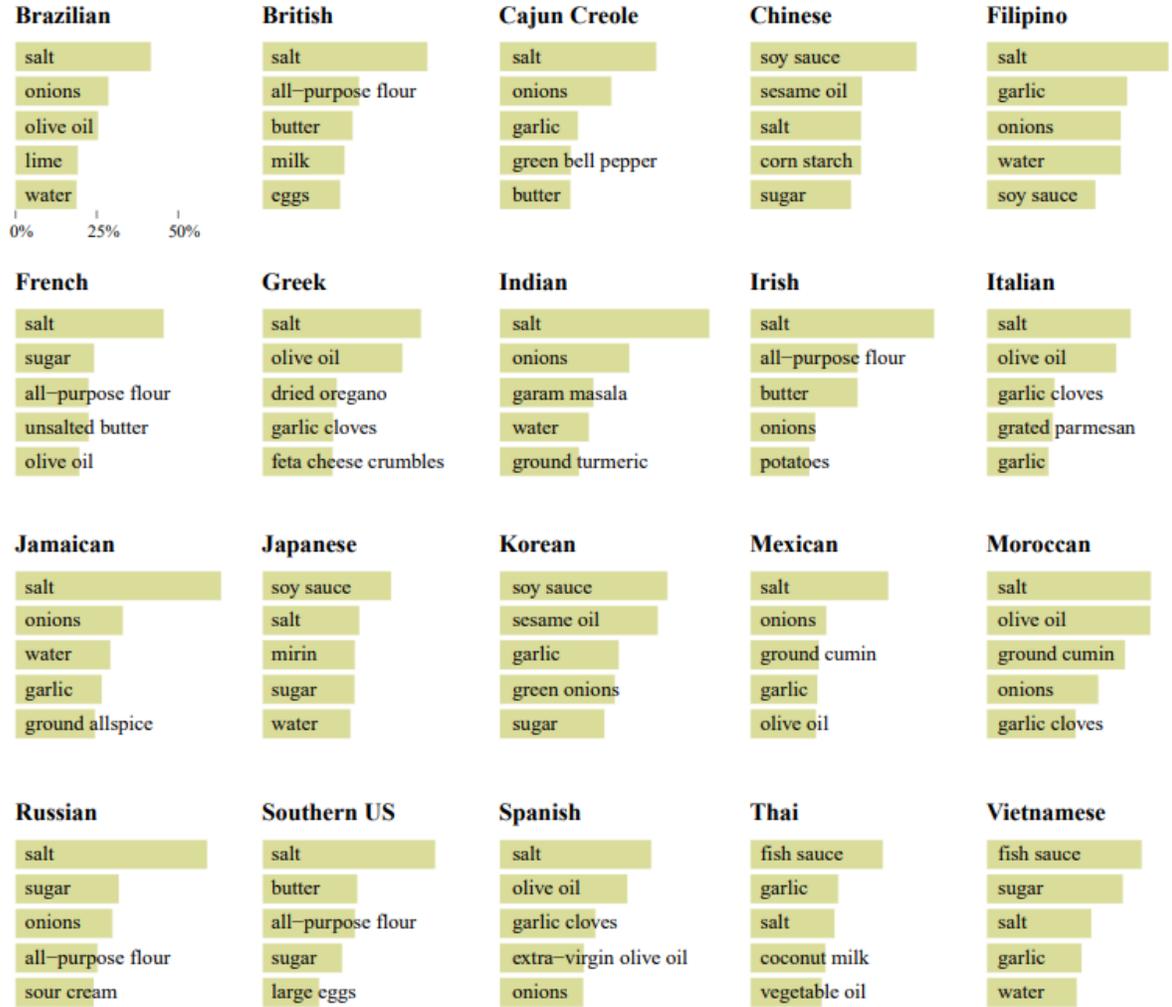


7/18/2024

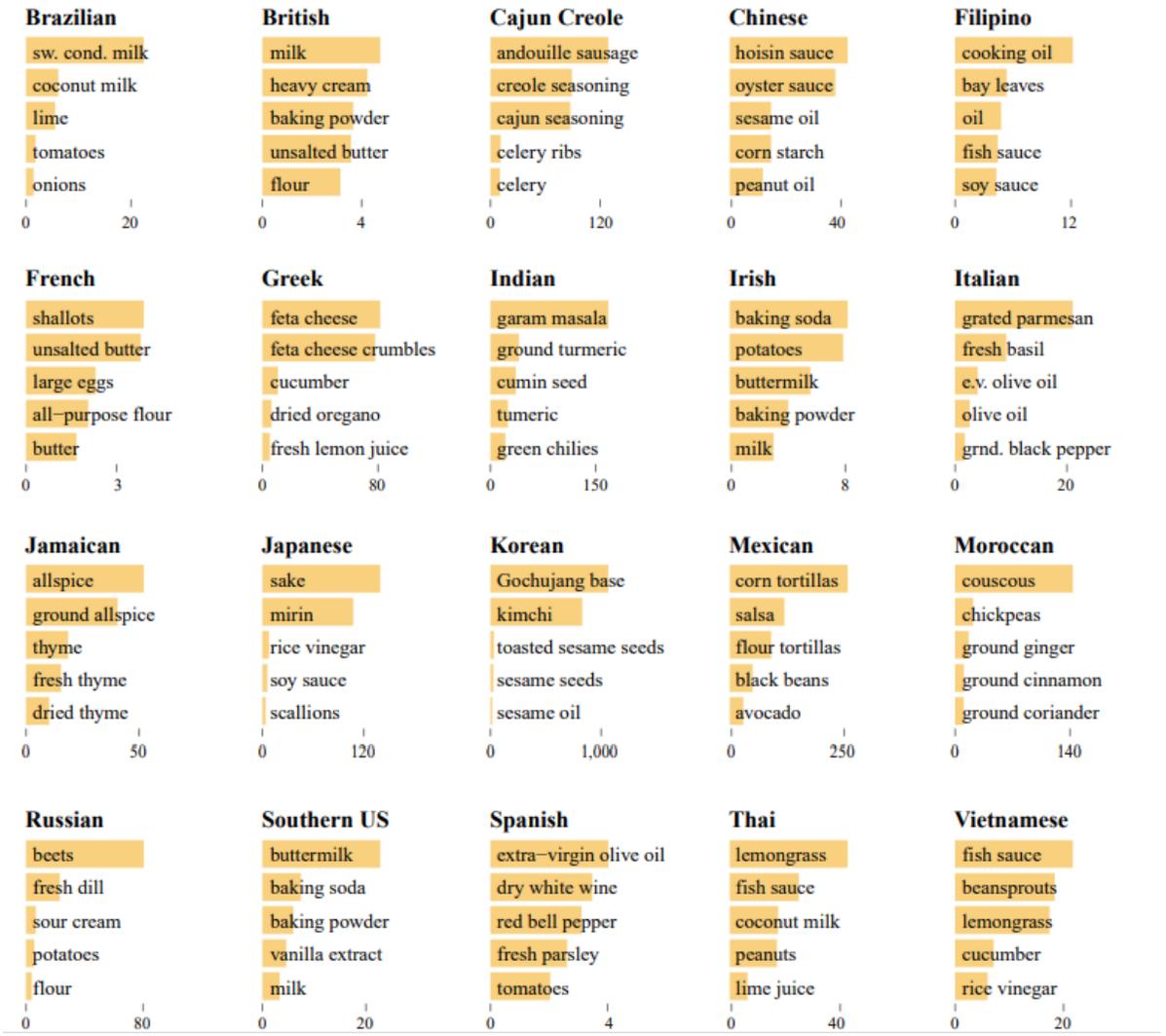
Indicatori: niveluri de măsurare

- Nivelul **nominal**: județul, numele de familie, zodia
 - Valorile sunt cuvinte
 - Valorile nu pot fi ordonate
 - Permite numărări, clasificări, dar nu operații algebrice (modul)
- Nivelul **ordinal**: satisfacția cu calitatea aerului, percepția corupției
 - Valorile sunt cuvinte: foarte mult, mult, puțin, foarte puțin (scala Likert)
 - Valorile pot fi ordonate (permite calculul medianei și al modului)
 - Uneori traduse în numere
- Nivelul **numeric**
 - Valorile sunt numere
 - Permite calcul algebric (medie)

Most Used Ingredients



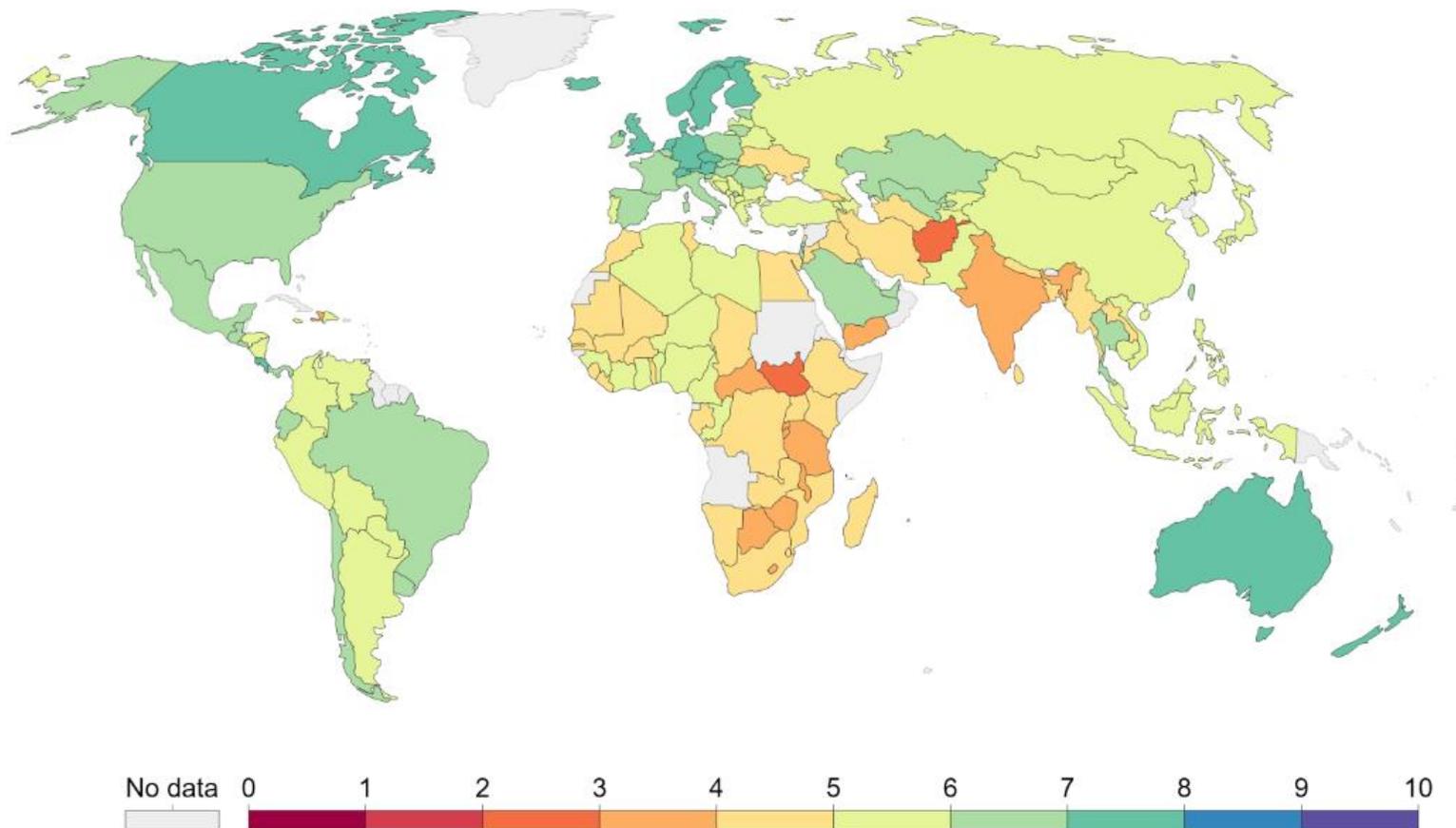
Most Cuisine-Specific Ingredients



Măsurare ordinală: satisfația cu viața

Self-reported Life Satisfaction, 2018

Life satisfaction is self-reported as the answer to the following question: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"



Source: World Happiness Report (2019)

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

Nominal: unde ai absolvit generala și liceul?

Liceul × Generala

			Generala			Total
			În București	Într-o localitate rurală	Într-un alt oraș	
Liceul	În București	Column %	98.6%	22.2%	3.8%	43.1%
	În altă localitate	Column %	1.4%	77.8%	96.2%	56.9%
Total		Column %	100.0%	100.0%	100.0%	100.0%

(Sondaj studenți Sociologie, 198 respondenți, 2023, eșantion de disponibilitate la licență, masterat, doctorat)

Ordinal – ordonat prin mărimea localității

Liceul * Scoala generala Crosstabulation

% within Scoala generala

		Scoala generala			Total
		.00 Într-o localitate rurală	1.00 Într-un alt oraș	2.00 În București	
Liceul	.00 În altă localitate	77.8%	96.2%	1.4%	56.9%
	1.00 În București	22.2%	3.8%	98.6%	43.1%
Total		100.0%	100.0%	100.0%	100.0%

Ordinal aproximat numeric {0, 1} și {0, 1, 2}

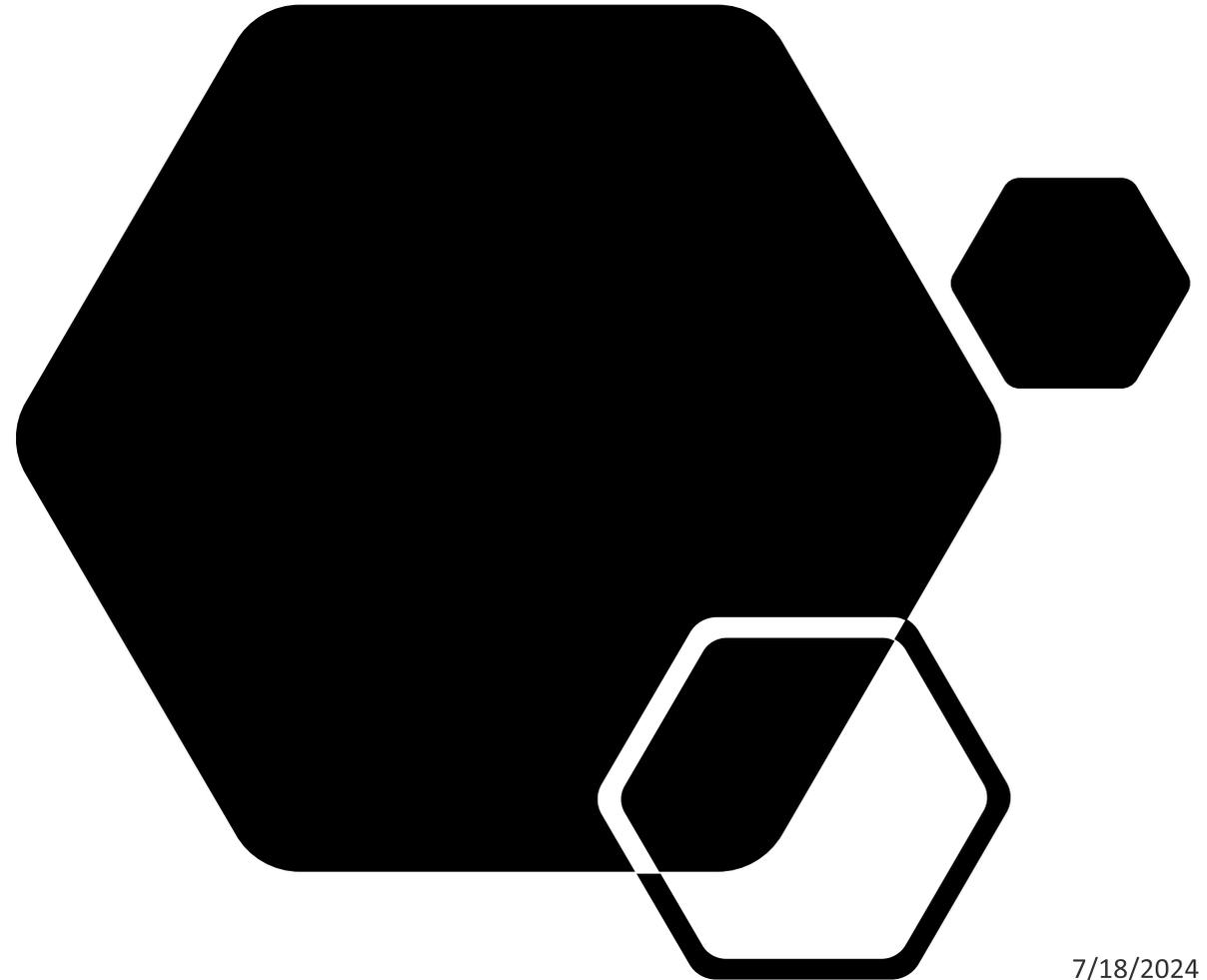
Correlations

		Scoala generala	Liceul
Scoala generala	Pearson Correlation	1.000	.672
	Sig. (2-tailed)		.000
	N	198	197
Liceul	Pearson Correlation	.672	1.000
	Sig. (2-tailed)	.000	
	N	197	197

Distribuții

Descrierea datelor

- Tendință centrală
- Variabilitate
- Cazuri excepționale
- Distribuția normală



Descrierea distribuțiilor prin măsurare

- Prin măsurare putem vedea:
 - **Tendința** centrală a fenomenului (cum este în linii mari)
 - Media, mediana și modul
 - **Variabilitatea**
 - Excepții, outliers
 - Percentile
 - Indici sintetici ai variabilității

Tendință centrală și variabilitate

- Fie $X = \{x_1, x_2, \dots, x_n\}$ și $n = |X|$
- Tendința centrală
 - **Medie:** $\mu_X = \frac{\sum_{i=1}^n x_i}{n}$
 - **Mediană:** separă 50% cazuri la stânga, 50% cazuri la dreapta
 - **Mod:** este cea mai frecventă valoare din distribuție
 - **Percentile (cuartile, cvintile etc):** separă x% din cazuri
- Variabilitatea: cât de mult variază față de medie
 - **Variația** $\sigma_X^2 = \frac{\sum_{i=1}^n (x_i - \mu_X)^2}{n}$
 - **Abaterea standard** $\sigma_X = \sqrt{\sigma_X^2}$ (sau deviația standard)

Distribuția normală

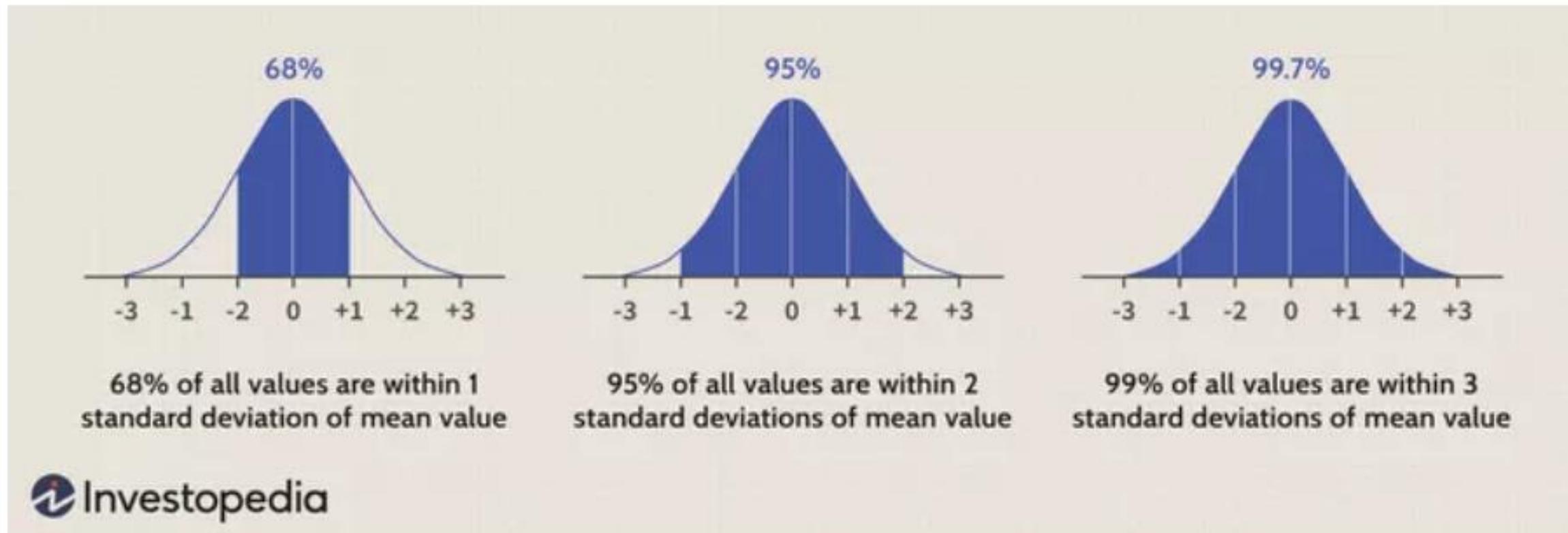
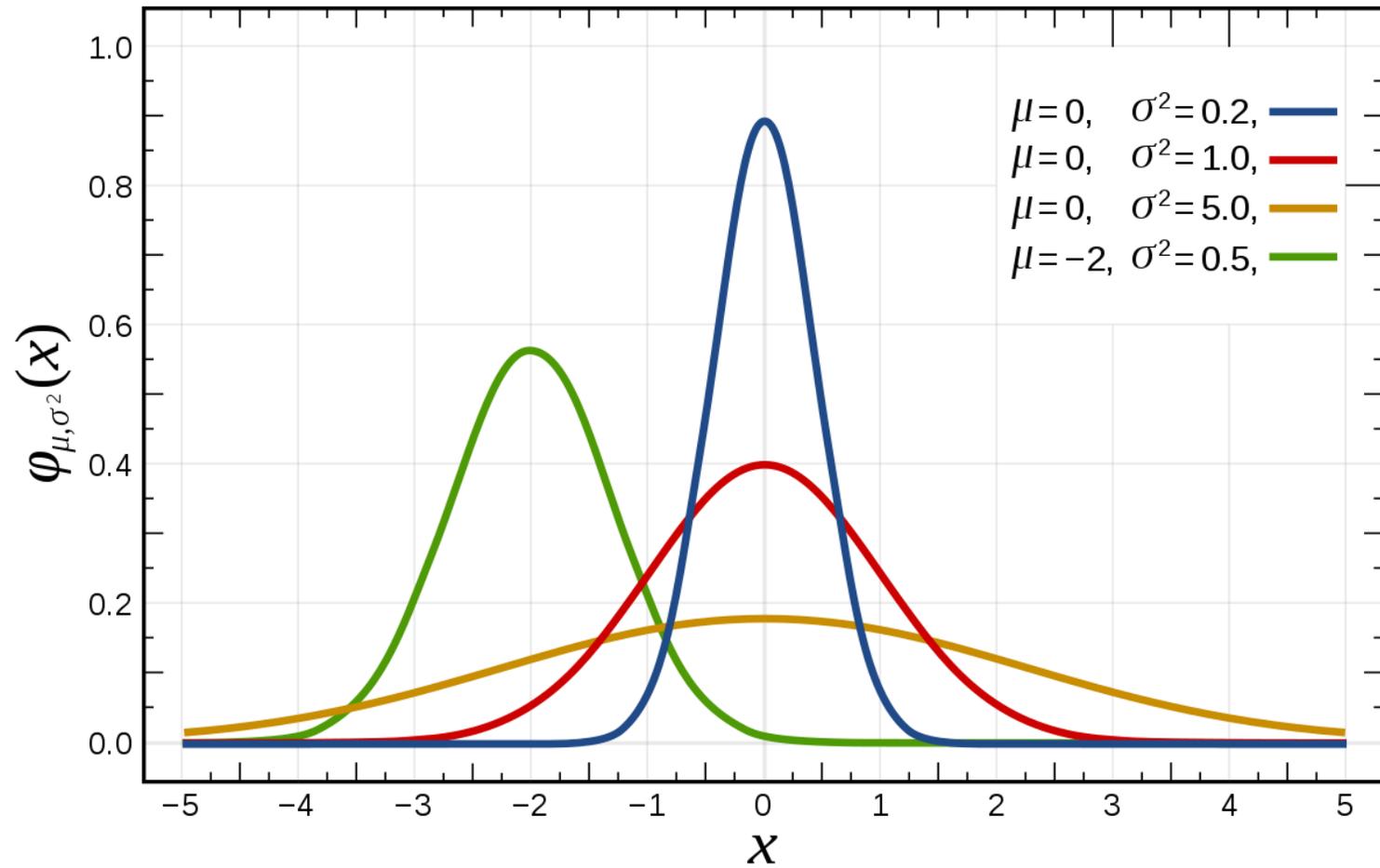


Image by Sabrina Jiang © Investopedia 2021

Distribuții diverse



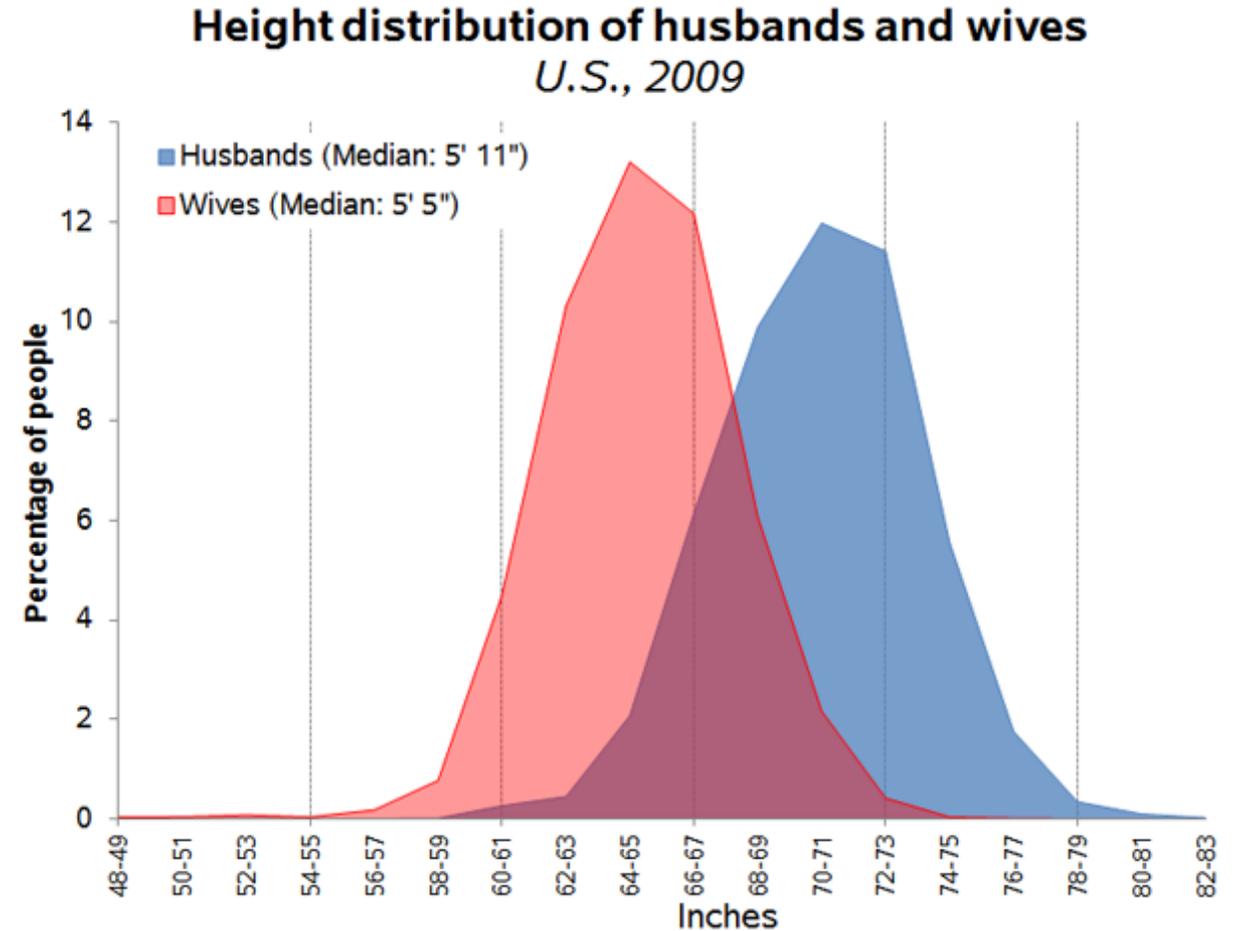


Stereotipuri vs. măsurare

- Stereotipurile sunt adesea și corecte, și greșite
 - „Bărbații sunt mai înalți decât femeile”
 - „Femeile sunt mai puțin agresive decât bărbații”

Înălțimea și genul

- Aproape întotdeauna soții sunt mai înalți decât soțiile
 - Cuplare selectivă
- Dar sunt întotdeauna bărbații mai înalți decât femeile?

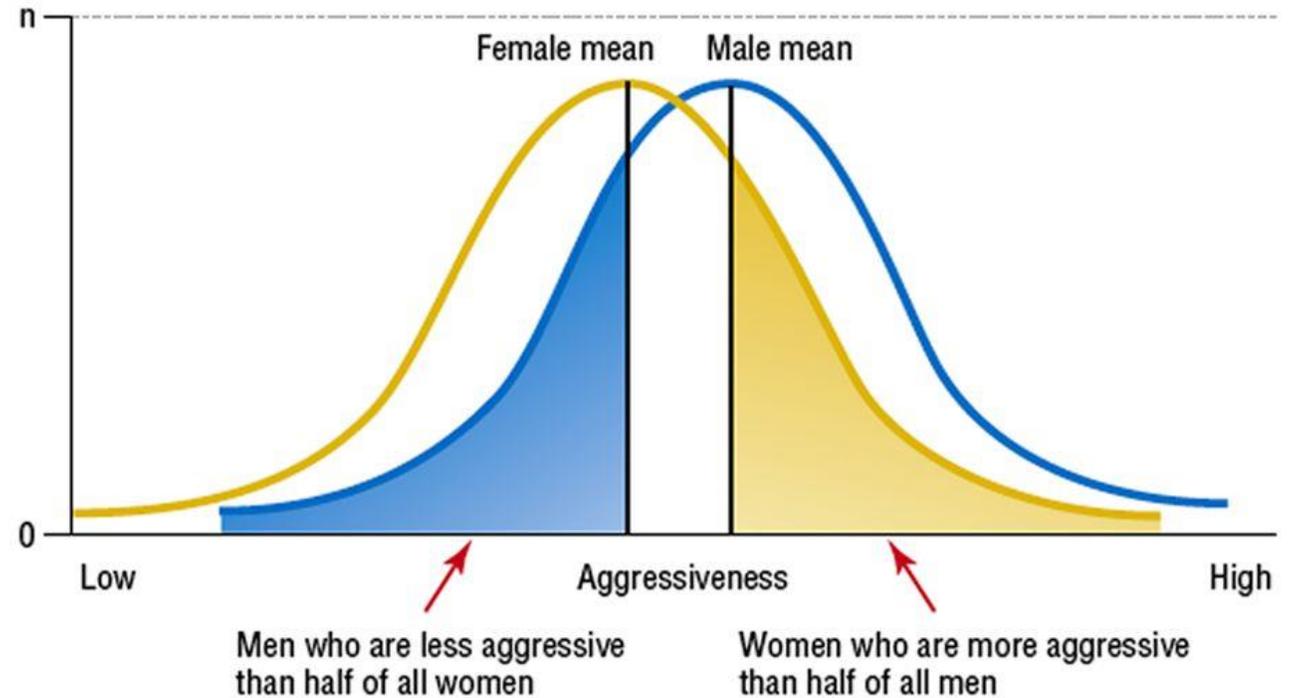


[The Atlantic](#)

Agresivitatea și genul

*Distribution of Aggressiveness among Men and Women

- Închisorile dețin mult mai mulți bărbați decât femei pentru infracțiuni violente
- Rezultă că femeile nu sunt violente?
 - Distribuția violenței extreme
 - Discriminări ce influențează „măsurarea” violenței prin sentințe



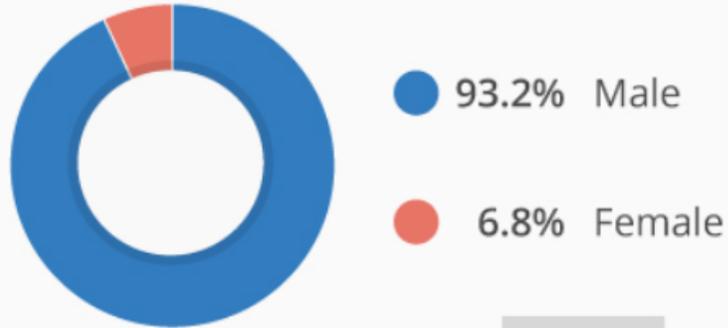
[Sursa](#)

Sursă: [Miller](#)

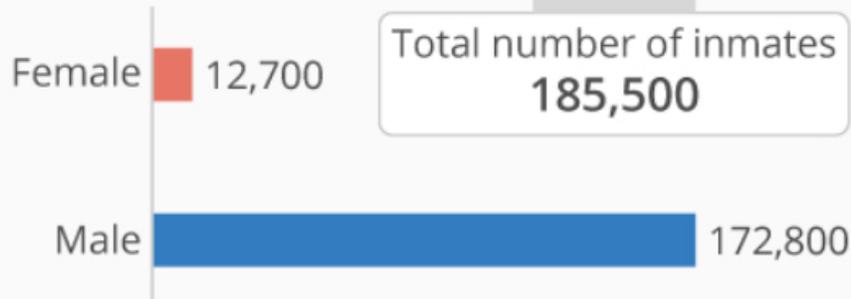
The Prison Gender Gap

Gender of inmates in U.S. federal prisons compared to the general population

Gender of federal inmates*



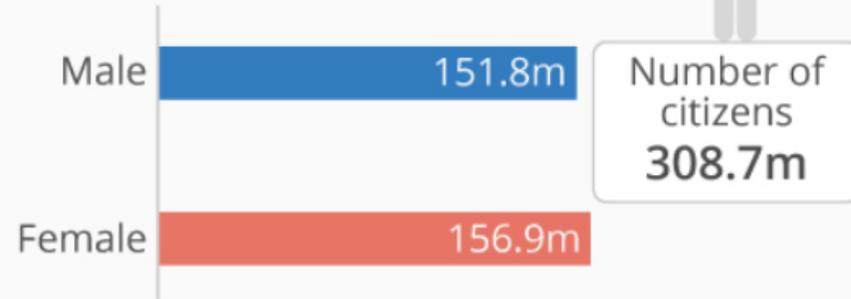
Number of federal inmates



Gender of all U.S. citizens**



Number of U.S. citizens (in m)



* As of September 23, 2017; no state prisons or local jails included

** 2010 Census

Source: BOP, U.S. Census Bureau

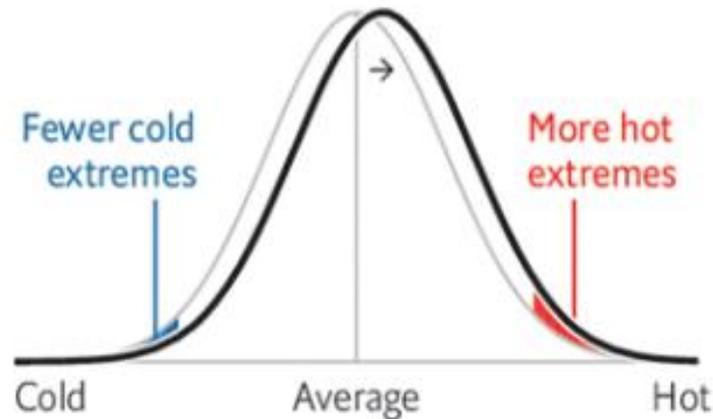


Schimbarea climatică și extremele

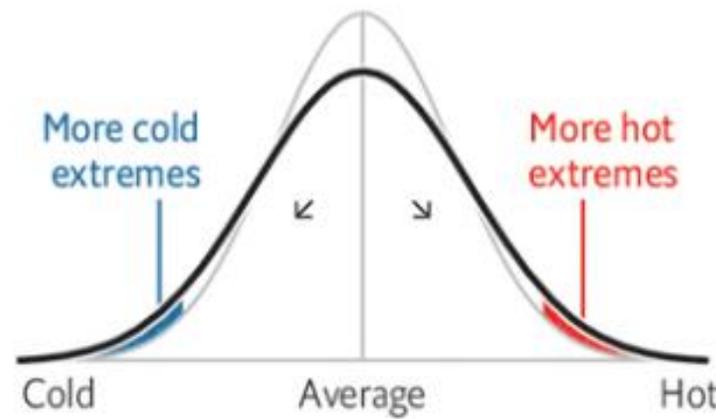
→ **A hotter planet is a more extreme one; some regions become unliveable**

Effect of changes in global temperature

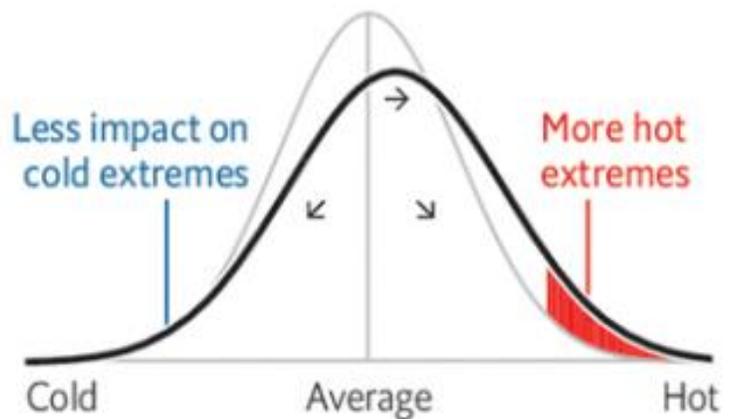
Increase in mean



Increase in variance



Increase in mean and variance



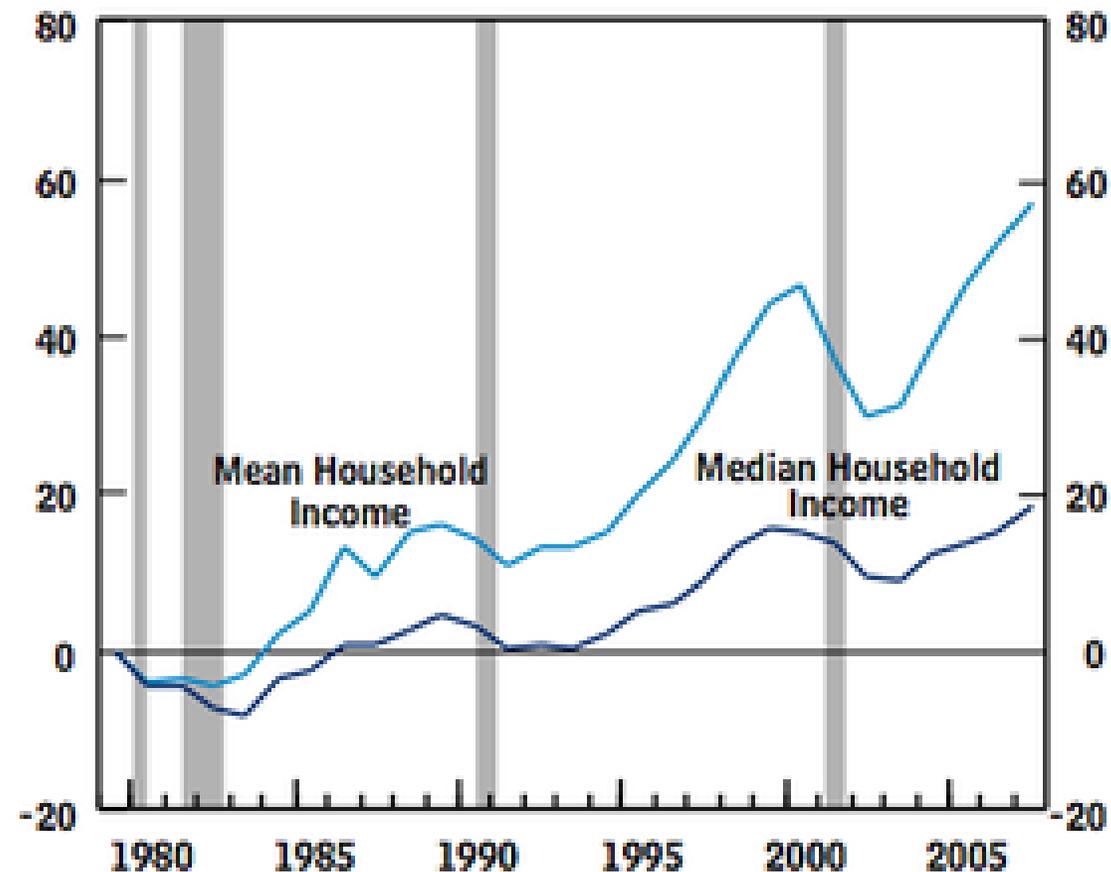
Source:
The Economist

Cumulative Growth in Mean and Median Household Market Income

(Percentage change in income since 1979, adjusted for inflation)

Tendința centrală:
medie vs. mediană

- Media este mai influențată de valorile extreme
- Extremele tip top 1% trag în sus media averii gospodăriilor

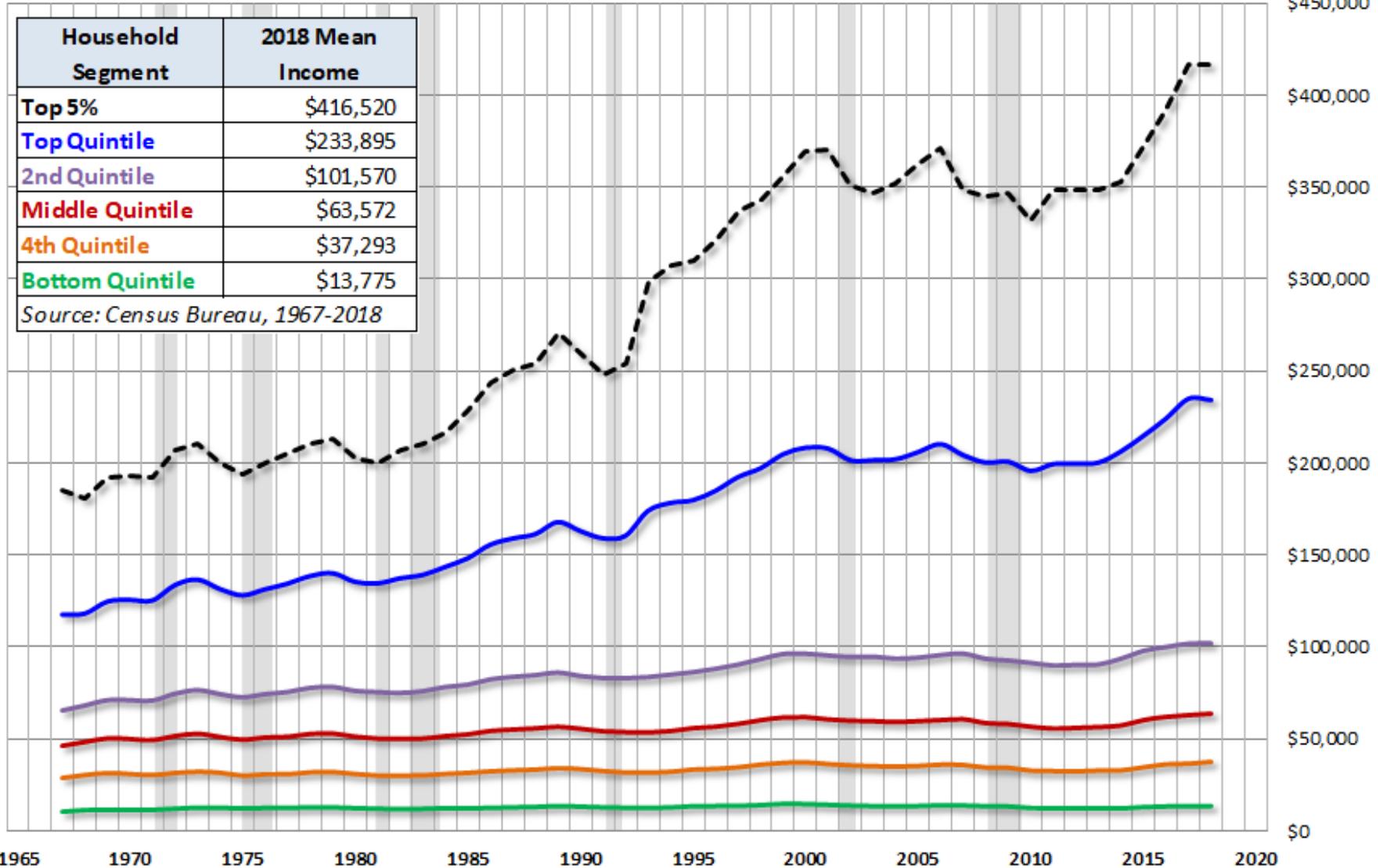


Source: Congressional Budget Office.

Real Average Household Income By Quintile and Top 5 Percent in 2018 Dollars

With Recessions Highlighted

Percentile:
Înțelegerea
dinamicilor
creșterii
venitului

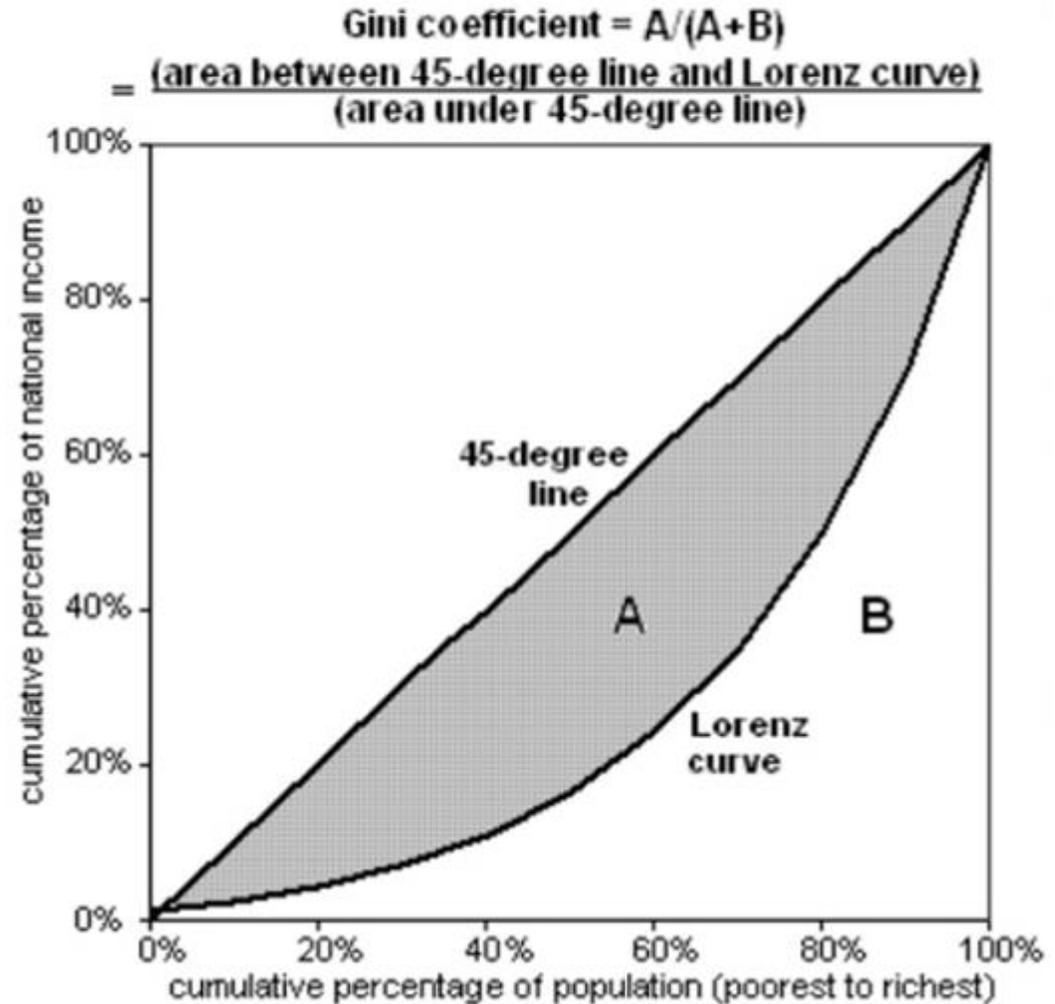


[Sursă](#)

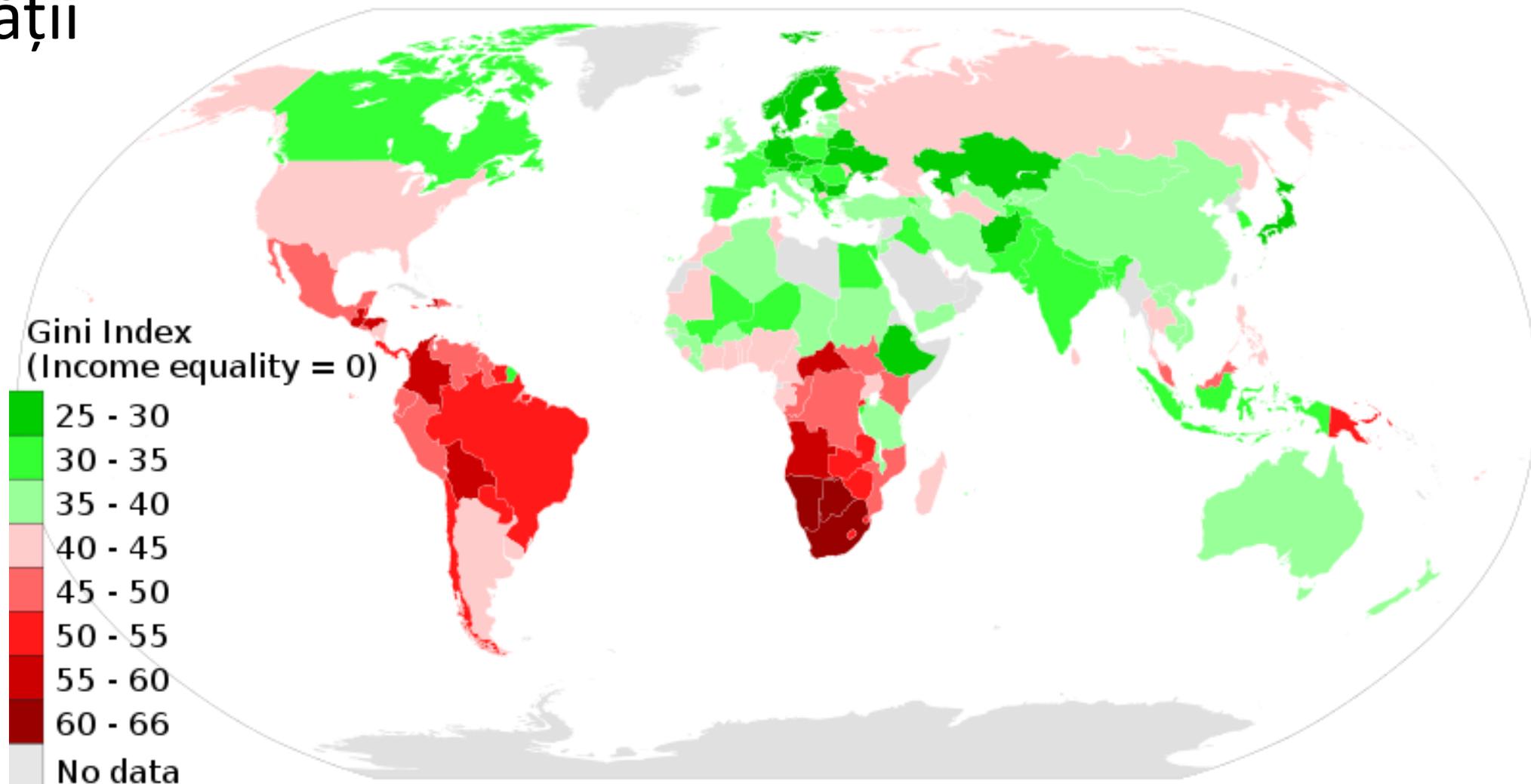
Variabilitatea
averii:

Coeficientul Gini

Gini Coefficient



Variabilitatea variabilității venitului

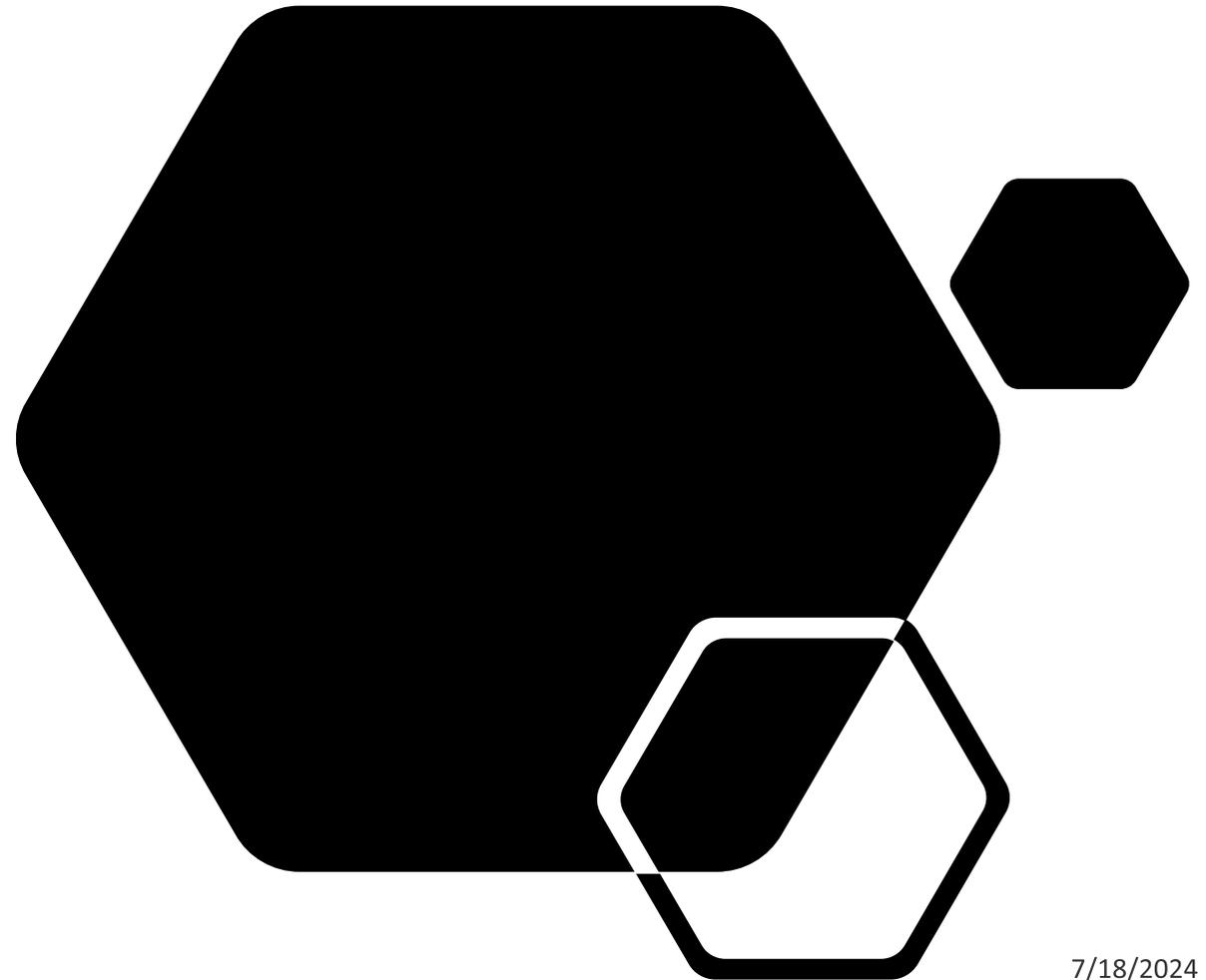


[Wikimedia](#)

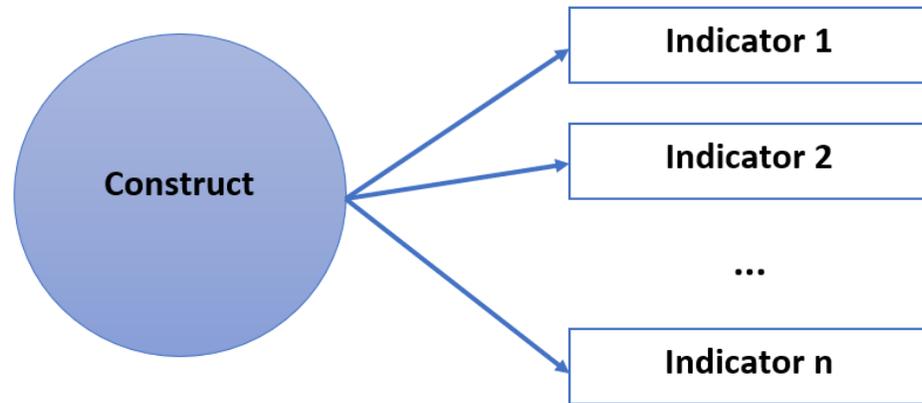
Modele de măsurare

Modelul reflectiv

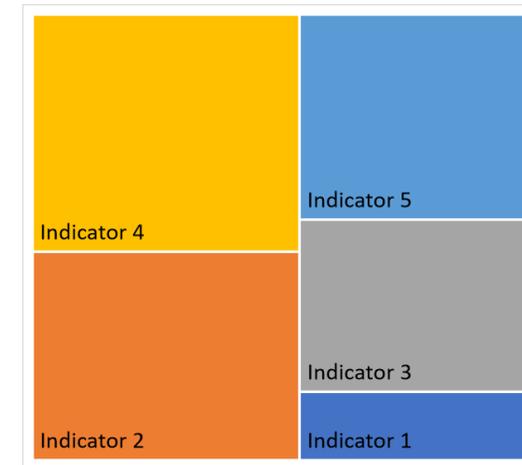
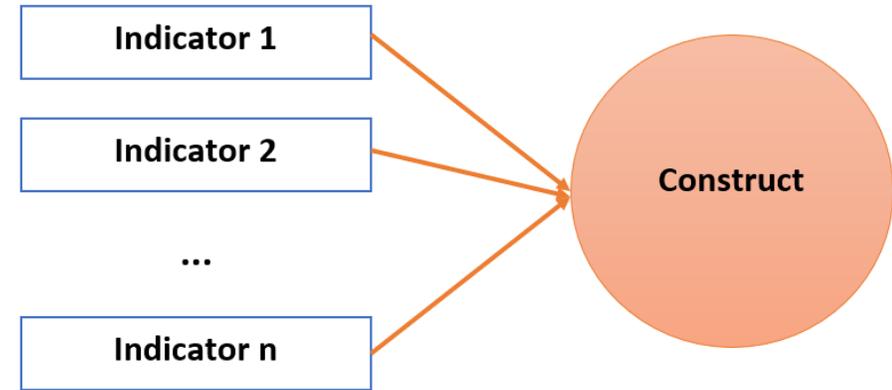
Modelul formativ



7/18/2024



Modelul reflectiv:
 Constructul este cauza indicatorilor
 Indicatorii corelează puternic
 Analiză factorială: identificarea factorului latent



Modelul formativ:
 Indicatorii cauzează / compun constructul
 Indicatorii pot să nu coreleze
 Sumă / medie ponderată etc

Exemple

- Ce măsoară notele? Trebuie notele să fie normalizate?

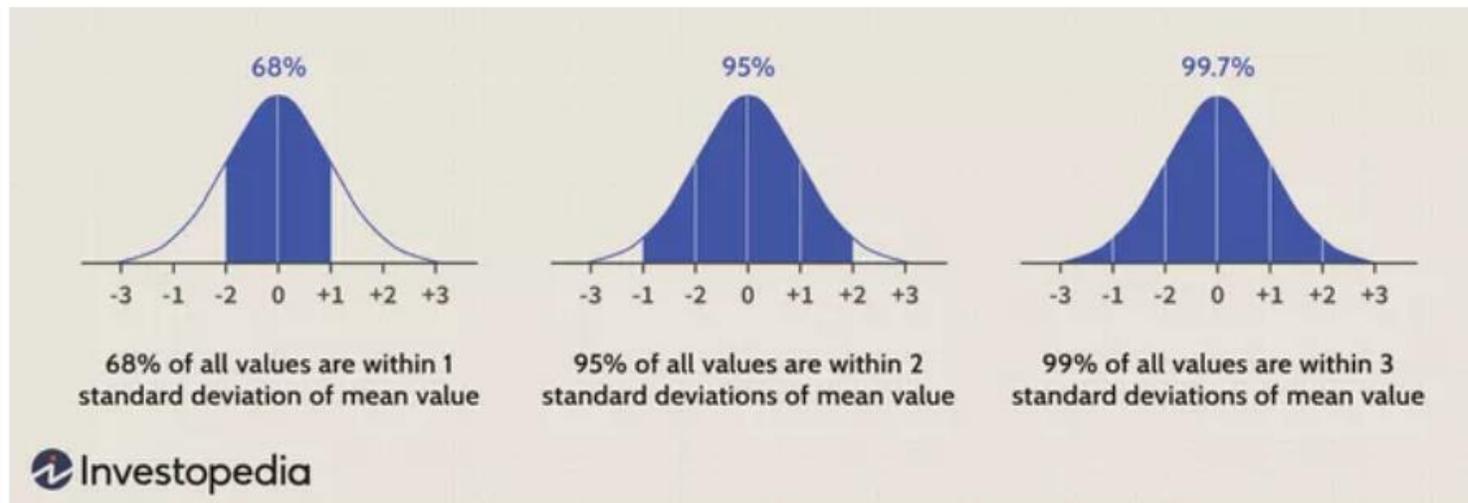
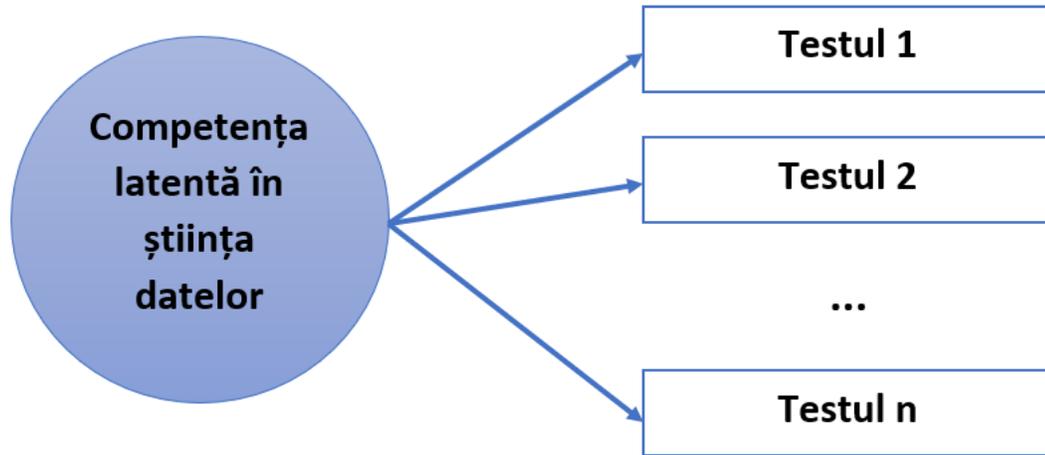
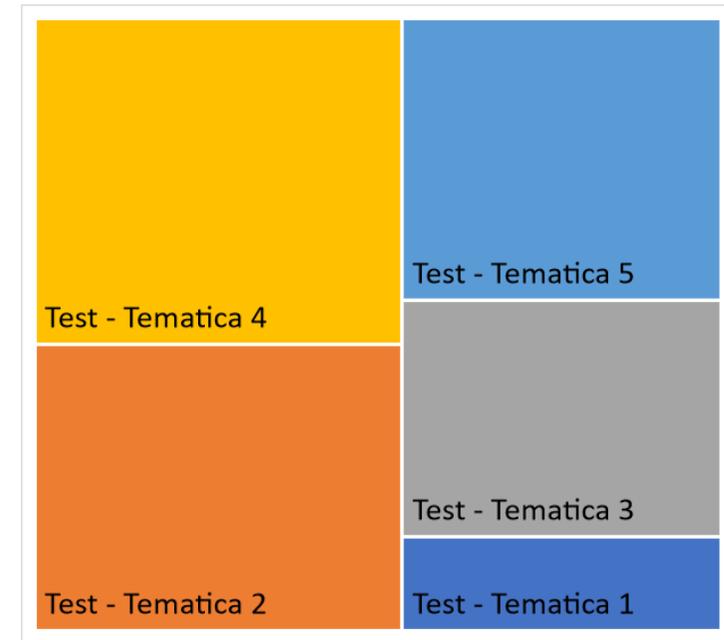


Image by Sabrina Jiang © Investopedia 2021



Model reflectiv – competența latentă

- Constructul este cauza indicatorilor
- Testele reflectă competența latentă (invizibilă) în știința datelor
- Toate testele măsoară imperfect aceeași competență
- Notele la teste sunt corelate (imperfect)
- Notele sunt distribuite normal



Model formativ – cunoștințe

- Indicatorii compun constructul
- Testele măsoară fiecare o altă tematică
- Cele 5 tematici acoperă cunoștințele de știința datelor
- Notele la teste pot să nu fie corelate
- Notele pot fi distribuite variabil (eg toți studenții au cunoștințele cerute)

Ce modele de măsurare au...

- Inflația
- Inteligența (IQ)



Ce este inflația?

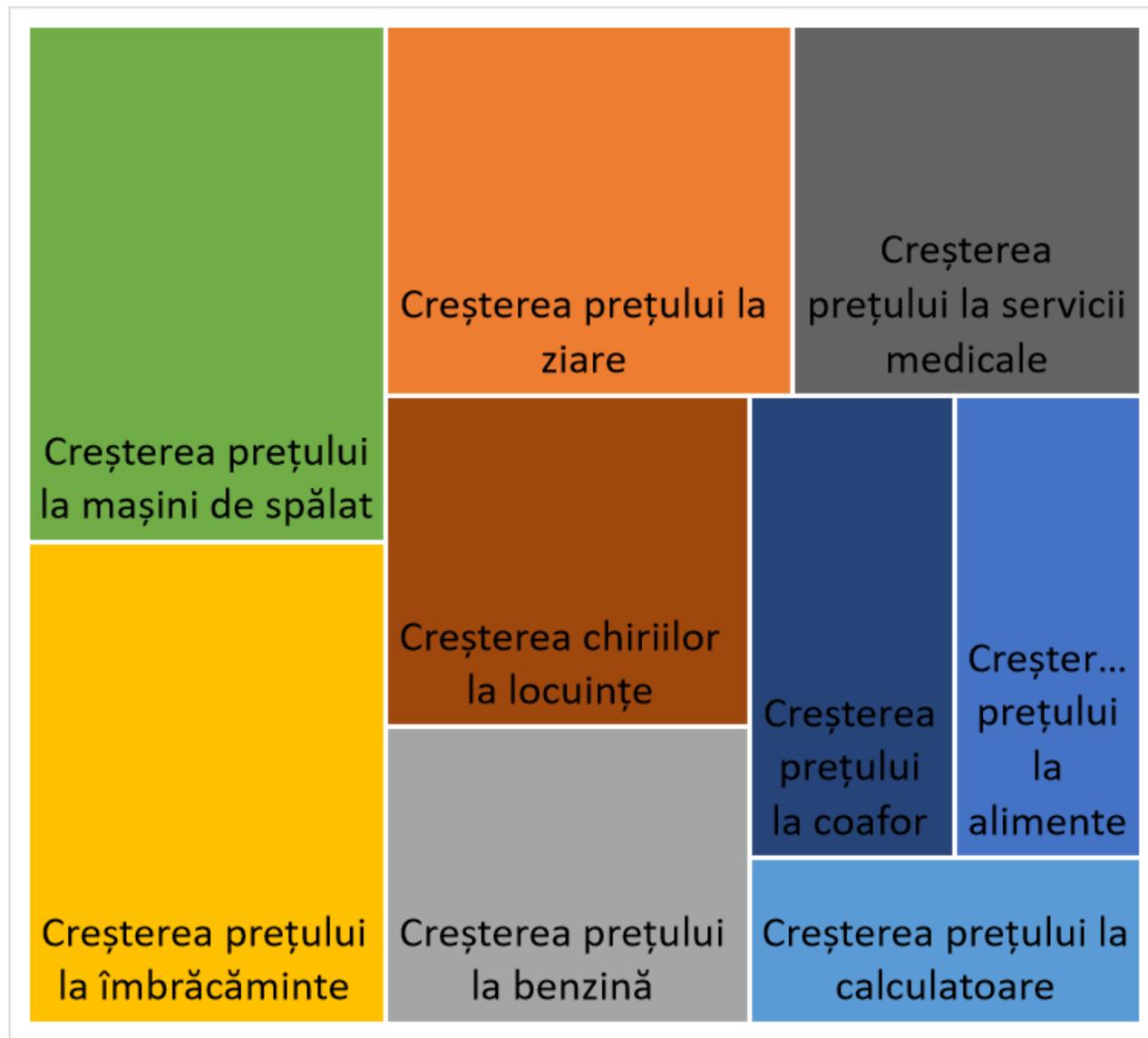
„Coșul” de bunuri de consum

Fiecare produs din acest coș are un preț care poate varia în timp.

Rata anuală a inflației este dată de prețul coșului integral într-o anumită lună comparat cu prețul acestuia în aceeași lună a anului precedent.

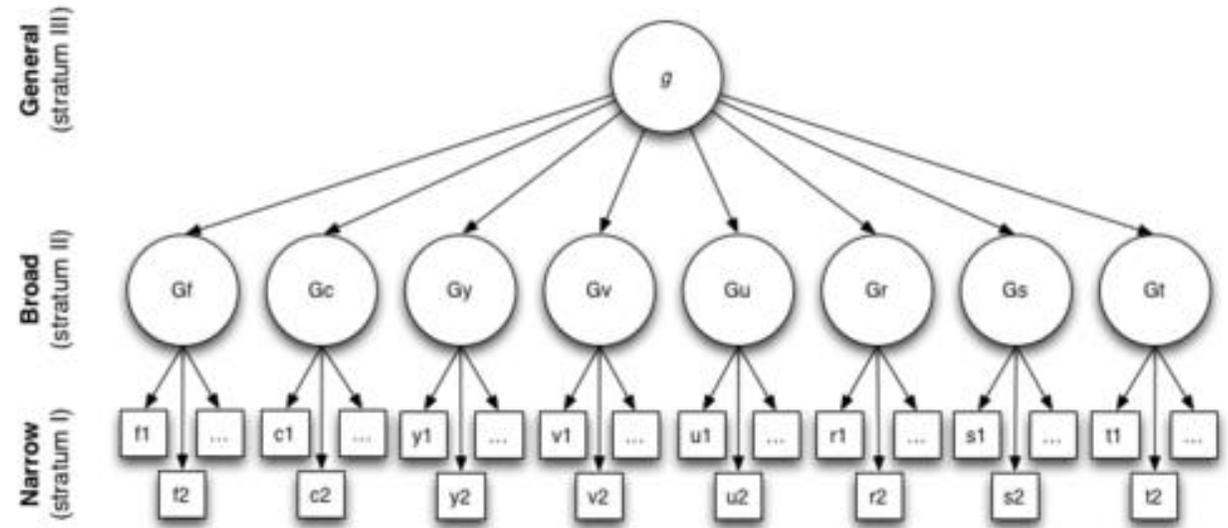
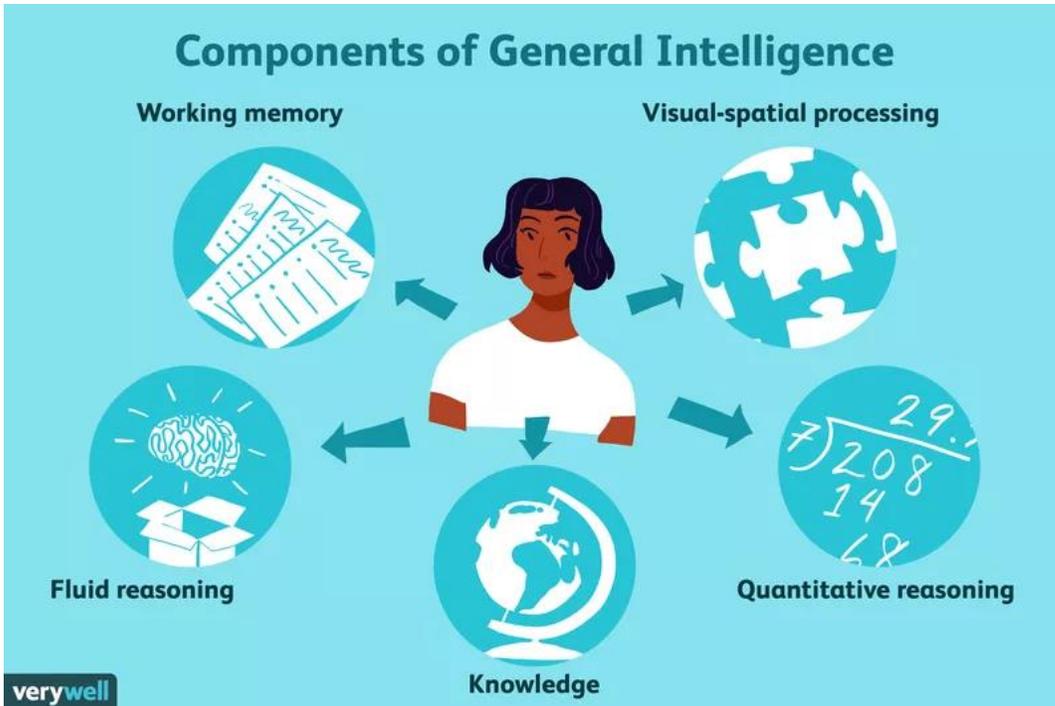
Ponderi diferite pentru categorii diferite

- Ponderi mari pentru produsele pentru care cheltuim mai mult



[Sursă](#)

Ce este IQ?

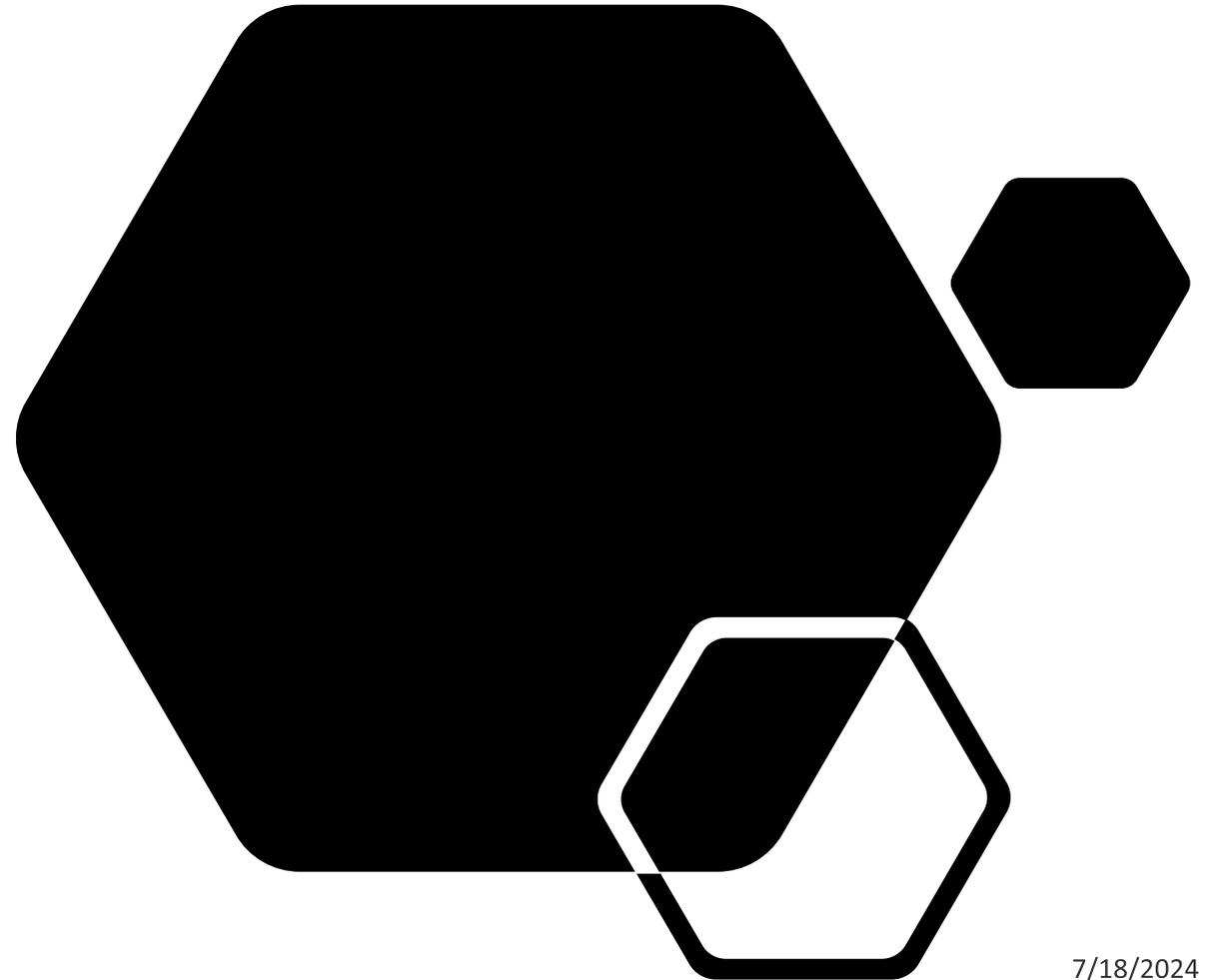


An illustration of [John B. Carroll's three stratum theory](#), an influential contemporary model of cognitive abilities. The broad abilities recognized by the model are fluid intelligence (G_f), crystallized intelligence (G_c), general memory and learning (G_y), broad visual perception (G_v), broad auditory perception (G_u), broad retrieval ability (G_r), broad cognitive speediness (G_s), and processing speed (G_t). Carroll regarded the broad abilities as different "flavors" of g .

Erori de măsurare

Erori sistematice și aleatorii

Validitate și fidelitate



7/18/2024

Calitatea măsurării

- Valoarea măsurată - valoarea reală = **eroare**
- Tipuri de erori în funcție de direcția lor:
 - Erori **aleatorii**
 - Erori **sistematice**: de ex., dezirabilitatea socială
 - Orientarea sexuală
 - Preferințele politice
- Măsurarea timpului:
 - Ochiometric, clepsidră: erori aleatorii
 - Un ceas care rămâne în urmă: erori sistematice



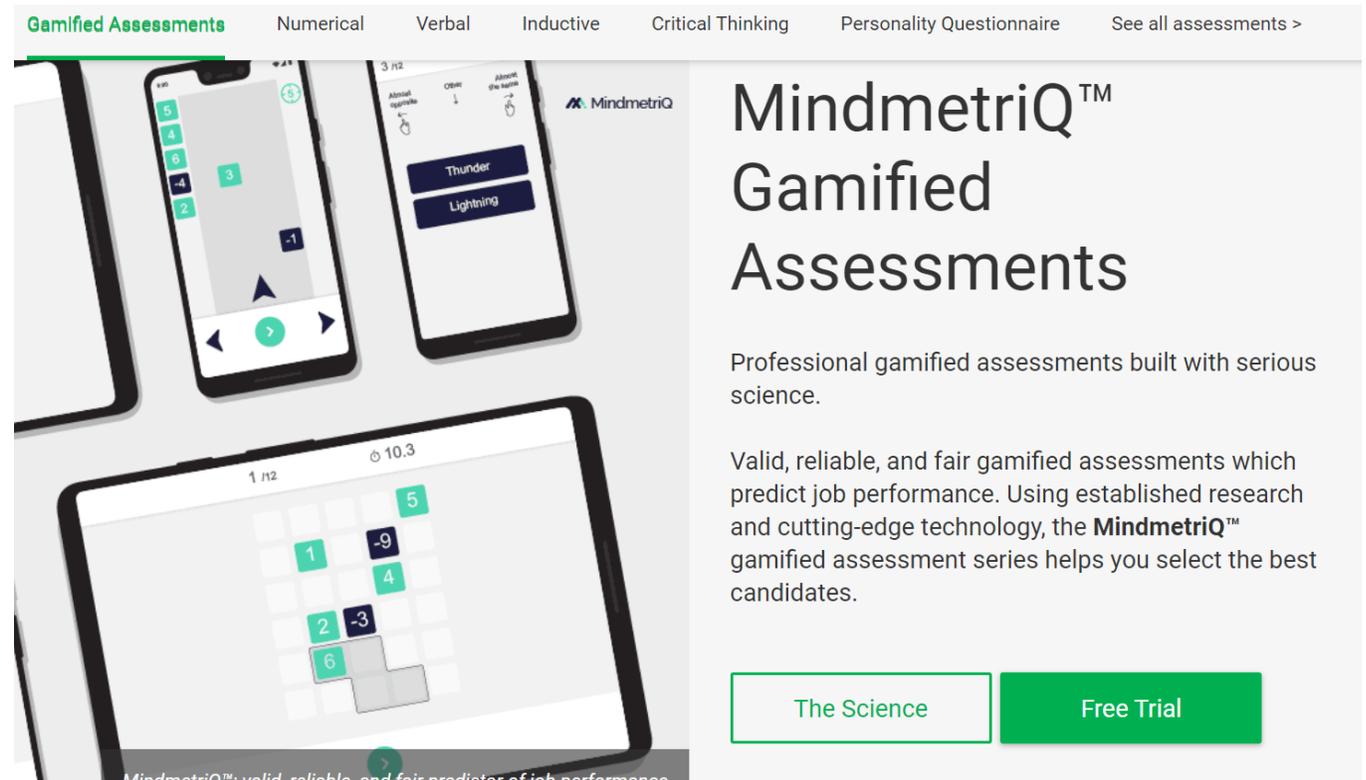
Ce fel de erori de măsurare?

- **Venitul** măsurat prin declarațiile depuse la Fisc
- **Temperatura** măsurată punând mâna pe frunte
- **Popularitatea** unui candidat politic măsurată prin sondaje de opinie
- **Talentul** unui copil măsurat întrebând părinții
- **Riscul** de accident măsurat prin teama pasagerilor într-un avion



Erori: Validitatea măsurării

- En. *validity*
- **Ce vrei să măsoari vs. ce măsoari de fapt?**
- Ex.:
 - Măsoară prețul calitatea produsului?
 - Măsoară un test gamificat abilitățile candidatului?



The screenshot displays the MindmetriQ website interface. At the top, there is a navigation bar with the following items: "Gamified Assessments" (highlighted in green), "Numerical", "Verbal", "Inductive", "Critical Thinking", "Personality Questionnaire", and "See all assessments >". Below the navigation bar, there are three mobile devices showing different assessment screens. One screen shows a grid of numbers (5, 4, 8, -4, 2, 3, -1) with a green arrow pointing to the number 3. Another screen shows a word puzzle with "Thunder" and "Lightning" as options. A tablet in the foreground shows a grid of numbers (1, 5, -9, 4, 2, -3, 6) with a green arrow pointing to the number 1. To the right of the devices, the text "MindmetriQ™ Gamified Assessments" is displayed. Below this, there is a paragraph: "Professional gamified assessments built with serious science." and another paragraph: "Valid, reliable, and fair gamified assessments which predict job performance. Using established research and cutting-edge technology, the **MindmetriQ™** gamified assessment series helps you select the best candidates." At the bottom right, there are two buttons: "The Science" (white with green border) and "Free Trial" (solid green).

Erori: Fidelitatea măsurării

- En. *reliability*
- Cât de **repetabil** este rezultatul?
- Măsurători repetate pentru același nivel dau aceeași valoare?
 - Măsurarea temperaturii scoțând capul pe geam
 - Măsurarea temperaturii cu un termometru



Altă conceptualizare: acuratețea și precizia

- **Acuratețea:** cât de apropiate sunt măsurătorile de valoarea adevărată
 - Similar validității
 - Erori sistematice
- **Precizia** se referă la cât de apropiate sunt măsurătorile între ele
 - Similar fidelității
 - Erori aleatorii



Scenarii

accurate
repeatable



Temperatura
[termometrul]
- **Erori aleatorii
redușe**

not accurate
repeatable



Ora [ceas care o ia
înainte]
Orientarea sexuală
[chestionar]
- **Erori sistematice dar
estimări stabile**

accurate
not repeatable



Ora
[capul pe geam]
- **Erori aleatorii mari**

not accurate
not repeatable



Propria inteligență
[auto-estimare]
- **Erori sistematice și
valori instabile**

[Sursă](#)

Măsurarea zodiei prin data nașterii

- Foarte precis
- Foarte repetabil
- Cât de valid?
 - Este data nașterii un indicator valid pentru constructul „zodie”?

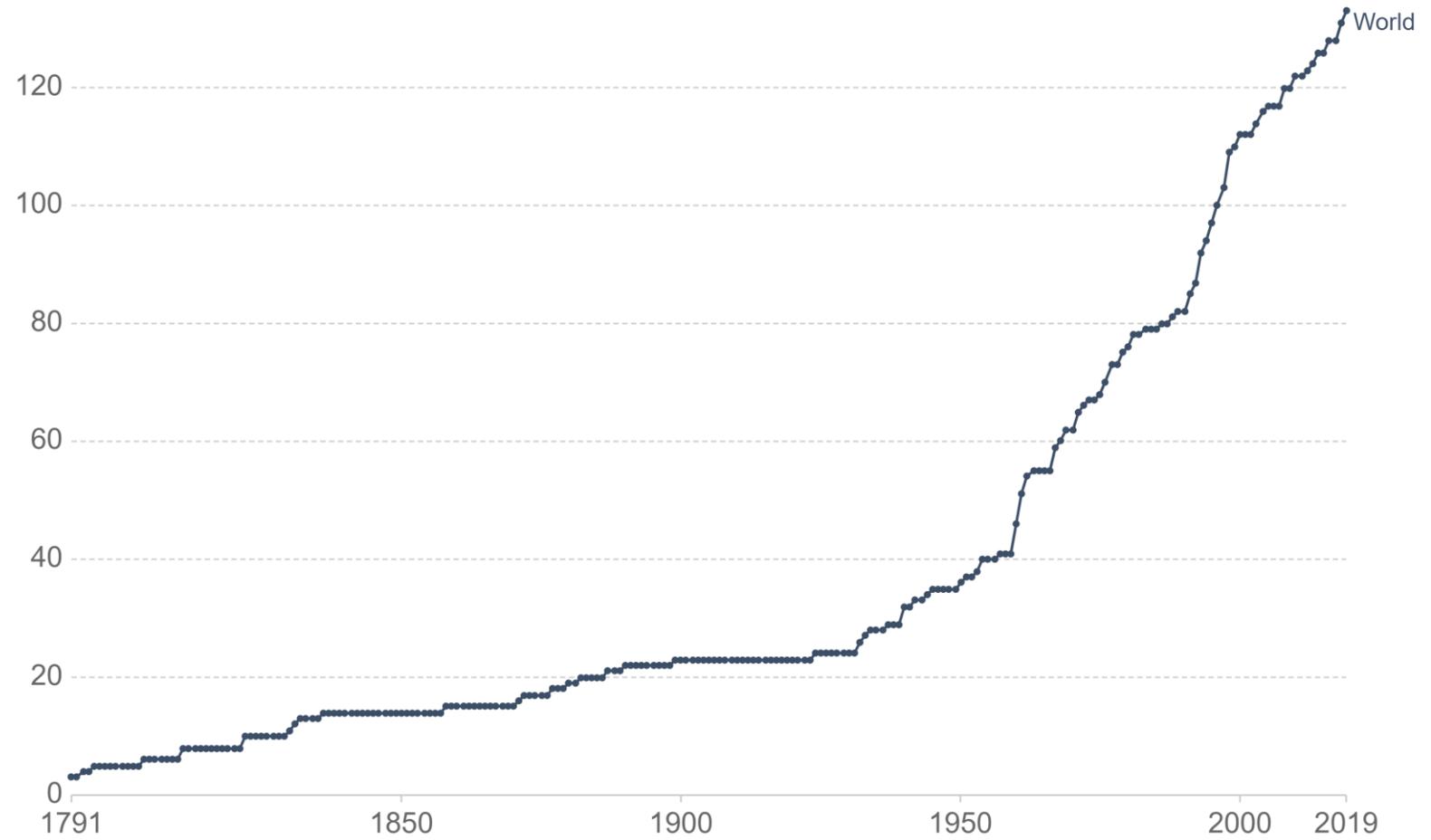


Măsurarea orientării sexuale prin chestionar

Number of countries where homosexuality is legal, 1791 to 2019



Countries where same-sex sexual acts are not considered a criminal offence. Some countries never contained a criminalising provision in the Penal Codes, while other consciously removed the criminalising provisions.



Source: OWID based on Kenny & Patel (2017)

CC BY

Facebook vs. variabilitate culturală

- Trăsături personale pot fi inferate pe baza like-urilor Facebook [[Kosinski et al.](#)]
 - Validare pe un lot de voluntari din SUA
- Cât de valide sunt inferențele pentru utilizatori din România?

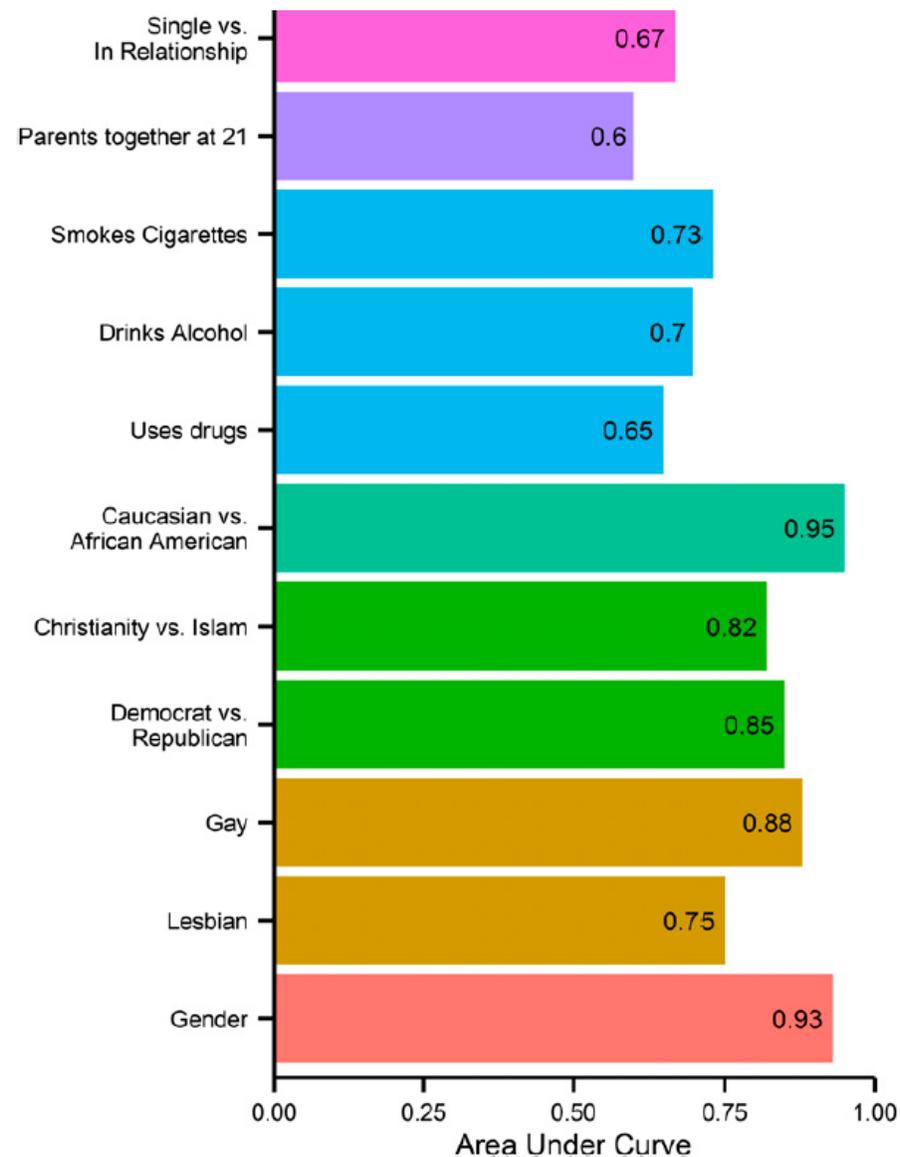
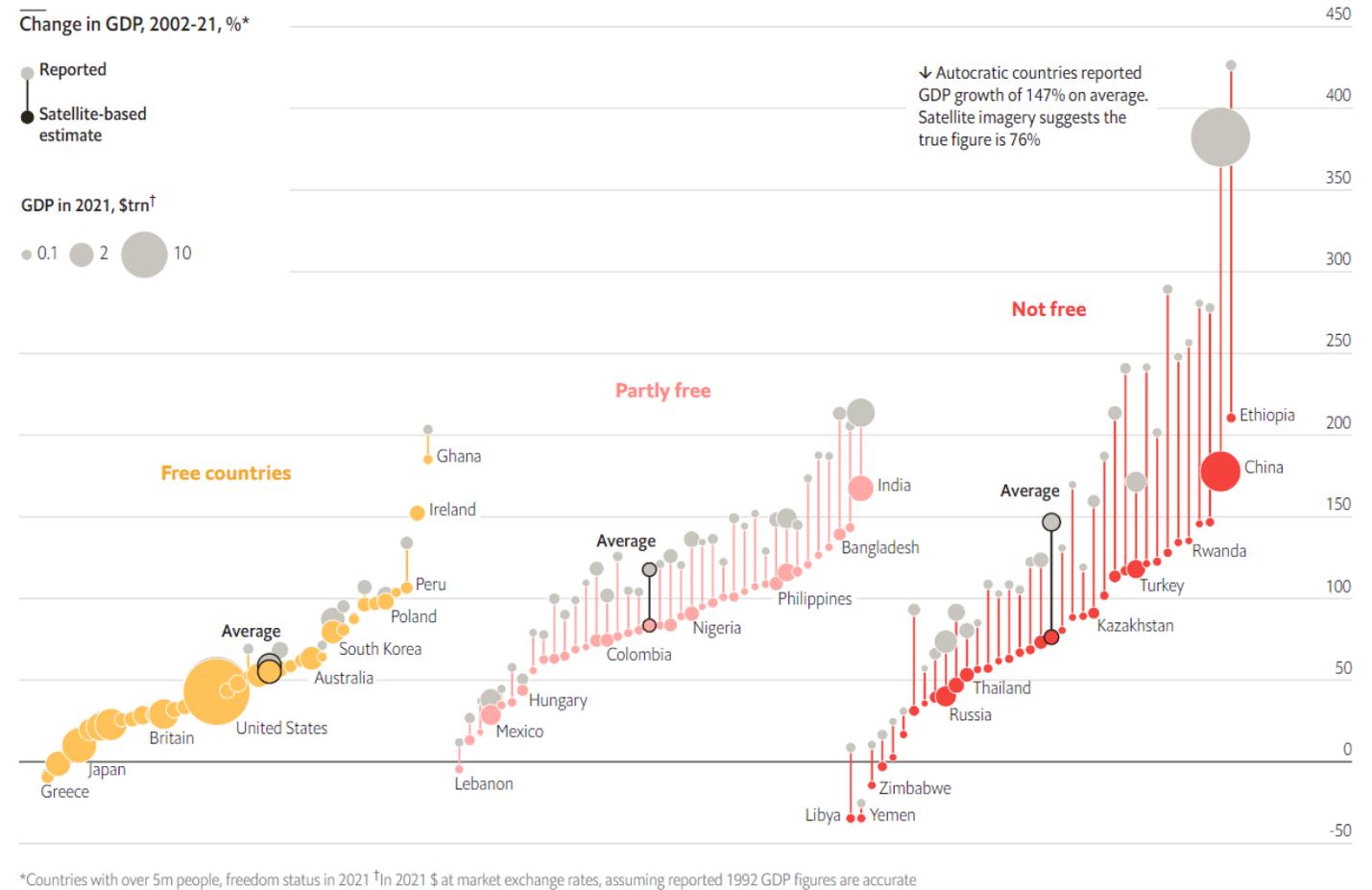


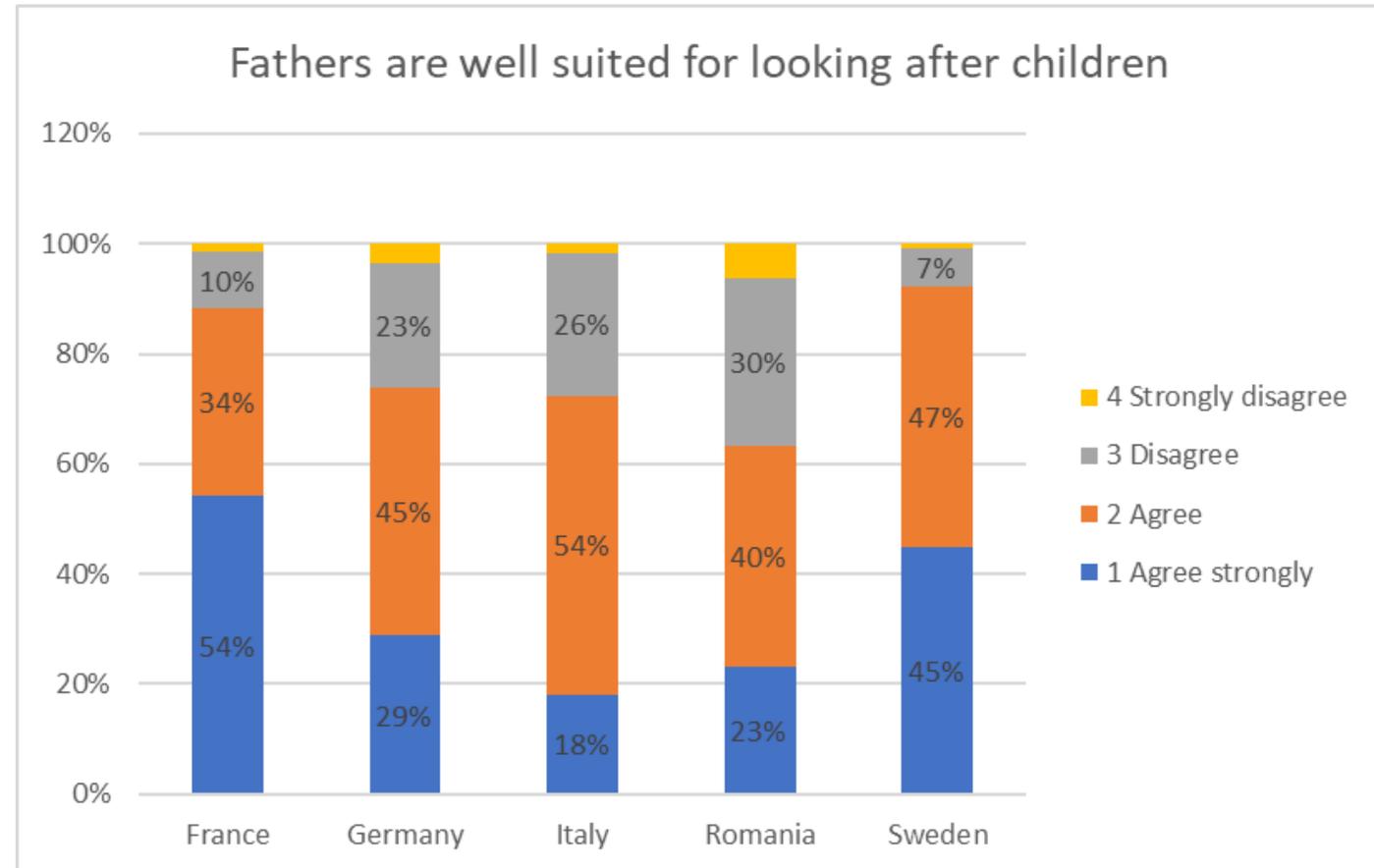
Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

Măsurarea creșterii GDP în state democratice sau autoritare



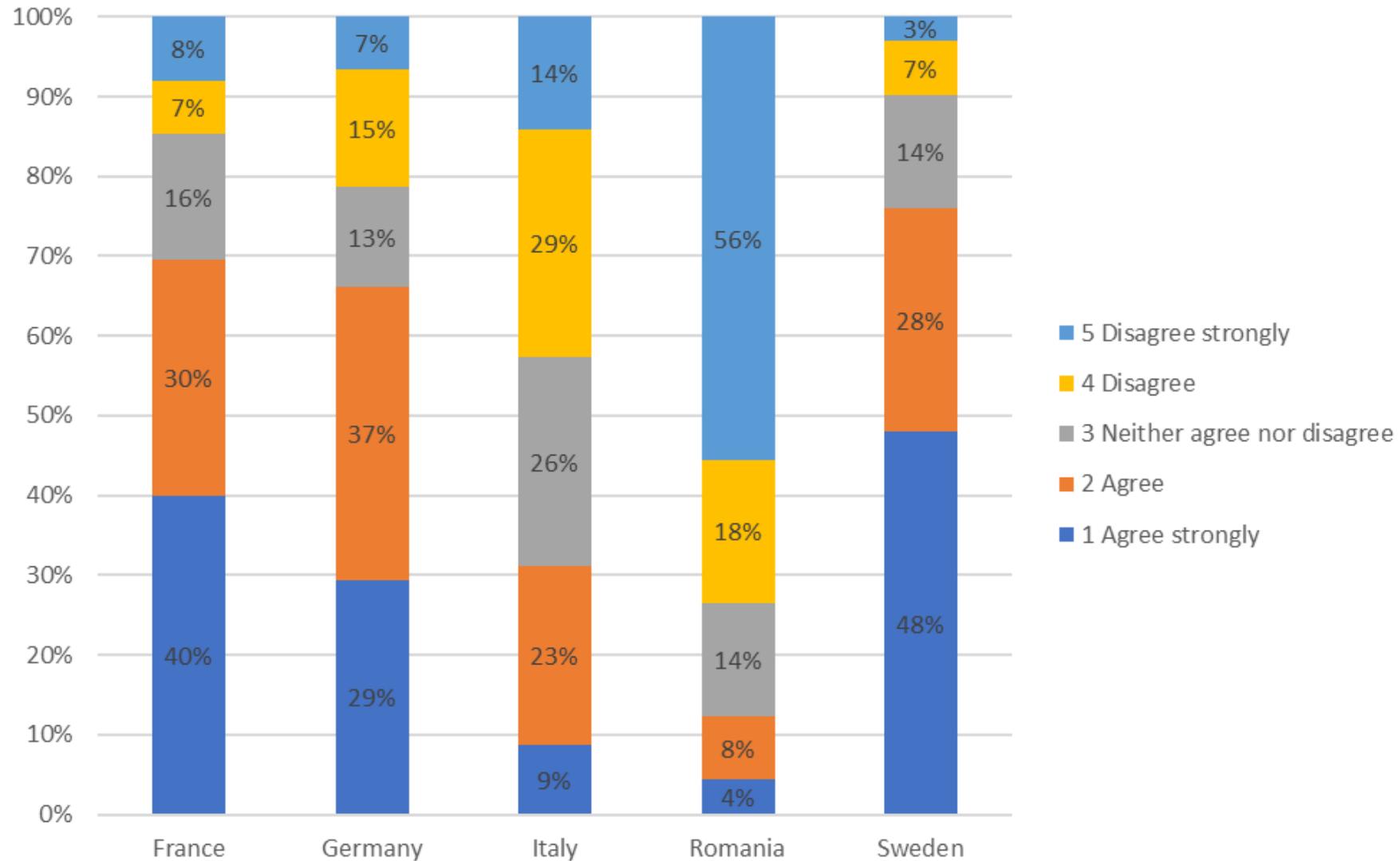
Alte tipuri de erori pentru măsurarea opiniei

- Stilul de răspuns
 - Preferința pentru a răspunde „Da” (agreeabilitate, aquiescență)
 - Preferința pentru a răspunde „Nu știu”
- Alegerea scalei
 - Cu punct de mijloc sau fără punct de mijloc?



World Value Survey, 2008-2010

Homosexual couples are as good parents as other couples



World Value Survey, 2017-2020

Concluzii

- Fenomenele sunt măsurate prin indicatori
- Diferență: fenomenul 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

