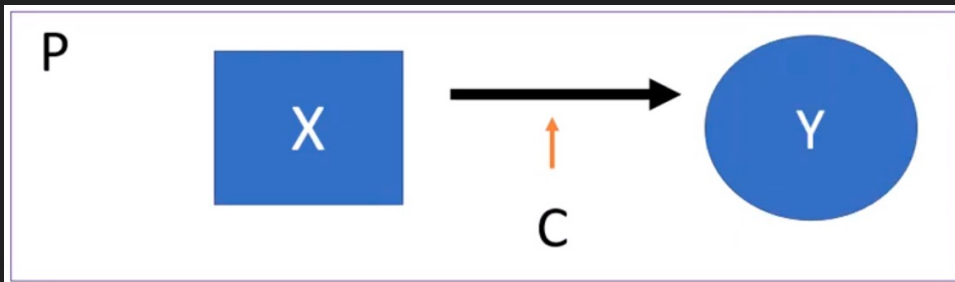# Empirical studies: validity, data collection and data type

IN4143: Data Analysis and Causal Inference

Previously ...

- Most common reasons why we cannot accurately say:
  - *X **causes** Y, under condition C, for population P*

  - *Versus "X **correlates** Y, under condition C, for population P"*
- We can classify studies according to
  - The method
  - The extent of intervention

Chocolate with High Cocoa Content as a Weight-Loss Accelerator
(Bohannon et al. 2015 - Global Journal of Medical Research)

Chocolate with High Cocoa Content as a Weight-Loss Accelerator (Bohannon et al. 2015 - Global Journal of Medical Research)

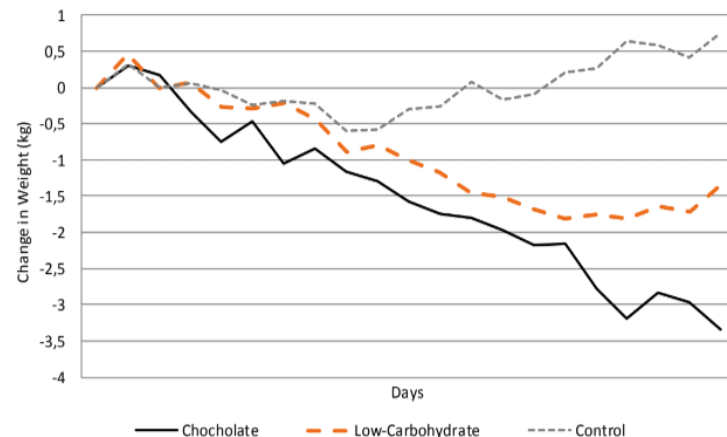| *Table 1* : Absolute changes in lipid levels, liver values, and albumin values in an analysis that include data on all subjects in the relevant groups.					
Variable	Chocolate Diet		Low-Carbohydrate		P-Value
Cholesterol (mg/dl)					
Day 21	-12,2	± 26,7	2,3	± 15,9	0,19
Triglycerides (mg/dl)					
Day 21	-22,6	± 85,7	3,0	± 41,3	0,55
LDL cholesterol (mg/dl)					
Day 21	-17,4	± 22,8	-5,0	± 22,4	0,00
ALT (U/l)					
Day 21	-6,4	± 6,7	-11,5	± 3,6	0,11
GGT/GGTP (U/l)					
Day 21	-8,8	± 5,5	-2,0	± 0,0	0.23
Albumin (g/dl)					
Day 21	0,0	± 0,4	0,1	± 0,3	0.23
Plus-minus values are means ±. The chocolate group had 5 subjects, in the low-carbohydrate group only 4 subjects could be considered.					
P values are for the differences between the two groups.					

Figure 1 : Daily weight development by group

Importance of empirical knowledge

Source: https://gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800

Importance of empirical knowledge

I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.

By John Bohannon   Published May 27, 2015 | Comments (476)

Here's a dirty little science secret: If you measure a large number of things about a small number of people, you are almost guaranteed to get a "statistically significant" result. Our study included 18 different measurements—weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, etc.—from 15 people. (One subject was dropped.) That study design is a recipe for false positives.

$P(winning) = 1 - (1 - p)^n$

With our 18 measurements, we had a 60% chance of getting some "significant" result with $p < 0.05$. (The measurements weren't independent, so it could be even higher.) The game was stacked in our favor.

It's called p-hacking—fiddling with your experimental design and data to push p under 0.05—and it's a big problem. Most scientists are honest and do it unconsciously. They get negative results, convince themselves they goofed, and repeat the experiment until it "works." Or they drop "outlier" data points.
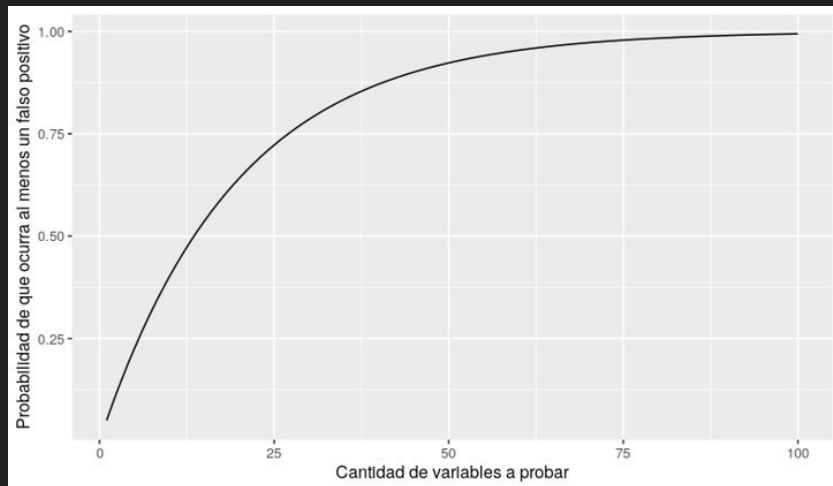
Today's class: Empirical research

- Validity
  - Definition
  - Types of validity
  - Ways to improve it

- Types of data
  - Organization
  - Modes of measurement

What could have gone wrong?

LT LATERCERA

Crecen delitos cometidos en la vía pública entre las 6 y 8 horas

Víctor Rivera, Santiago
21 JUN 2015 01:31 AM
Tiempo de lectura: 5 minutos

Diputados y un experto en seguridad atribuyen índice a la oscuridad generada por no cambiar el horario de verano.

Mediante la Ley de Transparencia, La Tercera accedió a cifras de Carabineros sobre el registro de robos con intimidación, robos con sorpresa, robos con violencia y hurto a nivel nacional, entre las 6 y las 8 horas, entre el 1 de abril (estación otoño) y el 2 de junio, en 2014 y 2015.

De acuerdo a los datos entregados por Carabineros, en 2014 se produjeron 1.418 de estos delitos en el período consultado, en tanto, este año se han registrado 1.750 de estos ilícitos. En consecuencia, hubo un aumento de un 23,4% (ver infografía).

What could have gone wrong?

MEDICINA

Aumentan nacimientos prematuros entre mujeres latinas; la elección de Trump sería la causa, según un estudio

En los nueve meses a partir de noviembre de 2016, se produjeron entre 3,2% y 3,6% más de los nacimientos prematuros en mujeres latinas por encima de los niveles de nacimientos prematuros que se hubieran esperado si la elección no hubiera ocurrido, sugiere el estudio, publicado en la revista médica JAMA el viernes.

- LEE: Tener el primer hijo genera más estrés que divorciarse o ver morir a tu pareja, según estudio

El estudio tenía algunas limitaciones, entre ellas el hecho de que los investigadores no pudieron separar los datos sobre las mujeres latinas para determinar las diferencias entre las personas nacidas en el extranjero y las que nacieron en los Estados Unidos, y solo se encontró una asociación entre los nacimientos prematuros y La elección presidencial - no una relación causal.

What could have gone wrong?

Calidad de vida

Estudio: Cinco o más horas de uso de celulares por día pueden aumentar la obesidad

Por Meganoticias
28 jul. 2019 - 15:15 hrs.

Los estudiantes universitarios que usaron sus celulares más de cinco horas al día tiene un 43% de probabilidades de sufrir de obesidad. Así lo revela un estudio realizado por la Universidad de Simón Bolívar en Barranquilla, Colombia.

La investigación presentada en ACC Latin America Conference 2019, señala que los jóvenes que tienen más tiempo con los dispositivos inteligentes pierden tiempo para tener otros hábitos de estilo de vida que puedan evitar enfermedades cardíacas.

Los investigadores encontraron que el riesgo de obesidad aumentaba en un 43 por ciento si se usaba un teléfono inteligente cinco horas o más al día, ya que los estudiantes participantes tenían el doble de probabilidades de **tomar más bebidas azucaradas, comidas rápidas, dulces, bocadillos y tener menos actividad física.** El 26% de los sujetos con sobrepeso y el 4,6% obesos pasaron más de cinco horas utilizando su dispositivo.

Validity: Studies and statements

- When is a statement **valid**?
  - We are sure that the concusions drawn from a study are **accurate** and **precise**

- **Construct** validity
  - Does your variable <u>measure</u> the construct of interest?

- **Internal** validity
  - Can <u>causal</u> conclusions be drawn from the study?

- **External** validity
  - Are the results <u>generalizable</u> to the population of interest?

Construct validity

BBC NEWS MUNDO

Noticias   América Latina   Internacional   Hay Festival   Economía   Ciencia   Salud

Centroamérica Cuenta

Qatar, el país más rico del mundo pero no el más feliz
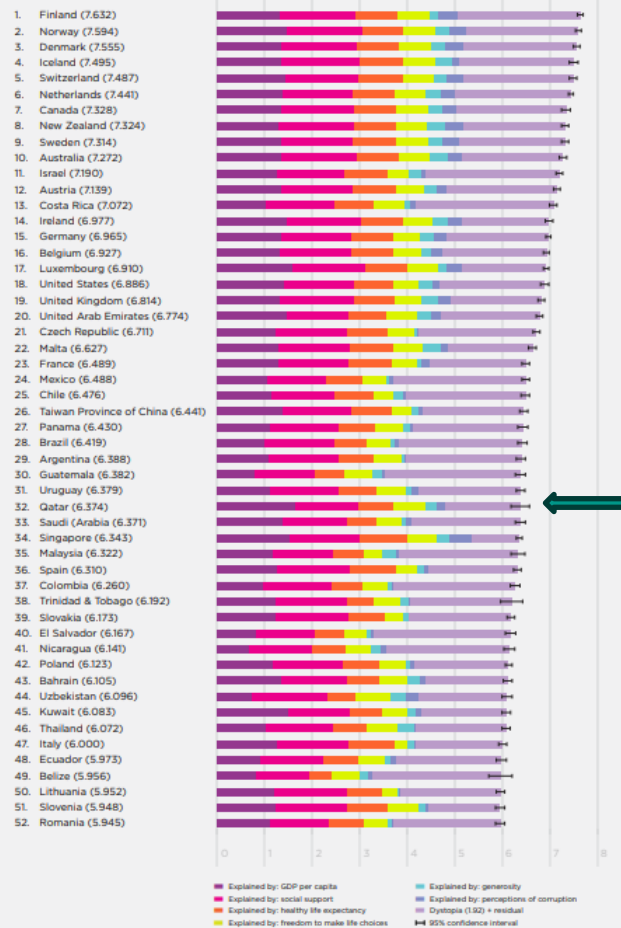
Matthew Teller
Doha, Qatar

29 abril 2014
Actualizado 30 abril 2014

> *Lo importante para mantenernos felices y saludables a lo largo de la vida es la calidad de nuestras relaciones*
>
> Robert Waldinger

Figure 2.2: Ranking of Happiness 2015–2017 (Part 1)

Internal validity

- [The Marshmallow Test](#)

The researchers followed each child for more than 40 years and over and over again, the group who waited patiently for the second marshmallow succeed in whatever capacity they were measuring. In other words, this series of experiments proved that the ability to delay gratification was critical for success in life.

- Kidd C, Palmeri H, Aslin RN. 2013
  - The first group was exposed to a series of unreliable experiences.
  - The second group had very reliable experiences.
    - The children in the unreliable group had no reason to trust that the researchers would bring a second marshmallow and thus they didn't wait very long to eat the first one. (James Clear)

External validity

● [Stanley Milgram Experiment](#) (Stanford students: +60% completed the exp.)
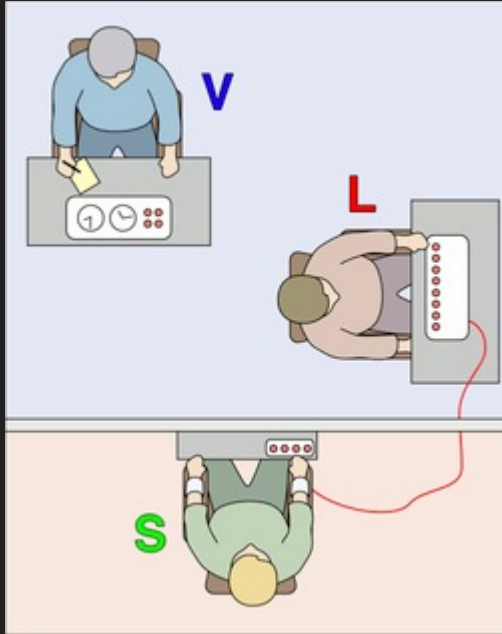
Source: Wikipedia

Some threads to (internal) validity

- Interactions between groups (SUTVA violation)

- Cross-overs (aka. Non-compliance)

- Attrition

- Participation itself affects participants behavior (Hawthorne effect, John Henry effect, etc)

Maximizing validity

"The way that a research study is designed and conducted has a great deal to do with the validity of the conclusions"

Brewer, M. B. (2000)

- **Construct** validity
  - Use different variables for the same construct

- **Internal** validity
  - Try random assignment or look for exogenous variations (RCTs vs quasi-experiments)
  - Show a clear logical path for causation (e.g., DAGs)

- **External** validity
  - Precision in measuring
  - Random sampling

Question: Is random sampling the same as random assignment?

Types of data: Organization

- Cross-sectional data
  - Subjects (people, households, firms, countries, etc.) are observed at a point in time (or time is not considered)
    - Example: Experiments (treatment vs. control)
- Time series
  - Observation over time (e.g. trends and seasonality)
  - Daily, weekly, quarterly, etc.
    - Example: exchange rate and stock prices
- Panel
  - Subjects are observed over time (combines features from cross-sectional and time series)
    - Example: Encuesta de Protección Social [EPS] (Chile)

Types of data: Organization
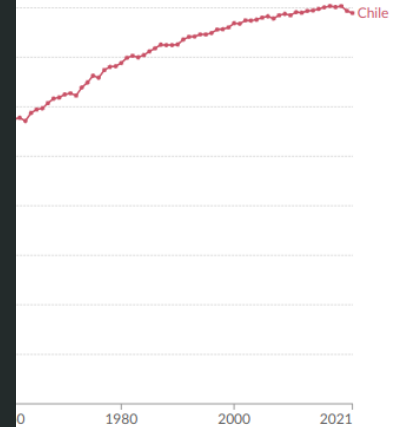
Types of data: Modes of measurement
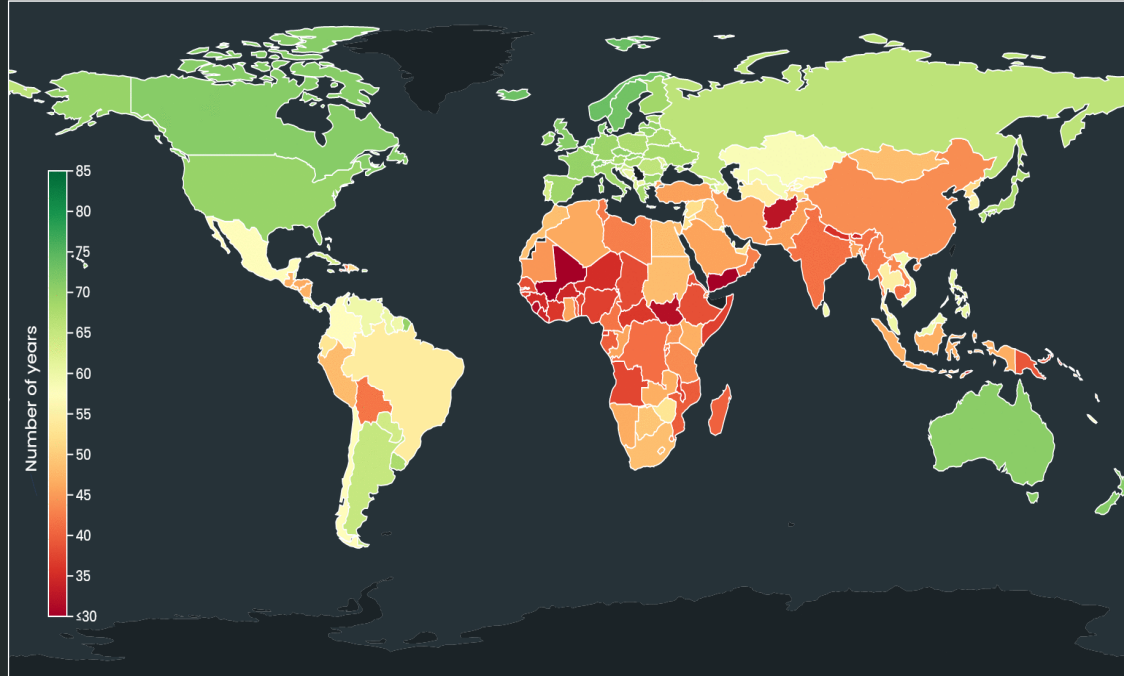
- Qualitative data
    - Interviews
    - Focus groups
    - Mental models (*what is the current journey/routine people are going through*)

- Quantitative data
    - Direct questioning (telephone / face-to-face / online survey)
    - Observation (e.g. video taping – computer vision)
    - Experiments
    - Administrative data

Coming soon

- Next class:
  - EDA
  - Data visualization
  - Ethics in data work

Thank you for coming! 😁