Data Mining

Intro. Data Science

Santiago Alonso-Díaz

Tecnólogico de Monterrey
EGADE, Business School

Photo: Dalle2

Table of contents

Prediction, Inference, Causality (PIC)

PIC

What does P(Rain|Sun) mean?

Correlation is causation?

Is conditional probability causality?

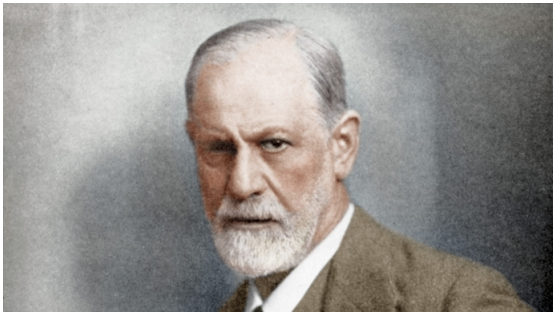
For example, if P(Rain|Sun) = 0, is there causality? Does sun causes no rain?

probability is sometimes non-intuitive

Class activity: Monty hall problem with 4 doors

PIC

Do citations depend on prestige or the other way around?



$$Citations = \beta_1 Prestige + \mu$$

PIC

Do grades depend on prestige or the other way around?

$$Notas = \beta_1 Prestige + \mu$$

PIC

Do commissions depend on prestige or the other way around?

$$Comisiones = \beta_1 Prestige + \mu$$

PIC

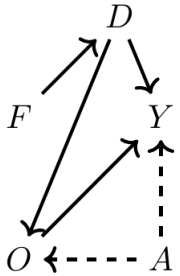
Salary discrimination by gender (Source: Cunningham, 2021)

Google stated that it does not discriminate by gender, once controlling for type of position, hours, and other job characteristics.

But what is the causal model? F: gender; D: Discrimination; Y: income; O: occupation; A: non-observable abilities

PIC

F: gender; D: Discrimination; Y: income; O: occupation; A: non-observable abilities

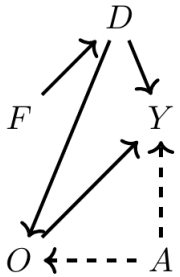


PIC

In the model, discrimination (D) occurs via occupational sorting (O) and that may be affecting (Y).

You have to control for occupation (O) and unobserved skills (A)

Let's look at the code in (Intro_DM_TEC.ipynb; DAG GOOGLE section).



PIC

We saw in the code that without a causal model, Google's occupation control is uninformative. It gives us biased estimators.

The problem of not having clear causal models even leads to paradoxes. Let's look at Simpson's paradox.

PIC

To medicate or not? Paradox: the group effect is different from the individual

Both	Cured	Not cured	Total	Ratio
Drug	20	20	40	**50%**
No drug	16	24	40	40%
	36	44	80	
Man	Cured	Not cured	Total	Ratio
Drug	18	12	30	60%
No drug	7	3	10	**70%**
	25	15	40	
Woman	Cured	Not cured	Total	Ratio
Drug	2	8	10	20%
No drug	9	21	30	**30%**
	11	29	40	

(Pearl & Mackenzie, The Book of Why, 2018)

PIC

A causal understanding solves the paradox

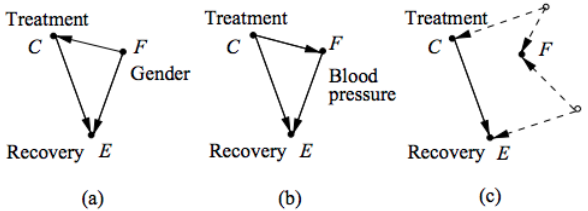


Figure 2: Three causal models capable of generating the data in Fig. 1. Model (a) dictates the use of the gender-specific tables, whereas (b) and (c) dictate use of the combined table.

(Pearl & Mackenzie, The Book of Why, 2018)

PIC

Activity: change in the table man -> psychologist ; woman -> economist ; Drug -> stock selling quota

Discuss in groups (max: 3)

PIC

Another example of the paradox.

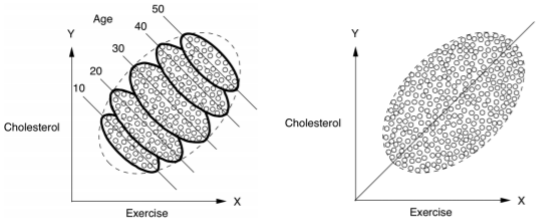


FIGURE 6.6. Simpson's paradox: exercise appears to be beneficial (downward slope) in each age group but harmful (upward slope) in the population as a whole.

(Pearl & Mackenzie, The Book of Why, 2018)

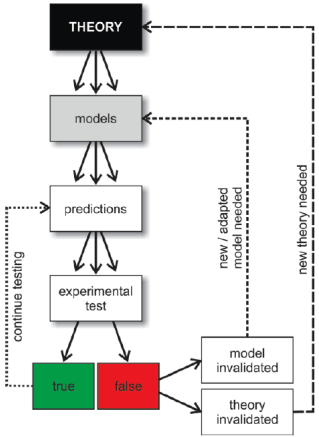
PIC

Another example of the paradox.

Click for thread in X about COVID vaccination

If it does not open or there is no connection, see document X thread by DadosLaplace (Simpsons paradox).pdf

PIC



Blohm et al., 2017

PIC

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Pearl (2018, https://arxiv.org/pdf/1801.04016.pdf)

Parametric and non-parametric approaches

- Parametric (most of this course): we assume the data has a distribution with parameters (e.g. *Mouse Pointer Position* $\sim$ *Normal*($coord_{center}$, 150px))
- Non-parametric: we do not assume the data has a distribution (e.g. sign test for pointer on the left vs right of screen)

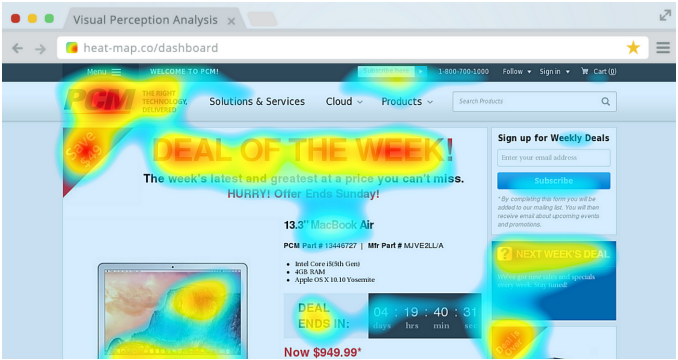


Figure: Source: Jesse Rowe, http://tinyurl.com/keabfe4p

Are weekly sales in our stores larger than the *population* median of 450?

Store	Weekly Sales	Sign (above pop. median = 450)
A	485	+
B	562	+
C	415	-
D	860	+
E	426	-
F	474	+
G	662	+
H	380	-
I	515	+
J	721	+

Anderson, D., Shoesmith, E., Sweeney, D., Anderson, D., & Williams, T. A. (2014). Statistics for business and economics 3e. Cengage Textbooks.

Non-parametric: agnostic on the data distribution

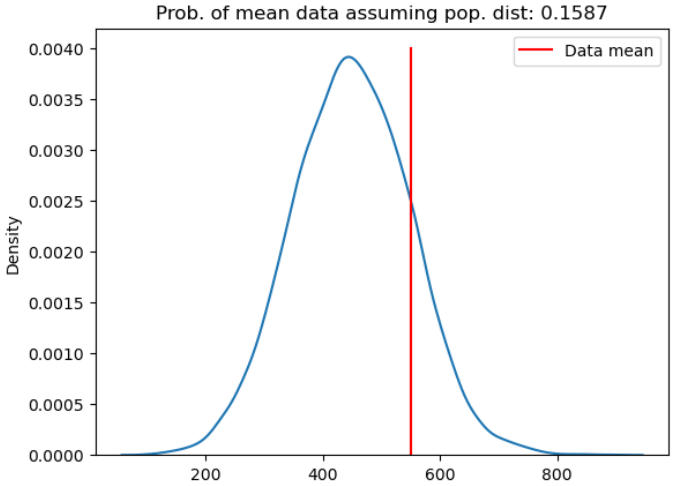
Prob. of positive signs? This table in (Intro_DM_TEC.ipynb; Binomial Probs).

Table: Binomial probabilities with n = 10 and p = 0.5

# of plus signs	Probability
0	.0010
1	.0098
2	.0439
3	.1172
4	.2051
5	.2461
6	.2051
7	.1172
8	.0439
9	.0098
10	.0010

Parametric: assume a data distribution

This graph in (Intro_DM_TEC.ipynb; Normal Probs)



When non-parametric?

- No a priori knowledge of data distribution
- Data does not follow proposed distributions (e.g. non-normal; strong outliers)
- Categorical or qualitative data (prone to counts or ranks)

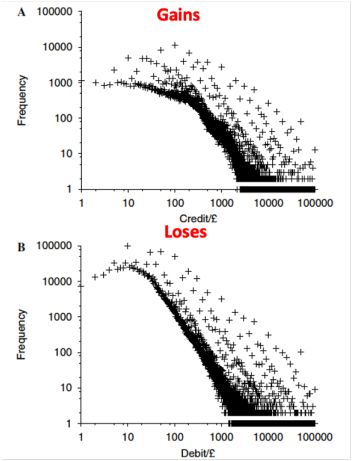
Do variables have distributions?

Zipf mystery (video)

here if it doesn't work

Do variables have distributions?

Yes, of course. Decisions by Sampling paper by Stewart et al (2006, 2019)

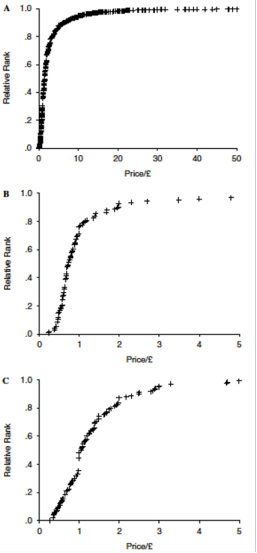


Deposits (gains) and withdrawals (loses) in bank accounts in UK

Power law (e.g. Zipf's law)

Stewart, et al, (2006)

Do variables have prob. distributions?

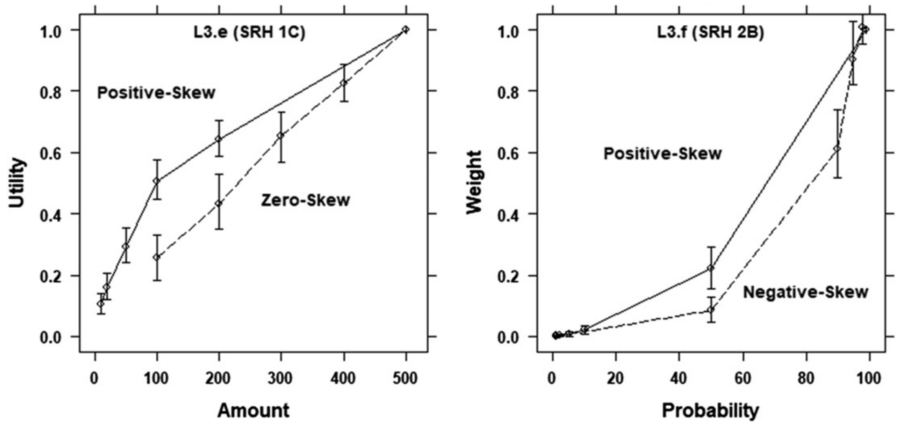


Not just banks

Stewart, et al, (2006)

Do variables have prob. distributions?

And the distributions/contexts change behaviors (utility and probability percepts)?



Alempaki et al., 2019

Do variables have prob. distributions?

Class activity: Benford's Law in first digit of student's documents (also ask friends in Whatsapp or social media) (e.g. ID, credit cards, etc)

Evidence based management

Discuss: evidence-based management and underdetermination of theory by data (UTD)

References

Alempaki, D., Canic, E., Mullett, T. L., Skylark, W. J., Starmer, C., Stewart, N., & Tufano, F. (2019). Reexamining how utility and weighting functions get their shapes: A quasi-adversarial collaboration providing a new interpretation. *Management Science*, *65*(10), 4841–4862.

Blohm, G., Schrater, P., & Kording, K. (2017). Cosmo 2017. *Cosmo*, *1*(1), 1.