Data Mining

Intro. Data Science

Santiago Alonso-Díaz

Tecnólogico de Monterrey
EGADE, Business School

Photo: Dalle2

Data & Business

Using data works for business

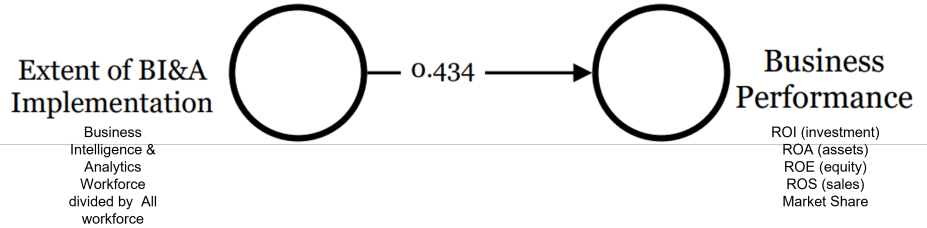


Figure: Daneshvar Kakhki and Palvia, 2016

Data & Business

Using data works for business. Even simple descriptives

Table 4

Meta-analytic results.

n	*N*	*k*	Est. eff.	CI 95 %
Descriptive analytics → firm performance				
14	3,133	20	0.422	0.317–0.517
Heterogeneity: Q-value = 147.078**, *df* (Q) = 13, I^2 = 91 %				
Predictive analytics → firm performance				
63	32,804	90	0.446	0.395–0.494
Heterogeneity: Q-value = 1311.202**, *df* (Q) = 62, I^2 = 95 %				
Prescriptive and autonomous analytics → firm performance				
8	1,833	10	0.504	0.359–0.625

Figure: Oesterreich et al., 2022

Data & Business

Using data works for business. Needs structure.

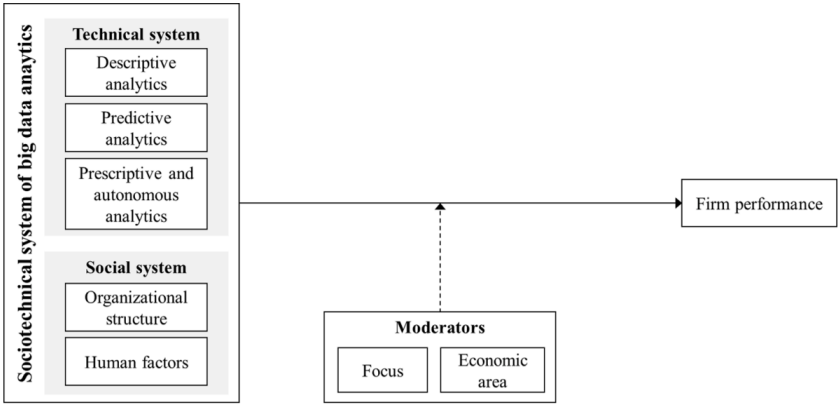


Figure: Oesterreich et al., 2022

Evidence based management

Discuss in class: evidence-based management and underdetermination of theory by data (UTD, ask an LLM or look in Google).

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?

PIC

Imagine a streaming service (Youtube, Netflix, Twitch, Tiktok)

What does P(Engagement|Watch Time) mean?

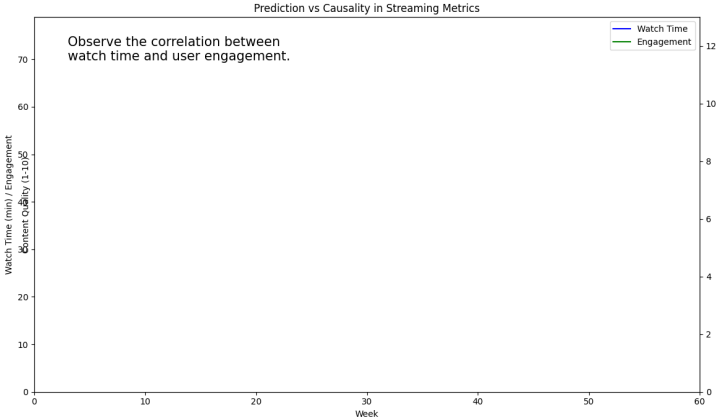
Correlation is causation?

Is conditional probability causality?

For example, if P(Engagement|Watch Time) = 1, is there causality? Does Watch Time causes Engagement?

PIC

Class activity: Youtube case (Cases/Recommendation_Algorithm.docx)



PIC

First assignment:

Write a 1-page report of class discussion + your own views

PIC

Salary discrimination by gender (Source: Cunningham, 2021)

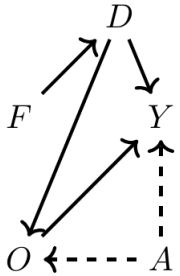
Google stated that it does not discriminate by gender, once controlling for type of position, hours, and other job characteristics.

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

Class activity: Draw in whiteboard

PIC

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

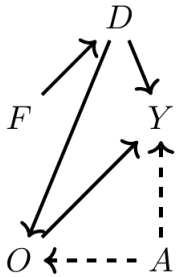


PIC

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 (group $\neq$ subgroup effect).

PIC

Same data, different conclusions!! Solution: what is the causal model, left or right graph? More exercise decreases or increases cholesterol?

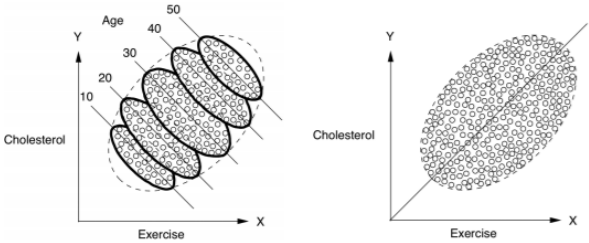


FIGURE 6.6. Simpson's paradox: exercise appears to be beneficial (downward slope) in each age group but harmful (upward slope) in the population as a whole

(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

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)

Evidence based management

Discuss again in class: evidence-based management and underdetermination of theory by data (UTD, ask an LLM or look in Google).

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: does not assume a distribution (e.g. sign test for pointer on the left vs right of screen)

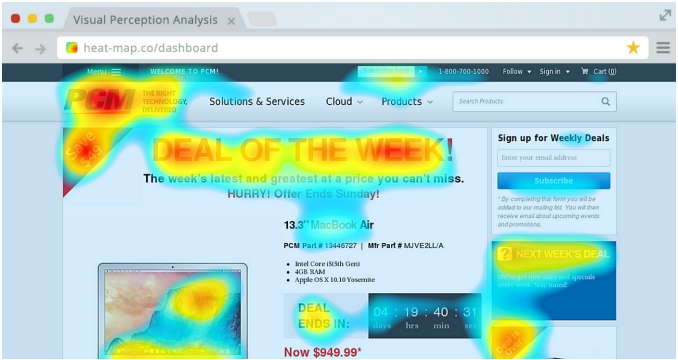


Figure: Source: Jesse Rowe, http://tinyurl.com/keabfe4p

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

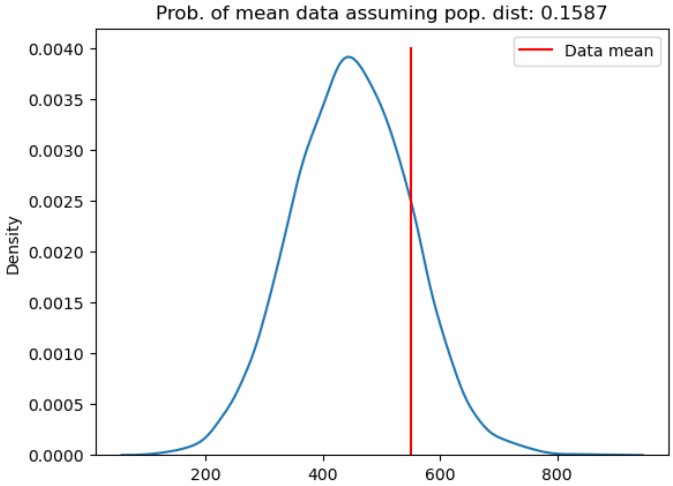
# of plus signs	Probability
0	.0010
1	.0098
2	.0439
3	.1172
4	.2051
5	.2461
6	.2051
7	.1172
8	.0439
9	.0098
10	.0010

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)

Parametric: assume a data distribution

What is x? What is y?



Do variables have prob. distributions?

Class activity: Benford's Law in first digit of documents. Ask GPT to generate random IDs of 9 digits. Then ask it to do an histogram of the first digits.

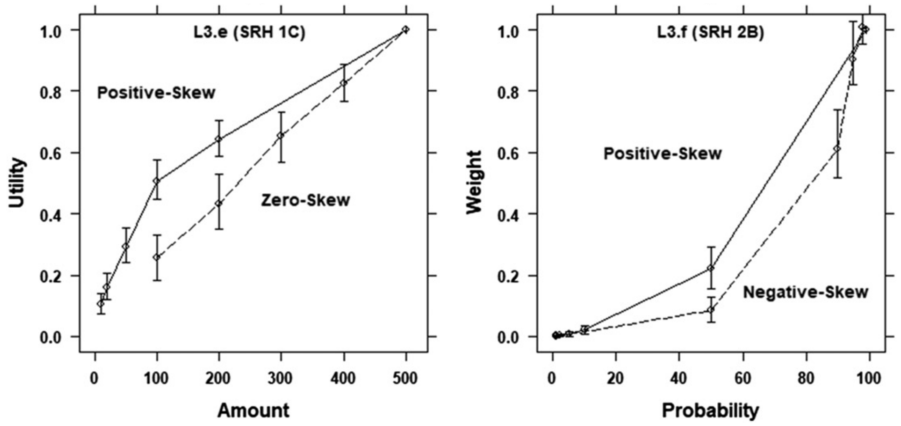
Do variables have distributions?

Zipf mystery (video)

here if it doesn't work

Do variables have prob. distributions?

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



Alempaki et al., 2019

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.

Daneshvar Kakhki, M., & Palvia, P. (2016). Effect of business intelligence and analytics on business performance.

Oesterreich, T. D., Anton, E., Teuteberg, F., & Dwivedi, Y. K. (2022). The role of the social and technical factors in creating business value from big data analytics: A meta-analysis. *Journal of Business Research*, *153*, 128–149.