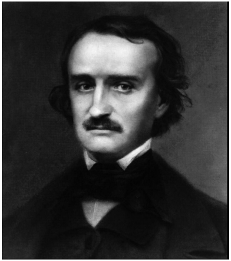
Missing Data & Data Quality

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069
Applied Data Science
for Social Scientists

Spring 2023
Columbia University



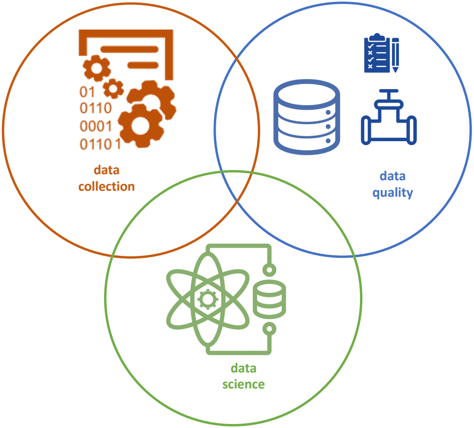
“To observe attentively is to remember distinctly; [...] it is in matters beyond the mere rule that the skill of the analyst is evinced. He makes, in silence, a host of observations and inferences. So, perhaps, do his companions; and the difference in the extent of the information obtained, lies not so much in the validity of the inference as in the quality of the observation. ***The necessary knowledge is of <u>what</u> to observe.***”

The Murders in the Rue Morgue, Edgar Allan Poe (1841)

"You consider that to be important?" [Inspector Gregory] asked.
"Exceedingly so."
"Is there any point to which you would wish to draw my attention?"
"To the curious incident of the dog in the night-time."
"The dog did nothing in the night-time."
*"**That was the curious incident**," remarked Sherlock Holmes.*

Silver Blaze, Sir Arthur Conan Doyle (1892)

A roadmap for this week



Missing Data

Social Science's **secret sauce** for Data Science

- Social Scientists are **critical thinkers** about their data
- concerned — among other things — with **identification problems** in the data
- **identification problems** can have equally **severe consequences** for **inference** and **prediction**
- **identification $\neq$ inference**

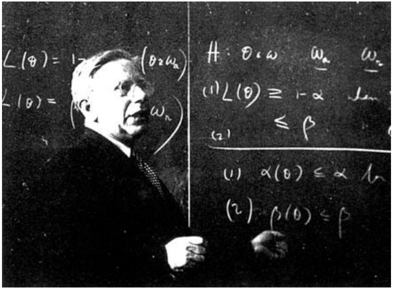
“***Statistical inference*** *seeks to characterize how sampling variability affects the conclusions that can be drawn from samples of limited size.*” — Manski (2003)

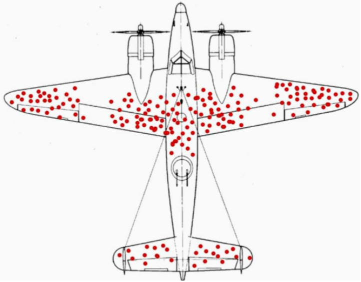
What is **identification**?

> “*Studies of **identification** determine the conclusions that could be drawn if a researcher were able to observe a data sample of unlimited size.*” — Manski (2003)

- key questions to address:
  - **where did the data come from?**
    - not about data-generating mechanisms, but...
    - about how data was “collected”
  - **what “parts” of the data are we not getting?**
    - is my data a “weak sample” of the population?
    - fundamental: do I / can I know what the population is?
  - **how does this affect inference and prediction?**
- understanding **identification** is foundational for **critical thinking** about data

Wald and the Statistical Research Group at Columbia


$L(\theta) = 1-$
$(\theta \varepsilon \omega_a)$
$L(\theta) =$
ω_r
$H: \theta \varepsilon \omega$
ω_a
ω_r
$(1) L(\theta) \geq 1-\alpha$
$\leq \beta$
(2)
$(1) \alpha(\theta) \leq \alpha$
$(2) \beta(\theta) \leq \beta$



identification: what inferences and predictions can be made given the **assumptions** and **data** at hand?

ignoring identification problems: more common than you'd think...

THIS IS WHERE YOU LOST YOUR WALLET?
NO, I LOST IT IN THE PARK. BUT THIS IS WHERE THE LIGHT IS.

a (slightly) more formal view of the problem

Assume a **population** ($n = 10$) for which we have full data (D):

n	Y	X_i	X_j	X_k
unit	**purchases**	**income**	**evaluation**	**education**
1	922	25	3	14
2	340	20	-2	12
3	284	300	0	16
4	189	220	1	20
5	276	10	-1	11
6	922	180	2	13
7	389	50	-3	16
8	741	98	4	8
9	329	131	0	12
10	642	600	1	11

Average 503
$E(Y)$

usually we only observe a “part” of the data

$$D = \begin{bmatrix} 1 & 922 & 25 & 3 & 14 \\ 2 & 340 & 20 & -2 & 12 \\ 3 & 284 & 300 & 0 & 16 \\ 4 & 189 & 220 & 1 & 20 \\ 5 & 276 & 10 & -1 & 11 \\ 6 & 922 & 180 & 2 & 13 \\ 7 & 389 & 50 & -3 & 16 \\ 8 & 741 & 98 & 4 & 8 \\ 9 & 329 & 131 & 0 & 12 \\ 10 & 642 & 600 & 1 & 11 \end{bmatrix} \qquad M = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

where

$D = \{D^{obs}, D^{miss}\}$
D^{miss} = **missing** data
D^{obs} = **observed** data
$M = \{1, 0\}$: missingness vector

not too problematic when the sample is **random**

- if the “sample” — governed by M — is random $\implies E(Y) = E(Y^{obs})$

n	Y	X_i	X_j	X_k	M
1	?	?	?	?	1
2	?	?	?	?	1
3	?	?	?	?	1
4	189	220	1	20	0
5	276	10	-1	11	0
6	922	180	2	13	0
7	389	50	-3	16	0
8	741	98	4	8	0
9	?	?	?	?	1
10	?	?	?	?	1
$\bar{y}$	503				

- random in the sense that $P(M|D) = P(M)$

very problematic when the sample is **not random**

- if the “sample” only includes wealthy subscribers (income (X_i) >100) who self-selected to use our expensive service $\Longrightarrow E(Y) \neq E(Y^{obs})$

n	Y	X_i	X_j	X_k	M
1	?	?	?	?	1
2	?	?	?	?	1
3	284	300	0	16	0
4	189	220	1	20	0
5	?	?	?	?	1
6	922	180	2	13	0
7	?	?	?	?	1
8	?	?	?	?	1
9	329	131	0	12	0
10	642	600	1	11	0
$\bar{y}$	473				

- non-random in the sense that $P(M|D) \neq P(M)$

A crucial question for identification: **what governs data missingness**?

- Y^{obs} is a random sample if the data — observed (D^{obs}) or unobserved (D^{miss}) — does not inform missingness (M)

$$P(M|D) = P(M|D^{obs}, D^{miss}) = P(M)$$

- when that holds

$$E(Y) = E(Y^{obs}) = E(Y^{miss})$$

- ...and Y^{obs} can be safely used to infer/predict on the whole population since it is “equivalent” to the unobserved portion (Y^{miss})

why do we worry about identification?

..and data missingness?

- **Fundamental concerns:**
  - using only **available** information but not all **possible** information
  - observed distributions $\neq$ true distrbutions because of **missingness mechanism**
- **Empirical concerns:**
  - most algorithms **assume no missingness** in the data and “handle” it in different ways
- **Consequences:**
  - **valid inferences/predictions** for the **wrong population**
  - **invalid inferences/predictions** for **unknown population segments**
  - **underestimated variances** (relevant on inferential problems)

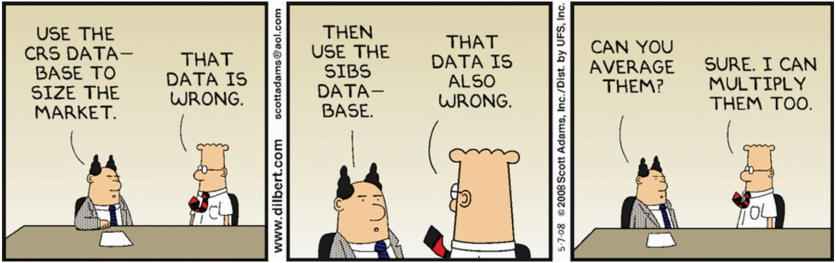
it only gets more complicated with M as a matrix...

$$D = \begin{bmatrix} 1 & 33 & 25 & 3 & 14 \\ 2 & 22 & 20 & -2 & 12 \\ 3 & 50 & 300 & 0 & 16 \\ 4 & 30 & 220 & 1 & 20 \\ 5 & 18 & 10 & -1 & 11 \\ 6 & 45 & 180 & 2 & 13 \\ 7 & 76 & 50 & -3 & 16 \\ 8 & 29 & 98 & 2 & 14 \end{bmatrix} \quad M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

where

$D = \{D^{obs}, D^{miss}\}$
D^{miss} = **missing** data
D^{obs} = **observed** data
$M = \{1, 0\}$: missingness matrix

a comment on identification in big data contexts


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.
www.dilbert.com
scottadams@aol.com
5·7·08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.

a comment on identification in big data contexts

- **belief:** missingness becomes less relevant as size tends towards **big data**
  - belief that **asymptotics** kick in and solve everything
  - belief that **large samples** are, by definition, unbiased
  - belief that **implementations of common algorithms** handle missing data natively (and thus appropriately)
- **problem:** missingness may be generating a **biased sample** of observed data... regardless of size
- **consequences:**
  - training and testing sets $\neq$ general population
  - changes in missingness parameters “change” data when true distributions remain unchanged

WELCOME
TO Fabulous
Asymptopia

hello Large Sample Distribution Theory!

- **limiting distributions and convergence in probablilty:**

  $$\lim_{n\to\infty} Prob(|x_n - x| > \varepsilon) = 0 \quad \forall\, \varepsilon > 0$$

  - **intuition:** as n increases, values that are not close to x become extremely unlikely

- the assumption that **any (big data) sample** converges to *the* population implies that

  $$\lim_{n\to\infty} Prob(|\{Y|M\}_n - Y| > \varepsilon) = 0$$

  - this would only happen when **missingness is ignorable**

    $$P(M|D) = P(M)$$

  - correction is known to be possible if data missingness mechanism and/or true population are known

Data Quality

why is data quality important?

why is data quality important?

“Why is the quality of data - the most important element in a data warehouse -so often ignored until it is too late? Is it because it is too complex to solve, or is it simply not our business? Answering this fundamental question is tough, but part of the answer may be the lack of solid methodology to deal with data quality. ”

Integrating Data Quality into Your Data Warehouse Architecture, Jean-Pierre Dijcks

think big, start small

- define where your data is going to be used
- identify the potential impact of missing data and wrong data
- identify the potential stakeholders to validate assumptions of missing data
- define the subject and elements of your quality checks
- define your methodology in handling data quality
- **remember**: data quality is a difficult thing to measure
- **remember**: the issue could be from the source

what is the reality?

- **first steps:**
  - explore your data
  - identify your most important columns or attributes
  - explore your most important columns and attributes
- **next steps:**
  - document the quality issues identified
  - review with data owners and stakeholders
  - identify methodology to tackle the quality issues
  - select the technology/algorithm/code to solve the quality issues
  - implement the fix for data quality issues
  - document the fix and columns that have been resolved or affected

types of data quality checks

- **Lack of integrity constraints**
  - Rule that defines the consistency of a given data or dataset in the database (e.g., Primary key, uniqueness)
  - **Example of uniqueness violation:** Two customers having the same SSN number customer 1= (name="John", SSN="12663") , customer 2= (name="Jane", SSN="12663")
- **Poor schema design**
  - Imperfect schema level definition
  - **Example 1:** Attributes names are not significant: FN stands for First Name and Add stands for Address
  - **Example 2:** Source without schema: "John;Doe;jd@gmail.com;USA"
- **Embedded values**
  - Multiple values entered in one attribute
  - **Example:** name="John D. Tunisia Freedom 32".

types of data quality checks

- **Duplicate records**
  - Data is repeated. Misspellings, different ways of writing names and even address changes over time can all lead to duplicate entries
  - Another form of duplication is the conflicts of entities when inserting a record having the same id as an existing record
- **Missing values**
  - Data in one field appears to be null or empty
- **Variety of data types**
  - Different data types between the source and the target schema
- **Naming conflicts**
  - If we have two data sources which have two synonymous attributes (e.g., gender/sex) then the union of the aforementioned sources requires schema recognition

types of data quality checks

- **Syntax inconsistency (Structural conflicts)**
  - There are a different syntactic representations of attributes whose type is the same
  - **Example 1:** French date format (i.e., dd/mm/yyyy) is different from that of the US format (i.e., mm/dd/yyyy)
  - **Example 2:** Gender attribute is represented differently in the two data sources, e.g., 0/1, F/M.
- **Wrong mapping of data**
  - Linking a data source to the wrong destination results in the spread of wrong data
- **Variety of data types**
  - Different data types between the source and the target schema
- **Naming conflicts**
  - If we have two data sources which have two synonymous attributes (e.g., gender/sex) then the union of the aforementioned sources requires schema recognition

Missing Data & Data Quality

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069
Applied Data Science
for Social Scientists

Spring 2023
Columbia University