Missing _ 2: Missing Harder

Data Science, Spring 2021

Before we start...

Before we start...

1. Our mod of the day.

Before we start...

1. Our mod of the day.
2. Projects

Our moderator

Our moderator

1. None. Be gentle.

Project 1 discussion

Project 1 discussion

1. Deadlines, not livelines.

Project 2

[1] Notice the timezone

Project 2

1. It's out.

[1] Notice the timezone

Project 2

1. It's out.
2. Due March 26th, 2021

[1] Notice the timezone

Project 2

1. It's out.
2. Due March 26th, 2021
3. Nothing will be accepted after 11:59pm EDT[1]

[1] Notice the timezone

Missing Data

There are three main ‘types’ of missing data

Missing Data

There are three main ‘types’ of missing data

1. Missing Completely at Random (MCAR)

Missing Data

There are three main 'types' of missing data

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)

Missing Data

There are three main 'types' of missing data

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing Not at Random (MNAR)

Imagine a world where we could loiter outside CS

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades and you start asking for CMSC131 grades:

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades and you start asking for CMSC131 grades:

1. You use no coercion

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades and you start asking for CMSC131 grades:

1. You use no coercion
2. You write down their ...

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades and you start asking for CMSC131 grades:

1. You use no coercion
2. You write down their ...
    2.1 Response

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades and you start asking for CMSC131 grades:

1. You use no coercion
2. You write down their ...
   - 2.1 Response
   - 2.2 Height (>6ft or not)

Imagine a world where we could loiter outside CS

Students just got their mid-semester grades and you start asking for CMSC131 grades:

1. You use no coercion
2. You write down their ...
   - 2.1 Response
   - 2.2 Height (>6ft or not)
   - 2.3 Their hair color

You get this:

Hair	Tall?	Grade
Red	Y	A
Brown	N	A
Black	N	B
Black	Y	A
Brown	Y	
Brown	Y	
Brown	N	
Black	Y	B
Black	Y	B
Brown	N	A
Black	N	
Brown	N	C
Red	N	A
Brown	Y	A
Black	Y	A

Missing data matrix:

Hair	Tall?	Grade
0	0	0
0	0	0
0	0	0
0	0	0
0	0	1
0	0	1
0	0	1
0	0	0
0	0	0
0	0	0
0	0	1
0	0	0
0	0	0
0	0	0
0	0	0

MCAR

The probability of having a ‘1‘ in a column is not dependent on any of the data, observed or unobserved.

MCAR

The probability of having a ‘1‘ in a column is not dependent on any of the data, observed or unobserved.
You can test this using conditional probability for the observed data.

MCAR

The probability of having a ‘1‘ in a column is not dependent on any of the data, observed or unobserved.

But how realistic is this?

MCAR: Not very realistic

When it comes to data we've collected, especially from people, completely random mechanisms are very unlikely:

MCAR: Not very realistic

When it comes to data we've collected, especially from people, completely random mechanisms are very unlikely:

1. Older folks less like to answer questions online

MCAR: Not very realistic

When it comes to data we've collected, especially from people, completely random mechanisms are very unlikely:

1. Older folks less like to answer questions online
2. In long-term studies of non-trivial sample sizes, people will die before study is complete

MCAR: Not very realistic

When it comes to data we've collected, especially from people, completely random mechanisms are very unlikely:

1. Older folks less like to answer questions online
2. In long-term studies of non-trivial sample sizes, people will die before study is complete
3. People/Institutions are often reluctant to reveal financial information

MAR

The probability of having a ‘1‘ in a column is dependent on observed data.

MAR

The probability of having a ‘1‘ in a column is dependent on observed data.
This allows us to model the mechanism for when the data is missing.

MAR

The probability of having a ‘1‘ in a column is dependent on observed data.

We use observed data as input into our model.

MAR: Pretty realistic

Can you think of examples?

MAR: Pretty realistic

Can you think of examples?

Because we can model for it, we can compensate for it!

MNAR

The probability of having a ‘1‘ in a column is dependent on unobserved data.

MNAR

The probability of having a ‘1‘ in a column is dependent on unobserved data.
You can’t ignore this.

MNAR

The probability of having a ‘1‘ in a column is dependent on unobserved data.

Any analysis including MNAR data must model and guess what the missing data is, otherwise what’s the point?

MNAR

Can you think of some examples?

MAR vs MNAR

I've got some bad news.

MAR vs MNAR

I've got some bad news.

1. Whether something is MAR or MNAR is not testable without getting the missing data.

MAR vs MNAR

I've got some bad news.

1. Whether something is MAR or MNAR is not testable without getting the missing data.
2. You have to understand your data.

Back to our questions

Hair	Tall?	Grade
Red	Y	A
Brown	N	A
Black	N	B
Black	Y	A
Brown	Y	
Brown	Y	
Brown	N	
Black	Y	B
Black	Y	B
Brown	N	A
Black	N	
Brown	N	C
Red	N	A
Brown	Y	A
Black	Y	A

Get some more data

Hair	GPA	Tall?	Grade
Red	3.4	Y	A
Brown	3.6	N	A
Black	3.7	N	B
Black	3.9	Y	A
Brown	2.5	Y	
Brown	3.2	Y	
Brown	3.0	N	
Black	2.9	Y	B
Black	3.3	Y	B
Brown	4.0	N	A
Black	3.65	N	
Brown	3.4	N	C
Red	2.2	N	A
Brown	3.8	Y	A
Black	3.67	Y	A

Thanks for your time!

:)