How To Handle Missing Data

Matt Arthur

University of California, Riverside

August 17, 2022

Objectives and Scope

Big Picture

1. Have an experiment and collect data.
2. Given data, make inference on an underlying aspect of the population

Problem: Not all the data were collected, or were available.

Objectives and Scope

Types of Missing Data

- Non- or partial response in sample surveys.
- Dropout and noncompliance in clinical trials.
- Surrogate measurements and missingness by design.

Non or Partial Response in Surveys

- Conduct a survey
- Non-Response: Do not get responses back
- Partial Response: ‘Decline to state’, ‘Don’t know’, etc.

Dropout and Noncompliance

- A clinical trial with two treatments.
- Subjects are recruited and are randomly assigned to one of the treatments.
- Subjects return to clinic weekly for check ups.
- Some subjects may fail to show up for any clinic visit beyond a certain point, "dropping out" of the study. Some miss clinic visits sporadically.

Surrogate Measurements and Missingness by Design

- Missingness is deliberate.
- Some studies might be costly or burdensome for subjects. Separate subjects into a surrogate group, and a validation group.
- Check surrogate measurements with validation group.

Objectives and Scope

Problem

- The objective was to carry out inference in on the full data.
- Without the full data, the analysis maybe be compromised
  - May have bias in parameter estimates and standard errors
  - Variability could be underestimated
  - Can lead to inefficient use of the data
  - Can inflate Type 1 and Type 2 errors

Objectives and Scope

- One possible approach is to just ignore the problem and analyze the observed data that were collected as if they were the intended, full data.
- The can lead to misleading conclusions in many situations.
- Statistical methods are required to attempt to correct (somehow) for most missing data problems.
- Although missing data has been around a long time, methods to handle missing data are fairly new.

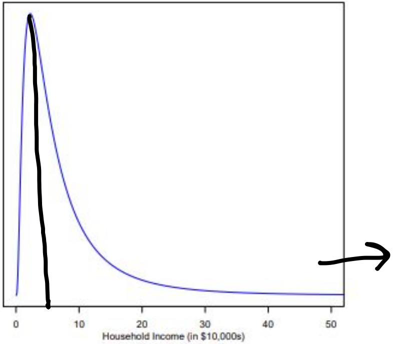
Objectives and Scope

Objectives

- Formal framework for thinking about missing data, associated terminology.
- Naive methods and their drawbacks
- Methods for drawing valid inferences in the presences of missing data under certain assumptions.

Example: Household Income

- Suppose a survey is conducted to learn about the financial status of households in the US population.
- A random sample of 1,000 individuals selected.
- Many participants fail to answer some (or all) of the survey questions.
- Plot below shows the true density, which we do not know in practice.



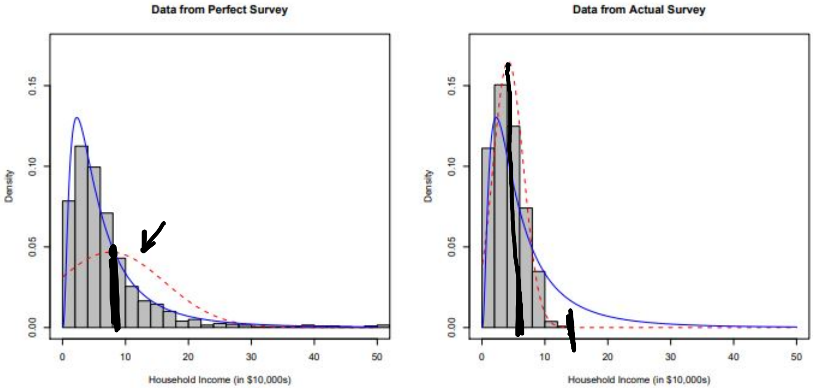
Example: Household Income

We only had 661 responses (out of 1000) regarding household income. We have no way of knowing if this sample is representative. So we must make an assumption about the missing data.

1. Suppose we assume there is no difference between people who respond and do not respond to the question.
   - We thus would have a random sample of 661 people that is still representative.
2. Suppose we assume that wealthy people were less likely to answer the household income question.
   - Sample is no longer representative. Must be careful with inference.

(We will limit ourselves to these two possible assumptions, although many other assumptions are possible)

Example: Household Income



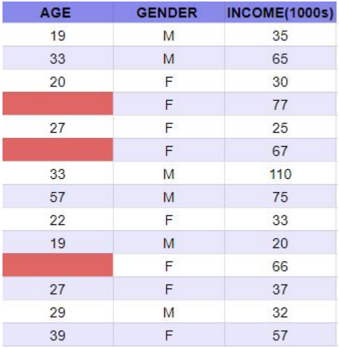
Left panel: Approximate histogram under assumption 1.

Right panel: Approximate histogram under assumption 2.

Red line is normal approximation of the density. Blue line is the true density.

Example: Salary Study

- Interested in salary of general adult population.
- Want to compare salary between males and females.
- Only sampled 14 people, some missing values occurred.



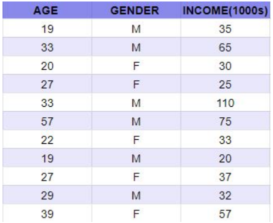
AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
🟥	F	77
27	F	25
🟥	F	67
33	M	110
57	M	75
22	F	33
19	M	20
🟥	F	66
27	F	37
29	M	32
39	F	57

Example: Salary Study

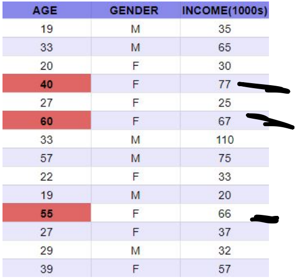
Suppose again that we limit ourselves to two different assumptions (many more are possible)

1. Suppose we assume there is no difference between people who respond and do not respond to the question.
2. Suppose we assume women are less likely to report their age.
   - Sample is no longer representative. Must be careful with inference. (Use educated guesses from a similar, more extensive study).

Example: Salary Study



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
27	F	25
33	M	110
57	M	75
22	F	33
19	M	20
27	F	37
29	M	32
39	F	57



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
40	F	77
27	F	25
60	F	67
33	M	110
57	M	75
22	F	33
19	M	20
55	F	66
27	F	37
29	M	32
39	F	57

Under assumption 1 (left data), the average salary for males is 56.2, and females is 36.4.

Under assumption 2 (right data) is that the average salary for males is 56.2, and females is 49.

Summary of Examples

These examples helps illustrate that

- Different assumptions about non-response will lead to different inferences.
- Assumptions about non-response are unverifiable from the data alone.
- Standard technique of deleting missing observations blindly can be problematic.

Assumptions

The assumptions we make for why some of our data is missing is critical. Any assumption that is made can fall into one of three classes of assumptions. Our analysis is dependent on what we assume.

- MCAR: Missing completely at random
- MAR: Missing at random
- MNAR: Missing not at random

MCAR: Missing completely at random

- One value is just as likely to be missing as another
- No relationship between the missing data and the other measured variables
- Probability for missing data is the same across units – considered “ignorable”

Examples:

- Someone does not report income level because they accidentally skipped a line on the survey.
- Someone moved before a health study was completed.

MAR: Missing at random

- Extent that missingness is correlated with other variables that are included in the analysis
- Allows missing data to depend on things that are observed, but not on things that are not observed

Examples:

- Women are less likely to report their weight on a survey.
- Low income people less likely to report their education level.

MNAR: Missing not at random

- Likelihood of a piece of data being missing is related to the value that would have been observed
- Most problematic type – considered nonignorable missing data

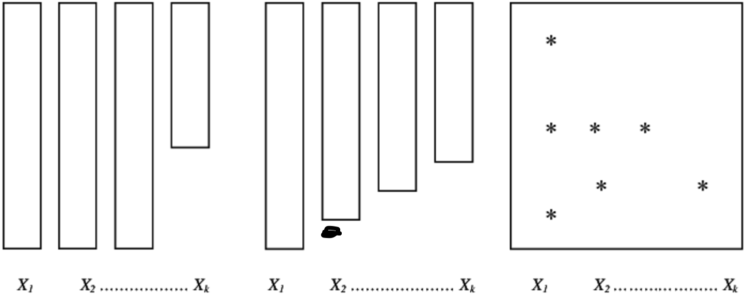
Examples:

- Wealthy households less likely to report their income.
- Students who struggle with division more likely to skip division problems.

Tools to Help Determine Missingness Relationship

- In practice, we do not know what type of missingness we have.
- There is almost no way to verify if our assumptions are correct.
- Must use context of the problem and tools to determine what type of missingness we have.

Tools to Help Determine Missingness Relationship

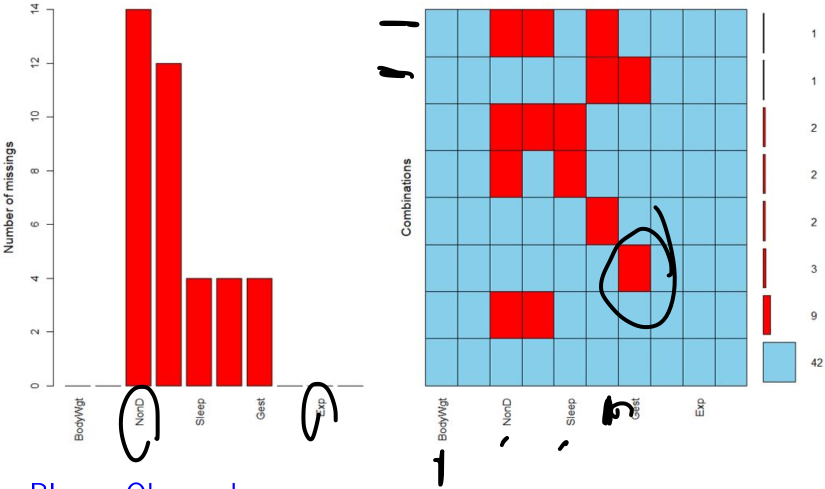

X_1
X_2 .... X_k
X_1
X_2 .... X_k
X_1
X_2 .... X_k
*
* * *
* *
*

(a) Univariate Pattern

(b) Monotone Pattern

(c) Arbitrary Pattern

Tools to Help Determine Missingness Relationship



Red = Missing, Blue = Observed

General Framework

- We let Z denote the full data intended to be collected on each individual, which can be partitioned into K components.

$f_z(z;\theta)$

$$\underline{Z = (Z_1, Z_2, ..., Z_K)}$$

- We let R denote a vector of indicators that corresponds to each component Z_j

$$R = (R_1, R_2, ..., R_K)$$

$$R_j = \begin{bmatrix} 1 & Z_j \text{ observed} \\ 0 & \text{otherwise} \end{bmatrix}$$

- Any components in Z can be missing or observed.

Methods to Address Missing Data

- Naive Methods
    - Deletion (Listwise and Pairwise)
    - Average Imputation
    - Hot-Deck Imputation
    - Regression Substitution
- Observed Data Likelihood
- Multiple Imputation

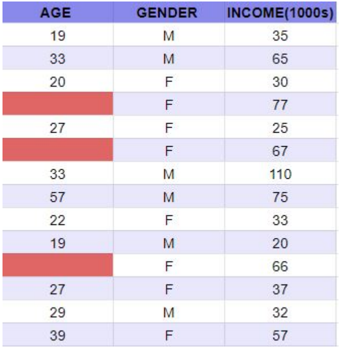
Deletion (Listwise vs Pairwise)

Delete observations that contain missing values.

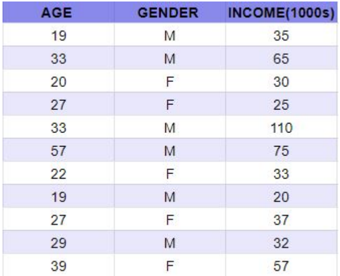
- **Pros**
  - Complete removal of data with missing values results in robust and highly accurate model
  - Deleting a particular row or a column with no specific information is better, since it does not have a high weight.
- **Cons**
  - Loss of information and data
  - Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset
  - Need a enough full observations for a good estimate.
  - Observations must be missing completely at random (MCAR).

Deletion - Example

Recall our previous salary example. Red cells are missing values. Second graph uses deletion.



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
🟥	F	77
27	F	25
🟥	F	67
33	M	110
57	M	75
22	F	33
19	M	20
🟥	F	66
27	F	37
29	M	32
39	F	57



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
27	F	25
33	M	110
57	M	75
22	F	33
19	M	20
27	F	37
29	M	32
39	F	57

Average Imputation

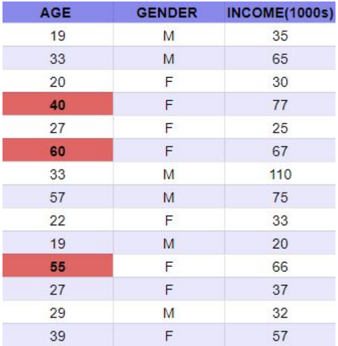
$$\left\{ \overline{(x_1 - \bar{x}) + (x_2 - \bar{x}) \cdots + (x_n - \bar{x}}) \right\} / n-1$$

Replace missing observations with the average (mean, median, mode) value for that variable.

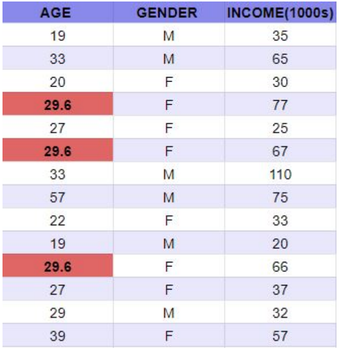
- **Pros**
    - This is a better approach when the data size is small.
    - It can prevent data loss which results in removal of the rows and columns.
    - Can still result in unbiased estimates.
- **Cons**
    - Decreases variance in data.
    - Works poorly compared to other methods.
    - Estimates statistically invalid.

Average - Example

Recall our previous salary example. Red cells are missing values. Second graph uses average imputation.



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
40	F	77
27	F	25
60	F	67
33	M	110
57	M	75
22	F	33
19	M	20
55	F	66
27	F	37
29	M	32
39	F	57



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
29.6	F	77
27	F	25
29.6	F	67
33	M	110
57	M	75
22	F	33
19	M	20
29.6	F	66
27	F	37
29	M	32
39	F	57

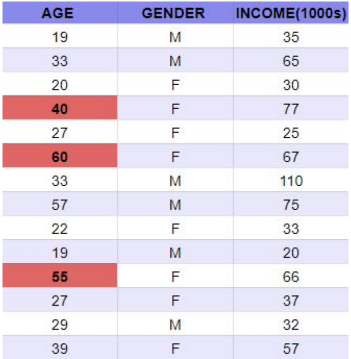
Hot Deck Imputation

Missing data point is filled in with a value from a similar observation in the current data set – also known as “matching”

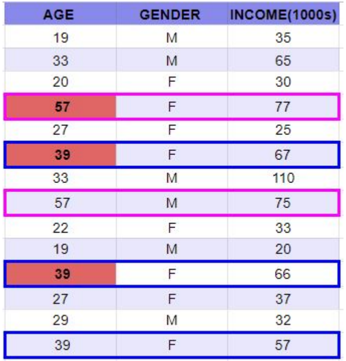
- **Pros**
  - This is a better approach when the data size is small.
  - It can prevent data loss which results in removal of the rows and columns.
  - Can still result in unbiased estimates.
- **Cons**
  - Decreases variance in data.
  - Becomes much more difficult as variables with missing data increase
  - Estimates statistically invalid.

Hot Deck Imputation - Example

Recall our previous salary example. Red cells are missing values, with the true unknown values filled in. Second graph uses hot deck imputation.



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
40	F	77
27	F	25
60	F	67
33	M	110
57	M	75
22	F	33
19	M	20
55	F	66
27	F	37
29	M	32
39	F	57



AGE	GENDER	INCOME(1000s)
19	M	35
33	M	65
20	F	30
57	F	77
27	F	25
39	F	67
33	M	110
57	M	75
22	F	33
19	M	20
39	F	66
27	F	37
29	M	32
39	F	57

The true average age for females is 36.25, average age under hot deck imputation is 33.75, under deletion is 27, under mean imputation is 28.

Regression Substitution

Replace missing values with predicted values from a regression equation. Use complete case data to make equation, then impute predicted values.

- **Pros**
  - Uses more information then previous data.
  - Gives better estimates.
- **Cons**
  - Correlates estimates
  - Decreases variability in data

Regression Substitution - Example

$$Age = \underbrace{\beta_0 + \beta_{GenderF}X_{GenderF} + \beta_{Salary}X_{Salary}}$$

Observed Data Likelihood

$$\boxed{f_Z(z_i;\theta)}$$

$$R_i = \begin{cases} 1 & \text{if } Z_i \text{ obs} \\ 0 & \text{o.w.} \end{cases}$$

Notation:

$$\overbrace{Z = (Z_1, \dots, Z_k)} \quad \overbrace{R = (R_1, \dots, R_k)}$$

Let $Z_{(r)}$ denote the subset of Z observed when $R = r$, Then the observed data are:

$$\boxed{(R, Z_{(R)})}$$

$\left[\text{where } Z_{(R)} = \sum_r \mathbb{I}(R = r) Z_{(r)}\right]$

$$K = 4 \quad Z = (Z_1, Z_2, Z_3, Z_4)$$

$$\left[\begin{array}{l} R = (1,0,1,0) \\ Z_{(R)} = (Z_1, Z_3) \end{array}\right]$$

Observed Data Likelihood

$(R, Z_{(R)})$ want $f_{R, Z_{(R)}}(r, z_r ; \theta)$

$$f_{R,Z}(r, z) = \underbrace{f_{R|Z}(r|z;\psi)}_{\text{missingness mechanism}} \overbrace{f_Z(z;\theta)}^{\text{full data density}}$$

$$f_{R, Z_{(R)}}(r, z_r; \psi, \theta) = \int \underbrace{f_{R|Z}(r|(z_r, z_{r'})\overset{\psi}{\downarrow})}_{\circledast} \underbrace{f_Z((z_r, z_{r'});\theta)}\, dz_{r'}$$

under MAR: $f_{R|Z} = f_{R|Z_r}$

$$\hookrightarrow f_{R, Z_{(R)}}(r, z_r, \psi, \theta) = \boxed{f_{R|Z_r}(r|z_r;\psi)} f_{Z_r}(z_r;\theta)$$

Observed Data Likelihood

$$L_i(\theta) = \prod_r \left[f_{Z_r}\left(z_r^{(i)}; \theta\right) \right]^{I(R_i = r)}$$

$$L(\theta) = \prod_{i=1}^{n} \prod_r \left[\underbrace{f_{Z_r}\left(z_r^{(i)}; \theta\right)} \right]^{I(R_i = r)}$$

Observed Data Likelihood

Observed Data Likelihood

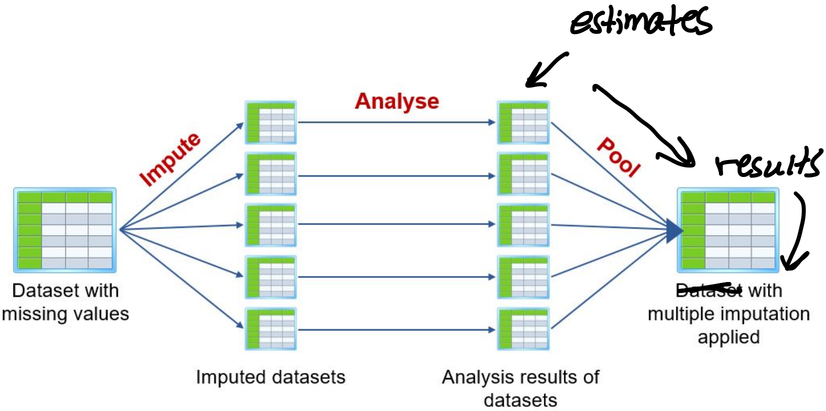
Observed Data Likelihood

Multiple Imputation

Each missing value is replaced with multiple plausible values. In general, we want 3-5 imputed values for each missing observation.

- **Pros**
  - Having multiple values reduces bias by addressing the uncertainty
  - Statistically valid (Woo!)
- **Cons**
  - Difficult to implement.
  - Only works when data are MAR.

Multiple Imputation



estimates

results

Impute

Analyse

Pool

Dataset with missing values

Imputed datasets

Analysis results of datasets

~~Dataset~~ with multiple imputation applied

Multiple Imputation

How do we get the multiple imputed values?

- Multiple different regression imputations
- Random For~~r~~est
- Etc.

Multiple Imputation

More formally, we sample from the distribution of the full data *given* the observed data:

$$f_{Z(R,Z_{(R)})}(z \mid r, z_{(r)}) = \underline{f_{Z|Z_{(r)}}(z \mid z_{(r)}, \theta)}$$

Can do this using a bayesian or frequentist approach.

- With a frequentist approach, we need an initial value for θ.

$f_{z|z_r} =$

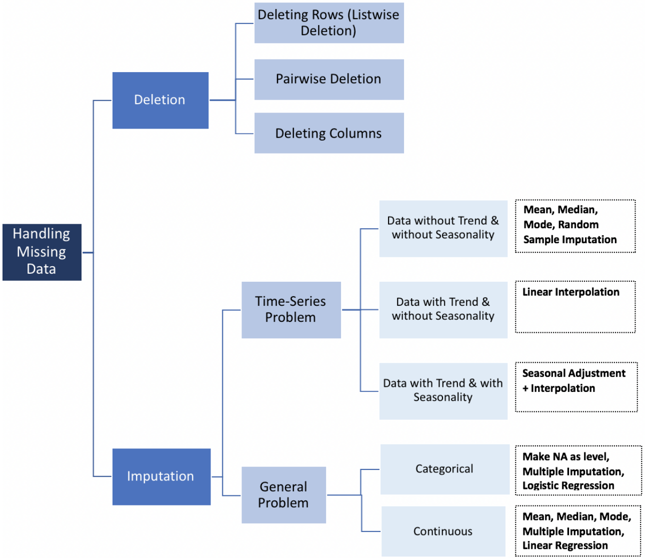
Softward for Multiple Imputation

- R: MICE, Amelia, missForrest, Hmisc, mi
- SPSS
- STATA

General Steps

- Investigate the data set. Where are your missing data? How much? What is the context of where your data came from? Are there any important noteworthy patterns?
- Make an assumption about the missingness mechanism. Are the data MCAR, MAR, or MNAR. If MCAR and the data set is large you can usually simply drop observations that have missing values. If MAR, you can usually conduct a multiple imputation procedure. If MNAR, investigate more advanced methods.
- If imputed data sets have been created, investigate if they seem plausible.
- If the imputed data sets seem plausible, conduct your statistical procedure on each imputed data set.
- Suppose you have 'm' imputed data sets. Thus you have 'm' imputed results. Pool these results for your final value you that you report.
- If appropriate, conduct further sensitivity analysis by investigating extreme situations.

Warning



Handling Missing Data

Deletion

Deleting Rows (Listwise Deletion)

Pairwise Deletion

Deleting Columns

Imputation

Time-Series Problem

Data without Trend & without Seasonality

Mean, Median, Mode, Random Sample Imputation

Data with Trend & without Seasonality

Linear Interpolation

Data with Trend & with Seasonality

Seasonal Adjustment + Interpolation

General Problem

Categorical

Make NA as level, Multiple Imputation, Logistic Regression

Continuous

Mean, Median, Mode, Multiple Imputation, Linear Regression

Sensitivity Analysis

- After running your imputed data you should conduct sensitivity analysis.
- **Sensitivity analysis** is the study of how the uncertainty in the output of a mathematical model or system.
- That is, we should see how well our model preforms under extreme cases.

Example - Data Context

- **Titanic Data**
- We have 1309 passengers
- Data comes from the R `Titanic` package
- We have the variables:
  - Survived: Yes = 1, No =0
  - Pclass: Passenger class
  - SibSp: Number of siblings or spouses on board
  - Parch: Number of parents or children on board
  - Embarked: What port they embarked on
  - Age, Sex, Fare
- Some data are missing.

Example - Explore the Data

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22.00	1	0	7.25	S
1	1	female	38.00	1	0	71.28	C
1	3	female	26.00	0	0	7.92	S
1	1	female	35.00	1	0	53.10	S
0	3	male	35.00	0	0	8.05	S
0	3	male		0	0	8.46	Q
0	1	male	54.00	0	0	51.86	S
0	3	male	2.00	3	1	21.07	S
1	3	female	27.00	0	2	11.13	S
1	2	female	14.00	1	0	30.07	C

Example - Explore the Data

- First we begin by separating our data into a test set (418) and a training set (891)
- Next we begin by using our tools to see what type of missingness we have.
- We begin by making a missingness plot using the `VIM` package in R.

Example - Assess Type of Missingness

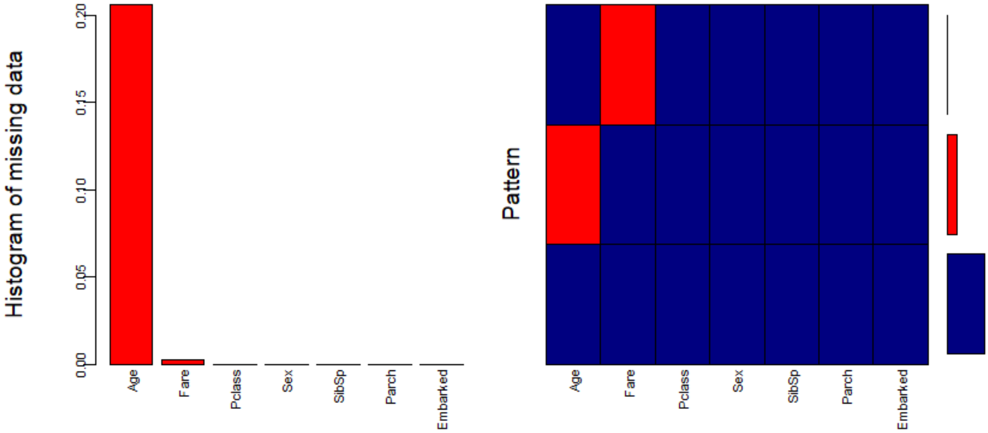


Figure: Missingness Maps Training Data

Example - Make an Assumption

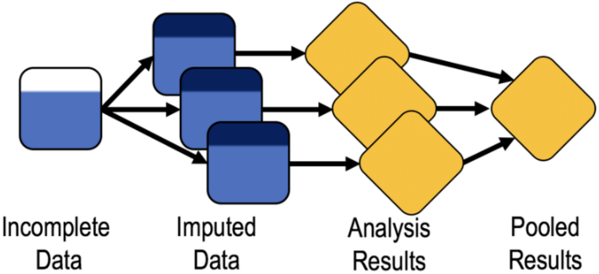
- For illustration purposes, lets assume MAR, and conduct a multiple imputation procedure with the `mice` package in R.

Example - Generate Imputed Values

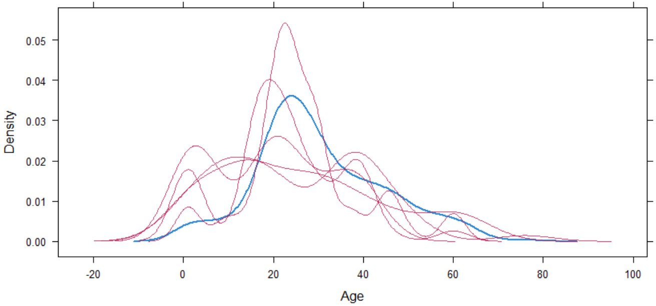
	Imputations				
Passenger	1	2	3	4	5
429	23.00	17.00	27.00	19.00	36.50
441	23.00	55.00	60.00	32.50	45.00
448	18.00	22.00	1.00	31.00	39.00
452	22.00	16.00	1.00	22.00	21.00
455	0.83	60.50	30.00	19.00	0.83
458	0.83	9.00	30.00	21.00	22.00
460	60.00	63.00	60.00	39.00	49.00
466	23.00	0.83	22.00	1.00	36.50
473	25.00	20.00	18.00	26.00	20.00
477	16.00	27.00	36.50	0.17	18.00

The output shows the imputed data for each observation (first column left) within each imputed dataset (first row at the top). These are the imputed value for Age for 10 different subjects.

Example - Create Imputed Data Sets



Example - Do these Imputed Data Sets Seem Reasonable?



The density of the imputed data for each imputed dataset is showed in magenta while the density of the observed data is showed in blue. Again, under our previous assumptions we expect the distributions to be similar. We can create a density plot like this for each variable with missing values.

Example - Continue to Statistical Analysis

- If we believe are imputed data sets are acceptable, we can continue to the next step.
- Next we would conduct our analysis on each set of imputed data sets.
- We then pool our results together for a single analysis.

Summary

- Missing data analysis depends entirely on the context in which the data came from, and what type of analysis is being conducted.
- Although we have tools that work in a variety of situations, there is no simple fill-in-the-blank procedure for missing data analysis.
- Analyst should use caution, and acknowledge the limitations of their analysis.

Thank you!

Sources

- https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf
- Based on “How to Handle Missing Data”, delivered for GradQuant by Rebecca Kurtz-Garcia, 2020-02.