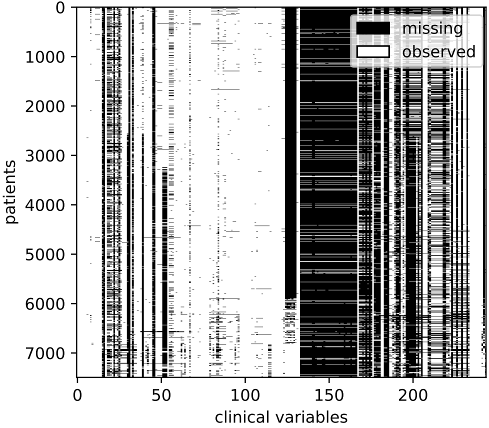
Introduction to missing values

Marine Le Morvan – Soda, INRIA

Incomplete data is ubiquitous in many fields



traumabase.eu

Statistical inference

Suppose we are given a data matrix $X \in \mathbb{R}^{n \times d}$ with n samples and d variables.

$$X = \begin{pmatrix} 1.2 & 0.7 & 0.2 & -0.4 \\ -0.3 & 0.1 & -0.1 & -0.9 \\ 1.5 & 0.4 & 2.3 & -0.5 \\ -2.8 & 1.9 & 1.6 & 2.2 \\ -0.4 & 1.7 & -2.4 & -1.5 \end{pmatrix}$$

We wish to make **inference** on some aspects of the distribution of X.

- Non-parametric inference: estimate the **mean**, the **covariance**, **variance**, ...
- Parametric inference: assume the data was drawn from a given probability distribution (e.g. Gaussian distribution) and estimate the parameters of the distribution.

Statistical inference

Suppose we are given a data matrix $X \in \mathbb{R}^{n \times d}$ with n samples and d variables.

$$X = \begin{pmatrix} 1.2 & \texttt{na} & 0.2 & -0.4 \\ -0.3 & \texttt{na} & \texttt{na} & -0.9 \\ \texttt{na} & 0.4 & \texttt{na} & \texttt{na} \\ -2.8 & 1.9 & 1.6 & 2.2 \\ \texttt{na} & 1.7 & -2.4 & \texttt{na} \end{pmatrix}$$

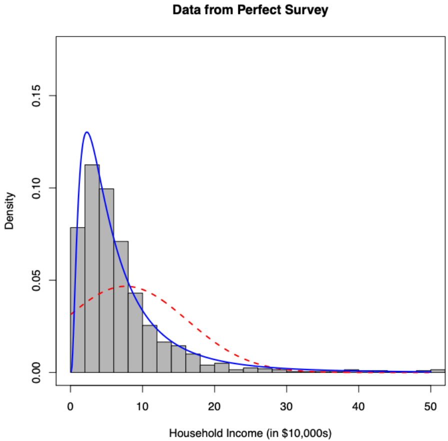
How to make valid inference using observed data only?

We wish to make **inference** on some aspects of the distribution of X.

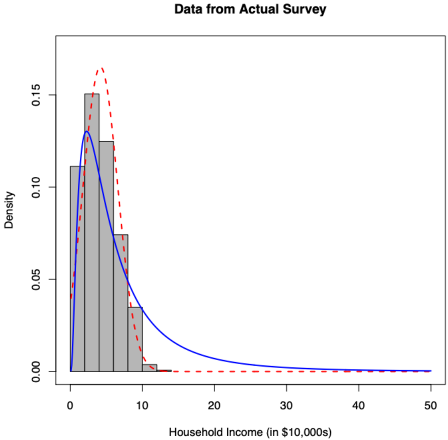
- Non-parametric inference: estimate the **mean**, the **covariance**, **variance**, ...
- Parametric inference: assume the data was drawn from a given probability distribution (e.g. Gaussian distribution) and estimate the parameters of the distribution.

The importance of the missing data mechanism

Data from the US Census Bureau for 2012 indicate that median annual household income in the US is approximately 51,000$ (distribution represented by the blue curve).



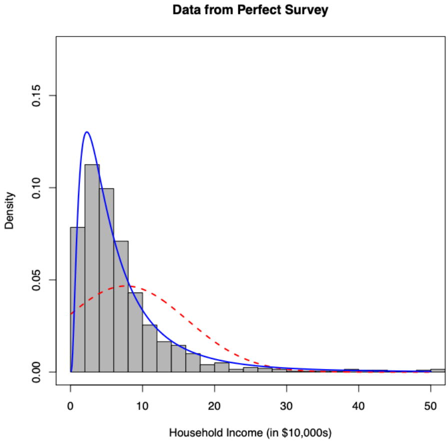
100% of survey participants responded
Sample median: 50,556$

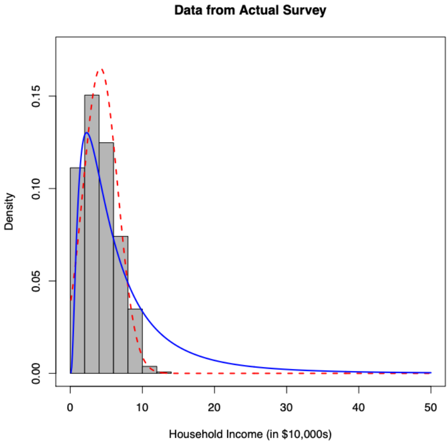


66% of survey participants responded
Sample median: 38,625$

The importance of the missing data mechanism

Data from the US Census Bureau for 2012 indicate that median annual household income in the US is approximately 51,000$ (distribution represented by the blue curve).



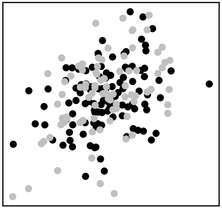


Different assumptions about the missing data mechanism leads to different inferences about the true distribution.

The missing data mechanisms: MCAR, MAR, and MNAR.

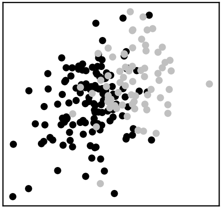
- In 1976, Rubin (1976) has formalized 3 missing data mechanisms:
  - Missing Completely at Random: $P(M = m|X) = P(M = m)$
  - Missing at Random: $P(M = m|X) = P(M = m|X_{obs(m)})$
  - Missing Non At Random: all other cases.

MCAR



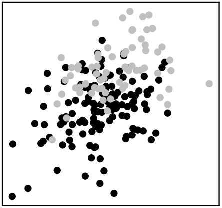
$P(M_2 = 1|X) = 0.5$

MAR



$P(M_2 = 1|X) = \sigma(X_1)$

MNAR



$P(M_2 = 1|X) = \sigma(X_2)$

The missing data mechanisms: MCAR, MAR, and MNAR.

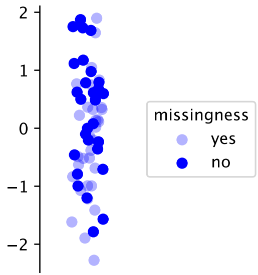
- The data analyst must make an assumption about the mechanism.
  - Difficulty for inference: MCAR $\leq$ MAR $\leq$ MNAR.
  - MCAR unrealistic in general.
  - Most developments for inference since the 70s require the MAR assumption to hold.
  - Fundamental challenge: the assumption cannot be verified from the observed data.
  - MAR can be justified using domain knowledge.
  - Guideline: collect rich additional variable, ideally always observed, to render MAR plausible.

The literature on missing values

Since the 70s, an abundant literature on missing data has flourished, mainly focused on **inference** and **imputation** tasks.

Inference

- e.g. estimating means and variances.



- Ad-hoc methods:
  - Complete-case analysis.
  - Single imputation.
- Principled methods:
  - Inverse Probability Weighting (IPW).
  - Likelihood-based inference on $P(M, X)$.
  - Multiple imputation.

Naive method 1: complete-case analysis

- **Complete-case analysis**: only keep the observations without missing values, and proceed with the complete subsample.

$$\begin{pmatrix} 1.2 & 3.2 & 0.2 & -0.4 \\ \texttt{na} & 0.4 & \texttt{na} & \texttt{na} \\ -2.8 & 1.9 & 1.6 & 2.2 \\ \texttt{na} & 1.7 & -2.4 & \texttt{na} \end{pmatrix} \implies \begin{pmatrix} 1.2 & 3.2 & 0.2 & -0.4 \\ -2.8 & 1.9 & 1.6 & 2.2 \end{pmatrix}$$

- **Example with inference of the mean**:

  Let $X \in \mathbb{R}^n$ be a variable with missing values.
  Let $M \in \{0,1\}^n$ be its missing indicator (1 means observed).

$$\hat{\mu}^{cc} = \frac{\sum_{i=1}^n M_i X_i}{\sum_{i=1}^n M_i}$$

- **Drawbacks:**
  - The loss of information can be considerable (d = 20, missing rate 10% $\implies$ proba of 0.13 to observe a complete case).
  - Dangerous when the data is not MCAR $\implies$ biased estimator.

Naive method 2: Single imputation methods

- **Single imputation analysis**: fill in the missing values and carry on with the completed dataset.
  - Unconditional imputation: use the mean of observed values.
  - Conditional imputation: learn a regressor to predict missing values from observed variables.

$$X = \begin{pmatrix} 1.2 & 3.2 & 0.2 & -0.4 \\ \texttt{na} & 0.4 & \texttt{na} & \texttt{na} \\ -2.8 & 1.9 & 1.6 & 2.2 \\ \texttt{na} & 1.7 & -2.4 & \texttt{na} \end{pmatrix} \implies \hat{X} = \begin{pmatrix} 1.2 & 3.2 & 0.2 & -0.4 \\ -0.8 & 0.4 & -0.2 & 0.9 \\ -2.8 & 1.9 & 1.6 & 2.2 \\ -0.8 & 1.7 & -2.4 & 0.9 \end{pmatrix}$$

- **Example with inference of the mean**:

  Let $X \in \mathbb{R}^n$ be a variable with missing values, and $\hat{X} \in \mathbb{R}^n$ its imputed version. Let $M \in \{0,1\}^n$ be its missing indicator (1 means observed).

$$\hat{\mu}^{imp} = \frac{1}{n}\sum_{i=1}^{n} M_i X_i + (1 - M_i)\hat{X}_i$$

Naive method 2: Single imputation methods

- **Single imputation analysis**: fill in the missing values and carry on with the completed dataset.
  - Unconditional imputation: use the mean of observed values.
  - Conditional imputation: learn a regressor to predict missing values from observed variables.

$$X = \begin{pmatrix} 1.2 & 3.2 & 0.2 & -0.4 \\ \texttt{na} & 0.4 & \texttt{na} & \texttt{na} \\ -2.8 & 1.9 & 1.6 & 2.2 \\ \texttt{na} & 1.7 & -2.4 & \texttt{na} \end{pmatrix} \implies \hat{X} = \begin{pmatrix} 1.2 & 3.2 & 0.2 & -0.4 \\ -0.8 & 0.4 & -0.2 & 0.9 \\ -2.8 & 1.9 & 1.6 & 2.2 \\ -0.8 & 1.7 & -2.4 & 0.9 \end{pmatrix}$$

- **Drawbacks**:
  - Biased inference unless MCAR in most cases.
  - Distorted measures of uncertainty in most cases (e.g. confidence intervals, variances, ...)

Inverse Probability Weighting - 1/2

- **Intuition**: Reweight samples so that those that are more likely to be missing are given larger weights to account for the similar samples that are missing.

- **Example with inference of the mean**:
  Let $X \in \mathbb{R}^n$ be a variable with missing values.
  Let $M \in \{0,1\}^n$ be its missing indicator (1 means observed).
  Let $V \in \mathbb{R}^n$ be a completely observed variable.
  Suppose that the proba of X_i being observed is given by $P(M_i = 1 | V_i) = \pi(V_i)$.

$$\hat{\mu}^{IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i M_i}{\pi(V_i)}$$

- **Need to estimate** $\pi(V_i)$
  Usually estimated with a logistic regression where M_i are the labels, and V_i the features.

- Unbiased under MAR and if $\pi(V_i)$ is well specified.

Inverse Probability Weighting - 2/2

- **Drawbacks**:
  - Difficult to know whether $\pi(V_i)$ is well specified.
  - Instabilities when $\pi(V_i)$ is too small.
  - Data inefficiency because only complete cases are used.
- Subsequent works have focused on improving the IPW:
  - The Augmented IPW (AIPW) solves the inefficiency issue:

$$\hat{\mu}^{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i M_i}{\pi(V_i)} - \frac{M_i - \pi(V_i)}{\pi(V_i)} \mathbb{E}\left[X_i | V_i\right] \right)$$

  See how $\mathbb{E}\left[X_i | V_i\right]$ replaces X_i when X_i is missing.

Likelihood-based methods - 1/3

- Review of maximum likelihood estimation with complete data.

- We need to assume a **parametric model** for the probability density of X, with parameters θ:

$$X \sim p_\theta(X).$$

  For example, we can assume the data follows a Gaussian distribution with parameters $\theta = (\mu, \Sigma)$.

- The goal is to find the parameters $\hat{\theta}$ that maximize the **likelihood** $\mathcal{L}$ of the data:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$$

  $\hat{\theta}$ is called a maximum likelihood estimator (MLE).

- Standard techniques for maximizing the likelihood cannot be applied when it is not fully observed.

Likelihood-based methods - 2/3

- We wish to estimate the parameter θ of the full data distribution $p_\theta(X)$ using only the observed data (M, X_{obs}).

$$\begin{aligned}
p_{\theta,\phi}(M, X_{obs}) &= \int p_{\theta,\phi}\left(M, X_{obs}, X_{mis}\right) dX_{mis} && \text{Marginalisation} \\
&= \int p_\phi\left(M|X_{obs}, X_{mis}\right) p_\theta\left(X_{obs}, X_{mis}\right) dX_{mis} && \text{Bayes rule} \\
&= \int p_\phi\left(M|X_{obs}\right) p_\theta\left(X_{obs}, X_{mis}\right) dX_{mis} && \textbf{MAR hyp.} \\
&= p_\phi\left(M|X_{obs}\right) \int p_\theta\left(X_{obs}, X_{mis}\right) dX_{mis} \\
&= p_\phi\left(M|X_{obs}\right) p_\theta\left(X_{obs}\right)
\end{aligned}$$

where $p_\phi(M|X)$ is called the missingness mechanism and $p_\theta(X_{obs})$ the observed data likelihood.

- **Ignorability**: The missingness mechanism can be ignored. It does not need to be modeled for the purpose of calculating the MLE $\hat{\theta}$.

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} p_\theta(X_{i,obs})$$

Likelihood-based methods - 3/3

Question:How do we solve this optimization problem?

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} p_\theta(X_{i,obs})$$

- Usually, we posit a model for the complete data distributuon: $p_\theta(X)$.
- It is not always analytically tractable to compute

$$p_\theta(X_{obs}) = \int p_\theta \left(X_{obs}, X_{mis}\right) dX_{mis}.$$

Answer: use the **Expectation-Maximization (EM)** algorithm.

- EM algorithm: maximizes objective functions in the presence of latent variables (in our case, the missing values).
- In practice, the `norm` package in R implements EM with missing values (assuming p_θ is multivariate Gaussian).

Multiple Imputation - 1/2

As for likelihood-based methods, we assume a parametric model $P_\theta(X)$, and the goal is inference on θ.

- Multiple Imputation involves 3 steps:
  1. **Impute** missing values R times to create R completed datasets:
  $$X^{(r)} = \left(X_{obs}, X_{mis}^{(r)}\right) \text{ for } r = 1, \ldots, R$$
  2. Carry out the **full data analysis**. For example, using the MLE:
  $$\hat{\theta}^{(r)} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta(X_{i,obs}, X_{i,mis}^{(r)})$$
  3. **Combine** the M results into the final one by averaging:
  $$\hat{\theta} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}^{(r)}$$
- Again, Multiple Imputation is only valid under **MAR** assumption.

Multiple Imputation - 2/2

Concerning the imputation (step 1):

- In theory, draw $X_{mis}^{(r)}$ from $p_{\theta^{(init)}}\left(X_{mis}|X_{obs}, M\right)$ (improper imputation).
- In practice, two widespread options to draw imputations:
  - Draw imputations assuming p_θ is a Gaussian distribution.
  - use the MICE imputation algorithm.
- Specialized packages in `R` (implement the 3 steps):
  - `norm`, `amelia` draw imputations from a multivariate Gaussian.
  - `mice` draw imputations based on the MICE algorithm.
- No specialized package in Python: use MICE + Maximum Likelihood estimation (MLE).
  - for MICE, `scikit-learn`'s `IterativeImputer`.
  - for MLE, `mvem` package, or define the likelihood yourself and use `scipy.optimize`.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–590.