# Test of missing data mechanisms

*An alternative to the Little test based on regression*

**Serguei Rouzinov, supervisor: André Berchtold**
NCCR LIVES - University of Lausanne, SSP
Serguei.Rouzinov@unil.ch

UNIL | Université de Lausanne

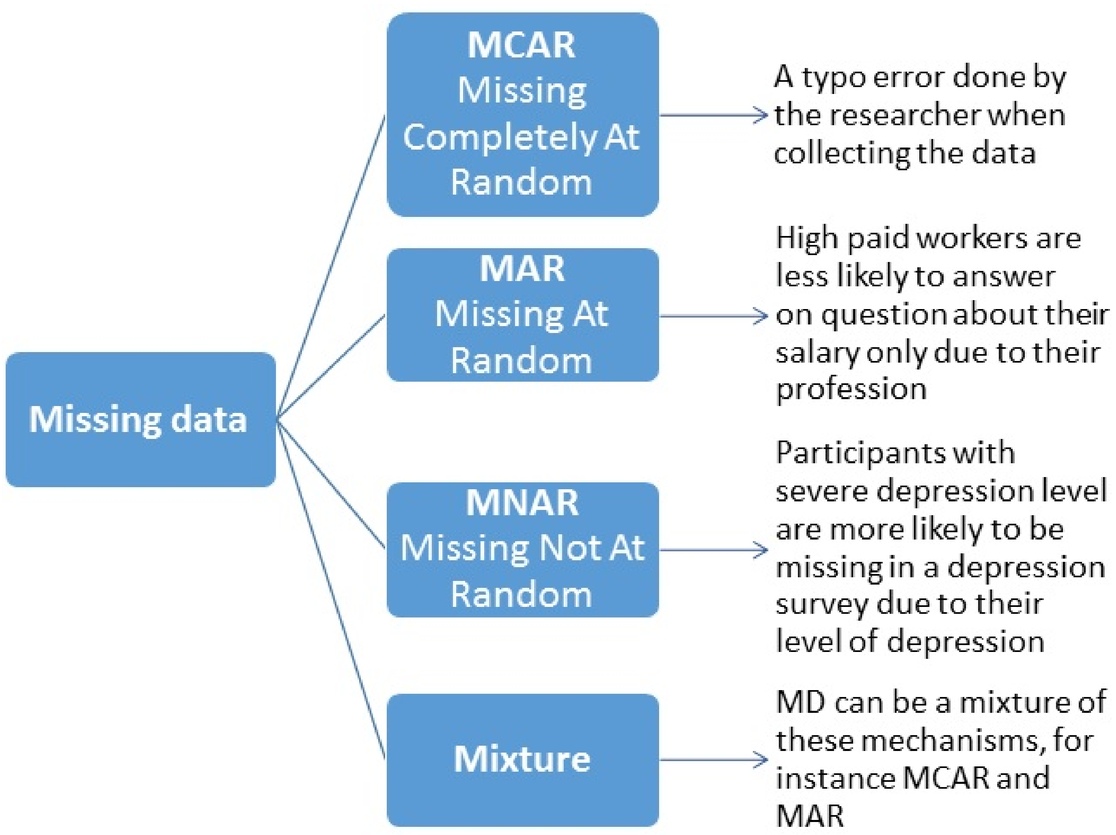
LIVES Swiss National Centre of Competence in Research

## Introduction

- Missing Data (MD): Data whose collect was planned, but not realized ⇒ Biased and inconsistent estimators
- How to handle MD?
  1. Test the MD mechanism
  2. Apply a method for handling MD AFTER knowing the mechanism

## Definition of MD mechanisms

Examples



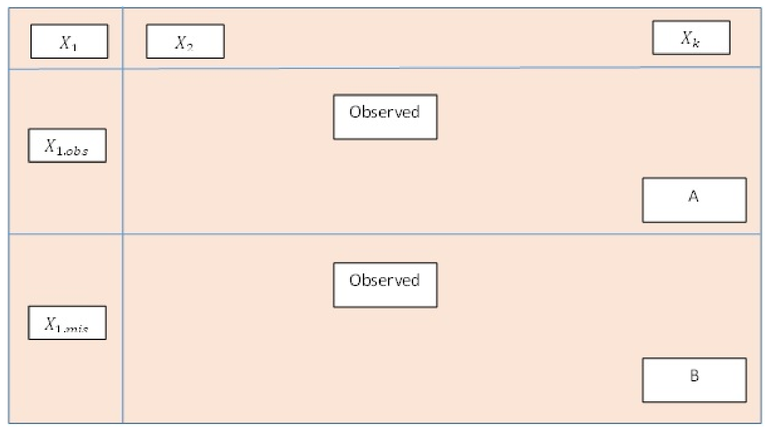

## Existing methods for testing MD mechanism

Observed and missing values are separated and a mean or distribution comparison test is conducted between the two groups. If they are significantly the same, then MD are MCAR. If not, MD are not MCAR [1], [2]:

$$\begin{cases} H_0 : MCAR \\ H_1 : \overline{MCAR} \end{cases}$$

## Alternative test of MD mechanisms

Database




1. Regression and prediction model
   - $X_{1,obs} \sim f(X_2^A, X_3^A, ..., X_k^A) + u$, $u \sim N(0,1)$
   - $\widehat{\beta}_A$ and $A$ to predict $X_{1,obs} \Rightarrow \widehat{X_{1,obs}}$ ; $\widehat{\beta}_A$ and $B$ to predict $X_{1,mis} \Rightarrow \widehat{X_{1,mis}}$
2. Testing process
   - Distribution test between $\widehat{X_{1,mis}}$ and $\widehat{X_{1,obs}}$
   - If $\widehat{X_{1,mis}} \overset{d}{\sim} \widehat{X_{1,obs}}$, then MD are MCAR

## Simulation results of numerical data

- 10,000 × 10 observations and 1,000 replications
- Big regression: all covariates are included in the model
- Small regression: only covariates which determine the MD mechanism are included in the model

Table 1. Single mechanisms, % of $H_0$ acceptance. L for Little [3] and A for Alternative approach

| | MCAR | | MAR1, Big | | MAR1, Small | |
|---|---|---|---|---|---|---|
| | L | A | L | A | L | A |
| 50% | 96.4 | 92.5 | 0 | 3.1 | 0 | 0 |
| 10% | 94.9 | 95.4 | 0 | 8.8 | 0 | 0 |
| 5% | 95.5 | 95.1 | 0 | 13.9 | 0 | 0 |
| 4% | 95.0 | 95.1 | 0 | 15.1 | 0 | 0 |
| 3% | 94.3 | 95.3 | 0 | 16.7 | 0 | 0 |
| 2% | 94.4 | 95.2 | 0 | 20.5 | 0 | 0 |
| 1% | 95.1 | 95.2 | 0 | 28.1 | 0 | 0 |

Table 2. Mixed mechanisms (MAR2+MCAR), % of $H_0$ acceptance. L for Little and A for Alternative approach

| | Big | | Small | |
|---|---|---|---|---|
| | L | A | L | A |
| 40% MAR2, 10% MCAR | 0 | 4.6 | 0 | 0 |
| 10% MAR2, 40% MCAR | 0 | 35.8 | 0 | 0.6 |
| 5% MAR2, 45% MCAR | 0 | 62.1 | 0 | 15.0 |
| 3% MAR2, 47% MCAR | 25.9 | 82.0 | 7.9 | 35.5 |
| 2% MAR2, 48% MCAR | 67.3 | 89.7 | 39.3 | 63.7 |
| 1% MAR2, 49% MCAR | 89.7 | 93.8 | 83.2 | 89.6 |

## References

[1] Joseph W Dixon. *BMDP Statistical Software Manual: To Accompany the... Software Release*. University of California Press, 1988.

[2] Mortaza Jamshidian and Ke-Hai Yuan. Examining missing data mechanisms via homogeneity of parameters, homogeneity of distributions, and multivariate normality. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(1):56–73, 2014.

[3] Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.