Generating missing values with ampute
And why on earth you would do that

Rianne Schouten

1. PhD Candidate, Utrecht University
2. Developer Data & Analytics, Samen Veilig Midden-Nederland

September 1, 2018
Why on earth would you ampute?

For simulation purposes:

1. Generate complete data (or use real dataset)
2. Generate missing values in complete data
3. Apply missing data method
4. Perform analysis and compare with complete data

But also for:

- Planned missing data survey designs
- Investigating measurement errors
- Reproducing your missing data situation
Missing data in your dataset

\texttt{head(inc\_data)}

\begin{verbatim}
## Income WorkingYears Age
## 1 -0.08877292 1.5343721 1.33739681
## 2 NA NA -0.41593616
## 3 -1.75818833 NA 0.06295286
## 4 NA NA 1.73468904
## 5 -0.38850735 NA -1.22025110
## 6 -1.81223387 0.0950749 0.44283715
\end{verbatim}

\texttt{require(mice)}
\texttt{md.pattern(inc\_data)}

\begin{verbatim}
## Age Income WorkingYears
## 206 1 1 1 0
## 412 1 1 0 1
## 382 1 0 0 2
## 0 382 794 1176
\end{verbatim}
### Generation of missing values

1. \( Y_1 \)
Generation of missing values

1. $Y_1$
2. $Y_2$

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$\cdots$</th>
<th>$Y_l$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$\cdots$</th>
<th>$X_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Generation of missing values**

1. $Y_1$
2. $Y_2$
3. $\ldots$
4. $Y_l$

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$\ldots$</th>
<th>$Y_l$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$\ldots$</th>
<th>$X_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table above represents the generation of missing values for the variables $Y_1, Y_2, \ldots, Y_l, X_1, X_2, \ldots, X_m$. Each entry can be replaced with a question mark to indicate a missing value.
Multivariate amputation with ampute
Multivariate Amputation with ampute

Explanation of the method:

Rianne Margaretha Schouten, Peter Lugtig & Gerko Vink (2018)
Multivariate amputation with ampute

```r
amp <- ampute(data,
    patterns = matrix(c(1, 0, 1,
                        1, 0, 0,
                        0, 0, 1),
    nrow = 3, byrow = TRUE),
    freq = c(0.6, 0.2, 0.2),
    prop = 0.5,
    mech = "MAR")
md.pattern(amp$amp)
```

```
## Income Age WorkingYears
## 501 1 1 1 0
## 300 1 1 0 1
## 99 0 1 0 2
## 100 1 0 0 2
## 99 100 499 698
```
Multivariate amputation with ampute

```r
require(mice)
?ampute

ampute(data, prop = 0.5, patterns = NULL, freq = NULL, mech = "MAR", weights = NULL, cont = TRUE, type = NULL, odds = NULL, bycases = TRUE, run = TRUE)

Explanation of all the arguments in vignette:
https://rianneschouten.github.io/mice_ampute/vignette/ampute.html
```
Missing data mechanisms

Missing Completely At Random (MCAR):
Missingness is not related to any variable
Pr(Income = missing) = 0.5

Missing At Random (MAR):
Missingness is related to an observed variable
Pr(Income = missing) = Age

Missing Not At Random (MNAR):
Missingness is related to the missingness itself or to an unobserved variable
Pr(Income = missing) = Income
Missing data mechanisms: Missing Completely At Random

Age

Income

MCAR
Missing data mechanisms: Missing At Random

[Graph showing the relationship between Age and Income with 'MAR' indicating the area of interest.]
Missing data mechanisms: Missing Not At Random
Multivariate amputation with ampute

```r
require(mice)
?ampute

ampute(data, prop = 0.5, patterns = NULL, freq = NULL, mech = "MAR", weights = NULL, cont = TRUE, type = NULL, odds = NULL, bycases = TRUE, run = TRUE)
```

Explanation of all the arguments in vignette:

https://rianneschouten.github.io/mice_ampute/vignette/ampute.html
Contact information

Rianne Schouten: r.m.schouten@uu.nl

Follow my work: rianneschouten.github.io
Missingness types

Standardized weighted sum scores

Probability to be missing according pattern

0.2
0.4
0.6
0.8

−3 −2 −1 0 1 2 3

LEFT MID RIGHT TAIL

Probability to be missing according pattern

Standardized weighted sum scores
Multivariate Amputation: Weighted sum scores

- Missing values in multiple variables

\[
P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}
\]

- Based on multiple variables

\[
wss_i = w_{1,1} \cdot y_{1i} + w_{1,2} \cdot y_{2i} + w_{1,3} \cdot x_i \text{ if case } i \text{ is in pattern 1}
\]

\[
W_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 5 \end{bmatrix}
\]
Mean imputation

![Scatter plot showing mean imputation](image)
Regression imputation

![Graph showing a scatter plot with two groups: Complete and Imputed. The plot displays data points for Age and Income, with a trend line indicating the relationship between the two variables.]