Simulation with Missing data

Anaranya Basu

November 2023

$\hat{\mu}^{cc}$ is unbiased under Missing Completely at Random case but may be biased under Missing at Random Case.

Solution:

Consider observations $y_i = (y_{i1}, y_{i2}...., y_{ip})'$
Define R_i where $R_{ij} = I(y_{ij} \in y_i^{obs})$
i.e. R_{ij}=**1** if y_{ij} is observed and **0** if missing.
Now, we have generated 100 random samples of observations(y) from Bernoulli distribution with mean(μ) as 0.5
Observations are as follows :

```
[1] 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 1 1 1 1 0 0 0 1 1 0 0
[29] 1 1 0 0 1 0 1 0 0 1 1 1 1 1 0 0 1 1 1 0 0 1 0 0 1 1 0 0
[57] 0 1 0 1 0 1 0 0 1 0 1 1 0 0 0 1 0 1 1 1 0 1 0 0 1 1 0 0
[85] 0 1 1 0 1 0 1 0 1 0 0 1 1 0 1 1
```

Now, we have generated another random sample (op) of 100 binary values with a lower mean (0.1) to simulate a binary indicator for missingness.

```
[1] 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[33] 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[65] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
[97] 0 0 0 0
```

Now creating a new vector y_1 that retains the values from y where (op) is not equal to 1 (indicating observed data) and assigns NA to y_1 where op is 1 (indicating missing data).

```
 [1]  0  1  0  1  0  0 NA  0  1 NA  1  0  0  0  1  0  1  1  1  1  1
 [22]  0  0  0  1  1  0  0  1  1  0  0  1  0  1  0 NA  1  1  1  1  1
 [43] NA  0  1  1  1  0  0  1  0  0  1  1  0  0  0  1  0  1  0  1  0
 [64]  0  1 NA  1  1  0  0  0  1  0  1  1  1  0  1  0  0  1  1  0  0
 [85] NA NA  1  0  1  0  1  0  1  0  0  1  1  0  1  1
```

Using the summary(y_1) in R we have detected 7 positions which are completely missing at random.The output is as below-

```
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
0.0000 0.0000 1.0000 0.5161 1.0000 1.0000 7
```

Finally, we calculated the mean of the modified data set y_1, but this time, the **na.rm = TRUE** argument is used to exclude the missing values when computing the mean i.e.

$$\hat{\mu}^{MCAR} = 0.516129 \approx 0.5(\mu^{cc})$$

Now using **Missing at Random** mechanism where the **missingness depends on the observed values** we will see:
similarly we generated random samples from Bernoulli distribution with mean (0.5)

```
[1] 1 1 1 0 0 1 1 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 1 1 0 1 0 0 0 0
[33] 1 0 1 0 1 1 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 1 1 0 0 0 1 1 0 0 1
[65] 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 1 1
[97] 1 0 1 0
```

Now generated (op1) with a probability that depends on the logistic function (**plogis**) of the difference between y and 0.5. This simulates a MAR mechanism.

```
[1] 1 0 1 1 0 1 0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 1 1 1 0 0 0 0 0 0 1 0
[33] 1 1 1 1 0 0 1 0 0 1 1 0 0 1 0 1 1 0 0 0 0 0 0 1 1 1 1 1 0 1 1 0 1
[65] 1 1 1 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 1 1 0 1 1 1 0 1 0 0 1 1
[97] 0 1 1 0
```

Now we created y2 with missing data by setting elements to NA where op1 is equal to 1 and we get -

```
[1] NA NA  1  0 NA  1 NA NA  0 NA  0  0  0  1  0  0  0 NA  0  0 NA
[22]  1  0  1  1 NA NA NA  0  0 NA  0 NA  0 NA NA NA NA  0 NA  0  1
[43] NA  0 NA NA  0  0 NA NA  0 NA NA NA  1 NA  0  0  0 NA NA  0  0
[64]  1 NA NA  0 NA  1 NA  1 NA NA  0 NA NA  0  0  0  0  0 NA NA  0
[85] NA NA NA  0 NA NA  0 NA  0 NA NA  1  1  0 NA  0
```

Now the true mean(μ) is 0.5 and the estimated mean for complete case analysis($\hat{\mu}^{cc}$) came out as 0.25 i.e.

$$E(\hat{\mu}^{cc}) \neq \hat{\mu}^{MAR}$$

Appendix

```
##Missing completely at random case###
set.seed(1345)
y = rbinom(100, 1, 0.5)
mu = mean(y)
mu
op = rbinom(100, 1, 0.1)
op
y1 = y
y1[op == 1] = NA
summary(y1)
mu_mcar = mean(y1, na.rm = TRUE)
mu_mcar
##Missing at random case##
y <- rbinom(100, 1, 0.5)
op1 <- rbinom(100, 1, plogis(y - 0.5))
y2 <- y
y1[op1 == 1] <- NA
true_mean <- mean(y, na.rm = TRUE)
mean_y2 <- mean(y2, na.rm = TRUE)
```