Workshop 1 – Solutions

Vanda Inacio

1. Note that, depending on how you argue, you can come up with a different missing data mechanism than the one stated below.

(a) The data on income and employment of five-sixth households are missing because those households are not asked about that. However, since those households whose information on income and employment is available are chosen at random, the missing data are MCAR.This can also be interpreted as a planned missing design situation.

(b) The data on volunteer time of those younger than 65 years old are missing. Because the missing depends only on age, which is known from the first round of response, the missing data are MAR.

(c) The data about employment of those people who refused to give any information are missing. It is reasonable to think that those people are different from the ones who did respond to the survey; e.g., people may be reluctant to answer if they have encountered some problem at work. Thus, the missing data are MNAR.

(d) The missing data of the concentration of bacillus in the rats is due to the death of rats. It is reasonable that the death of rats in the placebo group may be due to a high level of concentration of baccilus. Thus, the missing data for the concentration of bacillus are MNAR.

(e) The missing data is due to the transport system, and not related to the performance of the blades themselves or any other factor, and thus the missing data are MCAR.

2.

(a) MAR, since it depends on Y_1 and Y_3 which are fully observed.
(b) MCAR, since it is constant, thus not related to the data, observed or not, at all.
(c) MAR, since it is missing for all subjects whose Y_1 value exceeds y^* and Y_1 is fully observed.
(d) MNAR, since it depends on both Y_1 (fully observed) and Y_2 (which has missing values).

3.

(a) To start the exercise, we need to simulate $n = 500$ observations from a trivariate standard normal distribution with all correlations equal to 0.5, i.e., a trivariate normal distribution with mean vector and covariance matrix, respectively, given by

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

This can be implemented in `R` using the command `mvrnorm` (from the package `MASS`). Type `help(mvrnorm)` to know more about this function.

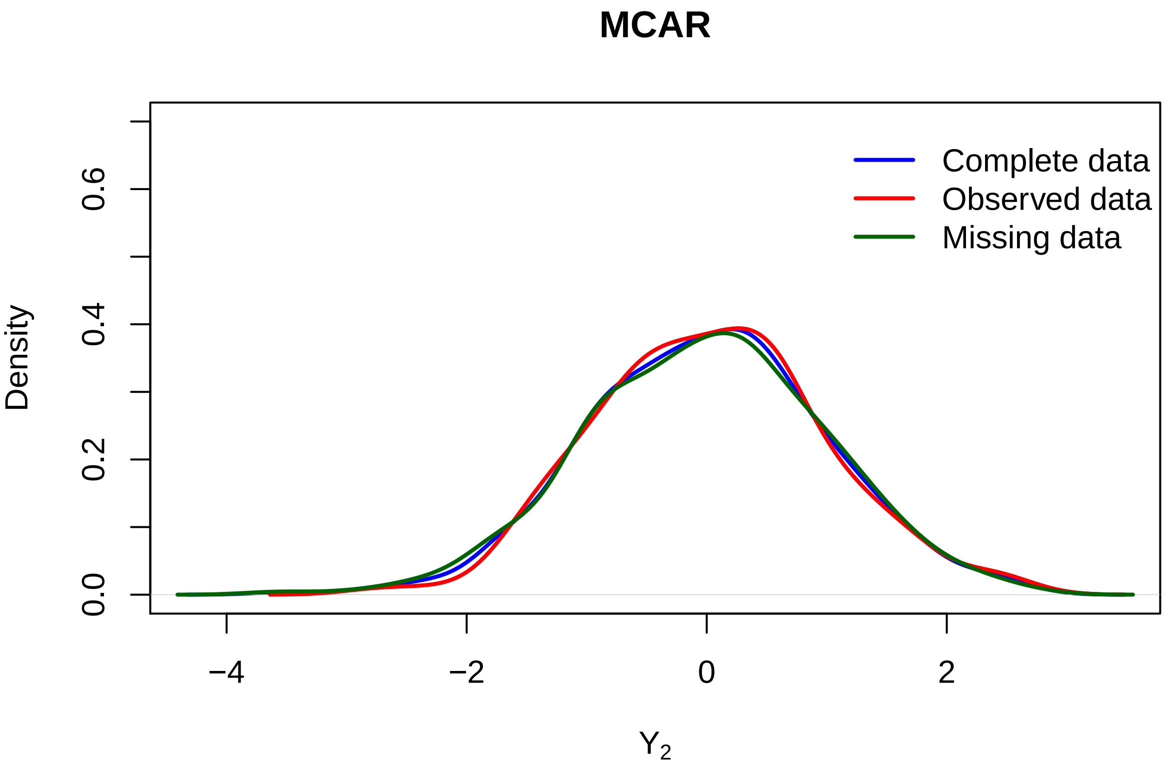
```r
require(MASS)
n <- 500
#defining the covariance matrix
Sigma <- matrix(c(1, 0.5, 0.5, 0.5, 1, 0.5, 0.5, 0.5, 1), nrow = 3, byrow = T)
#simulating the data
set.seed(1)
Y <- mvrnorm(n = n, mu = c(0,0,0), Sigma = Sigma)
```

```
#storing each column of Y in a separate variable
Y1 <- Y[,1]; Y2 <- Y[,2]; Y3 <- Y[,3]
```

After having generated the data, missingness will then be imposed on Y_2. For this part of the question, we know that $\Pr(R = 0 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = 0.65$ and so we can easily generate our missing indicator variable from a Bernoulli distribution with probability of success equal to $1 - 0.65 = 0.35$. In `R` this can be accomplished by using the `rbinom` function, which stands for the binomial distribution, setting the argument `size` (number of trials) equal to 1.

```
set.seed(1)
r_mcar <- rbinom(n, 1, 0.35)
#extracting the index associated to the observed values (1)
ind_mcar <- which(r_mcar == 1)
#storing the observed and missing values in new variables
Y2_MCAR_obs <- Y2[ind_mcar]
Y2_MCAR_mis <- Y2[-ind_mcar]

#plotting the densities
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
     main = "MCAR", ylim = c(0, 0.7))
lines(density(Y2_MCAR_obs), lwd = 2, col = "red")
lines(density(Y2_MCAR_mis), lwd = 2, col = "darkgreen")
legend(1, 0.7, legend = c("Complete data", "Observed data", "Missing data"),
       col = c("blue", "red", "darkgreen"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```



As can be observed, and as hypothesised under MCAR, the three (complete, observed, and missing values) distributions are similar.

(b) In this part, we do the same as in part (a) but we impose a MAR mechanism. Specifically, we are told that

$$\Pr(R = 0 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 Y_3}}{1 + e^{\beta_0+\beta_1 Y_3}},$$

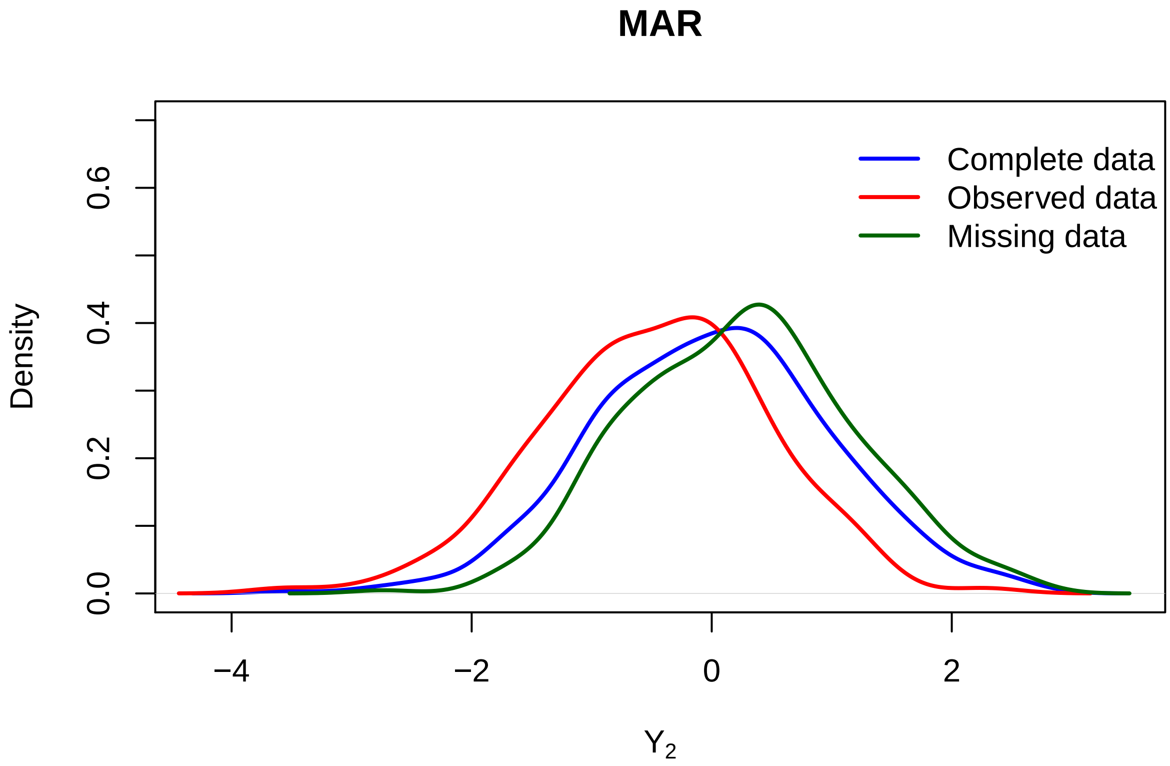
which implies that

$$\Pr(R = 1 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{1}{1 + e^{\beta_0+\beta_1 Y_3}}.$$

The R code is, with minor modifications, very similar to the one in (a).

```
beta0 <- 1.5; beta1 <- 3
set.seed(1)
r_mar <- rbinom(n, 1, 1/(1+exp(beta0+beta1*Y3)))
ind_mar <- which(r_mar == 1)
Y2_MAR_obs <-Y2[ind_mar]
Y2_MAR_mis <-Y2[-ind_mar]

#plotting the densities
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
     main = "MAR", ylim = c(0, 0.7))
lines(density(Y2_MAR_obs), lwd = 2, col = "red")
lines(density(Y2_MAR_mis), lwd = 2, col = "darkgreen")
legend(1, 0.7, legend = c("Complete data", "Observed data", "Missing data"),
       col = c("blue", "red", "darkgreen"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```



We can observe now that under the MAR mechanism, the three distributions are not as similar as they were in the MCAR case.

(c) Finally we will simulate MNAR data, according to

$$\Pr(R = 0 \mid Y_1, Y2, Y_3, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 Y_1+\beta_2 Y_2}}{1 + e^{\beta_0+\beta_1 Y_1+\beta_2 Y_2}}.$$

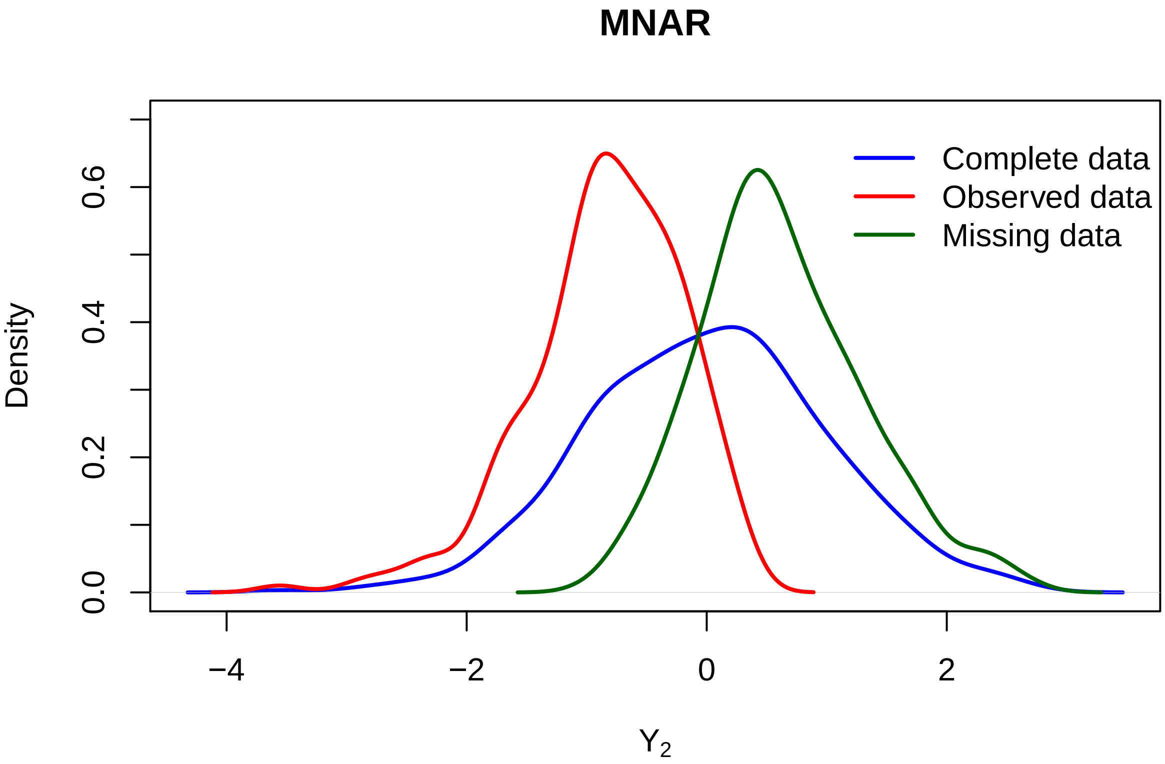
which implies that

$$\Pr(R = 1 \mid Y_1, Y_2, Y_3, \boldsymbol{\beta}) = \frac{1}{1 + e^{\beta_0 + \beta_1 Y_1 + \beta_2 Y_2}}.$$

We repeat the same process as before in R.

```
beta0 <- 1.5; beta1 <- 3; beta2 <- 5
set.seed(1)
r_mnar <- rbinom(n, 1, 1/(1+exp(beta0+beta1*Y1+beta2*Y2)))
ind_mnar <- which(r_mnar == 1)
Y2_MNAR_obs <-Y2[ind_mnar]
Y2_MNAR_mis <-Y2[-ind_mnar]

#plotting the densities
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
     main = "MNAR", ylim = c(0, 0.7))
lines(density(Y2_MNAR_obs), lwd = 2, col = "red")
lines(density(Y2_MNAR_mis), lwd = 2, col = "darkgreen")
legend(1, 0.7, legend = c("Complete data", "Observed data", "Missing data"),
       col = c("blue", "red", "darkgreen"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```



As expected, MNAR case is even more extreme in terms of the dissimilarities between the three distributions.

```
sessionInfo()
```

```
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS  10.15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
```

```
##
## locale:
## [1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] MASS_7.3-51.4
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.6.0  magrittr_1.5    tools_3.6.0     htmltools_0.4.0
##  [5] yaml_2.2.0      Rcpp_1.0.1      stringi_1.4.3   rmarkdown_2.1
##  [9] knitr_1.23      stringr_1.4.0   xfun_0.8        digest_0.6.19
## [13] rlang_0.4.8     evaluate_0.14
```