Incomplete Data Analysis

Missing Mechanisms—Examples

School of Mathematics, University of Edinburgh

V. Inácio de Carvalho & M. de Carvalho

Here we will simulate data under different missingness mechanisms. Because we are working with simulated data we can actually do comparisons between the complete, observed, and missing data distributions (something that when working with real data we, obviously, cannot).

Blood pressure simulation example

This first example is adapted from Schafer and Graham (2002, *Psychological Methods*). Suppose that the systolic blood pressure (SBP) of n individuals is recorded both in January (coded in variable Y_1) and in February (coded in variable Y_2). We will impose missing on Y_2 according to the three missingness mechanisms we have learned, that is, MCAR, MAR, and MNAR.

First of all, we need to generate the simulated data. We will simulate data from a bivariate normal distribution, with means $\mu_1 = \mu_2 = 120$, standard deviations $\sigma_1 = \sigma_2 = 20$, and correlation $\rho = 0.6$. This should generate reasonable SBP values. In `R`, we need to load the package `MASS`, which has the function `mvrnorm` that will allow us to simulate random numbers from a bivariate (in general from a multivariate) normal distribution.

```
require(MASS)
```

We will now simulate the complete data for Y_1 (measurement of systolic blood pressure in January) and Y_2 (measurement of systolic blood pressure in February). The function `mvrnorm` needs as input the number of random pairs we want to generate, the vector of means ($(\mu_{Y_1}, \mu_{Y_2})^T$), and the covariance matrix. For more information type `help(mvrnorm)` In the example, we are given the correlation between Y_1 and Y_2, denoted by ρ_{Y_1,Y_2}, and their respective standard deviations (σ_{Y_1} and σ_{Y_2}), and from this information, we can easily compute the covariance between Y_1 and Y_2, denoted by σ_{Y_1,Y_2}, by noting that

$$\rho_{Y_1,Y_2} = \frac{\sigma_{Y_1,Y_2}}{\sigma_{Y_1}\sigma_{Y_2}},$$

and therefore,

$$\sigma_{Y_1,Y_2} = \rho_{Y_1,Y_2}\sigma_{Y_1}\sigma_{Y_2},$$

Remember that the covariance matrix is given by

$$\Sigma = \begin{pmatrix} \sigma^2_{Y_1} & \sigma_{Y_1,Y_2} \\ \sigma_{Y_1,Y_2} & \sigma^2_{Y_2} \end{pmatrix}.$$

We can now generate the data, which we do as illustrated below. I will consider $n = 100$ and will also fix the seed (for the random number generator), so that the results are reproducible (i.e., we will all obtain the same values).

```
set.seed(1)
n <- 100
mu1 <- mu2 <- 120
```

```
sigma1 <- sigma2 <- 20
rho <- 0.6

#covariance matrix
Sigma <- matrix(c(sigma1^2, rho*sigma1*sigma2, rho*sigma1*sigma2, sigma2^2), 2, 2, byrow = T)

#generate Y=(Y1,Y2)
Y <- mvrnorm(n, mu = c(mu1, mu2), Sigma = Sigma)

#looking at the first 10 rows of the dataset
Y[1:10,]
```

```
##            [,1]      [,2]
##  [1,] 114.34238 103.24493
##  [2,] 122.90842 123.66181
##  [3,] 113.19935  96.90429
##  [4,] 147.12380 149.95070
##  [5,] 131.74920 120.03963
##  [6,]  89.51592 121.13011
##  [7,] 122.30897 135.12982
##  [8,] 125.06671 141.34840
##  [9,] 126.86363 133.73615
## [10,]  99.49121 129.58289
```

```
#storing and rounding the simulated values in two variables
Y1 <- round(Y[,1]); Y2 <- round(Y[,2])
mean(Y1); mean(Y2); sd(Y1); sd(Y2)
```

```
## [1] 122.27
```

```
## [1] 121.61
```

```
## [1] 18.1986
```

```
## [1] 18.17863
```

Let us start imposing missingness on Y_2 under a MCAR mechanism. One way to do that is to simply select, say 30 individuals out of the 100 in our sample.This mechanism is clearly MCAR.

```
set.seed(1)
ind <- sample(1:n, size = 30, replace = F)
Y2_MCAR_obs <- Y2[ind]
ind; Y2_MCAR_obs
```

```
##  [1] 68 39  1 34 87 43 14 82 59 51 85 21 54 74  7 73 79 37 83 97 44 84 33 35 70
## [26] 96 42 38 20 28
```

```
##  [1] 133 134 103 105 152 118  75 126 118 131 135 132  91 103 135 135 131 110 143
## [20] 110 126  80 132  98 161 121 126 114 129  93
```

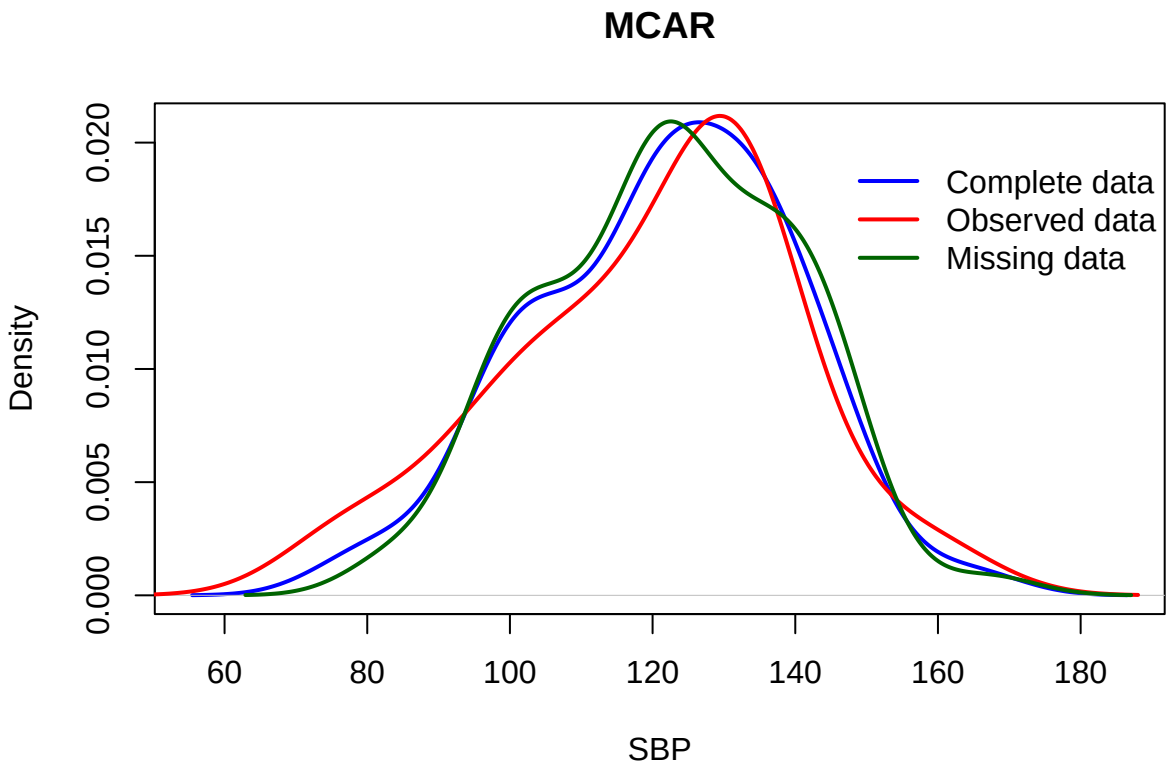
```
mean(Y2_MCAR_obs); sd(Y2_MCAR_obs)
```

```
## [1] 120
```

```
## [1] 20.10318
```

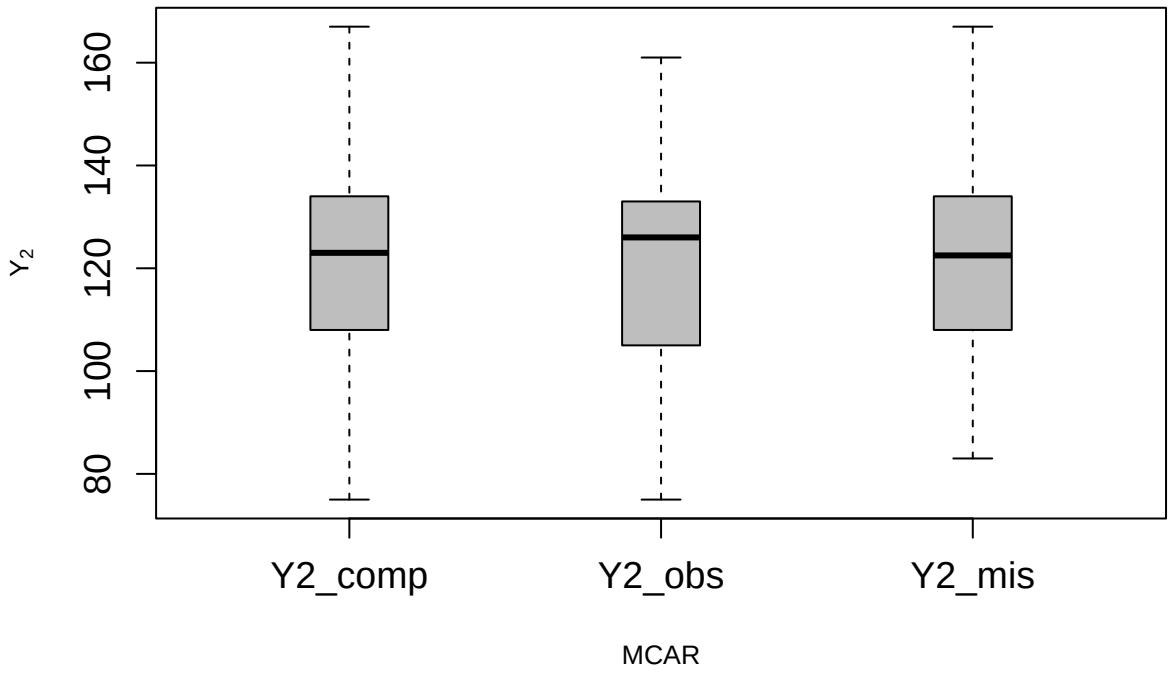
Let us know look at the density plots of the observed, complete, and missing data (remember that we can only do this because we are working with simulated data and so that we have access to the complete data and to the missing data as well).

```r
Y2_MCAR_mis <- Y2[-ind] #storing the "missing" Y2 values
plot(density(Y2), lwd = 2, col = "blue", xlab = "SBP", main = "MCAR")
lines(density(Y2_MCAR_obs), lwd = 2, col = "red")
lines(density(Y2_MCAR_mis), lwd = 2, col = "darkgreen")
legend(145, 0.02, 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")
```



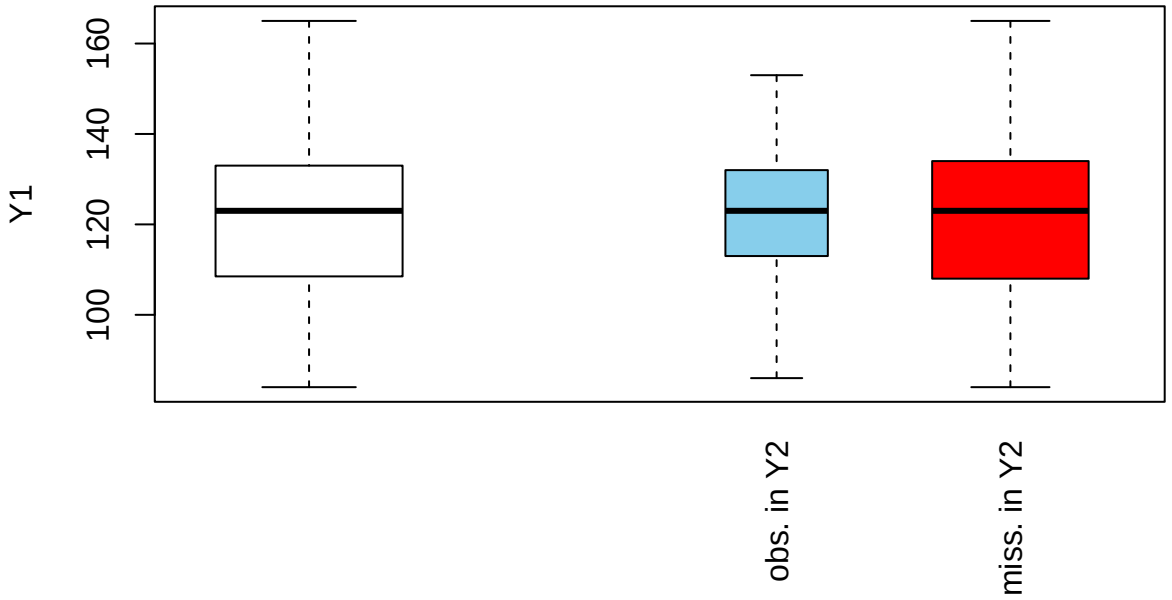
We can see that the three distributions are very similar, which makes sense, as the data were generated under the MCAR assumption. If you try to increase the sample size, e.g., from $n = 100$ to $n = 1000$ and select 300 rather than 30 individuals, you will notice that the distributions become much more similar. We could do a similar visualisation by using a boxplot.

```r
n_obs <- length(Y2_MCAR_obs)
n_mis <- length(Y2_MCAR_mis)
index <- rep("Y2_comp", n + n_obs + n_mis)
index[(n+1):(n+n_obs)] <- "Y2_obs"
index[(n+n_obs+1):(n+n_obs+n_mis)] <- "Y2_mis"
index1 <- factor(index, levels = c("Y2_comp","Y2_obs","Y2_mis"))
Y2boxmcar <- c(Y2, Y2_MCAR_obs,Y2_MCAR_mis)
boxplot(Y2boxmcar ~ index1, boxwex = 0.25, col = "grey", cex.lab = 0.8,
        cex.axis = 1.2, ylab = expression(Y[2]), xlab ="MCAR")
```



We can use also the function `pbox` in `VIM` as illustrated below, where Y_1 is stratified on the basis of the missingness in Y_2. Again, there is no evidence that the two induced groups are different.

```
Y2_MCAR_NA <- c(Y2_MCAR_obs, rep(NA, n_mis))
df_NA <- data.frame("Y1" = Y1, "Y2" = Y2_MCAR_NA)
require(VIM)
pbox(df_NA, pos = 1)
```



We will now, only as an example, perform a t-test to check the plausibility of the MCAR assumption. The null hypothesis is that the means in the two Y_1 groups formed by stratifying on the basis of missingness in Y_2 are equal. We therefore should expect not to reject the null hypothesis in this case. But again, I reiterate, we need to acknowledge that we are working with a small sample size, and so results may not turned out as expected. So, for the above generated MCAR data we will now conduct the unpaired t-test. For more information, type `help(t.test)`.

```r
g1 <- Y1[-ind]; g2 <- Y1[ind]
mean(g1); mean(g2)
```

```
## [1] 122.2286
```

```
## [1] 122.3667
```

```r
t.test(g1, g2, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  g1 and g2
## t = -0.032984, df = 49.542, p-value = 0.9738
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.549290  8.273099
## sample estimates:
## mean of x mean of y
##  122.2286  122.3667
```

We obtain a p-value much larger than 0.05 and thus we do not reject the null hypothesis.

We will now impose a MAR mechanism, and we will do that in the following way: those who have measurements in February are those whose January's measurement exceed 140 (i.e., Y_1>140), a threshold used for diagnosing high blood pressure or hypertension.

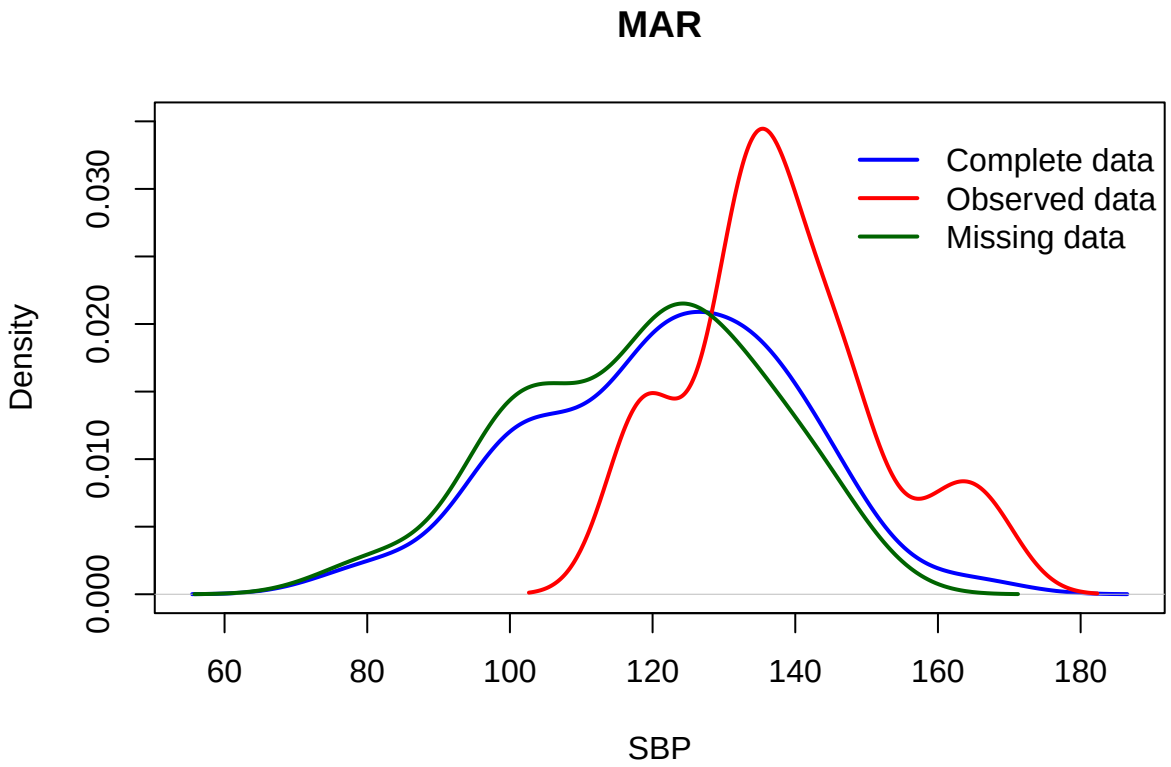
```r
Y2_MAR_obs <- Y2[Y1 > 140]
mean(Y2_MAR_obs); sd(Y2_MAR_obs)
```

```
## [1] 138
```

```
## [1] 13.79372
```

We can now also plot the densities of the complete, observed, and missing data. And as can be appreciated below the complete and observed data distributions are now quite different under data generated according to the MAR mechanism.

```r
Y2_MAR_mis <- Y2[Y1 < 140]
plot(density(Y2), lwd = 2, col = "blue", xlab = "SBP", main = "MAR", ylim = c(0,0.035))
lines(density(Y2_MAR_obs), lwd = 2, col = "red")
lines(density(Y2_MAR_mis), lwd = 2, col = "darkgreen")
legend(145, 0.035, 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")
```



With only an illustrative purpose, we conduct now again a t-test, and given that the missing values were now generated under the MAR assumption, we expect to reject the null hypothesis of equality of means.

```
ind_MAR <- which(Y1 > 140)
g1_MAR <- Y1[ind_MAR]; g2_MAR <- Y1[-ind_MAR]
mean(g1_MAR); mean(g2_MAR)
```

```
## [1] 151.125
```

```
## [1] 116.7738
```

```
t.test(g1_MAR, g2_MAR, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  g1_MAR and g2_MAR
## t = 14.158, df = 37.691, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  29.43803 39.26435
## sample estimates:
## mean of x mean of y
##  151.1250  116.7738
```

The p-value is much smaller than 0.05 and therefore we reject the null hypothesis of equality of means.

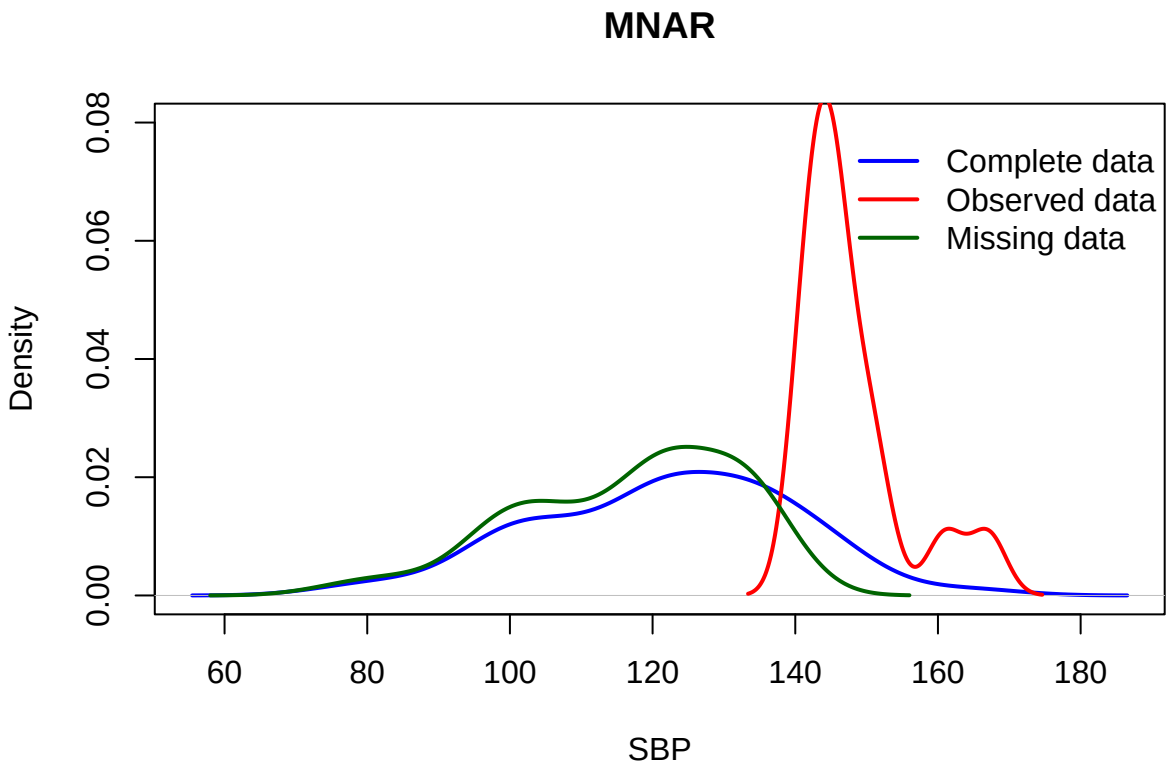
Finally, to induce a MNAR mechanism, the following was implemented: all individuals with measurements in February are those whose February measurement itself exceeded 140. This could happen, for instance, if all individuals had their measurements in February but the staff person only recorded it in case it was in the hypertensive range. MNAR could be induced in other ways, e.g., February measurement only recorded if it is substantial different from the January one. This is an example where the mising values depend both on the missing and observed values (because they depend on the difference between Y_1 and Y_2).

```r
Y2_MNAR_obs <- Y2[Y2 > 140]
mean(Y2_MNAR_obs); sd(Y2_MNAR_obs)
```

```
## [1] 147.7333
```

```
## [1] 7.41106
```

```r
Y2_MNAR_mis <- Y2[Y2 < 140]
plot(density(Y2), lwd = 2, col = "blue", xlab = "SBP", main = "MNAR", ylim = c(0,0.08))
lines(density(Y2_MNAR_obs), lwd = 2, col = "red")
lines(density(Y2_MNAR_mis), lwd = 2, col = "darkgreen")
legend(145, 0.08, 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")
```



In this example we can notice that as we move from MCAR to MAR to MNAR, the observed Y_2 values become an increasingly select and unusual group relative to the complete data. Although this phenomenon is not a universal feature of MCAR, MAR, and MNAR, it does happen in many realistic examples.

Numerical example from van Buuren (2nd edition, p. 37)

The aim of this example is also to simulate data from MCAR, MAR and MNAR mechanisms and it is, in essence, quite similar to the previous one. Let the data $Y = (Y_1, Y_2)$ be simulated from a standard bivariate normal distribution with correlation $\rho_{Y_1,Y_2} = 0.5$. Missing data are created in Y_2 using the missing data model

$$\Pr(R = 0 \mid Y_1, Y_2, \psi) = \psi_0 + \frac{e^{Y_1}}{1 + e^{Y_1}}\psi_1 + \frac{e^{Y_2}}{1 + e^{Y_2}}\psi_2,$$

with different parameter settings for $\psi = (\psi_0, \psi_1, \psi_2)$. For MCAR we set $\psi_{\text{MCAR}} = (0.5, 0, 0)$, for MAR we set $\psi_{\text{MAR}} = (0, 1, 0)$, and for MNAR we set $\psi_{\text{MNAR}} = (0, 0, 1)$. Thus, we obtain the following models:

- MCAR: $\Pr(R = 0 \mid Y_1, Y_2) = 0.5$.

- MAR: $\Pr(R = 0 \mid Y_1, Y_2) = \frac{e^{Y_1}}{1+e^{Y_1}}$ or, equivalently, $\text{logit}\{\Pr(R = 0)\} = Y_1$.
- MNAR: $\text{logit}\{\Pr(R = 0 \mid Y_1, Y_2)\} = Y_2$.

Here, $\text{logit}(p) = \log\{p/(1-p)\}$, for $0 < p < 1$, is the logit function. Note also that since only one variable has missing values, one missingness indicator suffices.

We start by generating the data and following the author we will generate 300 observations.

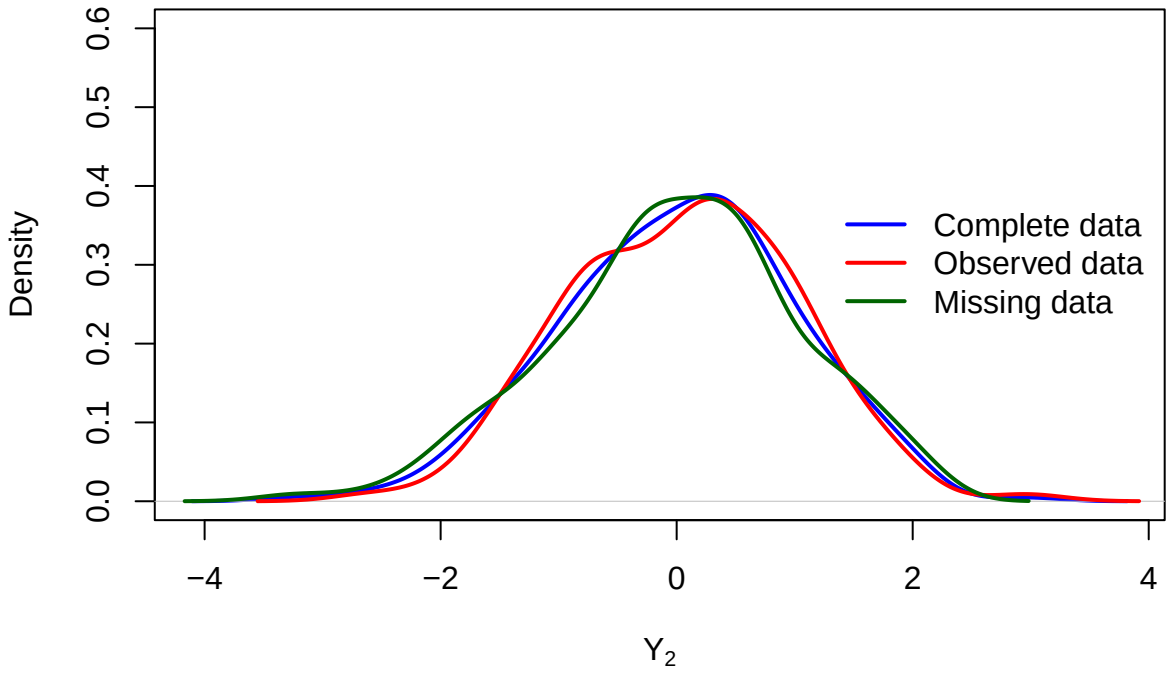
```
require(MASS)
set.seed(1)
n <- 300
Sigma <- matrix(c(1,0.5,0.5,1), nrow = 2, byrow = T)
Y <- mvrnorm(n = n, mu = c(0,0), Sigma = Sigma)
Y1 <- Y[,1]; Y2 <- Y[,2]
```

We will generate the missing data indicator R (which can only take the value 0 or 1) from a Bernoulli distribution. In `R` we can do that through the function `rbinom` (that generates random variables from a Binomial distribution) by letting `size=1`. We need to pass as input the probability of success (roughly speaking the probability of obtaining a one), which in our case, would correspond to $\Pr(R = 1 \mid Y_1, Y_2)$. For more information type `help(rbinom)`. Note that we are given the 'complementary' probability $\Pr(R = 0 \mid Y_1, Y_2)$. Since $\Pr(R = 1 \mid Y_1, Y_2) + \Pr(R_2 = 0 \mid Y_1, Y_2) = 1$, we then have $\Pr(R = 1 \mid Y_1, Y_2) = \frac{1}{1+e^{Y_1}}$. We will now impose missingness on Y_2 according to the three mechanisms stated above.

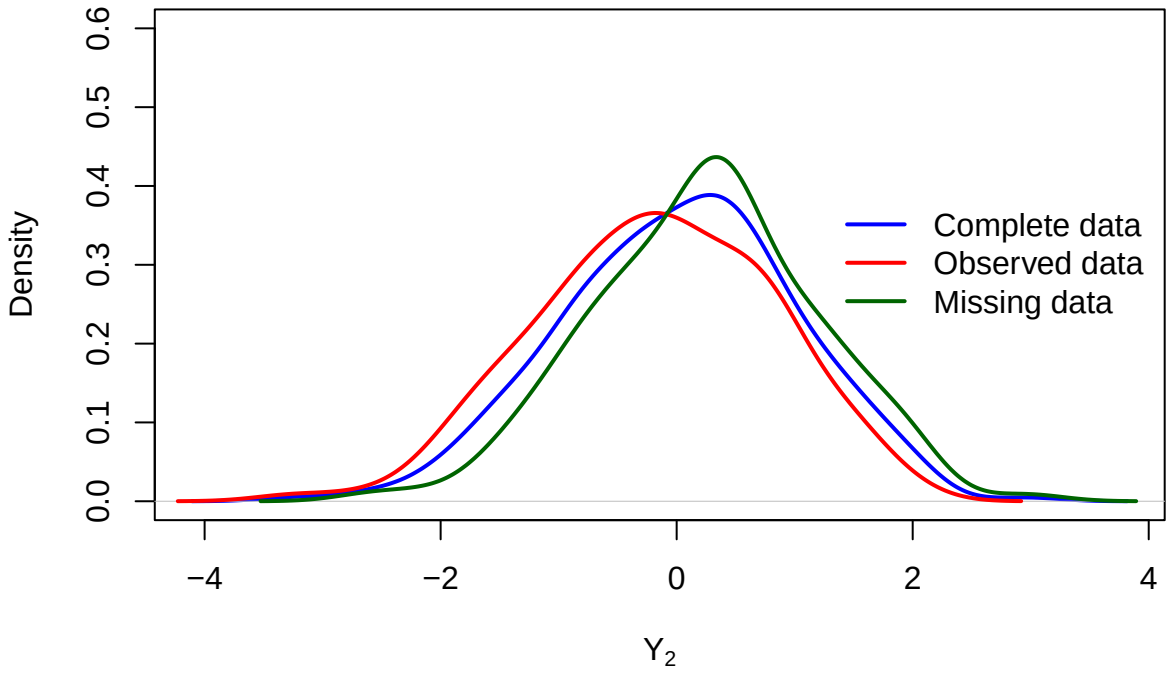
```
set.seed(1)
r_mcar <- rbinom(n = n, size = 1, prob = 0.5)
set.seed(1)
r_mar <- rbinom(n = n, size = 1, prob = 1/(1+exp(Y1)))
set.seed(1)
r_mnar <- rbinom(n = n, size = 1, prob = 1/(1+exp(Y2)))
```

We will now, as in the blood pressure example, plot the density of the complete, observed, and missing data for each mechanism.

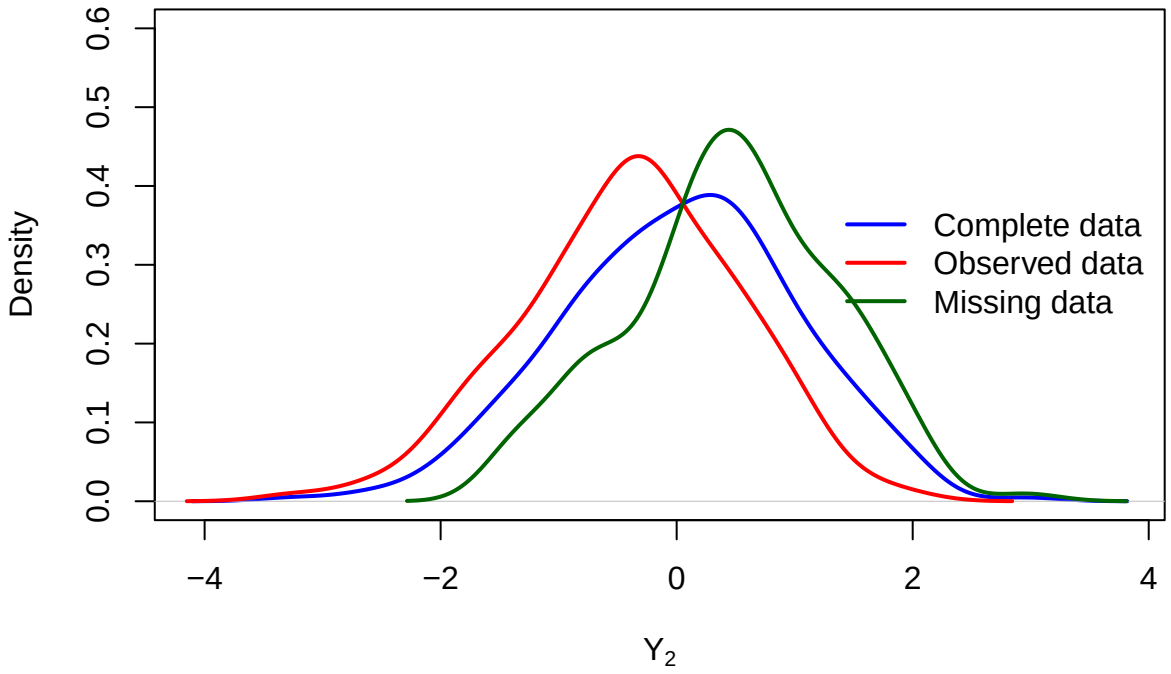
```
#MCAR
ind_mcar_obs <- which(r_mcar == 1)
Y2_MCAR_obs <- Y2[ind_mcar_obs]
Y2_MCAR_mis <- Y2[-ind_mcar_obs]
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "MCAR", ylim = c(0, 0.6))
lines(density(Y2_MCAR_obs), lwd = 2, col = "red")
lines(density(Y2_MCAR_mis), lwd = 2, col = "darkgreen")
legend(1.2, 0.4, 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")
```




```
#MAR
ind_mar_obs <- which(r_mar == 1)
Y2_MAR_obs <- Y2[ind_mar_obs]
Y2_MAR_mis <- Y2[-ind_mar_obs]
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "MAR", ylim = c(0, 0.6))
lines(density(Y2_MAR_obs), lwd = 2, col = "red")
lines(density(Y2_MAR_mis), lwd = 2, col = "darkgreen")
legend(1.2, 0.4, 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")
```




```
#MNAR
ind_mnar_obs <- which(r_mnar == 1)
Y2_MNAR_obs <- Y2[ind_mnar_obs]
Y2_MNAR_mis <- Y2[-ind_mnar_obs]
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "MNAR", ylim = c(0, 0.6))
lines(density(Y2_MNAR_obs), lwd = 2, col = "red")
lines(density(Y2_MNAR_mis), lwd = 2, col = "darkgreen")
legend(1.2, 0.4, 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 for the previous example, although not to the same extent, when moving from MCAR to MAR and to MNAR, the distributions of the complete, observed, and missing data become more distinct.

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin21.1.0 (64-bit)
## Running under: macOS Monterey 12.1
##
## Matrix products: default
## BLAS:   /usr/local/Cellar/openblas/0.3.19/lib/libopenblasp-r0.3.19.dylib
## LAPACK: /usr/local/Cellar/r/4.1.2/lib/R/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] VIM_6.1.1        colorspace_2.0-1 MASS_7.3-54
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6        knitr_1.33        magrittr_2.0.1    lattice_0.20-45
##  [5] rlang_0.4.11      fastmap_1.1.0     carData_3.0-5     highr_0.9
##  [9] stringr_1.4.0     car_3.0-12        tools_4.1.2       vcd_1.4-9
## [13] nnet_7.3-16       data.table_1.14.2 laeken_0.5.2      xfun_0.29
## [17] e1071_1.7-9       htmltools_0.5.2   class_7.3-19      lmtest_0.9-39
## [21] abind_1.4-5       yaml_2.2.1        digest_0.6.27     Matrix_1.3-4
```

```
## [25] robustbase_0.93-9 evaluate_0.14     rmarkdown_2.11    sp_1.4-6
## [29] proxy_0.4-26      stringi_1.6.2     ranger_0.13.1     compiler_4.1.2
## [33] DEoptimR_1.0-10   boot_1.3-28       zoo_1.8-9
```