Computer Assignment - Statistical methods

Laura Julià Melis

11/01/2019

1. Computer Exercises form Course's book

Do the following applet exercises from the book using R and write your comments on the results. Explain what you have learned from each exercise.

Exercice 4.84

First, we have to write the Gamma density function in R:

$$\left[\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right] y^{\alpha-1} e^{\frac{-y}{\beta}}$$

```
gamma_density <- function(alpha, beta, y){

  stopifnot(y>0)
  (1/gamma(alpha)*beta*alpha)*(y^(alpha-1))*(exp(1)^(-y/beta))

}
```

Then, we can create three vectors of 100 values such that each of them follows a gamma distribution of different parameters (the ones given in the exercice).

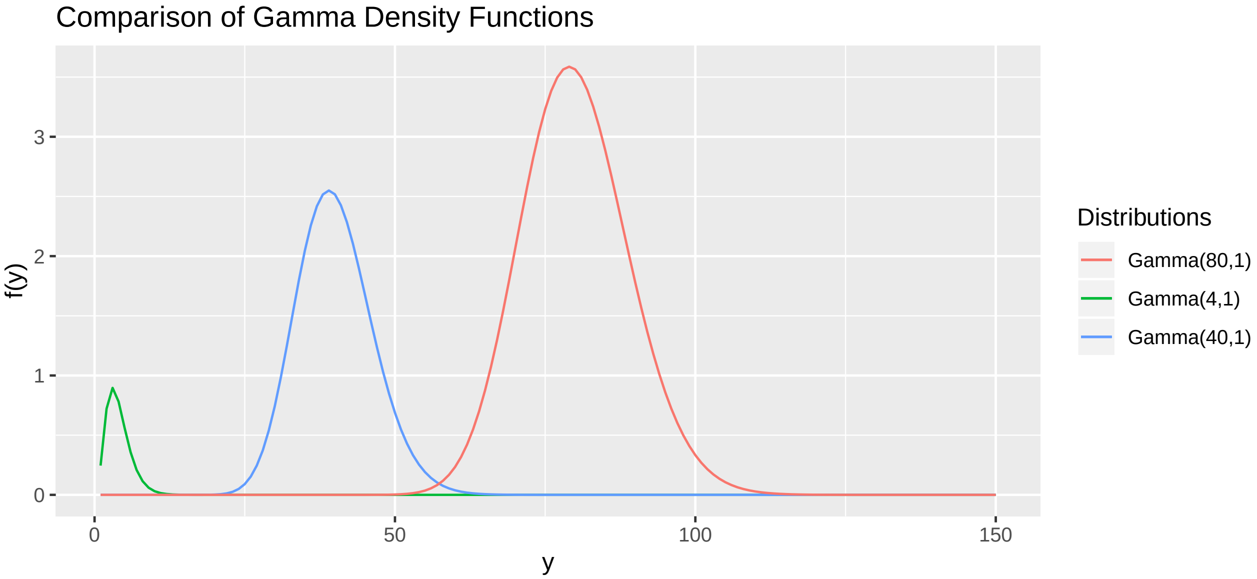
```
y <- 1:150

x1 <- gamma_density(alpha = 4, beta = 1, y)
x2 <- gamma_density(alpha = 40, beta = 1, y)
x3 <- gamma_density(alpha = 80, beta = 1, y)
```

Now that we have the points, we will plot them:

```
library(ggplot2)

df <-data.frame(x1, x2, x3)
ggplot(df) + geom_line(aes(x=y, y=x1, colour = "green"))
           + geom_line(aes(x=y, y=x2, colour = "red"))
           + geom_line(aes(x=y, y=x3, colour = "darkblue"))
           + ggtitle("Comparison of Gamma Density Functions")
           + ylab("f(y)")
           + scale_color_discrete(name = "Distributions",
                                  labels = c("Gamma(80,1)", "Gamma(4,1)", "Gamma(40,1)"))
```



QUESTIONS:

a. What do you observe about the shapes of these three density functions? Which are less skewed and more symmetric?

The shape of the distribution ploted in green (the one with the smallest value of α) is the most skewed distribution. Regarding the other two distributions, we can observe that the greater the value of α, the more symmetric the distribution is: the red line is the most symmetric density curve.

b. What differences do you observe about the location of the centers of these density functions?

For large values of α, the location of the center of the density functions is greater (right-sided).

c. Give an explanation for what you observed in part (b).

What has been observed in section (b) can be explained in terms of the expected value of a Gamma distribution. This is:

$$E[Y] = \alpha \cdot \beta$$

As the expected value is actually the center of the distribution, is understandable that in our distributions, the one with the greater α value is the one colored in red (in the right).

Exercice 4.117

The density function of a beta distribution is:

$$\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] y^{\alpha-1}(1-y)^{\beta-1}$$

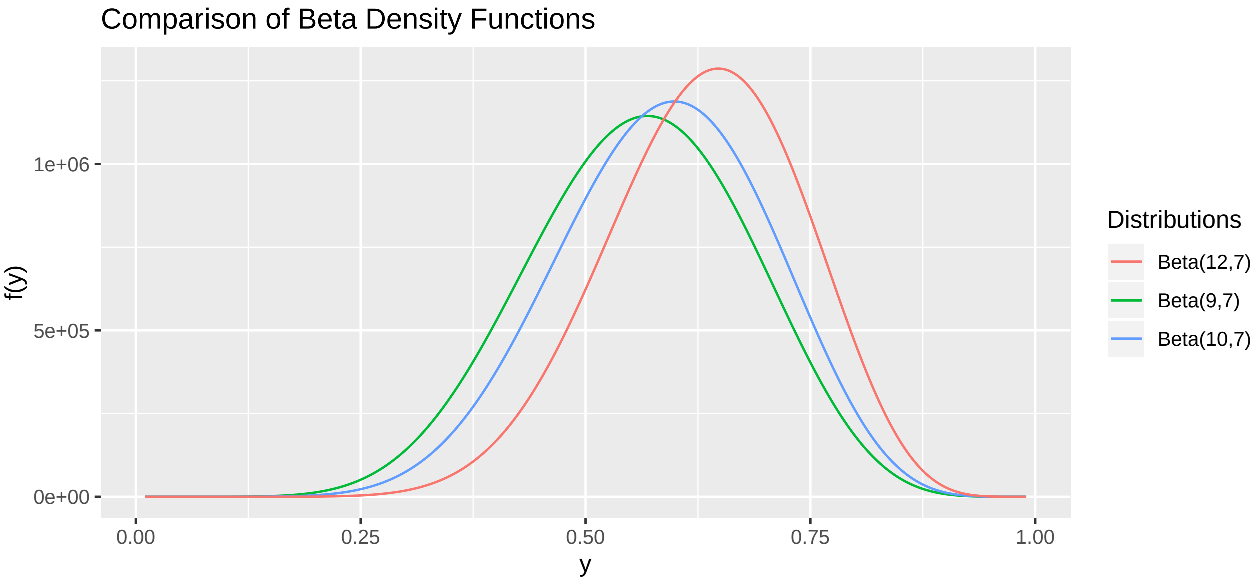
So first we will implement a function that calculates the density function.

```
beta_density_fun <- function(alpha, beta, y){
   stopifnot(y>0, y<1)
  (gamma(alpha+beta)/gamma(alpha)*gamma(beta))*(y^(alpha-1))*((1-y)^(beta-y))
}
```

And now, we can plot the different distributions:

```
y <- seq(0.01,0.99, by=0.001)
x1 <- beta_density_fun(alpha = 9, beta = 7, y)
x2 <- beta_density_fun(alpha = 10, beta = 7, y)
x3 <- beta_density_fun(alpha = 12, beta =7, y)
df <-data.frame(x1, x2, x3)
```

```
ggplot(df) + geom_line(aes(x=y, y=x1, colour = "green"))
           + geom_line(aes(x=y, y=x2, colour = "red"))
           + geom_line(aes(x=y, y=x3, colour = "darkblue"))
           + ggtitle("Comparison of Beta Density Functions")
           + ylab("f(y)")
           + scale_color_discrete(name = "Distributions",
                                  labels = c("Beta(12,7)", "Beta(9,7)", "Beta(10,7)"))
```



QUESTIONS:

a. Are these densities symmetric? Skewed left? Skewed right?

No, all of them are skewed to the left.

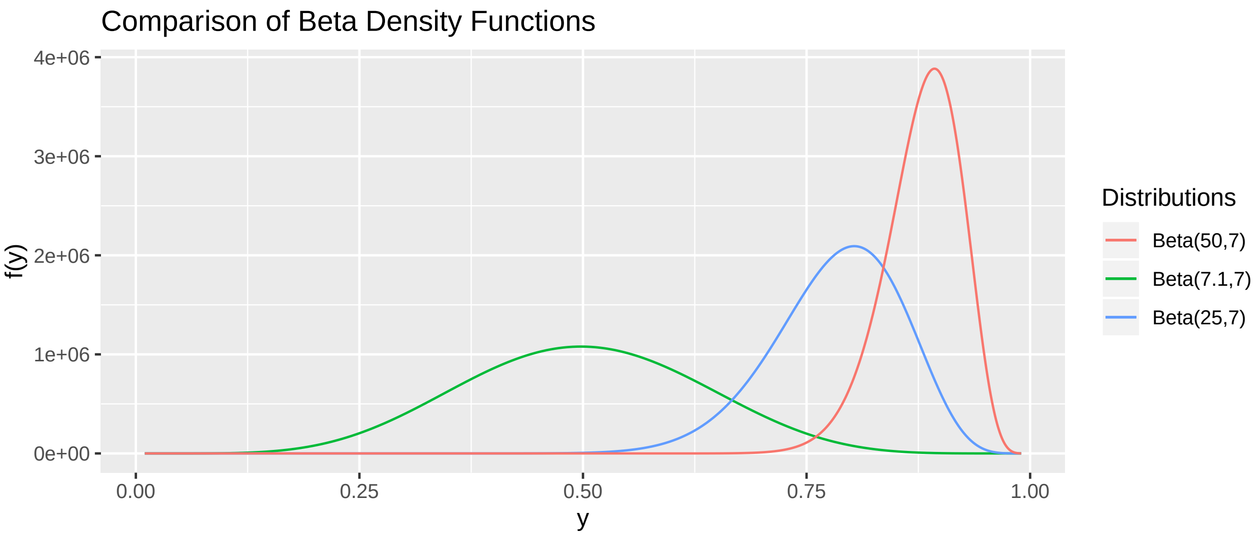
b. What do you observe as the value of α gets closer to 12?

Th closer the value of α is to 12, the more symmetric the density function is.

c. Graph some more beta densities with $\alpha > 1$, $\beta > 1$, and $\alpha > \beta$. What do you conjecture about the shape of beta densities with $\alpha > \beta$ and both $\alpha > 1$ and $\beta > 1$?

```
x1 <- beta_density_fun(alpha = 7.1, beta = 7, y)
x2 <- beta_density_fun(alpha = 25, beta = 7, y)
x3 <- beta_density_fun(alpha = 50, beta =7, y)
df <-data.frame(x1, x2, x3)
```

```
ggplot(df) + geom_line(aes(x=y, y=x1, colour = "green"))
           + geom_line(aes(x=y, y=x2, colour = "red"))
           + geom_line(aes(x=y, y=x3, colour = "darkblue"))
           + ggtitle("Comparison of Beta Density Functions")
           + ylab("f(y)")
           + scale_color_discrete(name = "Distributions",
                                  labels = c("Beta(50,7)", "Beta(7.1,7)", "Beta(25,7)"))
```



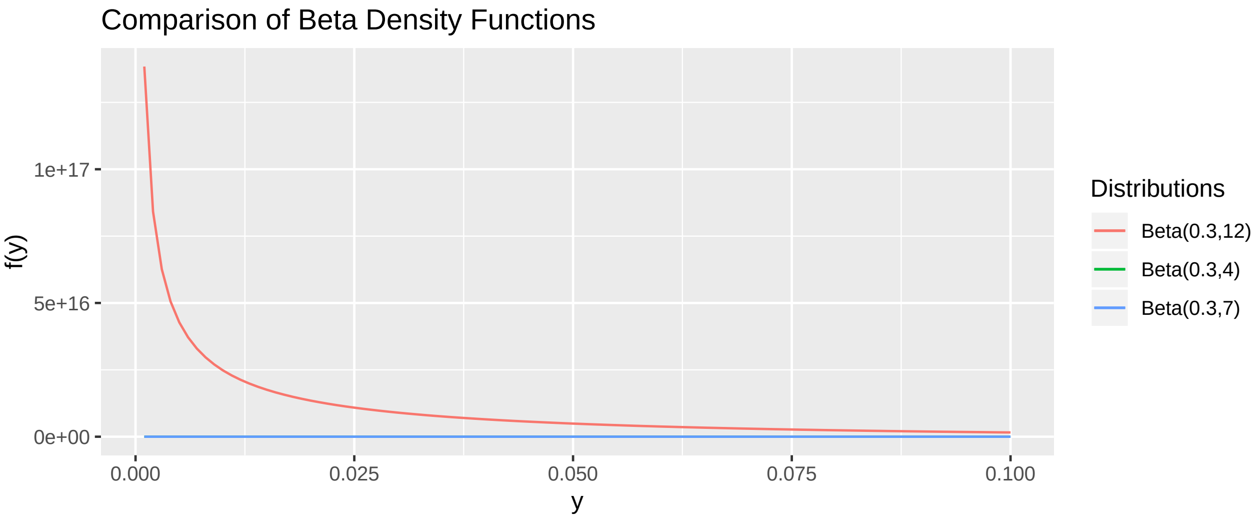
When the values of α and β are similar, the density curve is quite symmetrical (green curve) but when the difference between these values increases, the curve becomes increasingly left-skewed. So, when $\alpha > \beta$ $(\alpha > 1, \beta > 1)$ the shape is always skewed.

Exercice 4.118

In this exercise we will use the `beta_density_guntion()` function to create three new different distributions.

```
y <- seq(0.001,0.1, by=0.001)
x1 <- beta_density_fun(alpha = 0.3, beta = 4, y)
x2 <- beta_density_fun(alpha = 0.3, beta = 7, y)
x3 <- beta_density_fun(alpha = 0.3, beta =12, y)
df <-data.frame(x1, x2, x3)
```

```
ggplot(df) + geom_line(aes(x=y, y=x1, colour = "green"))
           + geom_line(aes(x=y, y=x2, colour = "red"))
           + geom_line(aes(x=y, y=x3, colour = "darkblue"))
           + ggtitle("Comparison of Beta Density Functions")
           + ylab("f(y)")
           + scale_color_discrete(name = "Distributions",
                                  labels = c("Beta(0.3,12)", "Beta(0.3,4)", "Beta(0.3,7)"))
```



QUESTIONS:

a Are these densities symmetric? Skewed left? Skewed right?

These densities are not symmetric, all of them are skewed right.

b What do you observe as the value of β gets closer to 12?

When β is close to 12 the points (y) around 0 are much more high compared to the value of the points when $\beta < 12$. Then, as β gets closer to 12, the decrease in the curve is much greater.

- Note that the limits of the y axis are so large that we can't see the green curve.

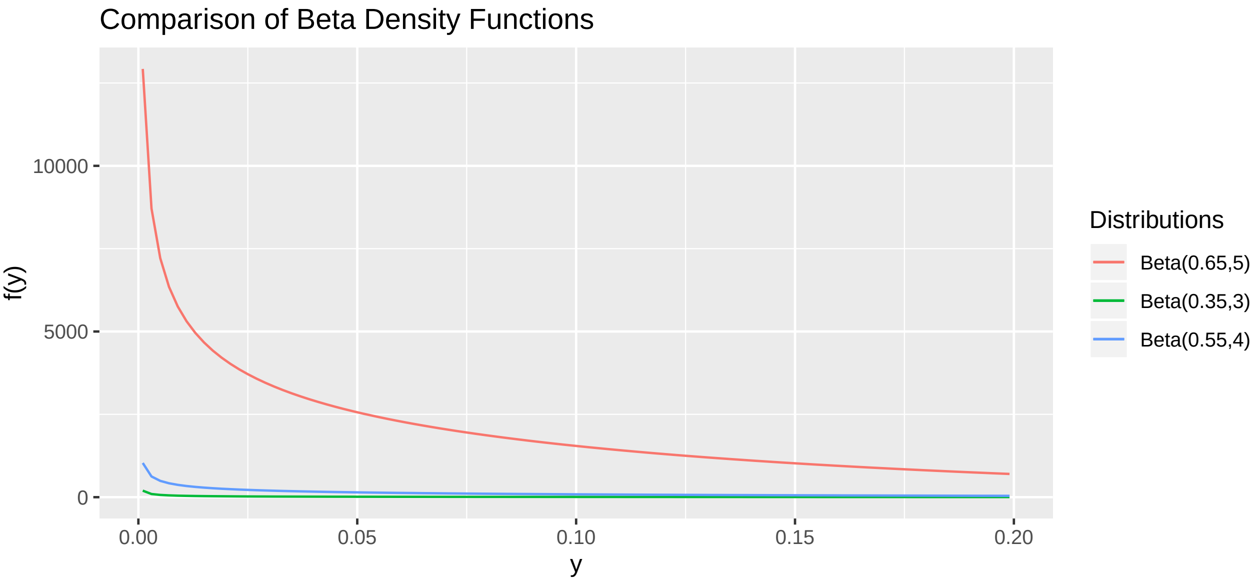
c Which of these beta distributions gives the highest probability of observing a value larger than 0.2?

The distribution with $alpha =0.3 $ and $\beta = 12$ has the highest probability:

d Graph some more beta densities with $alpha < 1$ and $\beta > 1$. What do you conjecture about the shape of beta densities with $alpha < 1$ and $\beta > 1$?

```
y <- seq(0.001,0.2, by=0.002)
x1 <- beta_density_fun(alpha = 0.35, beta = 3, y)
x2 <- beta_density_fun(alpha = 0.55, beta = 4, y)
x3 <- beta_density_fun(alpha = 0.65, beta =5, y)
df <-data.frame(x1, x2, x3)
```

```
ggplot(df) + geom_line(aes(x=y, y=x1, colour = "green"))
           + geom_line(aes(x=y, y=x2, colour = "red"))
           + geom_line(aes(x=y, y=x3, colour = "darkblue"))
           + ggtitle("Comparison of Beta Density Functions")
           + ylab("f(y)")
           + scale_color_discrete(name = "Distributions",
                                  labels = c("Beta(0.65,5)", "Beta(0.35,3)", "Beta(0.55,4)"))
```



When $alpha < 1$ and $\beta > 1$, the shape of the beta density function is always decreasing and skewed with a right tail.

Exercice 10.19

The hypothesis that we want to test are:

$$H_o : \mu = 130$$

$$H_a : \mu < 130$$

where μ represents the average output voltage for the electric circuit.

1. The data

```
n <- 40
mean <- 128.6
sd <- 2.1
```

2. The statistic of the test

As the n of our sample is large enough, we can approximate our sample distribution to a normal distribution ($\bar{Y} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$). Then, the formula to calculate the statistic of our test is the following:

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

```
z <- (mean-130)/(sd/sqrt(n))
z
```

```
## [1] -4.21637
```

3. The rejection region

In our test, H_a is a lower-tail alternative so the rejection region will be defined by:

$$z < -z_{alpha}$$

```
alpha <- 0.05 # given in the exercise
z_alpha <- qnorm(alpha,mean=0,sd=1, lower.tail = T)
z_alpha
```

```
## [1] -1.644854
```

4. Conclusion

As $z < -z_{alpha} \rightarrow -4.22 < -1.64$, we can reject H_0 and we conclude that the average output voltage for the electric circuit is less than 130.

Exercice 10.21

We have to test the following hypothesis:

$$H_o : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 denote the true mean of shear strength in the Type I and Type II soil, respectively.

1. The data

```
n1 <- 30
n2 <- 35
mean1 <- 1.65
mean2 <- 1.43
sd1 <- 0.26
sd2 <- 0.22
```

2. The statistic of the test

The formula to calculate the statistic of this test is:

$$z = \frac{\bar{y}_1 - \bar{y}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

wher D_0 in our case is 0.

```r
z <- (mean1-mean2-0)/(sqrt((sd1^2/n1)+(sd1^2/n1)))
z
```

```
## [1] 3.27714
```

3. The rejection region

In our test, H_a is a two-tailed alternative so the rejection region will be defined by:

$$|z| > z_{alpha/2}$$

```r
alpha <- 0.01 # given in the exercise
z_alpha <- qnorm(alpha/2,mean=0,sd=1, lower.tail = F)
z_alpha
```

```
## [1] 2.575829
```

4. Conclusion

As $|z| > z_{alpha/2} \rightarrow 3.28 > 2.57$, we can reject H_0 and we conclude that the soils do appear to differ with respect to average shear strength, at the 1% significance level.

Exercice 11.31

From the information given in the exercise, we may want to fit a linear regressionmodel:

$$Y = \beta_0 + \beta_1 \cdot X_1$$

where Y represents the peak current generated and X_1, the amount of nickel.

As we want to test if peak current increases as nickel concentrations increase, our hypothesis will be:

$$H_o : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

We will use the `lm()` function to test these hypothesis:

```r
x1 <- c(19.1, 38.2, 57.3, 76.2, 95, 114, 131, 150, 170)
y <- c(0.095, 0.174, 0.256, 0.348, 0.429, 0.500, 0.580, 0.651, 0.722)
model<- lm(y~x1)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0133264 -0.0042777 -0.0000231  0.0080557  0.0098107
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.875e-02  6.129e-03   3.059   0.0183 *
## x1          4.215e-03  5.771e-05  73.040 2.37e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008376 on 7 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9985
## F-statistic:  5335 on 1 and 7 DF,  p-value: 2.372e-11
```

In the 'Coefficients' table we can see that the p value associated to the x_1 variable is much more smaller than $\alpha = 0.05$ (and even smaller if we divide the p-value by 2, because we have a two-tailed alternative hypothesis) giving us reasons to conclude that the statistic is significant and to reject the null hypothesis. Then, we have enough evidences to say that peak current increases as nickel concentrations increase

Exercice 11.69

QUESTIONS:

a. Letting Y denote sales and x denote the coded year (-7 for 1996, -5 for 1997, through 7 for 2003), fit the model $Y = \beta_0 + \beta_1 x + \epsilon$.

```r
x <- seq(-7,7,2) # coded years
y <- c(18.5, 22.6, 27.2, 31.2, 33, 44.9, 49.4, 35)
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      32.725        1.812
```

The fitted model, calculated using the `lm()` function, is: $\hat{y} = 32.725 + 1.812x$

b. For the same data, fit the model $Y = \beta_0 + \beta_1 x + \beta_2 x2 + \epsilon$.

```r
x2 <- x^2
lm(y~x+x2)
```

```
##
## Call:
## lm(formula = y ~ x + x2)
##
## Coefficients:
## (Intercept)            x           x2
##     35.5625       1.8119      -0.1351
```

The fitted quadratic model (adding as a covariable the variable x to the power of 2) is: $\hat{y} = 35.5625 + 1.8119x - 0.1351x^2$

2. Imputations techniques

For the course 732A93 you shall read the chapter on the link: http://www.stat.columbia.edu/~gelman/arm/missing.pdf and include in your report the answers to the following:

1. Which type of missing mechanism do you prefer to get a good imputation?

In the *Missing-data imputation* document it is explained that there are four different missing mechanisms:

(i) Missingness completely at random.
(ii) Missingness at random.
(iii) Missingness that depends on unobserved predictors.
(iv) Missingness that depends on the missing value itself.

I would say that the best situation for which there are missing values, in order to get a good imputation, is the one in which the missings are completely at random. When this is the reason why there is missing data in a dataset, this means that the probability of missingness is the same for all cases with NA's. In other words, there is no relationship between cases with missing values and any other variable.

I think this is the ideal type of missing mechanism because when this happens it is possible to not take into account the cases with missing values and, still, obtain unbiased results.

However, it is important to mention that this is not the most common situation in real life problems and datasets.

2. Say something about simple random imputation and regression imputation of a single variable.

The two following imputation performances are two different approaches to random imputation.

- Simple random imputation.

This is the simplest method of imputation. It consists of randomly assigning the value of other cases (cases without missings) to all those cases with NA's. This is, basically, imputing missing values of a variable based on the observed information (the data) of that variable.

- Regression imputation of a single variable.

To use regression imputation one has to fit a regression model in which the dependent variable is the variable with missing values. Then, this model has to be used to make predictions that will be used to substitute the missing values in the variable.

3. Explain shortly what Multiple Imputation is.

Multiple Imputation is a method of imputation used when missing values are not at random. This technique replaces each missing value in the dataset with multiple imputed values (more than one) in order to reflect the uncertainty caused by the presence of missing data in the imputation model. These simulated values are obtained by making predictions of different models.

As we will obtain multiple imputed values, we will have multiple datasets and for this reason, to analyze the data and calculate inferences, it will be necessary to combine the estimations.