## Question 1 (33/100 points)

(a) **[8 points]** Consider the following data points

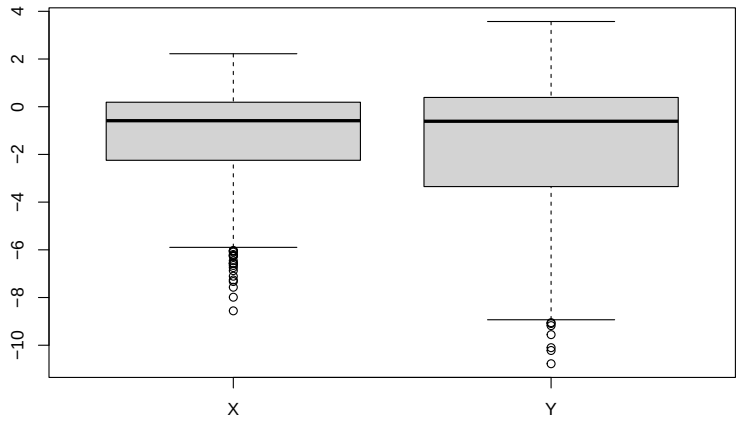
$$-6.7; \quad -2.5; \quad 3.2; \quad -2.1;$$

Obtain the sample mean and the sample variance.

**Solution:**

The sample mean is $-2.02$ and the sample variance is 16.46.

(b) **[15 points]** You have available a dataset that contains two variables x and y. For each variable, you have obtained the boxplot given below (boxplot of $x$ is on the left and boxplot of $y$ is on the right).




A colleague of yours makes the following statements:

(i) *"I expect both variables to have a negative skewness"*.

(ii) *"I expect the sample variance of x to be smaller than the sample variance of y"*.

(iii) *"I expect the kurtosis of y to be larger than the kurtosis of x"*.

(iv) *"I expect the variables to have a positive correlation $r_{xy}$"*.

For each statement, say whether you agree or not. Justify your answers.

**Solution:**

(i) I agree with the statement. The boxplot indicates that the left tail of both distributions is heavier than the right tail and therefore we can expect negative skewness.

(ii) I agree, we can see that the boxplots indicate that the dispersion of $y$ is larger than the dispersion of $x$. Therefore, we expect larger variance as the variance is a measure of dispersion.

(iii) I do not agree with the statement. The variable $x$ seems to have heavier tails as indicated by the larger number of extreme values (outliers) on the left tail.

(iv) I disagree, boxplots do not provide any information on the relationship between the two variables.

(c) **[10 points]** The R vectors "age" and "salary" contain the age and the monthly salary of 1000 individuals. The following R code is given:

```
n <- length(age)
out <- rep(0,n)
k <- 1
while(k <= n){
    if(age[k]>45){
        out[k] <- income[k]
        if(income[k]<=mean(income)){out[k] <- 0}
    }
  k <- k+1
}
```

Explain briefly what the R code is doing. What is contained in "out" after the *while loop*?

How would you write some R code that produces the same result but without using a loop? Sketch the code and explain what it does.

**Solution:**

First, the vector out is created setting all its elements to zero. Then, the R code runs a *while loop* for $k$ from 1 to $n$. Within the loop, if an individual has age larger than 45, the corresponding element of out is set to the income of the individual. Additionally, if the income of the individual is below the mean of the incomes, the corresponding element of out is set back to zero. The vector out will contain the income of individuals with age larger than 45 and with income larger than the average income. The remaining elements of out will be equal to zero.

The same result can be obtained using a logical operator as follows:

```
out <- rep(0,length(age))
select <- (age>45)&(income>mean(income))
out[select] <- income[select]
```

**Question 2 (34/100 points)**

(a) **[10 points]** You have available a dataset that contains the variables `math_score` and `country` for some high school students. The variable `math_score` reports the result of an international math test and the variable `country` indicates the country of residence of the student. The variable `country` takes 3 possible values: 0 if the student is a resident of Belgium, 1 if the student is a resident of The Netherlands, and 2 if the student is a resident of Germany. You are interested in regressing `math_score` on `country`. Write down the regression model you would consider. Justify your choice. Discuss the interpretation of the regression coefficients of the model you have proposed.

**Solution:**

The `country` is a categorical variable. I would therefore create a dummy variable `Netherlands` that is 1 when the individual is a resident of The Netherlands and 0 otherwise, and a dummy variable `Germany` that is 1 when the individual is a resident of Germany. I would then consider the following regression model:

$$\text{math\_score} = \beta_0 + \beta_1 \times \text{Netherlands} + \beta_2 \times \text{Germany} + \text{error}.$$

The coefficient $\beta_0$ indicates the expected math score of a student resident in Belgium. The coefficient $\beta_1$ indicates expected difference in math score between a student resident in The Netherlands and one resident in Belgium. Finally, the coefficient $\beta_2$ indicates expected difference in math score between a student resident in Germany and one resident in Belgium.

(b) **[14 points]** Available is a dataset with 2 variables and $n = 12$ observations for each of the 2 variables. Consider a linear regression model of the form $y_i = \beta_0 + \beta_1 x_i + u_i$. The OLS estimates of $\beta_0$ and $\beta_1$, the $R^2$ and the explained sum of squares (ESS) are obtained:

$$\hat{\beta}_0 = -6.5, \quad \hat{\beta}_1 = -3.1, \quad R^2 = 0.90, \quad ESS = 122.5.$$

(i) Obtain a prediction of $y$ given $x = 3.5$.

(ii) Obtain the standard error of the regression ($SER$).

(iii) Obtain the sample variances of the variables $s_x^2$ and $s_y^2$.

**Solution:**

(i) The prediction from the model is given by $-6.5 + 3.5 \times (-3.1) = -17.35$.

(ii) First, we obtain the TSS as follows

$$\text{TSS} = \text{ESS}/R^2 = 136.11.$$

Then, we get the RSS as

$$\text{RSS} = \text{TSS} - \text{ESS} = 136.11 - 122.5 = 13.61.$$

Finally, the SER is

$$\text{SER} = \sqrt{\frac{RSS}{n-2}} = \sqrt{1.361} = 1.17$$

(iii) The sample variance of $y$ is obtained as

$$s_y^2 = \frac{\text{TSS}}{n-1} = 136.11/11 = 12.37.$$

The sample variance of $x$ can be obtained from the equation $\hat{\beta}_1 = r_{xy}s_y/s_x$. In particular, we obtain that

$$s_x^2 = \frac{r_{xy}^2 s_y^2}{\hat{\beta}_1^2} = \frac{R^2 s_y^2}{\hat{\beta}_1^2} = \frac{0.9 \times 12.37}{9.61} = 1.16.$$

(c) **[10 points]** A colleague of yours has estimated the following regression models using a variable of interest $y_i$ and 2 regressors, $x_{1,i}$ and $x_{2,i}$, $i = 1, \ldots, n$.

(1) $y_i = \beta_0 + \beta_1 x_{1,i} + u_i$.
(2) $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + u_i$.
(3) $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i$.

Your colleague makes the following 2 statements:
(i) "*If the adjusted $R^2$ (adj-$R^2$) of model (1) is larger than the adj-$R^2$ of model (2), we can conclude that the relationship between y and $x_1$ is linear.*"
(ii) "*I have obtained that the $R^2$ of model (3) is larger than the $R^2$ of model (2). Instead, the adj-$R^2$ of model (2) is larger than the adj-$R^2$ of model (3). There must be an error since $R^2$ and adj-$R^2$ provide the same information.*"

Comment on each statement and say whether you agree or not. Justify your answers.

**Solution:**
(i) It is true that if the adj-$R^2$ of model (1) is larger than the adj-$R^2$ of model (2), then the quadratic term $x_1^2$ is not useful to better explain the relationship between the variable $y$ and $x_1$. However, this does not necessarily rule out the possibility of a nonlinear relationship.
(ii) It is true that there must be an error as the number of coefficients of model (2) and model (3) are the same and therefore if $R^2$ of model (3) is larger, then adj-$R^2$ has to be larger as well. This is due to the fact that the penalty term will be the same. However, in general, it is not true that $R^2$ and adj-$R^2$ provide the same information.

**Question 3 (33/100 points)**

(a) **[15 points]** We have an observation $x$ that we want to classify as a member of any of the three populations $\Pi_1$, $\Pi_2$ and $\Pi_3$. We know that population $\Pi_1$ has an exponential distribution[1] with rate $\lambda = 1$, population $\Pi_2$ has an exponential distribution with rate $\lambda = 2$ and population $\Pi_3$ has an exponential distribution with rate $\lambda = 4$.

(i) Obtain the discriminant regions $R_1$, $R_2$ and $R_3$ based on the Maximum Likelihood (ML) discriminant rule.

(ii) Obtain the probabilities of correct classification $p_{11}$, $p_{22}$ and $p_{33}$ of the ML rule.

**Solution:**

(i) The densities are $f_1(x) = e^{-x}$, $f_2(x) = 2e^{-2x}$ and $f_3(x) = 4e^{-4x}$. We obtain that $f_1(x) > f_2(x)$ for $x > \log(2)$, $f_2(x) > f_3(x)$ for $x > \log(2)/2$ and $f_1(x) > f_3(x)$ for $x > 2\log(2)/3$. Therefore,

$$R_1 = (\log(2), \infty), \quad R_2 = (\log(2)/2, \log(2)), \quad R_2 = (0, \log(2)/2).$$

(ii) The correct classification probabilities are

$$p_{11} = \int_{\log(2)}^{\infty} e^{-x} dx = e^{-\log(2)} = \frac{1}{2},$$

$$p_{22} = \int_{\log(2)/2}^{\log(2)} 2e^{-2x} dx = e^{-\log(2)} - e^{-\log(4)} = \frac{1}{4},$$

and

$$p_{33} = \int_{0}^{\log(2)/2} 4e^{-4x} dx = 1 - e^{-\log(4)} = \frac{3}{4}.$$

(b) **[8 points]** Consider the ML discriminant rule with two normal populations with means $\mu_1$ and $\mu_2$, $\mu_1 > \mu_2$, and the same variance $\sigma^2$. The discriminant regions are $R_1 = \{x : x > \frac{\mu_2+\mu_1}{2}\}$ and $R_2 = \{x : x \leq \frac{\mu_2+\mu_1}{2}\}$.

Show that the misclassification probabilities are given by

$$p_{12} = p_{21} = \Phi\left(-\frac{\mu_1 - \mu_2}{2\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

**Solution:**

See course material.

[1] The probability density function of an exponential distribution with rate $\lambda > 0$ is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

(c) **[10 points]** You have implemented for a given dateset the ML discriminant rule based on two normal populations with means $\mu_1$ and $\mu_2$, $\mu_1 > \mu_2$, and the same variance $\sigma^2$. A colleague of yours suggest to estimate the misclassification probabilities $p_{12}$ and $p_{21}$ as follows

$$\hat{p}_{12} = \hat{p}_{21} = \Phi\left(-\frac{\bar{x}_1 - \bar{x}_2}{2s}\right),$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample mean of the observations from populations 1 and 2, and $s$ is the sample standard deviation. Discuss potential advantages (if any) and disadvantages (if any) of the method proposed by your colleague. Could you present an alternative approach to estimate $p_{12}$ and $p_{21}$?

**Solution:**

The estimate proposed by the colleague is a reasonable way to estimate the misclassification probability as it plugs in the sample means and variance in the expression of the misclassification probability under normal assumption with equal variances. However, these estimates can lead to misleading results if the assumption of normal populations is not true as, for instance, it imposes that $\hat{p}_{12} = \hat{p}_{21}$. Instead, an alternative way to estimate the misclassification probabilities that does not rely on any assumption on the distributions is given by

$$\hat{p}_{12} = \frac{n_{12}}{n_1}, \quad \hat{p}_{21} = \frac{n_{21}}{n_2},$$

where $n_{12}$ and $n_{21}$ is the number of misclassified observations from populations 1 and 2, and $n_1$ and $n_2$ is the the total number of observations from populations 1 and 2.