Séance 7 - Couples et conditionnement

Semaine 7

Table des matières

Objectif

L' objectif est dans un premier temps de comprendre les définitions liées à la dépendance de variables aléatoires. Ensuite, il s'agit d'introduire des exemples permettant d'appréhender le lien entre les notions de corrélation et de variance expliquée (formule d'Eve).

Exercice 1

On considère un couple de variables aléatoires (X, Y) tel que la loi de X est la loi Gamma(2,1). Pour tout $x > 0$, la loi conditionnelle de la variable aléatoire Y de sachant $X = x$ est la loi exponentielle de paramètre x.

Question 1

Déterminer la densité de la loi du couple (X, Y).

Solution.

On rappelle que la loi Gamma(2,1) admet pour densité

$$f_X(x) = xe^{-x}\,, \quad x > 0\,.$$

Alors on obtient:

pour $x > 0$ et $y > 0$

$f(x, y) = f_X(x) f_Y(y/X = x) = xe^{-x} xe^{-xy} = x^2 e^{-x(1+y)}$

Question 2

Déterminer la densité de la loi de la variable Y.

Solution.

$$f_Y(y) = \frac{2}{(1+y)^3}$$

Question 3

Vérifier le calcul précédent à l'aide de la simulation d'un échantillon de taille $n = 1000$. Afficher un graphe quantile-quantile pour vérifier que la simulation est correcte.

Solution. Pour afficher un graphe quantile-quantile, on montre d'abord que la fonction de répartition de la loi de Y est égale à

$$\mathbb{P}(Y \leq t) = 1 - \frac{1}{(1+t)^2}\,, \quad t > 0\,.$$

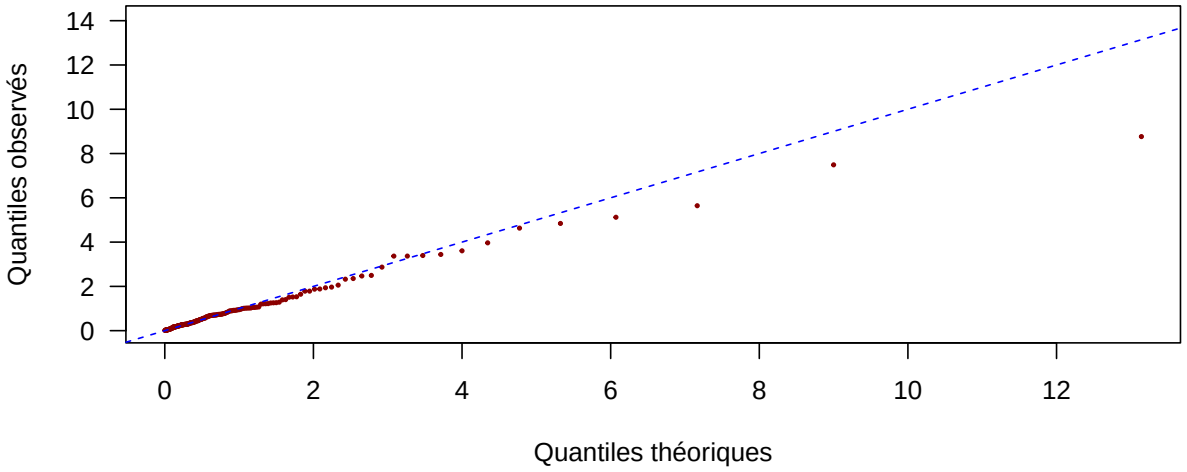
Puis on inverse cette fonction, pour obtenir la fonction quantile théorique.

```
set.seed(10101)
# taille de l'échantillon
n = 200

# simulation du couple (x,y)
# loi marginale
x = rgamma(n, 2, rate = 1)
# loi conditionnelle
y = rexp(n,x)

# fonction quantile
Q_Y <- function(u){
  return(1/sqrt(1-u)-1)
}

# on trace un graphe quantile_quantile pour vérifier si
#la distribution empirique de Y correspond à la loi théorique donnée par Q_Y.
u = (1:n)/n
plot(Q_Y(u), sort(y),
     xlab = "Quantiles théoriques",
     ylab = "Quantiles observés",
     pch = 19,
     col = "red4", las = 1,
     cex = .3)
abline(0,1, col = "blue", lty = 2)
```



Le graphe quantile-quantile montre que les quantiles empiriques sont proches des quantiles théoriques de la loi de Y, donc que la simulation est correcte.

Question 4

Calculer l'espérance conditionnelle $\mathbb{E}[Y|X]$. En déduire l'espérance de la variable $Z = 1/X$ à l'aide de la formule de l'espérance totale. Vérifier le calcul par une simulation et en calculant directement $\mathbb{E}[Z]$.

Solution.

Soit $x > 0$, la variable aléatoire $Y/X = x$ suit la loi exponentielle de paramètre x , donc $E(Y/X = x) = \frac{1}{x}$, alors en remplaçant x par X , on obtient:

$$E(Y/X) = \frac{1}{X}$$

On pose la variable aléatoire $Z = \frac{1}{X}$, alors $E(Z) = E(E(Y/X))$

En appliquant la formule de l'espérance totale à la variable aléatoire Y, on obtient :

$$E(Y) = E(E(Y/X)) = \int_0^{+\infty} \mathbb{E}(Y \mid X = x)\, f_X(x)\, dx = \int_0^{+\infty} e^{-x} dx = 1$$

Alors $E(Y) = E(Z) = 1$

```r
set.seed(10101)
# taille de l'échantillon
n = 10000

# simulation du triplet (x,y,z)
# loi marginale
x = rgamma(n, 2, rate = 1)
# loi conditionnelle
y = rexp(x)
```

```r
# simulation de z
z=1/x
cat("l'espérance de Y est",mean(y),"\n")
```

```
## l'espérance de Y est 0.9844573
```

```r
cat("l'espérance de Z est",mean(z),"\n")
```

```
## l'espérance de Z est 0.9818873
```

On remarque que l'espérance de Y et celle de Z sont proches de 1, donc le calcul est vérifié.

Exercice 2

Une population fictive de chats domestiques contient une proportion $1-p$ d'individus de race A (européenne, $X = 0$) et une proportion p d'individus de race B (maine coon, $X = 1$).

On suppose que le poids corporel d'un animal, Y, peut être modélisé par une loi normale $N(m_0, \sigma^2)$ pour la race européenne et par une loi normale $N(m_1, \sigma^2)$ pour la race main coon. Pour justifier les lois normales, on suppose que les coefficients de variation rendent les valeurs négatives improbables.

Un objectif de cet exercice est d'interpréter la variance expliquée et la variance non-expliquée en terme de variance inter-groupe et variance intra-groupe.

Question 0

Dans la formule de la variance totale, rappeler le terme correspondant à la variance de Y expliquée par X et le terme correspondant à la variance non-expliquée.

Solution.

$Var(Y) = Var(E(Y/X)) + E(Var(Y/X))$

Le premier terme correspond à la variance expliquée par X , le second terme correspond à la variance non expliquée.

Question 1

Montrer que la loi de la variable Y est une loi dont la densité est égale à

$$p_Y(y) = (1-p) \times p_0(y) + p \times p_1(y)\,, \quad y \in \mathbb{R}\,,$$

où p_0 est la densité de la loi normale $N(m_0, \sigma^2)$ et p_1 la densité de la loi normale $N(m_1, \sigma^2)$. Cette forme de loi est parfois appelée *loi de mélange* (de chats, dans ce cas).

Solution.

En utilisant la formule des probabilités totales, on a:

$P(Y \leq t) = P_{X=0}(Y \leq t) \times (1-p) + P_{X=1}(Y \leq t) \times p$

En dérivant , on obtient bien $p_Y(y) = (1-p) \times p_0(y) + p \times p_1(y)$

Question 2

Soit X une variable aléatoire de loi de Bernoulli ($X \in \{0,1\}$, $\mathbb{P}(X=1)=p$) et Y^* une variable aléatoire liée à X de la manière suivante

$$Y^* = a_0 + a_1 X + \epsilon$$

où ϵ est une variable aléatoire de loi normale $N(0,\sigma^2)$ indépendante de X.

Calculer l'espérance conditionnelle $\mathbb{E}[Y^*|X=0]$ et l'espérance conditionnelle $\mathbb{E}[Y^*|X=1]$. Sous quelles conditions, l'équation décrit-elle le même modèle que dans le début de l'énoncé.

Solution.

$E(Y^*/X=0) = a_0$

$E(Y^*/X=1) = a_0 + a_1$

Pour que l'équation décrit le même modèle que dans le début de l'énoncé, il faut que $a_0 = m_0$ et $a_0 + a_1 = m_1$ càd $a_1 = m_1 - m_0$

Question 3

On dispose d'un échantillon $(y_1,\ldots,y_n)$ de taille n de cette population de chats pour laquelle vous connaissez la race des chats $(x_1,\ldots,x_n)$. Proposer des estimateurs des paramètres a_0 et a_1 du modèle de la question précédente.

Solution.

Estimateur de a_0 :La moyenne des poids de chats de race 0.

Estimateur de a_1 : La moyenne des poids de chats de race 1 - La moyenne des poids de chats de race 0.

Question 4 - Variance intra-groupe

Dans le modèle décrit dans l'énoncé, quel paramètre décrit la variance du poids au sein d'une race de chat ? Quelle hypothèse a-t-on fait sur cette variance ?

Solution.

Le paramètre qui décrit la variance du poids au sein d'une race de chat est σ^2 . On a fait l'hypothèse que c'est la même valeur de variance de poids pour les deux races.

Question 5 - Variance inter-groupe

On considère une variable aléatoire G égale à m_0 avec la probabilité $(1-p)$ et égale à m_1 avec la probabilité p. Calculer la variance de G, que l'on appelle *variance inter-groupe*. Expliquer ce que signifie ce terme.

Solution. Pour indication, on peut écrire

$$G = m_0 + (m_1 - m_0)X\,.$$

$Var(G) = (m_1 - m_0)^2 p(1-p)$

Ce terme renseigne sur la variation du poids des chats d'une race à une autre, d'ou vient le nom "inter-groupe" .

Question 6

Calculer la variance de Y. Montrer que cette variance se décompose en la somme de la variance inter-groupe et de la variance intra-groupe. Réinterpréter le résultat en terme de variance expliquée par X.

Solution.

X suit la loi de Bernouli de paramètre p , les variables aléatoires $Y/X = 0$ et $Y/X = 1$ suivent respectivement les lois normales $N(m_0, \sigma^2)$ et $N(m_1, \sigma^2)$. Donc Y suit la même loi que Y^* en tenant compte des conditions décrites dans la question 2 .

On peut alors écrire $Y = m_0 + (m_1 - m_0)X + \epsilon$.

$Var(Y) = Var(G + \epsilon) = Var(G) + Var(\epsilon) + 2cov(G, \epsilon) = Var(G) + Var(\epsilon)$. (puisque G et ϵ sont indépendantes)

Or $V(G)$ correspond à la variance expliquée par X car G correspond à $E(Y/X)$. En effet, pour $X = 0$, $G = m_0 = E(Y/X = 0)$ et de même pour $X = 1$. Par ailleurs, $V(\epsilon) = \sigma^2$ est la variance non expliquée par X car $V(Y/X) = \sigma^2$ quelque soit X d'ou son espérance est égale à σ^2 .

Finalement , $V(Y) = p(1-p)(m_0 - m_1)^2 + \sigma^2$

Question 7

On suppose que le poids moyen des chats européens est $m_0 = 5$kg, et que le poids moyen des chats maine coon est $m_1 = 8$kg, $\sigma = 1$kg et la proportion de maine coon est $p = 0.2$. Calculer la part de la variance du poids expliquée par la race de chat. Vérifier que cela correspond au carré du coefficient de corrélation linéaire entre Y et X.

Solution. D'après la question précédente, on calcule

$$\text{Part de variance expliquée} = \frac{\mathbb{V}[G]}{\mathbb{V}[G] + \mathbb{V}[\epsilon]} = 0,59016$$

Rappel : le carré du coefficient de corrélation est égal à

$$\rho^2 = \frac{\text{Cov}^2(X, Y)}{\mathbb{V}[Y]\mathbb{V}[X]} .$$

$Cov(X, Y) = E(XY) - E(X)E(Y)$

$E(XY) = E(X(m_0 + (m_1 - m_0)X + \epsilon)) = m_0E(X) + (m_1 - m_0)E(X^2)$

$E(X)E(Y) = m_0E(X) + (m1 - m0)E^2(X)$

donc $Cov(X, Y) = E(XY) - E(X)E(Y) = (m_1 - m_0)V(X)$

ainsi

$\rho^2 = \frac{(m_1-m_0)^2V(X)}{V(Y)} = \frac{V(G)}{V(G)+V(\epsilon)}$

Le carré du coefficient de corrélation s'interprète donc la part de variation du poids dans la population expliquée par la race du chat.

Question 8

On suppose que les paramètres de la population sont égaux à $m_0 = 5$kg, $m_1 = 8$kg, $\sigma = 1$kg et $p = 0.2$. Montrer que le code suivant simule l'échantillonnage des poids de $n = 100$ individus de la population.

```r
set.seed(1789)

# taille de l'échantillon
 n = 100

# groupes/races 0 et 1
 x = rbinom(n, 1, p = .2)

# simulation des poids corporels
 y = 5 + 3 * x + rnorm(n, sd = 1)
```

Solution. Traduire le code en langage mathématique.

La taille de l'échantillon est $n = 100$, On pose les variables aléatoires indépendantes $X_1, ..., X_n$ qui suivent la loi de Bernoulli de paramètre $p = 0.2$, et les variables indépendantes $\epsilon_1, ..., \epsilon_n$ qui suivent la loi normale $N(0, 1)$.

Pour tout i de 0 à n , on pose la variable aléatoire $Y_i = 5 + 3 \times X_i + \epsilon_i$

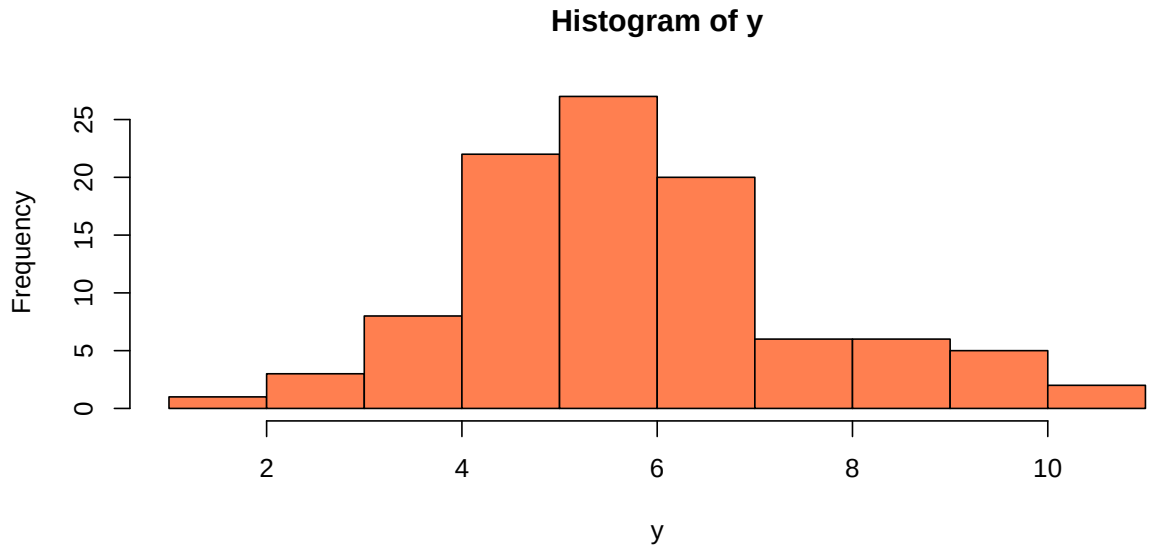
Les Y_i representent le poids du chat i en fonction de sa race, avec une modération qu'on la note ici ϵ , ainsi le code ci-dessus simule bien l'échantillonnage des poids de $n = 100$ individus de la population.

Question 9

Afficher un histogramme de l'échantillon. Afficher des boites a moustaches (`boxplot`) séparées pour les poids des deux races.

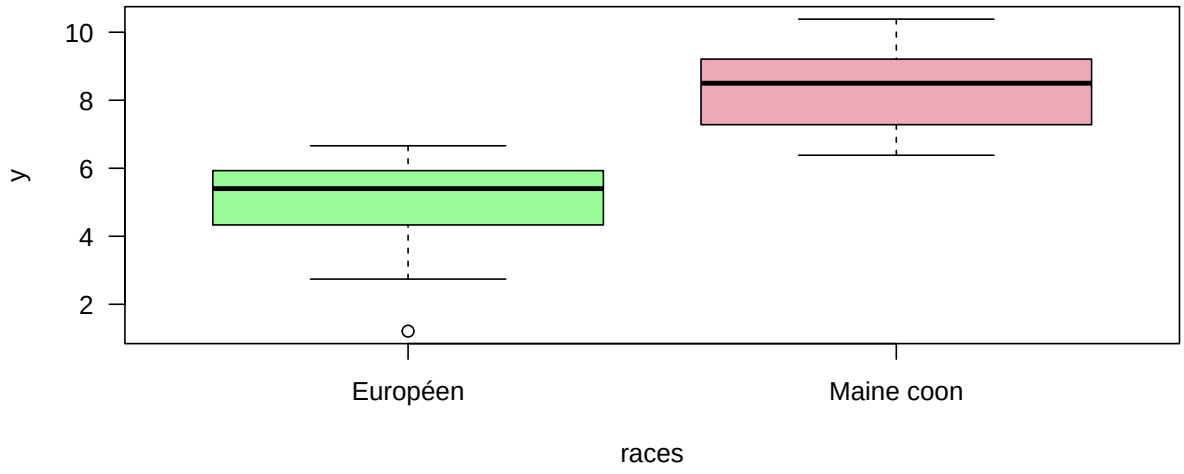
Solution. L'histogramme est

```r
hist(y , col = "coral")
```



On peut obtenir les boxplots des sous-échantillons de la manière suivante

```r
boxplot(y~x ,
        col = c("palegreen", "pink2"),
        las = 1,
        names = c("Européen", "Maine coon"),
        xlab = "races")
```



On voit bien sur les box-plots que les maine coons sont nettement plus lourds que les chats européens.

Question 10

On suppose que les paramètres intervenant dans ce problème sont inconnus. À partir des échantillons obtenus question 8 (`x` et `y`), estimer les paramètres m_0, m_1, a_0, a_1, ainsi que la part de variance du poids expliquée par la race du chat (commenter la valeur estimée à celle prédite théoriquement).

Solution. Compléter le bout de code en utilisant les questions précédentes.

```r
cat("Estimation de m_0 et a_0 : ", mean(y[x==0]),"\n")
```

```
## Estimation de m_0 et a_0 :  5.118687
```

```r
cat("Estimation de m_1 : ", mean(y[x==1]),"\n")
```

```
## Estimation de m_1 :  8.318352
```

```r
cat("Estimation de a_1 : ", mean(y[x==1])-mean(y[x==0]) ,"\n")
```

```
## Estimation de a_1 :  3.199664
```

Pour calculer la variance expliquée, insère un bout de code. N'oublie pas de commenter la valeur numérique obtenue.

```r
cat("estimation de la variance expliquée: ",mean(x)*(1-mean(x))*(mean(y[x==1])-mean(y[x==0]))**2,"\n")
```

```
## estimation de la variance expliquée:  1.756815
```

Le calcul théorique de la variance expliquée donne $V(G) = p(1-p)(m_1 - m_0)^2 = 1.44$ qui est proche de la valeur numérique obtenue.

Question 11

La différence entre les moyennes des deux races est elle significativement différente de zéro au seuil 1% ?

Solution.

On effectue le Test t de Student pour échantillons indépendants (on peut l'utiliser car les variances sont égales pour chacun des deux groupes) .

```r
t.test(y[x==1], y[x==0], var.equal = TRUE , conf.level = 0.99)
```

```
##
##  Two Sample t-test
##
## data:  y[x == 1] and y[x == 0]
## t = 12.227, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  2.512199 3.887129
## sample estimates:
## mean of x mean of y
##  8.318352  5.118687
```

la différence entre les moyennes des deux races est significativement supérieure à 0 au seuil 1% car 0 n'est pas compris dans l'intervalle de confiance obtenu pour la différence entre les moyennes de poids des deux races .

Exercice 3

Cet exercice est destiné aux plus tenaces de la semaine. On y apprend toutefois de belles choses sur les lois normales.

Soit $\rho \in (0,1)$. On considère un couple de variables aléatoires (X,Y) de densité jointe

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{q(x,y)}{2(1-\rho^2)}\right), \quad x,y \in \mathbb{R},$$

où $q(x,y)$ est la forme quadratique suivante

$$q(x,y) = x^2 - 2\rho xy + y^2 .$$

Question 1

Pour tout $x, y \in \mathbb{R}$, vérifier que

$$q(x,y) = (1-\rho^2)x^2 + (y-\rho x)^2 \,.$$

En déduire la loi marginale de X et la loi conditionnelle de Y sachant $X = x$.

Solution

On identifie les lois marginales et conditionnelles en factorisant la densité (et sa constante de normalisation). $X \sim \mathcal{N}(0,1)$ En effet, en factorisant q(x, y) et en ne gardant dans l'intégrale que le terme en $(y-\rho x)^2$, on reconnaît à un facteur près l'expression de l'intégrale de la densité d'une loi $\mathcal{N}(\rho x, 1-\rho^2)$, ce qui permet de déduire la valeur de l'intégrale car l'intégrale de la densité d'une loi de proba vaut 1. On trouve finalement $f_X(x) = \frac{1}{\sqrt{2\pi}} * \exp^{-x^2/2}$ d'où le résultat. Par ailleurs, $f_Y(y|X=x) = \frac{f(x,y)}{f_X(x)}$ d'où après calcul, $Y|X = x \sim \mathcal{N}(\rho x, 1-\rho^2)$

Question 2

Soit ϵ une variable aléatoire de loi normale $N(0, 1-\rho^2)$ indépendante de X. On définit la variable Y^* de la manière suivante

$$Y^* = \rho X + \epsilon \,.$$

Démontrer que les couples (X, Y) et (X, Y^*) ont même loi jointe.

Solution. Utiliser la question précédente et éviter les calculs (la solution ne demande qu'une espérance et une variance).

À cette occasion, on notera que les variables aléatoires sont toujours notées à l'aide d'une majuscule.

$Y^* = \rho X + \epsilon$ donc $[Y^*|X = x] = \rho x + \epsilon$.Comme ϵ est de loi normale $N(0, 1-\rho^2)$ alors $[Y^*|X = x]$ est de loi normale $N(\rho x, 1-\rho^2)$. $[Y^*|X = x]$ et $[Y|X = x]$ ont la même densité donc $(X, Y|X = x)$ et $(X, Y^*|X = x)$ ont la même loi jointe.

Question 3

En utilisant la formule de la variance totale, calculer la variance de la variable Y.

Solution. Suivre l'indication donnée dans l'énoncé.

$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}[Y|X]) + \mathbb{E}(\mathbb{V}[Y|X])$ or $\mathbb{V}(\mathbb{E}[Y|X]) = \mathbb{V}(\rho * X) = \rho^2$ d'après les calculs de la question 1 sur les espérances et variances des lois utilisées $\mathbb{E}(\mathbb{V}[Y|X]) = \mathbb{E}(1-\rho^2) = 1-\rho^2$

Commentaire final : Dans cette équation, la variance de Y expliquée par X est égale à ρ^2 et la variance non-expliquée est égale à $1-\rho^2$.

Question 4

Montrer que le coefficient de corrélation linéaire des variables X et Y est égal à

$$\mathrm{Cor}(X,Y) \equiv \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \rho \,.$$

Solution. D'après les questions précédentes, $Cor(X,Y) = Cor(X,Y^*) = \frac{cov(X,Y^*)}{\sqrt{Var[X]}\sqrt{Var[Y^*]}} = E[XY^*] - E[X]E[Y^*] = E[\rho X^2 + \epsilon X] = \rho E[X^2] + E[\epsilon X]$ or $Var[X] = E[X^2] - E[X]$ donc $E[X^2] = 1$ et $Cov(X, \epsilon) = E[\epsilon X] - E[X]E[\epsilon]$ et X et ϵ sont indépendants donc $E[\epsilon X] = 0$. Finalement $\mathrm{Cor}(X,Y) = \rho$

Question 5

On pose $\rho = .8$. Simuler un échantillon de taille $n = 100$ du couple (X, Y). Représenter le nuage de points et la courbe de l'espérance conditionnelle $x \mapsto \mathbb{E}[Y|X = x]$ dans un graphique. Commenter.

Solution.

```
rho = 0.8

# taille de l'échantillon
n = 100

# loi de x
x = rnorm(n)

# loi de y|x
y = rho*x + rnorm(n, sd = 1-rho^2)

plot(x, y, col = "red3", pch = 19)

# droite d'équation y=rho*x
abline(0, rho, col = "blue")
```