TP Maison : Découverte de R et analyse de données démographiques

Nicolas Fath, Pierre-André Horth

8 janvier 2019

Partie II: Analyse de données démographiques par pays, pour l'année 2000

1. Nous commençons tout d'abord par la définition de l'espace de travail et l'import des données.

```
> setwd('D:/EnvoiEleves')
> getwd()
[1] "D:/EnvoiEleves"
> donnees=read.csv("MSI_TPhome.csv",header=TRUE, sep=";")
```

Dans un second temps nous observons les différentes caractéristiques de l'objet *donnees* à l'aide des commandes suivantes.

```
>dim(donnees)
>summary(donnees)
>str(donnees)
>View(donnees)
```

Ensuite, nous avons décidé de rassembler l'analyse unidimensionnelle dans un objet de type $data.frame$. Nous avons calculé pour chaque variable l'étendue, le minimum, le maximum, la moyenne, la médiane, l'écart-type et la variance.

```
> names(donnees)=c("X","emission.co2","surface.foret", "pib.hab", "health.index",
"esperance.de.vie", "population", "mortalite.infantile")
> etendue=c();maxi=c();mini=c();mediane=c();variance=c();ecart_type=c();moyenne=c()
> for (i in 2:8) {etendue=c(etendue,diff(range(donnees[,i])));
    maxi=c(maxi,max(donnees[,i]));
    mini=c(mini, min(donnees[,i]));
    mediane=c(mediane, median(donnees[,i]));
    variance=c(variance,var(donnees[,i]));
    ecart_type=c(ecart_type, sd(donnees[,i]));
    moyenne=c(moyenne, mean(donnees[,i]))}
> Noms=names(donnees[2:8,])[2:8]
> analyse=data.frame(Noms,etendue, maxi, mini, mediane, variance, ecart_type,moyenne)
```

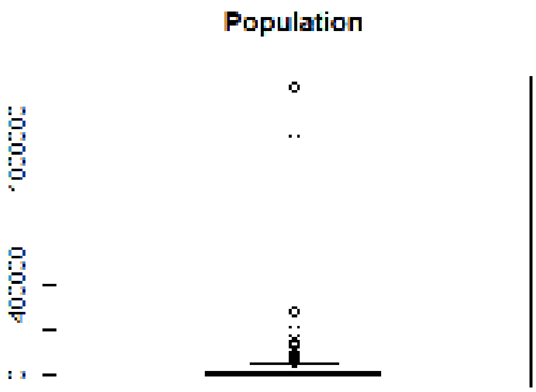
Nous obtenons ce tableau.

> analyse

	Noms	etendue	maxi	mini	mediane	variance	ecart_type	moyenne
1	emission.co2	206.400	206.500	0.100	7.800	6.952745e+02	2.636806e+01	1.735503e+01
2	surface.foret	809269.000	809269.000	0.000	2768.000	6.943829e+09	8.332964e+04	2.223246e+04
3	pib.hab	62077.000	62111.000	34.000	4666.000	1.523390e+08	1.234257e+04	9.792621e+03
4	health.index	0.653	0.965	0.312	0.792	2.704283e-02	1.644470e-01	7.342071e-01
5	esperance.de.vie	41.400	81.200	39.800	70.200	1.087191e+02	1.042684e+01	6.654734e+01
6	population	1269070.600	1269116.700	46.100	6517.800	1.698063e+10	1.303098e+05	3.508099e+04
7	mortalite.infantile	246.000	250.000	4.000	33.000	3.737399e+03	6.113590e+01	5.880473e+01

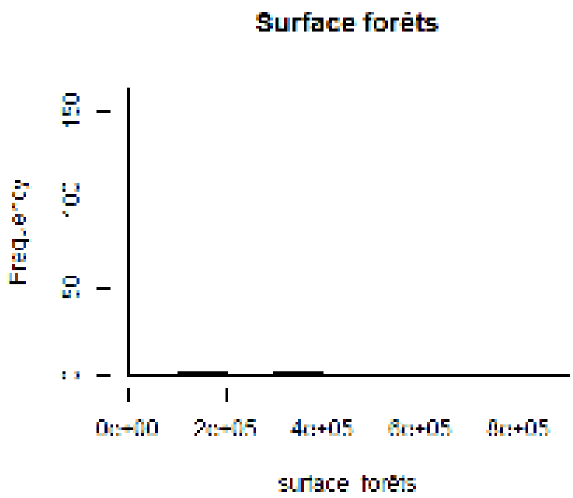
2. Nous construisons tout d'abord le diagramme en bâtons de la population à l'aide de ces lignes de commande et nous obtenons ce résultat.

```
1 > pop=donnees[,7]
2 > boxplot(pop, main="Population")
```



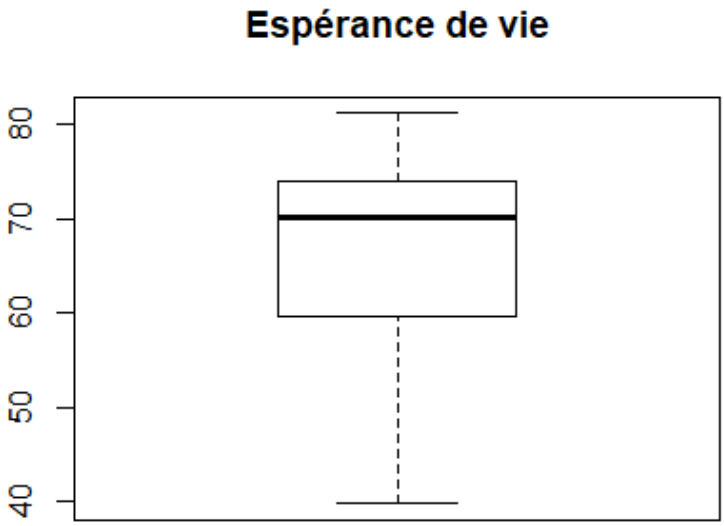
Nous construisons ensuite l'histogramme des hectares de forêt par pays à l'aide de ces lignes de commande et nous obtenons ce résultat.

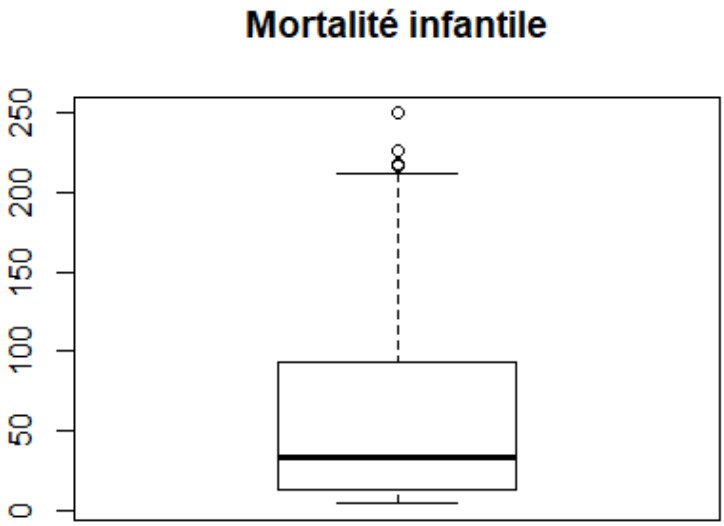
```
1 > surface_forêts=donnees[,3]
2 > hist(surface_forêts,main="Surface forêts")
```



3. Nous élaborons ensuite une boîte à moustaches de l'espérance de vie et du taux de mortalité des moins de 5 ans.

```
> esp_vie=donnees[,6]
> boxplot(esp_vie, main="Espérance de vie")
> mort_infant=donnees[,8]
> boxplot(mort_infant, main="Mortalité infantile")
```



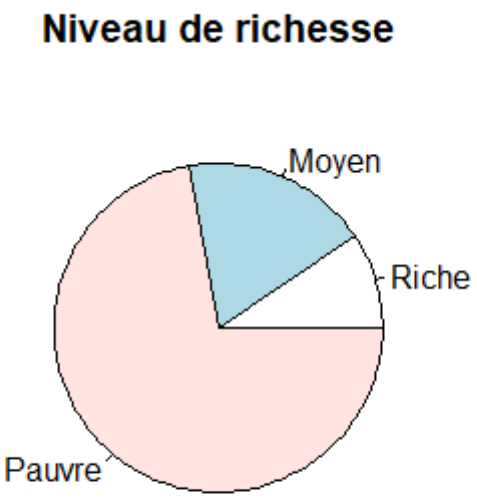


4. Nous créons la variable richesse de cette manière et nous l'ajooutons à une copie de *donnee* nommée *donnee*2.

```
> var_cam=data.frame(statut=c(rep(1,169)))
> for (i in 1:169)
    {if(donnees$pib.hab[i]>30000){var_cam[i,1]="Riche"};
    if(donnees$pib.hab[i]<10000){ var_cam[i,1]="Pauvre"};
    if (10000<=donnees$pib.hab[i] & donnees$pib.hab[i]<=30000) {var_cam[i,1]="Moyen"}}
> donnees2 = cbind(donnees, var_cam)
```

Nous réalisons ensuite le cammembert suivant.

```
> niv_rich=donnees2[,9]
> R=0; M=0; P=0
> for (i in 1:length(niv_rich))
    {if (niv_rich[i]=="Riche"){R=R+1};
    if (niv_rich[i]=="Moyen"){M=M+1};
    if (niv_rich[i]=="Pauvre"){P=P+1}}
> niv_rich2=c(R,M,P)
> niv_rich2
[1]  16  31 122
> pie (niv_rich2, labels=c("Riche","Moyen","Pauvre"),main="Niveau de richesse")
```

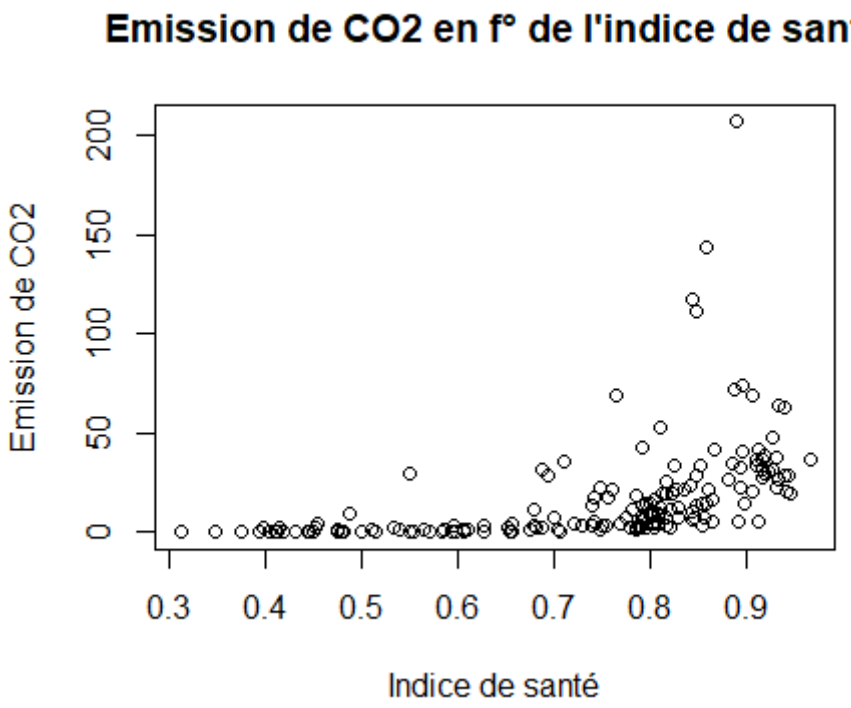


5. A l'aide de la commande *scale*, nous centrons et réduisons la variable population et nous la stockons dans la variable *Cinq*. Une variable centrée réduite possède une moyenne nulle et une écart-type de 1. Nous vérifions bien ces conditions ici (nous obtenons une valeur très proche de 0 pour la moyenne)

```
> Cinq=scale(donnees$population, center=TRUE, scale=TRUE)
> mean(Cinq)
[1] -3.87498e-18
> sd(Cinq)
[1] 1
```

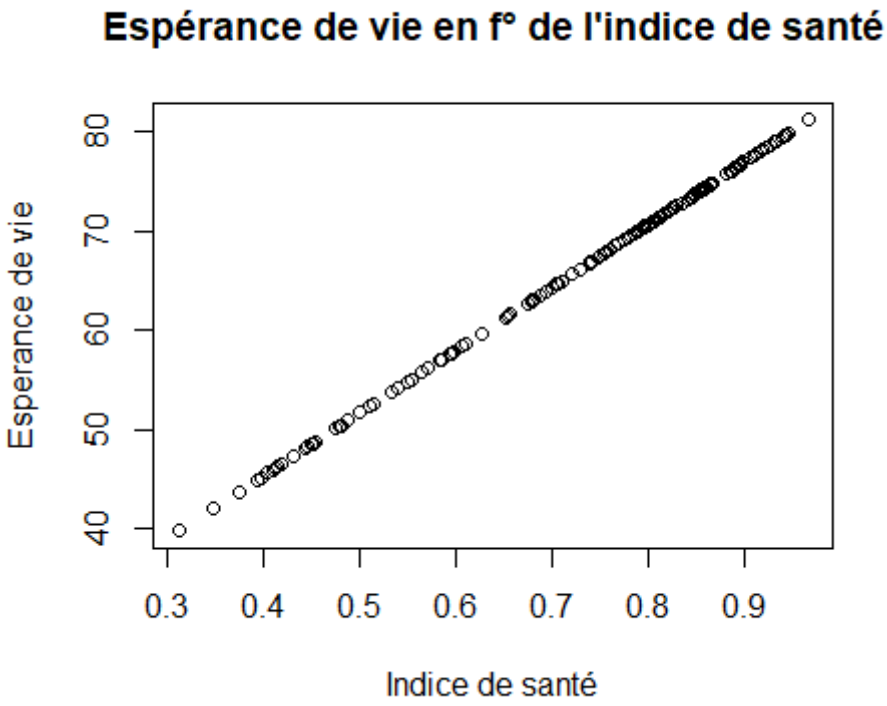
6. Nous traçons le nuage de points de la variable indice de santé et émission de CO2. Comme nous pouvons aussi le constater sur la figure, le coefficient de corrélation est mauvais.

```
> plot(donnees$health.index,donnees$emission.co2,xlab="Indice de santé",
ylab="Emission de CO2", main="Emission de CO2 en f° de l'indice de santé")
> cor(donnees$health.index,donnees$emission.co2
[1] 0.4588567
```



Nous traçons ensuite le nuage de points de la variable indice de santé et espérance de vie. En revanche, comme nous pouvons aussi le constater sur la figure, le coefficient de corrélation est bien meilleur.

```
> plot(donnees$health.index,donnees$esperance.de.vie,xlab="Indice de santé",
ylab="Esperance de vie", main="Espérance de vie en f° de l'indice de santé")
> cor(donnees$health.index,donnees$esperance.de.vie)
[1] 0.9999947
```



7. Nous effectuons une requête conditionnelle afin d'obtenir un tableau nous donnant les caractéristiques des 12 plus gros émetteurs de CO2 avec leurs caractéristiques.

```
> Sept=donnees[donnees[,2:8]>=rev(sort(donnees$emission.co2))[12],][1:12,]
```

```
> Sept
                      x emission_co2 surface_foret pib_hab health_index esperance_de_vie population mortalite_infantile
7              Australia        63.1         15492   29763        0.939           79.5    19164.4                 6
10               Bahrain       111.4             0   23294        0.847           73.7      638.2                13
21     Brunei Darussalam        71.8           397   47543        0.886           76.1      327.0                 8
27                Canada        64.2        310134   32447        0.932           79.1    30667.4                 6
85                Kuwait       117.0             5   33994        0.844           73.5     1940.8                13
93            Luxembourg        69.1            87   61059        0.906           77.5      435.5                 5
125               Qatar       206.5             0   62111        0.889           76.4      591.0                14
133          Saudi Arabia        52.5           977   19732        0.809           71.3    20045.3                23
137             Singapore        47.8             2   36793        0.926           78.7     3919.3                 4
155  Trinidad and Tobago        69.1           234    1387        0.764           68.5     1292.1                34
161 United Arab Emirates       143.1           310   44641        0.858           74.4     3033.5                11
163         United States        73.9        300195   39578        0.896           76.8   282496.3                 8
```

Grâce à cette manipulation nous pouvons aisément obtenir des informations sur ces pays comme la moyenne de leur PIB par habitant et de leurs émissions de CO2.

```
> mean(Sept$pib.hab)
[1] 36028.5
> mean(Sept$emission.co2)
[1] 90.79167
```