TP 2 Statistiques

Adib Habbou, Adel Kebli, Clark Ji

18/02/2022

Import data from a .csv file

```
summer_ozone = read.csv("summer_ozone.csv") # importation du fichier csv
str(summer_ozone) # structure
```

```
## 'data.frame':    491 obs. of  3 variables:
##  $ date2 : Factor w/ 491 levels "2015-05-01","2015-05-02",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ NEUIL : int  61 56 72 66 73 76 85 88 80 78 ...
##  $ RUR.SE: int  58 69 54 71 84 73 103 93 82 83 ...
```

```
head(summer_ozone) # 6 premières observations
```

```
##        date2 NEUIL RUR.SE
## 1 2015-05-01    61     58
## 2 2015-05-02    56     69
## 3 2015-05-03    72     54
## 4 2015-05-04    66     71
## 5 2015-05-05    73     84
## 6 2015-05-06    76     73
```

```
tail(summer_ozone) # 6 dernières observations
```

```
##          date2 NEUIL RUR.SE
## 486 2019-08-17    43     44
## 487 2019-08-18    51     47
## 488 2019-08-27   140    146
## 489 2019-08-28    89    116
## 490 2019-08-29    53     73
## 491 2019-08-30    83    104
```

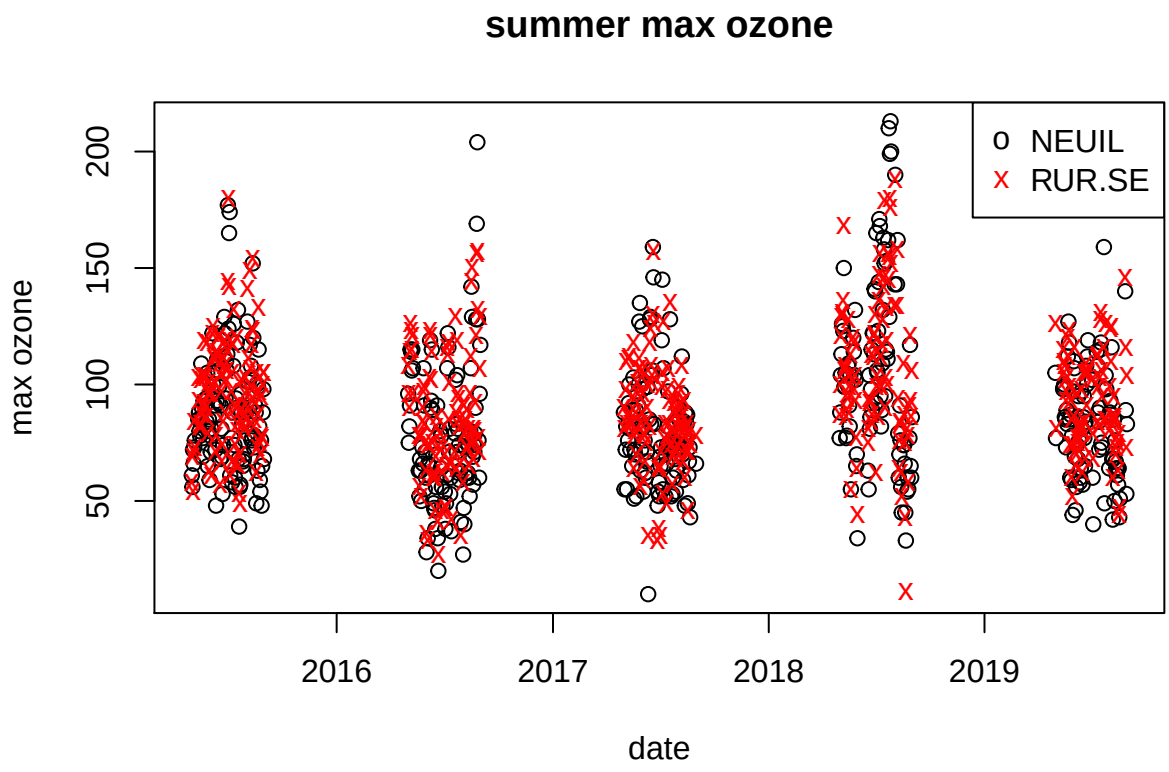
```
names(summer_ozone) # noms des variables
```

```
## [1] "date2"  "NEUIL"  "RUR.SE"
```

```
summary(summer_ozone) # min, max, mediane, quantile 1 et 3 de chaque variable
```

```
##         date2          NEUIL           RUR.SE
##  2015-05-01:  1   Min.   : 10.0   Min.   : 11.00
##  2015-05-02:  1   1st Qu.: 65.0   1st Qu.: 75.50
##  2015-05-03:  1   Median : 82.0   Median : 90.00
##  2015-05-04:  1   Mean   : 86.7   Mean   : 92.92
##  2015-05-05:  1   3rd Qu.:104.0   3rd Qu.:108.00
##  2015-05-06:  1   Max.   :213.0   Max.   :188.00
##  (Other)   :485
```

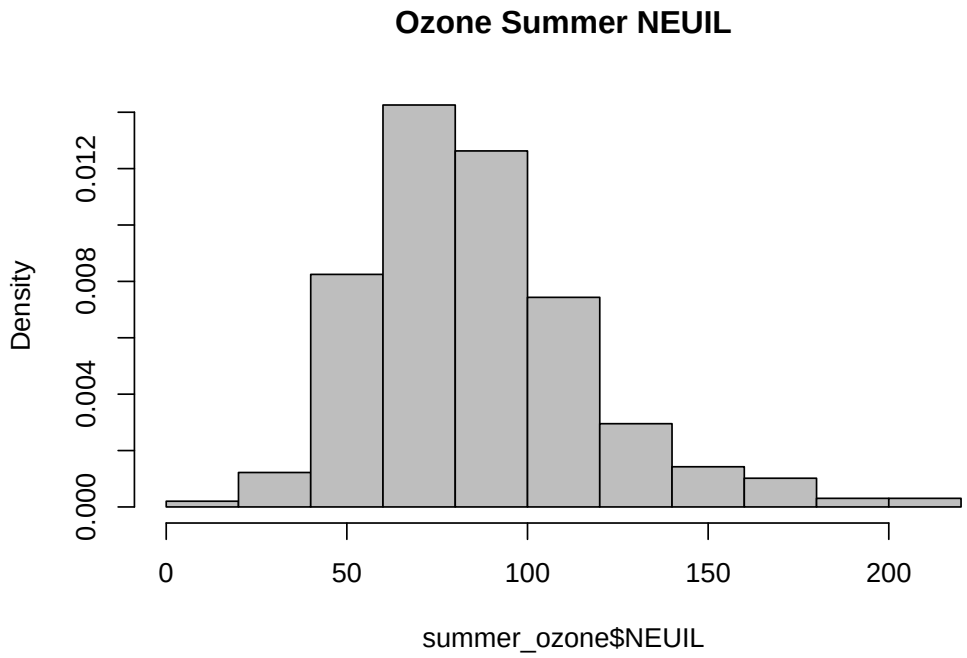
```
summer_ozone$date = as.Date(summer_ozone$date2) # convertit la chaîne de caractère en date
## plot
plot(summer_ozone$date, summer_ozone$NEUIL, xlab="date", ylab="max ozone", main="summer max ozone")
points(summer_ozone$date, summer_ozone$RUR.SE, col="red", pch = "x")
legend("topright", legend = c("NEUIL", "RUR.SE"), col=c("black","red"), pch=c("o","x"))
```



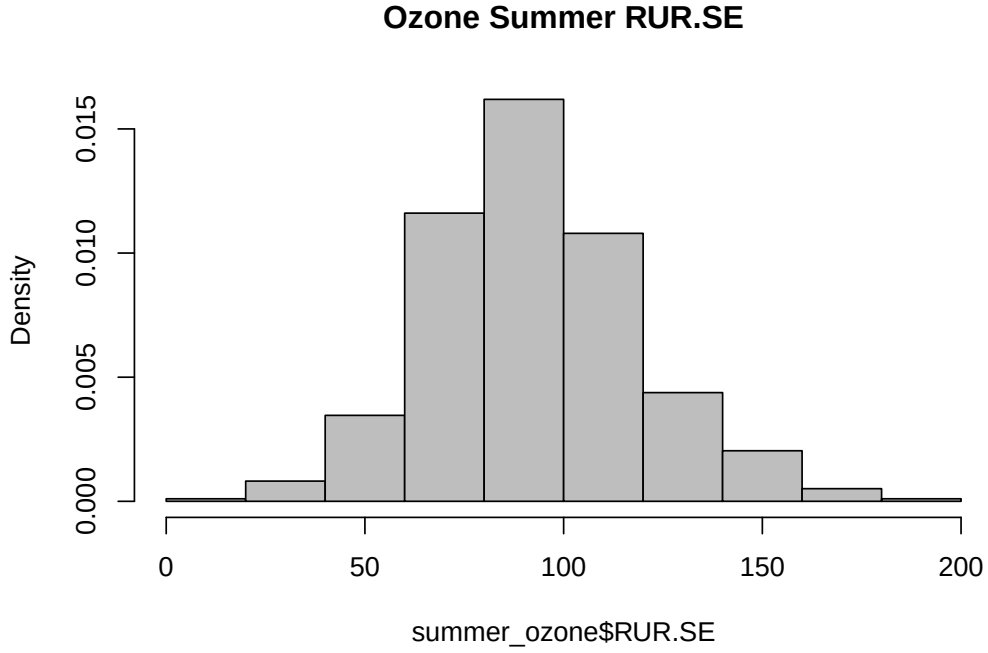
Question 1 :

Le fichier summer_ozone.csv contient 491 observations réparties sur 3 variables “date2”, “NEUIL” et “RUR.SE”. Il existe une différence entre les sites urbains et les sites ruraux. Cette différence apparaît principalement lorsque l’on regarde la valeur maximale pour chaque variable : les sites urbains affichent une valeur maximale supérieure à celle des sites ruraux. Toutefois, les sites ruraux ont tendance à avoir une moyenne plus élevée. D’après la figure tracée, il n’apparaît pas de différence notable entre chaque année, surtout lorsque l’on regarde la moyenne annuelle.

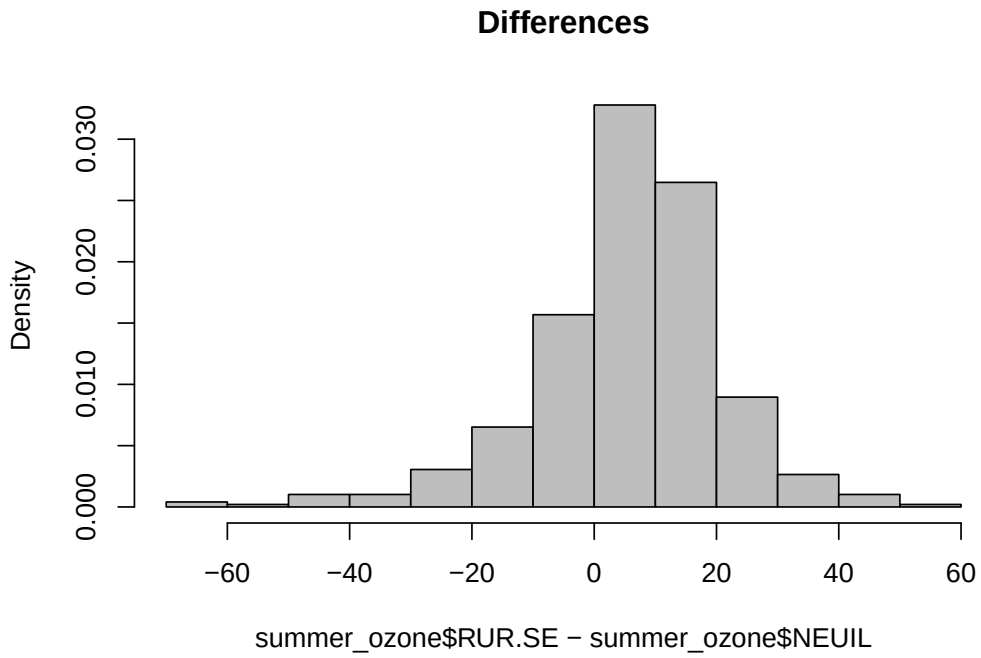
```
hist(summer_ozone$NEUIL, prob=TRUE, col="grey", main="Ozone Summer NEUIL")
```




```
hist(summer_ozone$RUR.SE, prob=TRUE, col="grey", main="Ozone Summer RUR.SE")
```



```
hist(summer_ozone$RUR.SE - summer_ozone$NEUIL, prob=TRUE, col="grey", main="Differences")
```



Question 2 :

On remarque que la plupart des différences de densité sont postives, ce qui témoigne du fait que les mesures sur les sites ruraux sont plus élevées que celles sur les sites urbains. Cette observation coïncide avec les attentes scientifiques que l'on peut avoir.

```
winter_ozone = read.csv("winter_ozone.csv") # importation du fichier csv
str(winter_ozone) # structure
```

```
## 'data.frame':    463 obs. of  3 variables:
##  $ date2 : Factor w/ 463 levels "2014-11-03","2014-11-04",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ NEUIL : int  42 51 31 36 58 40 32 33 60 57 ...
##  $ RUR.SE: int  70 62 51 62 72 67 51 66 66 64 ...
```

```
head(winter_ozone) # 6 premières observations
```

```
##        date2 NEUIL RUR.SE
## 1 2014-11-03    42     70
## 2 2014-11-04    51     62
## 3 2014-11-05    31     51
## 4 2014-11-06    36     62
## 5 2014-11-07    58     72
## 6 2014-11-08    40     67
```

```r
tail(winter_ozone) # 6 dernières observations
```

```
##          date2 NEUIL RUR.SE
## 458 2019-02-20    61     84
## 459 2019-02-21    25     82
## 460 2019-02-22    31     87
## 461 2019-02-23    51     60
## 462 2019-02-24    51     78
## 463 2019-02-25     3     50
```

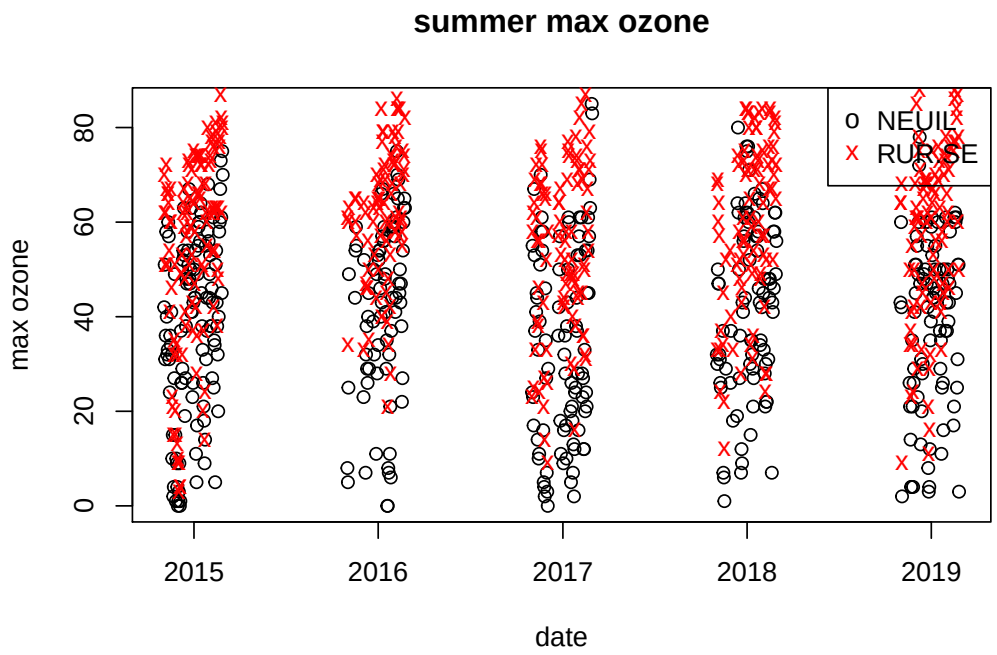
```r
names(winter_ozone) # noms des variables
```

```
## [1] "date2"  "NEUIL"  "RUR.SE"
```

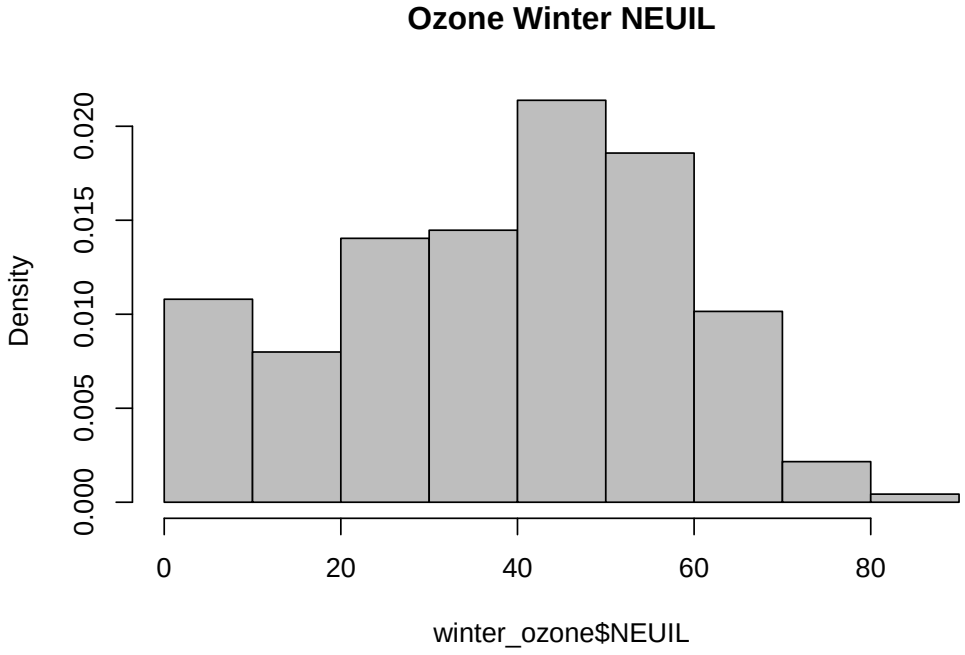
```r
summary(winter_ozone) # min, max, mediane, quantile 1 et 3 de chaque variable
```

```
##         date2         NEUIL           RUR.SE
##  2014-11-03:  1   Min.   : 0.00   Min.   :  3.00
##  2014-11-04:  1   1st Qu.:26.00   1st Qu.: 46.00
##  2014-11-05:  1   Median :42.00   Median : 61.00
##  2014-11-06:  1   Mean   :39.07   Mean   : 57.56
##  2014-11-07:  1   3rd Qu.:54.00   3rd Qu.: 72.00
##  2014-11-08:  1   Max.   :85.00   Max.   :108.00
##  (Other)   :457
```

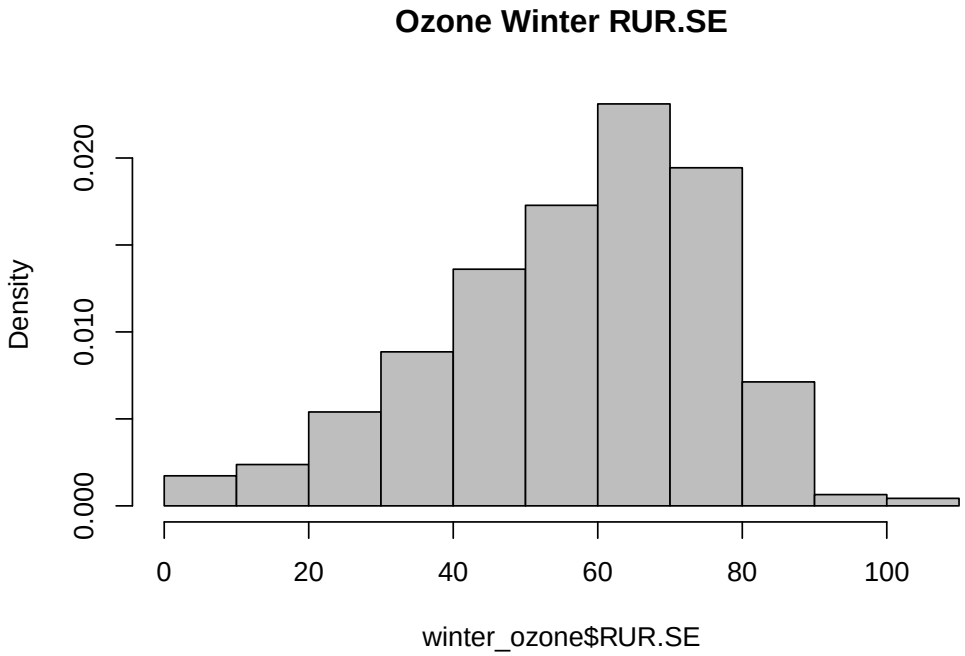
```r
winter_ozone$date = as.Date(winter_ozone$date2) # convertit la chaîne de caractère en date
## plot
plot(winter_ozone$date, winter_ozone$NEUIL, xlab="date", ylab="max ozone", main="summer max ozone")
points(winter_ozone$date, winter_ozone$RUR.SE, col="red", pch = "x")
legend("topright", legend = c("NEUIL", "RUR.SE"), col=c("black","red"), pch=c("o","x"))
```



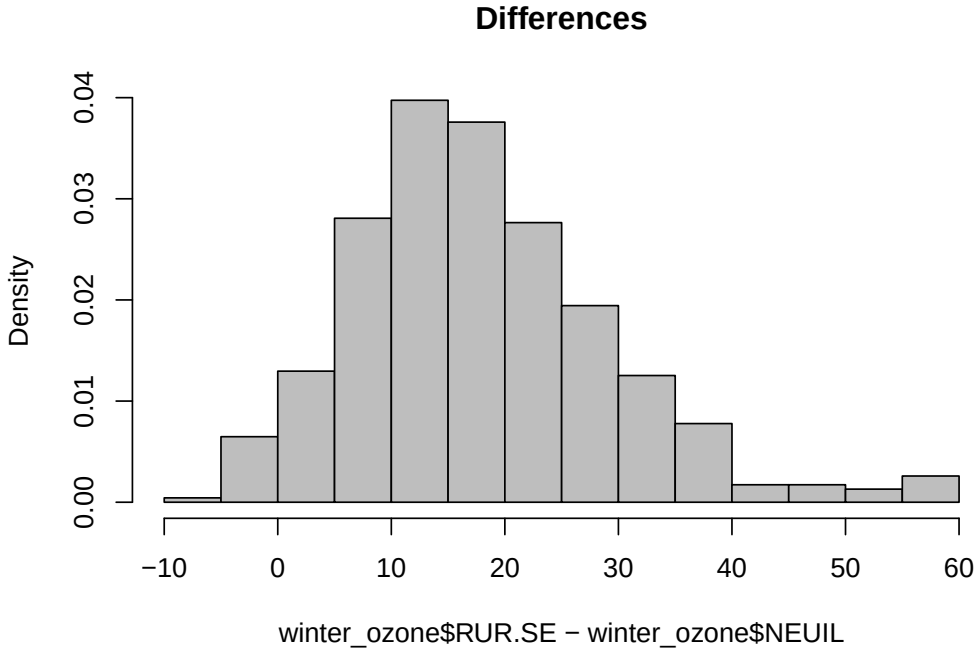

```
hist(winter_ozone$NEUIL, prob=TRUE, col="grey", main="Ozone Winter NEUIL")
```




```
hist(winter_ozone$RUR.SE, prob=TRUE, col="grey", main="Ozone Winter RUR.SE")
```




```
hist(winter_ozone$RUR.SE - winter_ozone$NEUIL, prob=TRUE, col="grey", main="Differences")
```



Question 3 :

Le fichier winter_ozone.csv contient 463 observations réparties sur 3 variables “date2”, “NEUIL” et “RUR.SE”. Il existe une différence entre les sites urbains et ruraux.

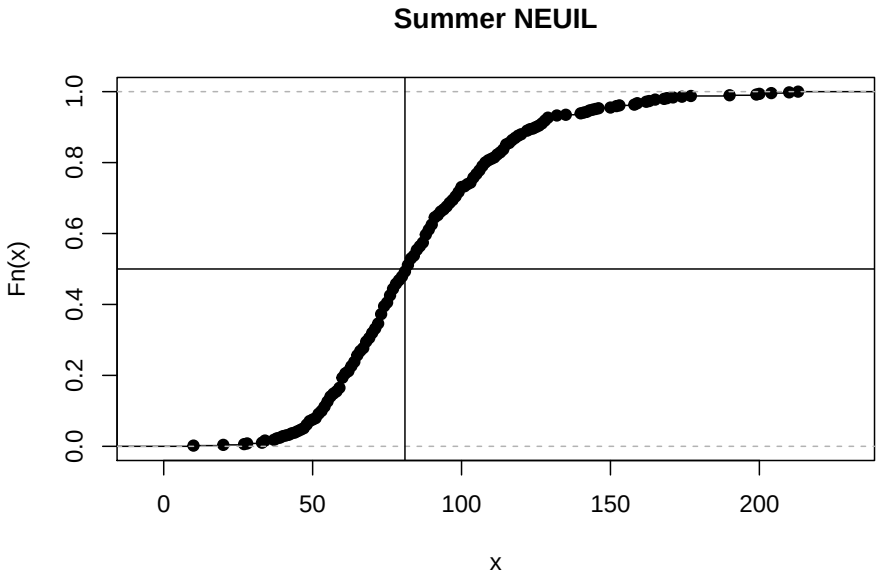
La moyenne est supérieure sur les sites ruraux, de même pour la valeur maximale. D’après la figure tracée, il n’apparaît pas de différence notable entre chaque année, surtout lorsque l’on regarde la moyenne annuelle.

On remarque que quasiment toutes les différences de densité sont postives, ce qui témoigne du fait que les mesures sur les sites ruraux sont bien plus élevées que celles sur les sites urbains.

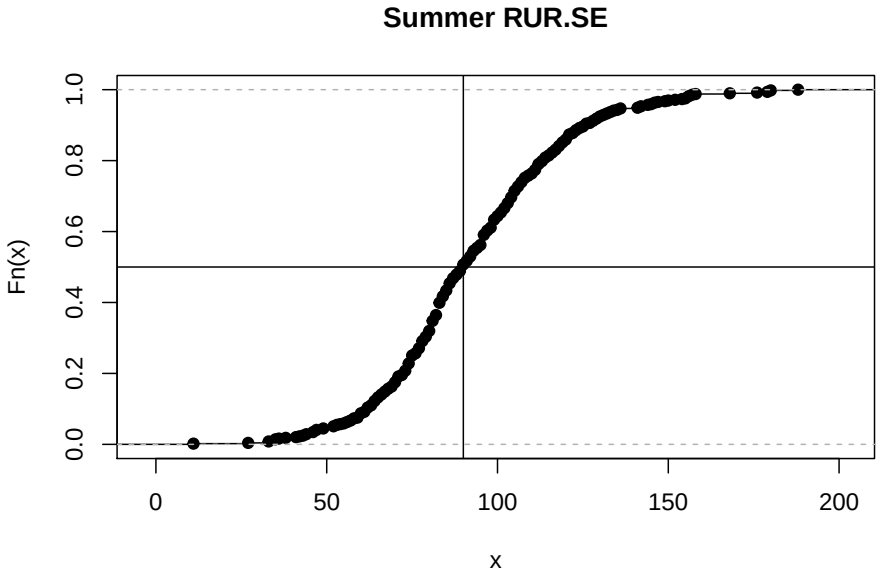
Cette observation coïncide parfaitement avec les attentes scientifiques que l’on peut avoir.

Empirical distribution function

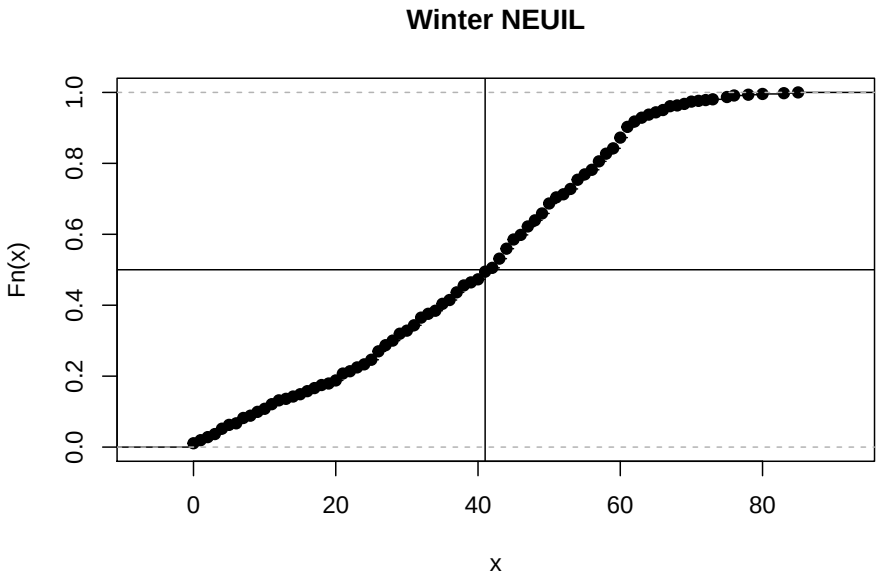
```
plot(ecdf(summer_ozone$NEUIL), main="Summer NEUIL")
abline(h = 0.5)
abline(v = 81)
```



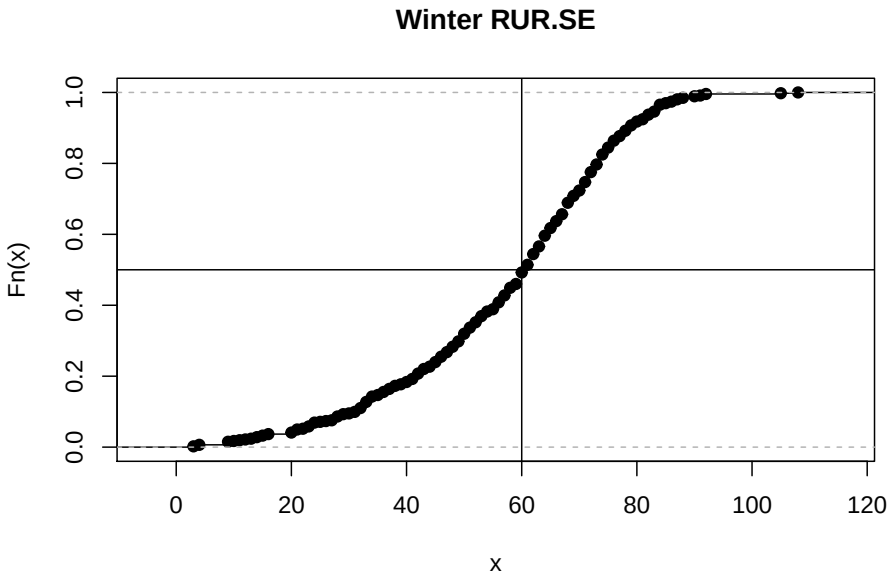

```
plot(ecdf(summer_ozone$RUR.SE), main="Summer RUR.SE")
abline(h = 0.5)
abline(v = 90)
```



```
plot(ecdf(winter_ozone$NEUIL), main="Winter NEUIL")
abline(h = 0.5)
abline(v = 41)
```




```
plot(ecdf(winter_ozone$RUR.SE), main="Winter RUR.SE")
abline(h = 0.5)
abline(v = 60)
```

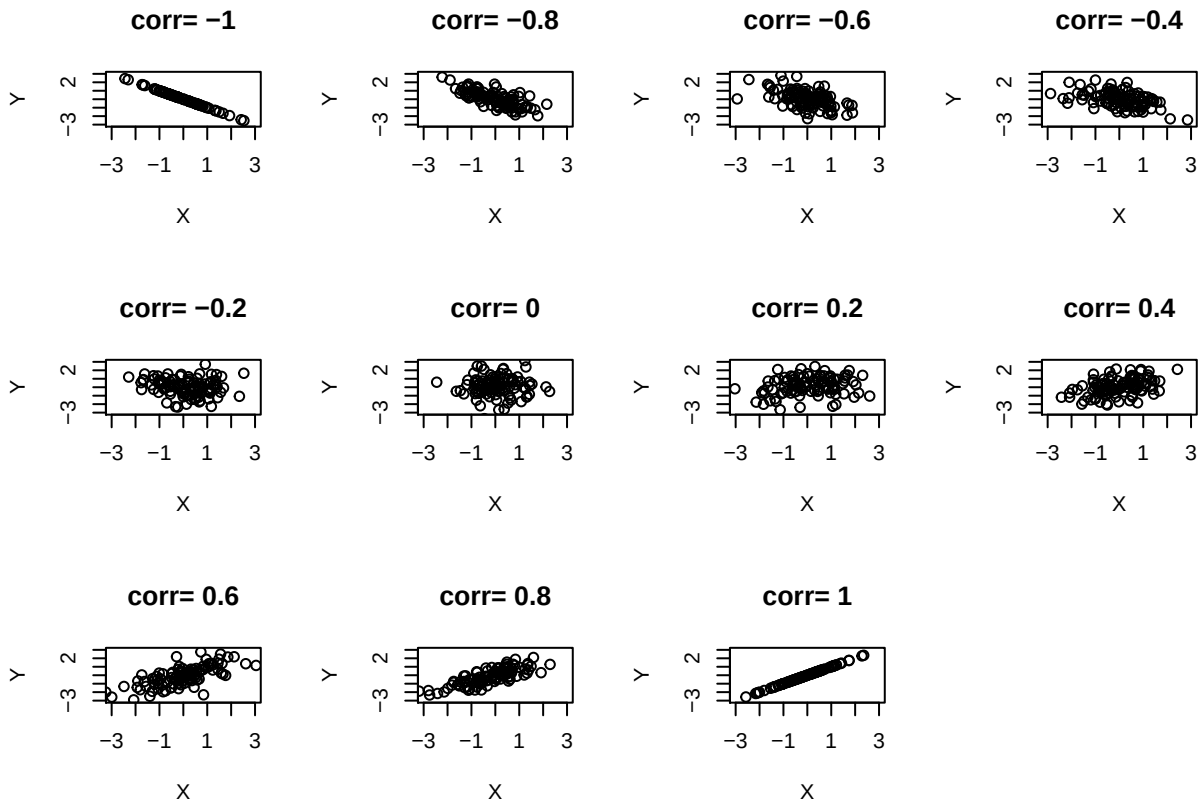


Question 4 :

On peut utiliser la fonction ecdf pour estimer la mediane en cherchant le point d'intersection entre la ligne horizontale telle que Fn(x) = 0.5 et la coubre. Ensuite on cherche la ligne verticale qui passe aussi par ce point et on lit sur l'axe des abscisses le résultat.

Sample covariance and sample correlation

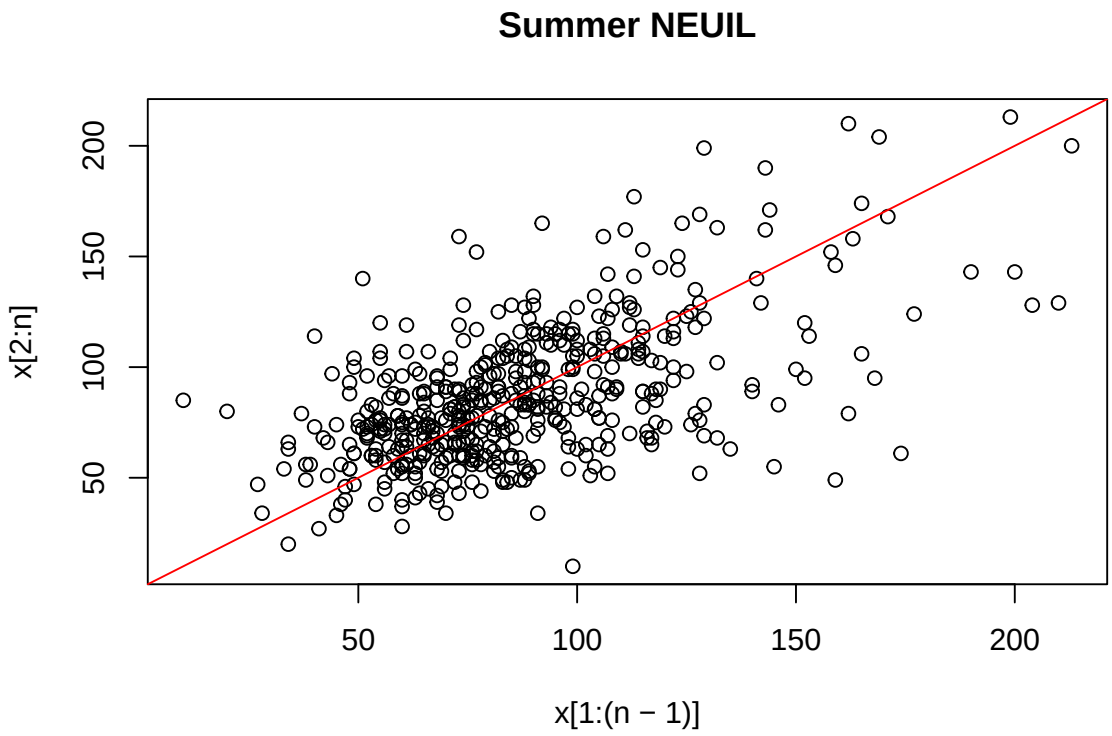
```
library("MASS")
vrho = seq(-1, 1, by = 0.2)
i = length(vrho)
xxlim = c(-3, 3)
par(mfrow = c(3, 4))
for (k in 1:i) {
  rho = vrho[k]
  xx = mvrnorm(n = 100, mu = c(0, 0), Sigma = matrix(c(1, rho, rho, 1), ncol = 2))
  plot(xx[,1], xx[,2], main = paste("corr=", signif(rho, 1)), xlim = xxlim, ylim = xxlim, xlab = "X", y
}
```



Question 5 :

Plus la corrélation est proche de la valeur 0 plus l'échantillon se disperse en un nuage de points, tandis que lorsqu'on se rapproche de 1 on observe l'apparition d'une droite de pente postive (X = Y) et lorsqu'on se rapproche de -1 on observe l'apparition d'une droite de pente négative (X = -Y).

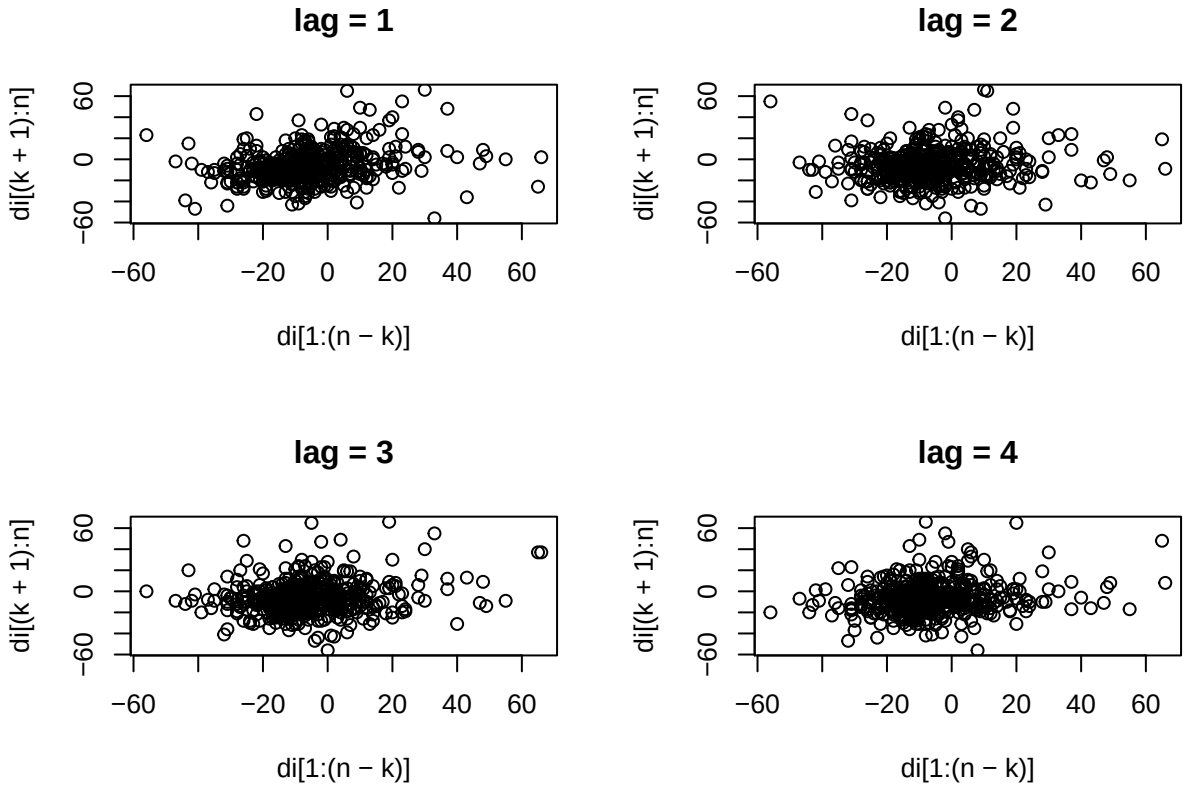
```r
x = summer_ozone$NEUIL
n = length(x)
plot(x[1:(n-1)], x[2:n], main = "Summer NEUIL")
abline(a = 0, b = 1, col = "red")
```



Question 6 :

On observe un nuage de points dispersés avec une concentration autour de la droite X = Y. En se basant sur les résultats de la question précédente, on estime le coefficient de corrélation à 0.4. Il existe donc une corrélation entre les valeurs à t-1 et celles à t.

```r
di = summer_ozone$NEUIL - summer_ozone$RUR.SE
n = length(di)
par(mfrow = c(2, 2))
for(k in 1:4) {
  plot(di[1:(n-k)], di[(k+1):n], main = paste("lag =", k))
}
```



Question 7 :

On ne remarque aucune corrélation significative entre les valeurs de la série différiencée. On en déduit qu'il n'existe pas de réelle dépendance entre les valeurs de concentration en ozone en zone rurale et urbaine en été.

Moyenne et phénomène de concentration

Question 8 :

L'inégalité de Bienaymé-Chebychef dans le cas Gaussien :

$$\mathbb{P}(|X - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}$$

L'inégalité de Bienaymé-Chebychef dans le cas Poisson :

$$\mathbb{P}(|X - \lambda| \geq \delta) \leq \frac{\lambda}{\delta^2}$$

Question 9-a :

On sait que :

$$\mathbb{P}(|X - \mu| > \delta) = \frac{1}{N} \sum_{i=1}^{N} 1_{[|X_i - \mu| > \delta]}$$

On pose la variable aléatoire :

$$\mathbb{Z} = \mathbb{I}_{[|X - \mu| > \delta]}$$

Par conséquent :

$$\mathbb{E}[\mathbb{Z}] = \mathbb{P}(|X - \mu| > \delta)$$

Question 9-b :

D'après la question précédente :

$$\mathbb{E}[\mathbb{Z}] = \mathbb{P}(|X - \mu| > \delta)$$

On trouve donc les estimations suivantes :

$$\mathbb{E}[\mathbb{Z}_{gauss}] \approx -0.034$$

$$\mathbb{E}[\mathbb{Z}_{pareto}] \approx 20.242$$

$$\mathbb{E}[\mathbb{Z}_{poisson}] \approx 10.121$$

La précision de cette estimation dépend de la valeur de N : plus N est grand plus l'estimation est bonne.

Question 9-c :

TODO

Question 9-d :

TODO

Question 10 :

```
gauss = rnorm(20, 0, 1)
poisson = rpois(20, 10)
```

Question 10-a :

TODO

Question 10-b :

TODO

Question 11 :

```
cauchy = rcauchy(20)
```

Question 11-a :

```
cauchy_20 = rcauchy(20)
moy_20 = sum(cauchy_20) / 20

cauchy_100 = rcauchy(100)
moy_100 = sum(cauchy_100) / 100

cauchy_1000 = rcauchy(1000)
moy_1000 = sum(cauchy_1000) / 1000

cauchy_10000 = rcauchy(10000)
moy_10000 = sum(cauchy_10000) / 10000
```

Question 11-b :

TODO

Question 11-c :

TODO