LMAFY1101 - Solutions - Série 6

Variables aléatoires: Quelques lois usuelles
PARTI II

Distribution d'échantillonnage et TCL

Exercice 1

Soit X_i le poids de l'individu i, avec $i = 1, \ldots, n$ et $n = 50$. Il s'agit de variables aléatoires i.i.d d'espérance $\mu = 80$ et d'écart-type $\sigma = 18$. Le poids total est donné par $T = \sum_{i=1}^{50} X_i$.

Si on suppose la normalité alors $T \sim N(50 \times \mu, 50 \times \sigma^2)$. Si on ne suppose pas la normalité, alors, par le TCL, $T \sim_a N(50 \times \mu, 50 \times \sigma^2)$.

1.

```
pnorm(4200, 50 * 80, sqrt(50) * 18, lower.tail = FALSE)
```

```
[1] 0.0581
```

2.

```
qnorm(0.9, 50 * 80, sqrt(50) * 18)
```

```
[1] 4163
```

Dans les deux cas, on effectue les calculs de la même façon, mais c'est l'interprétation qui change. En cas de normalité ces valeurs sont exactes, mais dans le cas contraire, il s'agit de simples approximations dont on ne peut jamais garantir l'exactitude et dont l'erreur peut être très grande.

Exercice 2

Commençons par formaliser le problème. Soit X_i le temps pour préparer la i-ème commande et $\overline{X}$ la moyenne de l'échantillon. Nous cherchons $P(\overline{X} \geq 71)$.

Pour calculer cette probabilité, nous avons besoin de la distribution de $\overline{X}$. Par le TCL, nous savons que si la taille d'échantillon n est suffisamment grande, $T = \sum_i X_i$ est approximativement distribué selon une $N(n\mu, n\sigma^2)$, avec $\mu = 70$ et $\sigma = 15$.

$$\Rightarrow \overline{X} = T/n \sim_a N(\mu, \sigma^2/n) = N(70, 15^2/100)$$

Dès lors, une valeur approximative de $P(\overline{X} \geq 71)$ est donnée par

```
pnorm(71, 70, 15/10, lower.tail = FALSE)
```

```
[1] 0.252
```

Exercice 3

Soit X = nombre de personnes présentes le jour du vol. On a $X \sim Bin(n = 160, p = 0.95)$. On cherche $P(X \leq 155)$.

- Calcul direct (exact)

```
pbinom(155, 160, 0.95)
```

```
[1] 0.906
```

- Par TCL

```
pnorm(155 + 0.5, 160 * 0.95, sqrt(160 * 0.95 * 0.05))
```

```
[1] 0.898
```

Excercice 4

On sait que $X_1 + X_2 + X_3 \sim N(3 \times \mu, 3 \times 12)$ et donc $\overline{X} \sim N((3 \times \mu)/3, (3 \times 12)/3^2) = N(\mu, 4)$.

$$\begin{aligned} P(|\overline{X} - \mu| < 3) &= P(|Z| < 3/\sqrt{4}) \\ &= P(-3/2 < Z < 3/2) = P(Z < 3/2) - P(Z \leq -3/2) \\ &= P(Z < 3/2) - P(Z \geq 3/2) = P(Z < 3/2) - (1 - P(Z < 3/2)) \\ &= 2P(Z < 3/2) - 1 \end{aligned}$$

```
2 * pnorm(3/2) - 1
```

```
[1] 0.866
```

Excercice 5

Commençons par formaliser le problème. Soit n la taille d'échantillon à trouver, X_i la taille du i-ème manchot et $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne de l'échantillon. On souhaite trouver n tel que

$$P(|\overline{X}_n - \mu| \leq 1) = 0.95$$

Grâce au TCL, on peut *approcher* la distribution de $\overline{X}_n$ par une $N(\mu, \sigma^2/n)$. En standardisant et en utilisant les propriétés de la normale, la condition mentionnée ci-dessus peut donc être réécrite comme

$$P(|Z| \leq z_n) = 0.95 \Leftrightarrow 2P(Z < z_n) - 1 = 0.95 \Leftrightarrow P(Z < z_n) = 0.975,$$

où $Z \sim N(0,1)$ et $z_n = \frac{\sqrt{n}}{\sigma}$.

Mais l'écriture $P(Z \leq z_n) = 0.975$ signifie aussi que z_n est le quantile d'ordre 0.975 d'un $N(0,1)$ qui vaut `qnorm(0.975)` $= 1.96 \rightarrow \frac{\sqrt{n}}{\sigma} = 1.96$.

$$\Rightarrow n \approx (1.96\sigma)^2 \approx 96.$$

Excercice 6

Commençons par formaliser le problème. Soit X_n le nombre d'individus soutenant la loi climat dans l'échantillon de taille n. Nous avons donc $X_n \sim Bin(n,p)$ où p est la probabilité qu'un individu pris au hasard dans la population soutienne la loi climat.

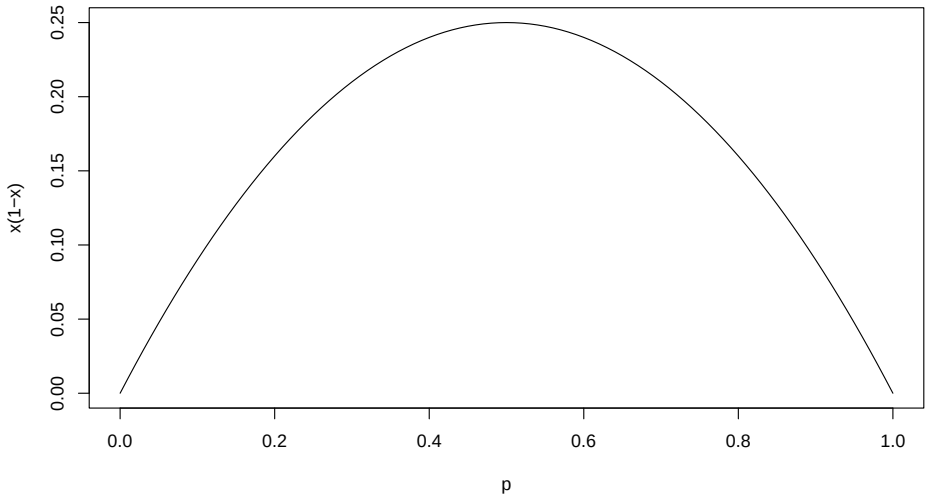
La proportion d'individus soutenant la loi climat dans l'échantillon est, elle, donnée par X_n/n. Par le TCL, on peut approximer la distribution de X_n par une $N(np, np(1-p))$ et donc approximer la distribution de X_n/n par une $N(p, p(1-p)/n)$. On cherche n tel que

$$P(|X_n/n - p| \leq 0.01) = 0.95 \Leftrightarrow P(Z \leq z_n) = 0.975,$$

où $Z \sim N(0,1)$ et $z_n = \frac{0.01}{\sqrt{p(1-p)/n}} = 1.96$.

$$\Rightarrow n \approx \frac{1.96^2}{0.01^2} \times p(1-p)$$

```
curve(x * (1 - x), 0, 1, xlab = "p", ylab = "x(1-x)")
```

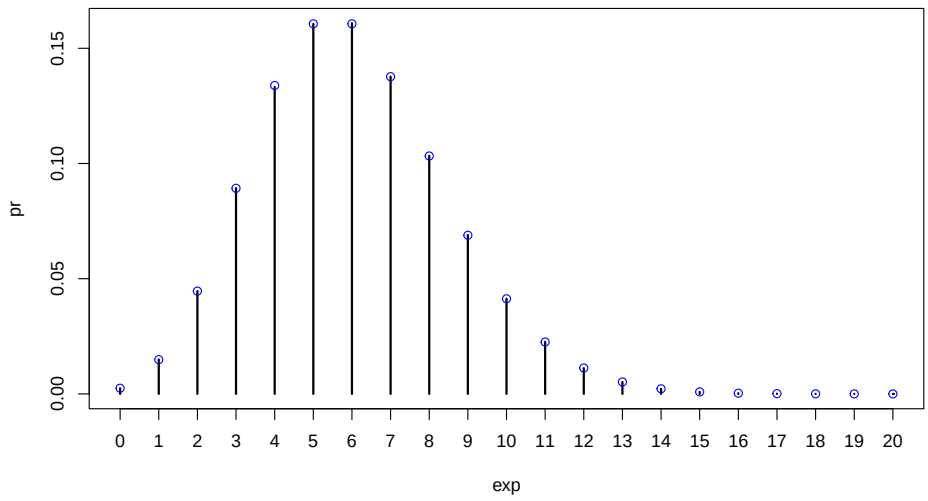


Il est facile de voir que cette fonction est maximale en $p = 0.5$. En utilisant cette valeur, nous obtenons $n \approx 9604$.

Excercice 7

1.

```
exp<-replicate(10^6, sum(rpois(3,2)))
pr<-table(exp) |> proportions()
plot(pr)
points(0:25, dpois(0:25,3*2), col="blue")
```



2.

- (a)-(b)

```
exp3<-replicate(10^6, sum(rpois(3,2)))
```

- (c)

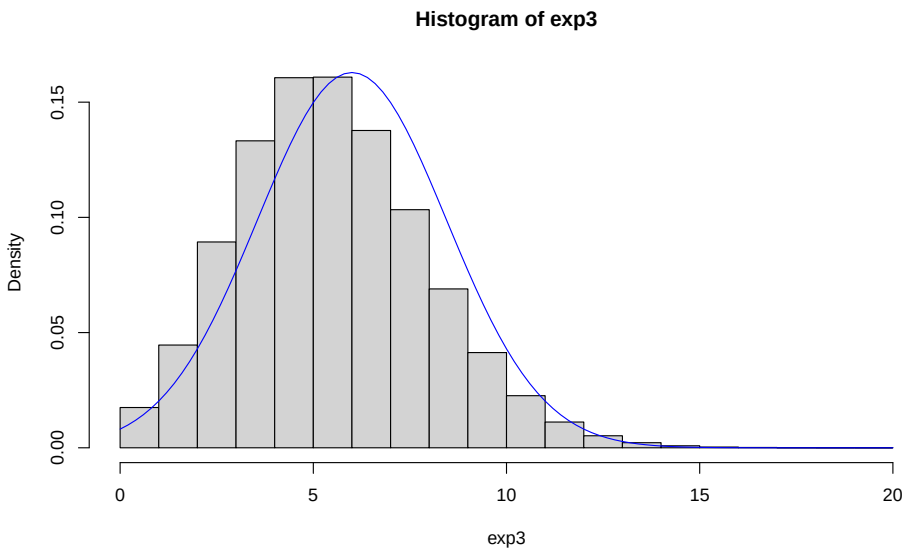
```
c(mean(exp3),var(exp3))
```

```
[1] 6 6
```

On peut constater que cette moyenne et cette variance sont très proches des valeurs théoriques (6 et 6, respectivement).

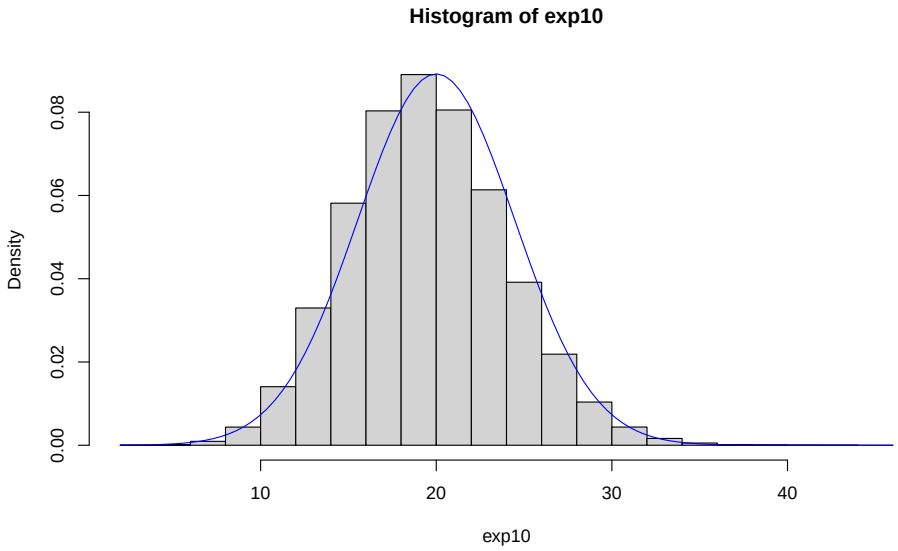
- (d)

```
hist(exp3, freq = FALSE)
curve(dnorm(x, mean = 2 * 3, sd = sqrt(2 * 3)), add = T, col = "blue")
```

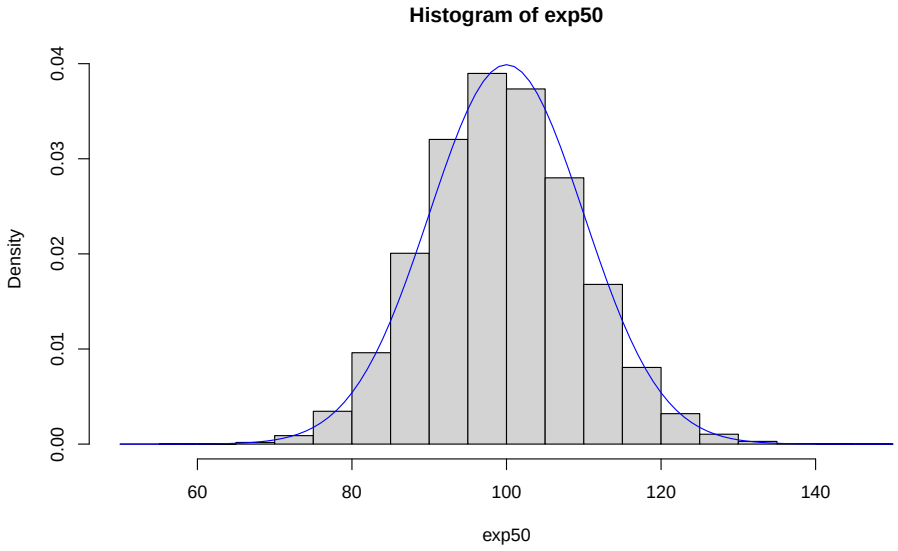


3.

```
exp10 <- replicate(10^6, sum(rpois(10, 2)))
hist(exp10, freq = FALSE)
curve(dnorm(x, mean = 10 * 2, sd = sqrt(10 * 2)), add = T, col = "blue")
```




```
exp50 <- replicate(10^6, sum(rpois(50, 2)))
hist(exp50, freq = FALSE)
curve(dnorm(x, mean = 50 * 2, sd = sqrt(50 * 2)), add = T, col = "blue")
```

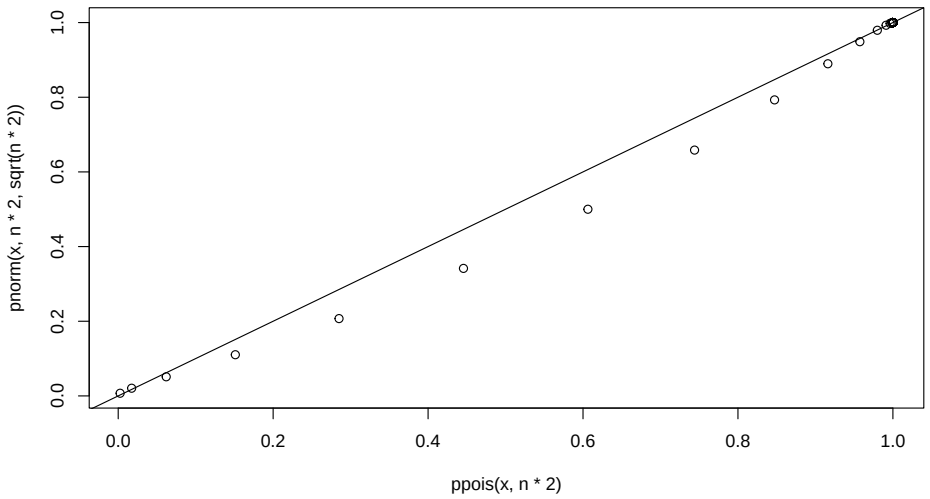


On constate que plus n est grand, plus la densité de T_n se rapproche de la normale. C'est le TCL!

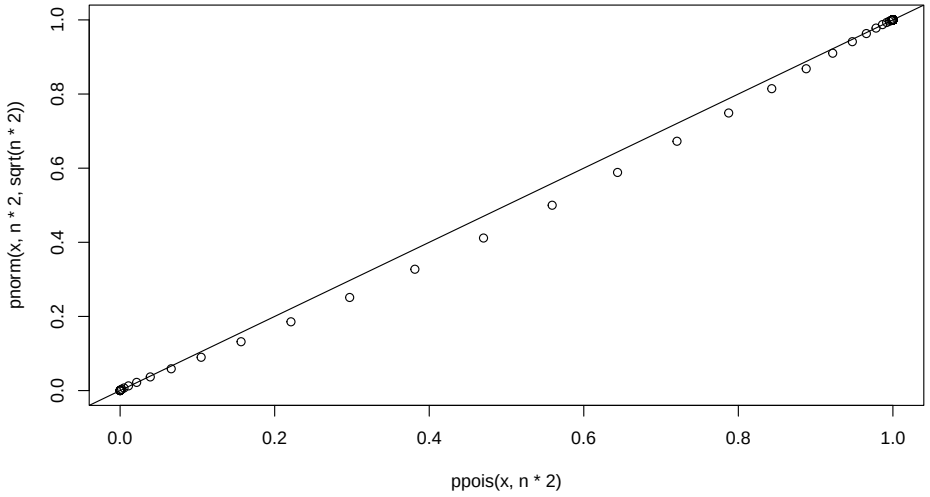
***Remarque**:*

On peut aussi confirmer que $Pois(2n) \approx P(N(2n, 2n)$ en comparant $x \mapsto P(Pois(2n) \leq x)$ à $x \mapsto P(N(2n, 2n) \leq x)$. Si nous traçons $P(N(2n, 2n) \leq x)$ vs $P(Pois(2n) \leq x)$, nous devons observer la droite $y = x$ quand n est grande.

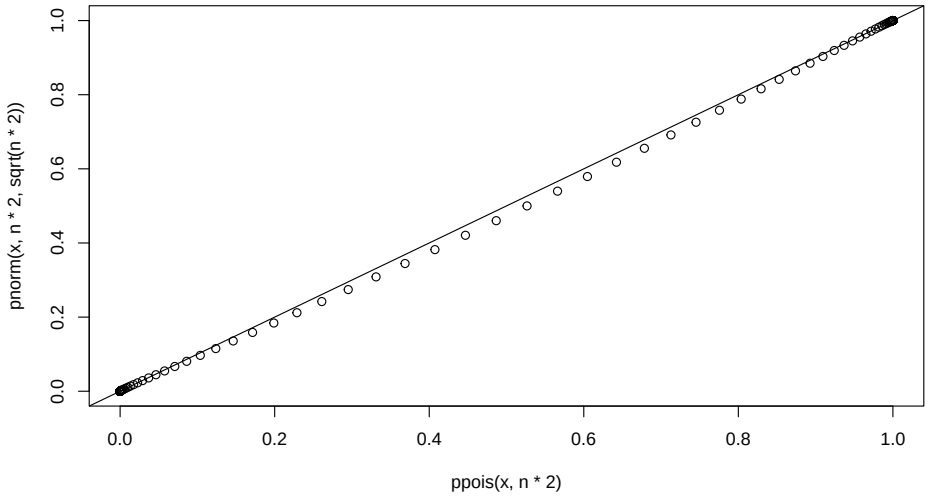
```
x <- 0:200

n <- 3
plot(ppois(x, n * 2), pnorm(x, n * 2, sqrt(n * 2)))
abline(a = 0, b = 1)  # Line y = 0 + 1*x
```




```
n <- 10
plot(ppois(x, n * 2), pnorm(x, n * 2, sqrt(n * 2)))
abline(a = 0, b = 1)
```




```
n <- 50
plot(ppois(x, n * 2), pnorm(x, n * 2, sqrt(n * 2)))
abline(a = 0, b = 1)
```



Conformité à la loi normale

Excercice 8

1.

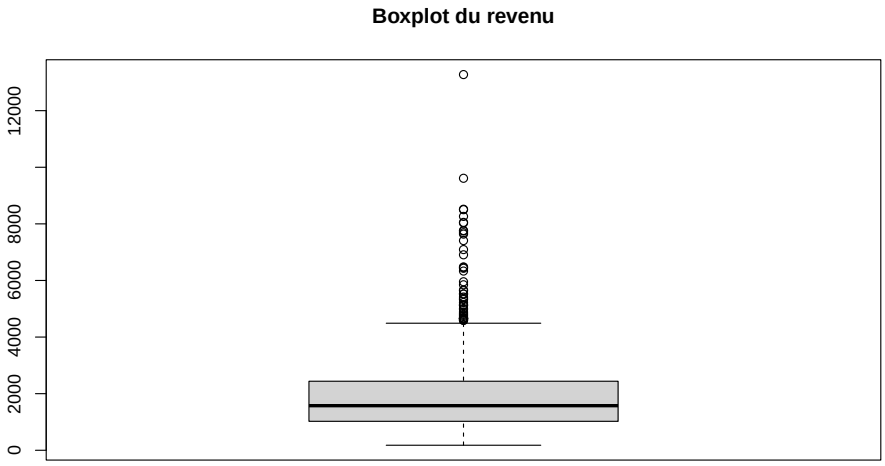
```
load(file = "Revenus.Rdata")
summary(Revenus)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    178    1024    1576    1928    2438   13275
```

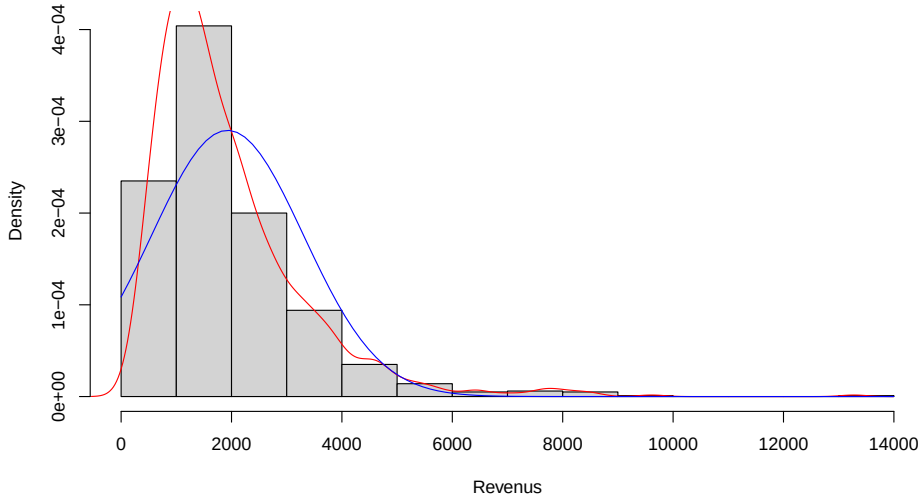
Le salaire médian est assez différent du salaire moyen, ce qui nous donne une première indication d'asymétrie. Étant donné que médiane $<$ moyenne, il s'agirait d'une asymétrie à droite càd. une surreprésentation (par rapport à une distribution symétrique) d'individus riches, ce qui se traduit par une moyenne supérieure à la médiane.

2.

```
boxplot(Revenus, main = "Boxplot du revenu")
```




```
hist(Revenus, freq = FALSE, main = "")
lines(density(Revenus), col = "red")
curve(dnorm(x, mean = mean(Revenus), sd = sd(Revenus)), add = T,
  col = "blue")
```



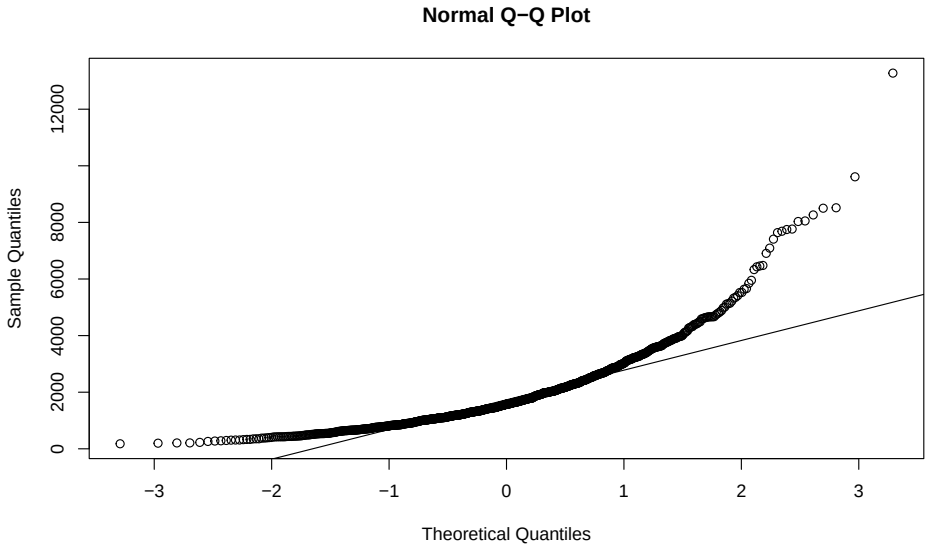
Ces deux graphiques confirment ce qui a déjà été relevé : présence non-négligeable d'individus très riches et une asymétrie à droite. Nous pouvons noter que

- Hétérogénéité assez forte chez les individus "riches" (i.e. ayant un revenu élevé, de plus de 2500 par exemple)
- Hétérogénéité assez faible chez les individus "pauvres" (i.e. ayant un revenu faible, de moins de 1200 par exemple)
- Présence d'individus très riches

3.

Vu l'asymétrie présente dans les données, nous pourrions d'ores et déjà conclure que la distribution normale n'est pas adéquate pour modéliser les revenus. Le code suivant permet de générer le qqplot des revenus.

```
qqnorm(Revenus)
qqline(Revenus)
```



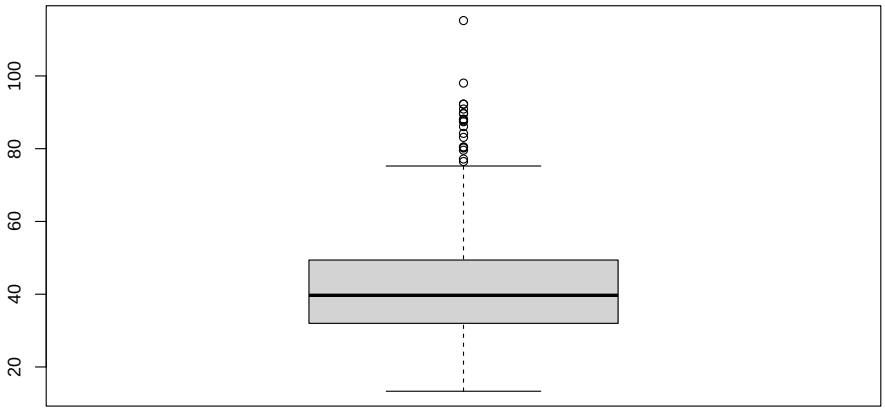
Ce graphique confirme que la distribution normale n'est pas adéquate pour modéliser ces revenus.

4.

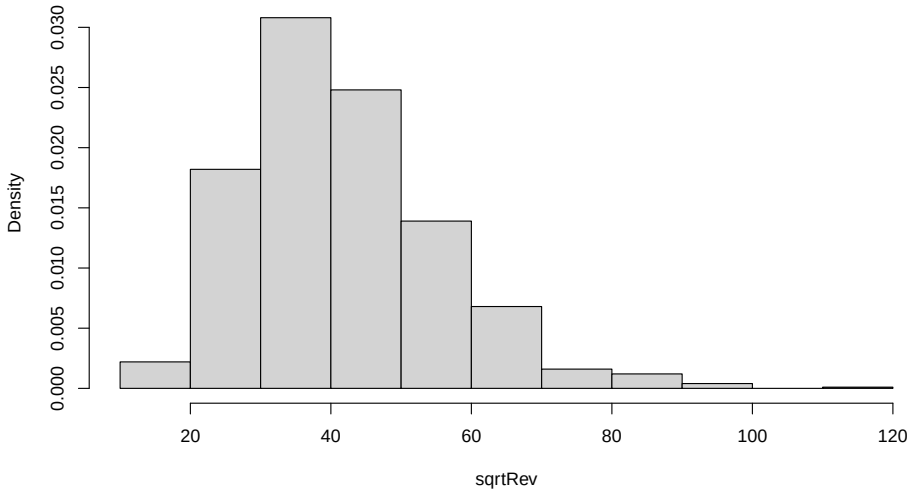
```
sqrtRev <- sqrt(Revenus)
summary(sqrtRev)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13.3    32.0    39.7    41.7    49.4   115.2
```

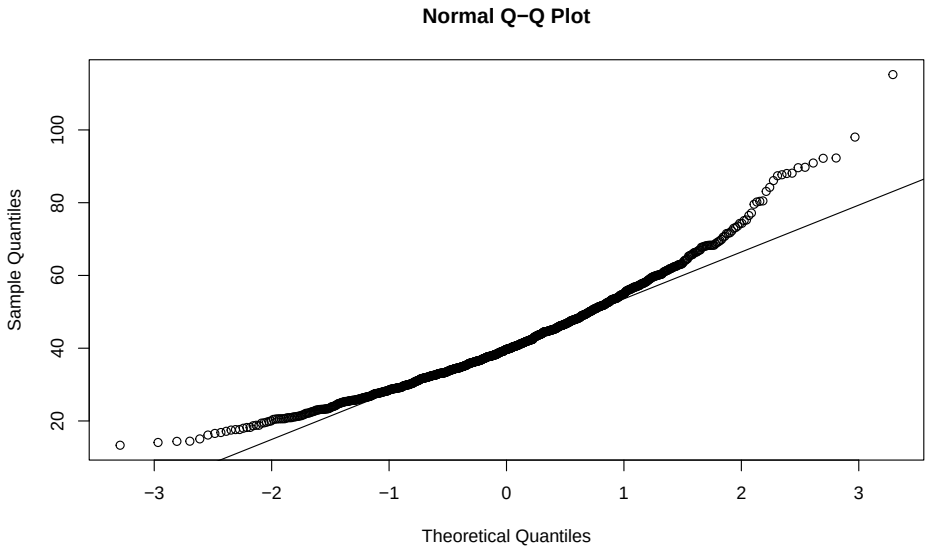
```
boxplot(sqrtRev)
```




```
hist(sqrtRev, freq = FALSE, main = "")
```




```
qqnorm(sqrtRev)
qqline(sqrtRev)
```

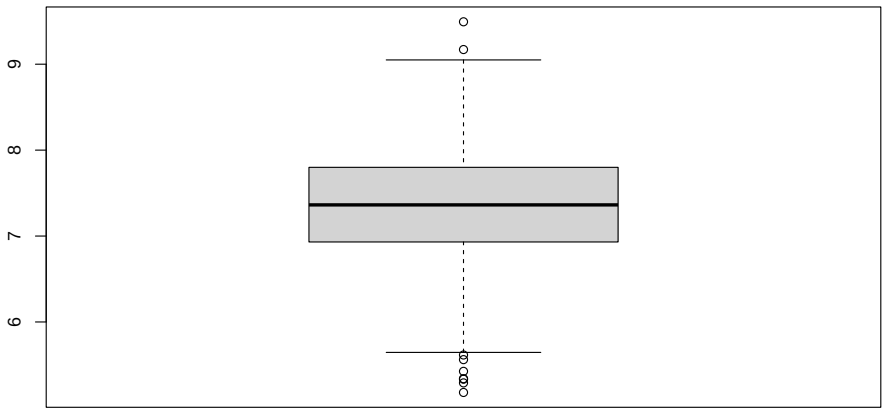


La distribution de $W = \sqrt{\text{Revenus}}$ n'est pas "parfaitement" symétrique. Cependant, l'asymétrie a clairement diminué. En effet, nous remarquons que la moyenne (41.65) et la médiane (39.7) sont très proches. Une inspection de l'histogramme et du boxplot confirme cette conclusion. Le qqplot permet de constater que la distribution de W n'est pas adéquatement modélisée par une normale, en particulier aux extrémités.

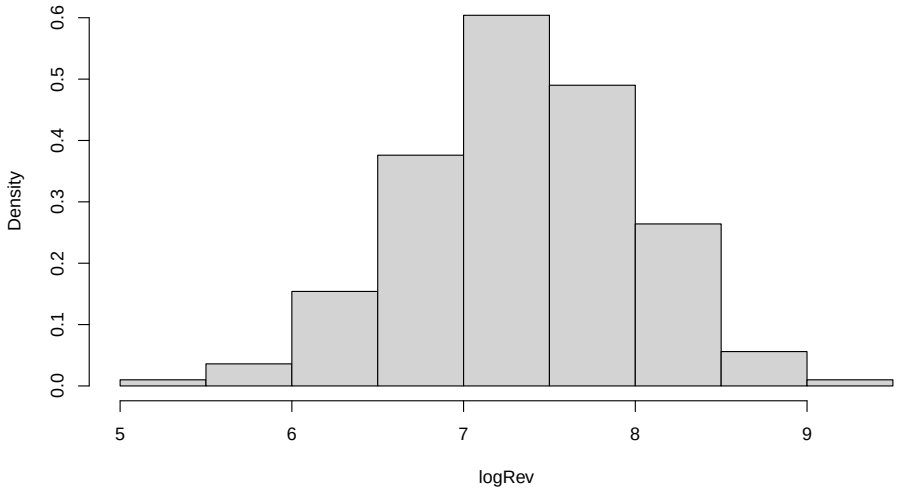
```
logRev <- log(Revenus)
summary(logRev)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.18    6.93    7.36    7.35    7.80    9.49
```

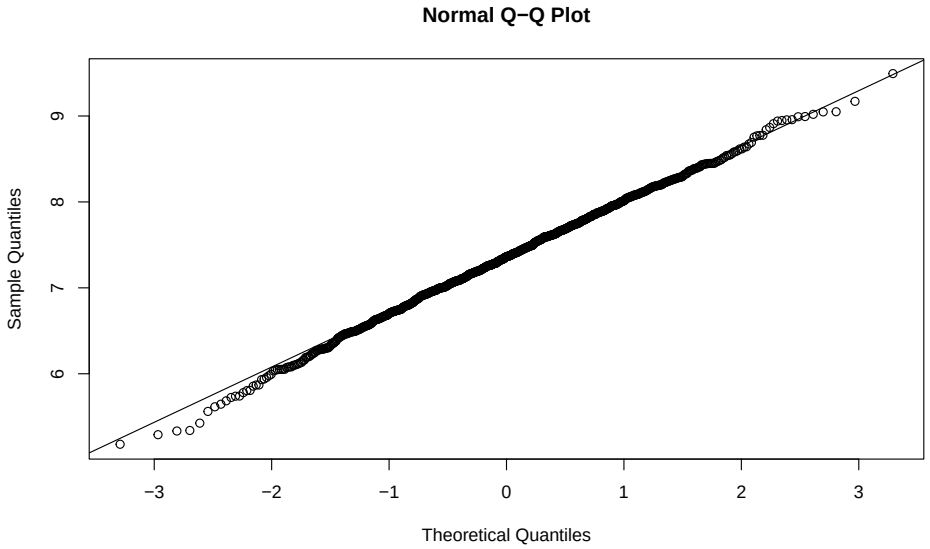
```
boxplot(logRev)
```



```r
hist(logRev, freq = FALSE, main = "")
```




```r
qqnorm(logRev)
qqline(logRev)
```



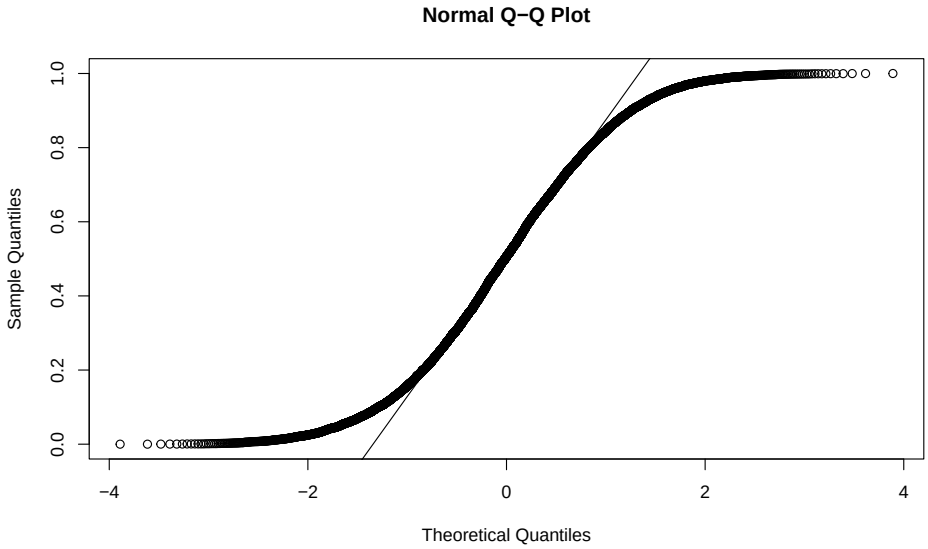
→ la distribution de log(Revenus) peut être adéquatement modélisée par une normale.

Excercice 9

1.

```r
xbars1 <- replicate(10^4, mean(runif(1)))  # ou tapez simplement runif(10^4)
```

```
qqnorm(xbars1)
qqline(xbars1)
```

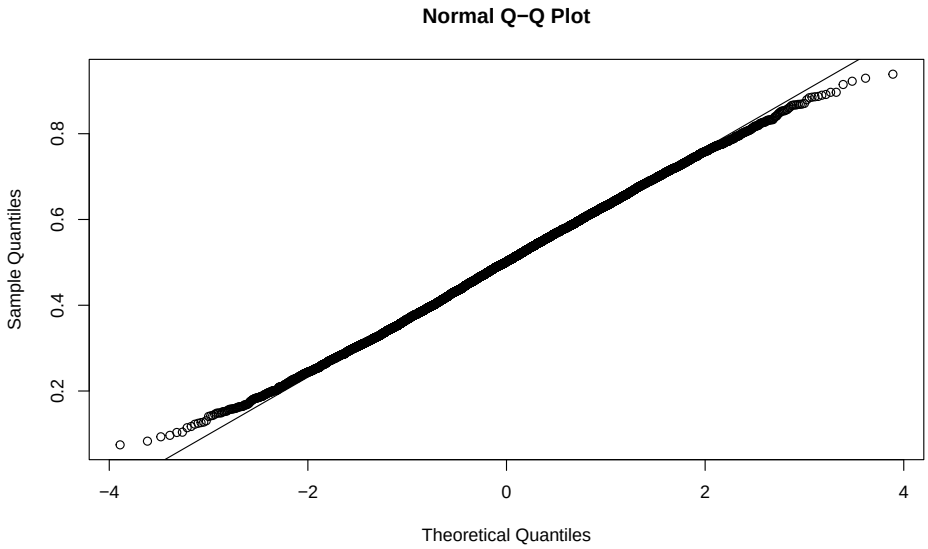


2.

- $n = 5$

```
xbars5 <- replicate(10^4, mean(runif(5)))
```

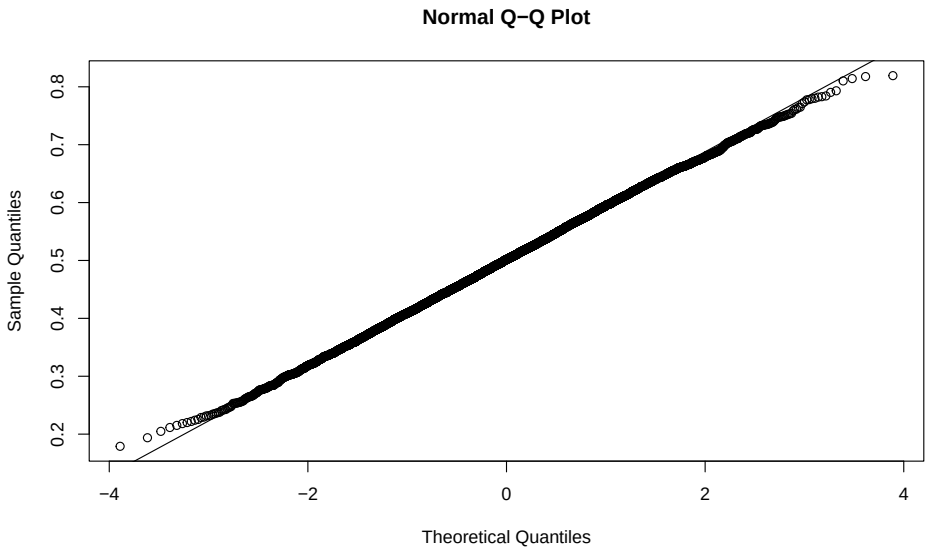
```
qqnorm(xbars5)
qqline(xbars5)
```



- $n = 10$

```
xbars10 <- replicate(10^4, mean(runif(10)))
```

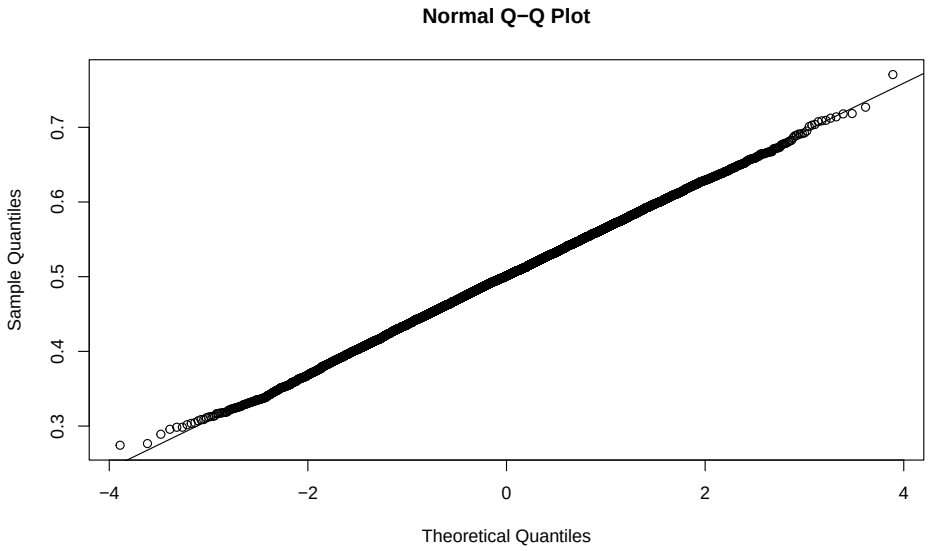
```
qqnorm(xbars10)
qqline(xbars10)
```



- $n = 20$

```
xbars20 <- replicate(10^4, mean(runif(20)))
```

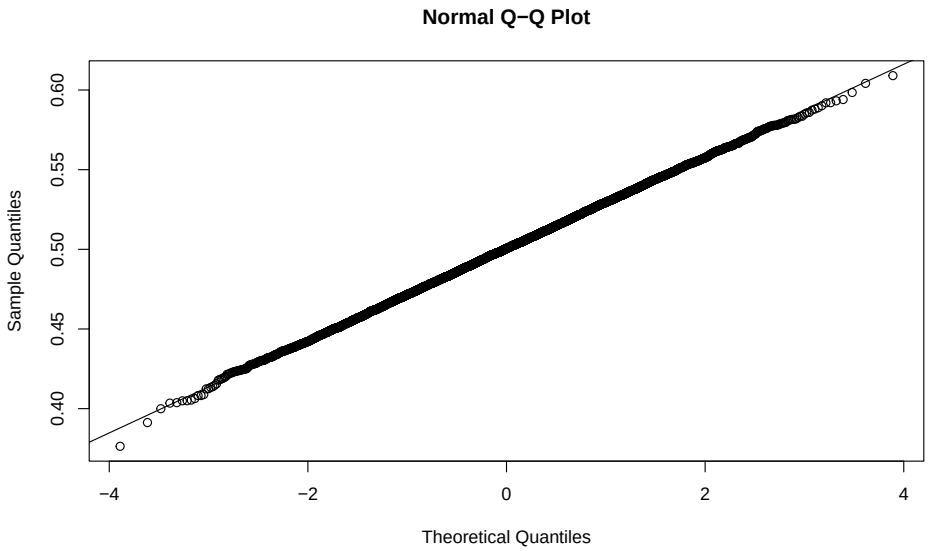
```
qqnorm(xbars20)
qqline(xbars20)
```



- $n = 100$

```r
xbars100 <- replicate(10^4, mean(runif(100)))
```

```r
qqnorm(xbars100)
qqline(xbars100)
```



On constate que plus n est grand, plus la densité de la moyenne se rapproche de la normale. C'est le TCL!

3.

Pour une variable $X \sim Unif(0,1)$, on a que $E(X) = 1/2$ et $Var(X) = 1/12$. Par le TCL, on sait que pour une échantillon i.i.d. de taille n qui provient d'une $Unif(0,1)$

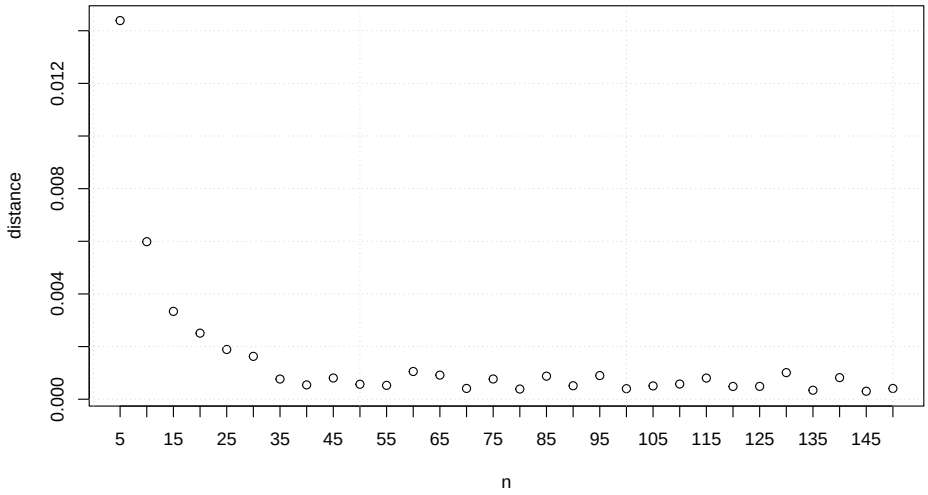
$$\overline{X}_n \sim_a N(1/2, 1/(12n))$$

Pour $n = 5$, par exemple, on peut mesurer numériquement la "distance" qui sépare les deux distributions de la façon suivante

```r
p <- seq(0.005, 0.995, by = 0.005)
{quantile(xbars5, p) - qnorm(p, mean = 1 / 2, sd = sqrt(1 / (12 * 5)))} |>
  abs() |> max()
```

```
[1] 0.00862
```

Par la suite on peut répéter cette analyse pour différentes valeurs de n et voir à partir de quel n on obtient une "bonne" approximation. La figure suivante montre les résultats obtenus pour $n = 5, 10, 15, \ldots, 150$



$\rightarrow$ avec $n = \pm 40$ on obtient déjà une très bonne approximation.