STID 1 - Programmation Statistique
TP4
Probabilités ; simulations

Antoine ROLLAND et Anthony SARDELLITTI

Mars 2023

Contents

```r
set.seed(2023)
```

1 Simulation d'une population

La taille moyenne des français est de 171cm avec un écart-type de 9 centimètres.

1. produire les tailles d'une population simulée de 10.000.000 de français répartis suivant une loi normale de moyenne 171 et d'écart-type 9. Stocker ces tailles dans un vecteur "population"

```r
moyenne_pop<-171
sd_pop<-9
population<-rnorm(1e7, mean=moyenne_pop, sd=sd_pop)
```

2. calculer la moyenne et l'écart-type de la population. Retrouvez-vous les valeurs attendues?

```r
mean(population)
```

```
## [1] 171.001
```

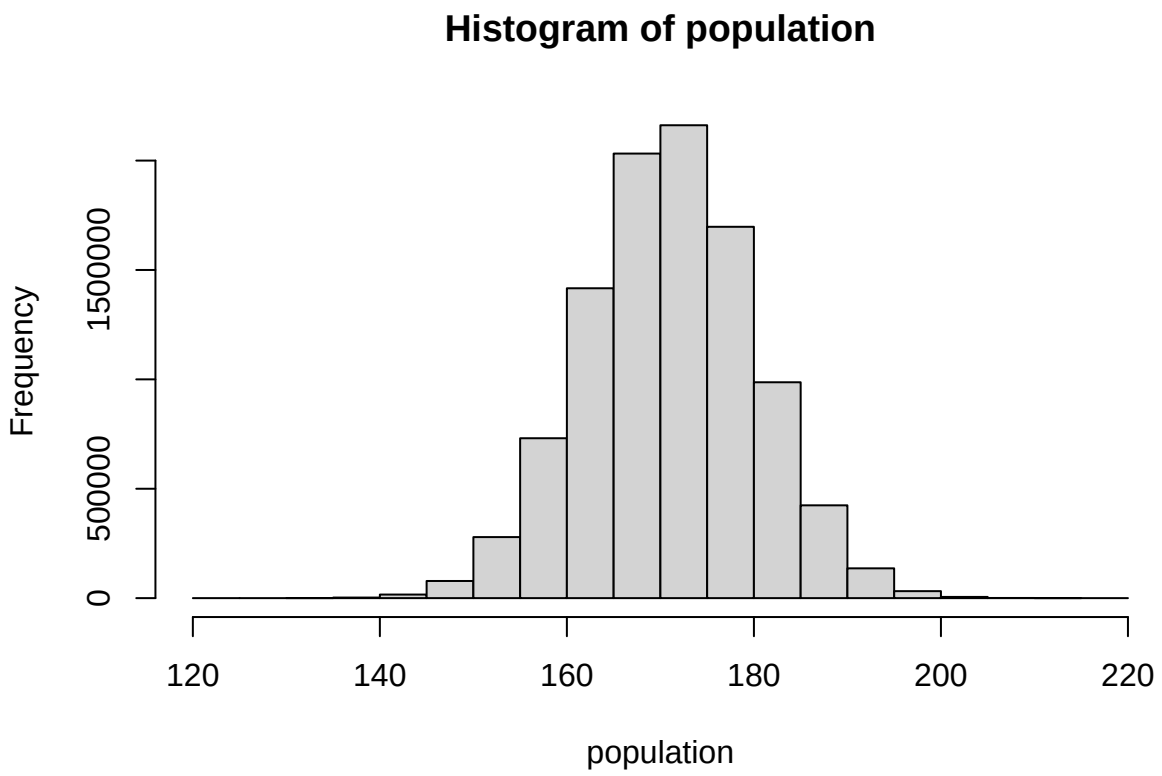
```r
sd(population)
```

```
## [1] 8.996877
```

La moyenne de la simulation de population est `{r} mean(population)`, qui est bien proche de 171, et l'écart-type `{r} sd(population)` qui est bien proche de 9.

3. établir l'histogramme de la taille. Retrouvez-vous la forme bien connues?

```
hist(population)
```



Quelle belle courbe en cloche :-)

4. Combien de personnes ont une taille supérieure à 190cm? Combien devriez-vous en trouver théoriquement?

```
sum(population>=190)
```

```
## [1] 174307
```

```
(1-pnorm(190, mean=moyenne_pop, sd=sd_pop))*1e7
```

```
## [1] 173813.8
```

Il y a 174307 personnes de plus de 190cm, et on s'attend en théorie à en trouver 1.7381381×10^5

5. Combien de personnes ont une taille inférieure à 144cm? Combien devriez-vous en trouver théoriquement?

```
sum(population<=144)
```

```
## [1] 13361
```

```
pnorm(144, mean=moyenne_pop, sd=sd_pop)*1e7
```

```
## [1] 13498.98
```

Il y a 13361 personnes de moins de 144cm, et on s'attend en théorie à en trouver 1.349898×10^4

2 Simulation d'échantillons

On va essayer d'estimer la taille moyenne de la population à partir d'un échantillon.

1. Tirez un échantillon de taille 100 dans la population initiale, à l'aide de la fonction sample. Quelle est la taille moyenne dans l'échantillon? Quelle est l'écart-type dans l'échantillon? Ces deux valeurs sont-elles proches de celles de la population?

```
taille_ech<-100
echantillon<-sample(population,taille_ech, replace = T)
mean(echantillon)
```

```
## [1] 171.5241
```

```
sd(echantillon)
```

```
## [1] 9.302921
```

2. à partir de l'écart-type de la population, calculez la largeur du demi-intervalle de fluctuation, puis les bornes inférieures et supérieures de l'intervalle de fluctuation (toujours à 95%)

```
largeur<-qnorm(0.975,mean=0,sd=1)*sd_pop/sqrt(taille_ech)
borne_inf<-moyenne_pop-largeur
borne_sup <-moyenne_pop+largeur
```

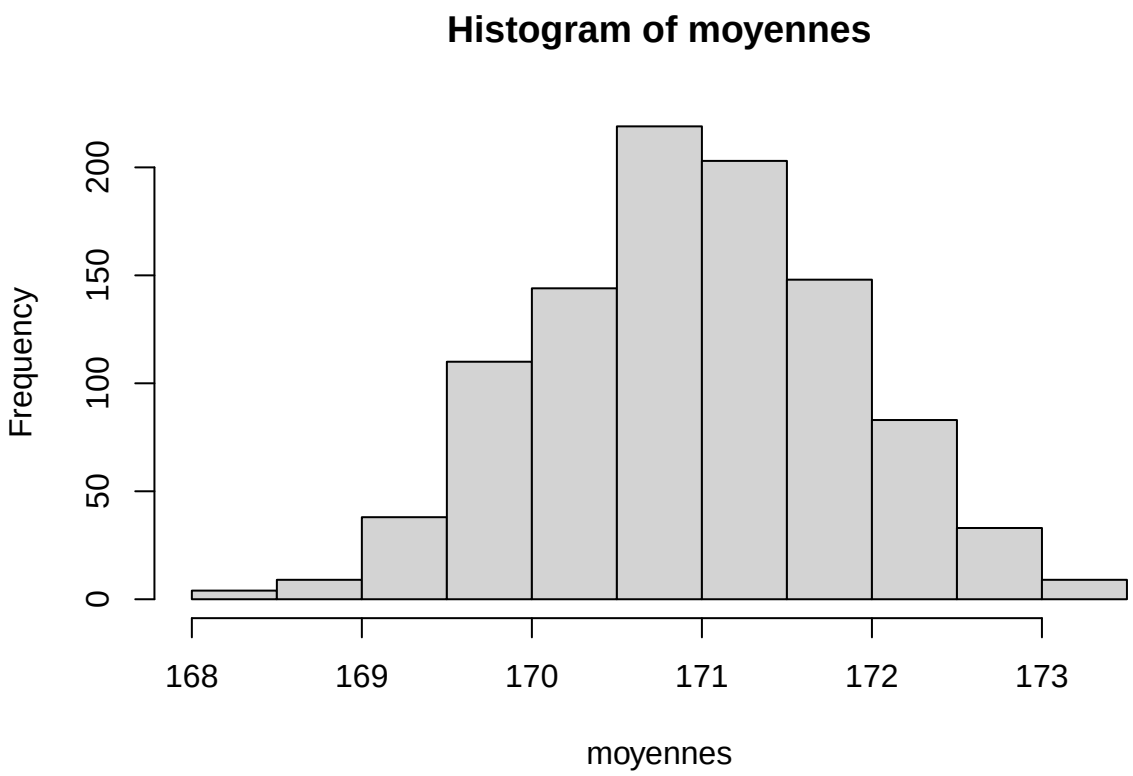
on a donc un intervalle de fluctuation à 95% qui est]169.2360324;172.7639676[.

3. A l'aide de la fonction replicate, tirez 1000 échantillons de taille 100. Stockez dans un dataframe la moyenne et l'écart-type de chaque échantillon (la fonction apply peut être utile).

```
taille_ech<-100
nb_replicat<-1000
echantillons<-replicate(nb_replicat, sample(population,taille_ech, replace = T))
moyennes<-apply(echantillons,2,mean)
ecart_types<-apply(echantillons,2, sd)
```

4. tracez l'histogramme des moyennes des échantillons. Retrouve-t-on une forme connue?

```
hist(moyennes)
```



5. Calculez la moyenne des moyennes des échantillons, ainsi que l'écart-type des moyennes des échantillons. Normalement la moyenne des moyennes doit être (à peu près) égale à la moyenne de la population : on dit que la moyenne est un estimateur non biaisé. De même l'écart-type des moyennes des échantillons doit être (à peu près) égal à 0,9 c'est-à-dire $\sigma/\sqrt{n}$.

```
mean(moyennes)
```

```
## [1] 170.9399
```

```
sd(moyennes)
```

```
## [1] 0.9039895
```

6. Combien d'échantillons ont une moyenne supérieure à 172,8cm? Quelle est le nombre théorique?

```
sum(moyennes>172.8)
```

```
## [1] 18
```

```
(1-pnorm(172.8, mean=moyenne_pop, sd=sd_pop/sqrt(taille_ech)))*nb_replicat
```

```
## [1] 22.75013
```

7. Pour chaque échantillon, calculez la largeur du demi-intervalle de confiance en utilisant l'estimation de l'écart-type calculée pour chaque échantillon, puis calculez les bornes inférieures et supérieures des intervalles de confiances (variables à rajouter dans votre dataframe).

```
largeur_est<-apply(echantillons, 2,function(x) pnorm(0.975)*sd(x)/taille_ech)
borne_inf_IC<-moyennes-largeur_est
borne_sup_IC<-moyennes+largeur_est
```

3 Effet taille de l'échantillon

1. créer une fonction "moyenne_echantillon" qui prend en entrée le vecteur V variable d'une population et une taille n d'échantillon et qui donne en sortie la moyenne d'un échantillon aléatoire de taille n tiré dans la population moyenne_echantillon <- function(V, n)

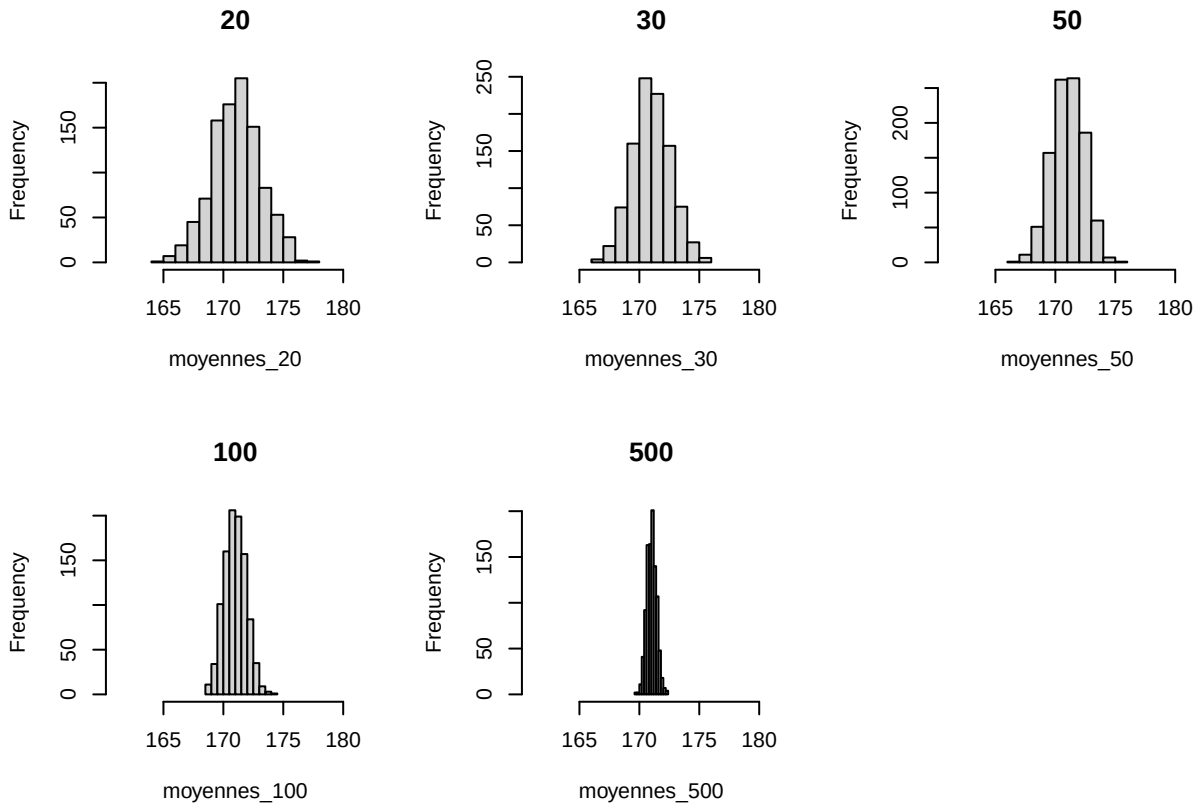
```
moyenne_echantillon<-function(V,n)
{moyenne_echantillon<-mean(sample(V,n, replace=T))
}
```

2. à l'aide de la fonction replicate, tirer 1000 échantillons pour chacune des tailles d'échantillons suivantes : 20, 30, 50, 100, 500, 1000 toujours à partir de la population initiale de 10.000.000 d'individus.

```
moyennes_20<-replicate(nb_replicat, moyenne_echantillon(population,20))
moyennes_30<-replicate(nb_replicat, moyenne_echantillon(population,30))
moyennes_50<-replicate(nb_replicat, moyenne_echantillon(population,50))
moyennes_100<-replicate(nb_replicat, moyenne_echantillon(population,100))
moyennes_500<-replicate(nb_replicat, moyenne_echantillon(population,500))
```

3. représentez les histogrammes des moyennes pour chaque taille d'échantillon en gardant les mêmes échelles des axes des abcisses et ordonnées. Que constate-t-on (spoiler alert : cela illustre le *théorème central limite*)?

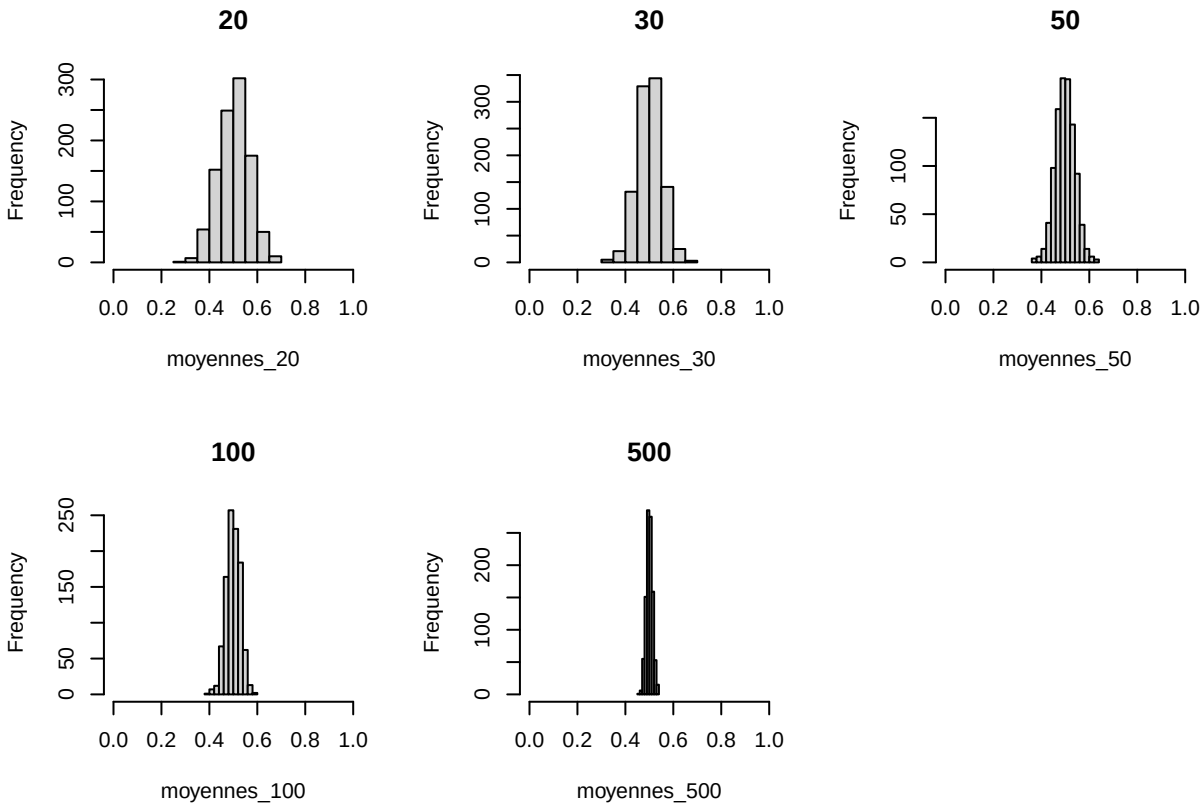
```
par(mfrow=c(2,3))
hist(moyennes_20, xlim=c(161,181), main="20")
hist(moyennes_30, xlim=c(161,181), main="30")
hist(moyennes_50, xlim=c(161,181), main="50")
hist(moyennes_100, xlim=c(161,181), main="100")
hist(moyennes_500, xlim=c(161,181), main="500")
```



4. reprendre les 3 questions précédentes avec une nouvelle population de 10.000.000 individus tirés à partir d'une loi uniforme sur [0,1]

```
population<-runif(1e7)

moyennes_20<-replicate(nb_replicat, moyenne_echantillon(population,20))
moyennes_30<-replicate(nb_replicat, moyenne_echantillon(population,30))
moyennes_50<-replicate(nb_replicat, moyenne_echantillon(population,50))
moyennes_100<-replicate(nb_replicat, moyenne_echantillon(population,100))
moyennes_500<-replicate(nb_replicat, moyenne_echantillon(population,500))
par(mfrow=c(2,3))
hist(moyennes_20, xlim=c(0,1), main="20")
hist(moyennes_30, xlim=c(0,1), main="30")
hist(moyennes_50, xlim=c(0,1), main="50")
hist(moyennes_100, xlim=c(0,1), main="100")
hist(moyennes_500, xlim=c(0,1), main="500")
```


20
Frequency
0
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
moyennes_20
30
Frequency
0
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
moyennes_30
50
Frequency
0
50
100
0.0
0.2
0.4
0.6
0.8
1.0
moyennes_50
100
Frequency
0
50
150
250
0.0
0.2
0.4
0.6
0.8
1.0
moyennes_100
500
Frequency
0
100
200
0.0
0.2
0.4
0.6
0.8
1.0
moyennes_500