SAE Estimation par échantillonnage

Binh Minh TRAN

2024-06-02

Introduction

Le projet vise à estimer les indicateurs statistiques dans des échantillons sans remise et avec remise.

En comprenant les estimations, vous pouvez identifier quelles sont les bonnes estimations pour les indicateurs statistiques ciblés.

Exercice 3 : Estimation des variances et intervalle de confiance.

On va simuler les estimations dans un contexte réel. Dans tous les cas réels, vous n'avez pas accès aux données de la population entière mais vous connaissez juste la taille de la population. Prenons encore le jeu de données `hdv2003.csv` et réaliser une estimation dans ce contexte réel.

```
hdv2003<-read.csv2(file="hdv2003.csv",sep=";",stringsAsFactors = TRUE,dec = ",")
hdv2003<-hdv2003[,-1] #supprimer premiere colonne
```

3A

Pour estimer la moyenne (dénotée par m) d'une variable quantitative sur toute la population, vous aller faire un sondage aléatoire simple (SAS) **avec remise**. Après cette étape, vous posséderez les données d'un échantillon de taille n. (pour ce projet, on fixe n=250 ou 500 à votre choix, ou faites les deux pour comparer). Vous ne pourrez utiliser que cet échantillon pour les étapes b)-d).

```
#un échantillon avec remise de taille n=500
SAS<-sample(1:nrow(hdv2003),500,replace = T)
ech<-hdv2003[SAS,]
```

3B

Vous calculez la moyenne (dénotée par $\hat{m}$) de la variable intéressée sur l'échantillon que vous possédez.Quelle théorie (théorème) vous permet de dire que m est une estimation de la vraie moyenne ?

On considère la variable `heures.tv` .

La fonction `EstimationM` représente le processus de prise d'échantillon avec remplacement et de calcul de sa moyenne.

```
EstimationM<-function(data,n,variable){
  Echantillon<-data[sample(1:nrow(data),n,replace = T),]
  EstimMoyenne<-mean(Echantillon[[variable]],na.rm=TRUE) #calculer la moyenne sans t
enir compte les NA.
  return(EstimMoyenne)
}
EstimationM(hdv2003,250,"heures.tv")#taille échantillon = 250
```

```
## [1] 2.434
```

```
(EstimMoyenne<-EstimationM(hdv2003,500,"heures.tv"))#taille échantillon = 500
```

```
## [1] 2.163454
```

On calcule la vraie moyenne de la population. Cette étape ne vise qu'à examiner la relation entre la moyenne de l'échantillon et la moyenne de la population. En réalité, cette étape ne sera pas réalisée car nous n'avons pas accès à la population.

```
# vérifier si la variable contient null.
any(is.na(hdv2003$heures.tv))
```

```
## [1] TRUE
```

```
#calculer la vraie moyenne de la variable heures.tv
(m<-mean(hdv2003$heures.tv,na.rm = TRUE))
```

```
## [1] 2.246566
```

On peut observer que la moyenne de l'échantillon tendra vers la moyenne de la population lorsque n devient suffisamment grand (normalement n>30 mais cela peut varier en fonction de chaque situation spécifique), ce qui peut être exprimé par le théorème central limite(TCL). Ainsi, la moyenne de l'échantillon sera une estimation non biaisée de la moyenne de la population.

> Le théorème central limite nous permet de dire que `EstimMoyenne` (denotée $\hat{m}$) est une estimation correcte de la vraie moyenne lorsque la taille de l'échantillon est grande.

3C

C'est également grâce au théorème central limite (TCL) que nous pouvons estimer la variance de la population à partir de l'échantillon.

On calcule la variance de l'échantillon non biaisée en utilisant la formule suivante :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Où :

- s^2 est la variance de l'échantillon non biaisée.
- n est la taille de l'échantillon.
- x_i sont les valeurs des observations dans l'échantillon.
- $\bar{x}$ est la moyenne des observations dans l'échantillon.

La raison pour laquelle nous utilisons $n-1$ est la suivante :

- Lorsque nous calculons la variance, nous mesurons la dispersion des observations par rapport à la moyenne de l'échantillon. La moyenne de l'échantillon elle-même est une estimation de la moyenne de la population, et cela introduit une contrainte supplémentaire dans notre estimation.
- Cette contrainte signifie que nous avons effectivement "perdu" un degré de liberté. En d'autres termes, après avoir utilisé une observation pour estimer la moyenne, nous n'avons que $n-1$ observations indépendantes restantes pour estimer la variance car la somme des écarts par rapport à la moyenne est toujours égale à zéro.

On trouve la variance de l'echantillon pour la variable `heures.tv` .

```
EstimationVar<-function(data,n,variable){
  Echantillon<-data[sample(1:nrow(data),n,replace = T),]

  #utiliser le formule au-dessus:
  EstimMoyenne<-mean(Echantillon[[variable]],na.rm=TRUE)
  EcartCarree<-(Echantillon[[variable]]-EstimMoyenne)^2
  var<-sum(EcartCarree,na.rm=TRUE)/(n-1)

  #utiliser var() pour calculer la variance sans tenir compte les NA.
  EstimVar<-var(Echantillon[[variable]],na.rm=TRUE)


  return(list("Estimation Variance par var()"=EstimVar,"Estimation Variance par le f
ormule"=var))
}
EstimationVar(hdv2003,250,"heures.tv")# taille échantillon = 250
```

```
## $`Estimation Variance par var()`
## [1] 3.155123
##
## $`Estimation Variance par le formule`
## [1] 3.129781
```

```
(EstimVariance<-EstimationVar(hdv2003,500,"heures.tv"))#taille échantillon = 500
```

```
## $`Estimation Variance par var()`
## [1] 4.322627
##
## $`Estimation Variance par le formule`
## [1] 4.305302
```

On suppose que la variance de l'échantillon est égale à la variance de la population.

On applique la formule suivante pour trouver la variance de l'estimation $\hat{m}$:

$$\mathrm{Var}(\hat{m}) = \frac{s^2}{n}$$

```
(Variance_EstimationR<-EstimVariance[[2]]/500)
```

```
## [1] 0.008610604
```

On calcule la vraie variance de la population. Cette étape ne vise qu'à examiner la relation entre la variance de l'estimation et la variance de la population. En réalité, cette étape ne sera pas réalisée car nous n'avons pas accès à la population.

```
#calculer la vraie variance de l'estimation
(varEstimateur<-var(hdv2003$heures.tv,na.rm = TRUE)/500)
```

```
## [1] 0.006307306
```

> On peut dire que la variance de l'estimation est toujours inferieur que la variance de la population.

3D

L'intervalle de confiance est défini comme suit :

$$\text{Estimation} \pm z \cdot SE$$

Où :

- **Estimation** est la valeur estimée d'un paramètre statistique,
- **z** est la valeur critique associée au niveau de confiance choisi.
- **SE** est l'erreur standard, qui mesure l'incertitude de l'estimation de la valeur du paramètre. SE = $\frac{\sigma}{\sqrt{n}} = \sqrt{Var(\hat{m})}$.

Pour un niveau de confiance de 95%, z est la valeur critique associée à la probabilité de 0.975 dans une distribution normale standard. Ainsi, z équivaut à **1.96** en se référant à une table de probabilités de la loi normale standard (https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf).

Supposons que σ **de la population** = σ **de l'échantillon**. On calcule l'estimation d'écart-type = $\sqrt{\text{variance}}$

> Dans ce cas, en supposant que l'écart type de la population (σ) est égal à l'écart type de l'échantillon, nous pouvons utiliser le Théorème Central Limite (Central Limit Theorem - CLT). Le CLT établit que la distribution de la moyenne d'un échantillon tend vers une distribution normale lorsque la taille de l'échantillon augmente, quelle que soit la distribution d'origine de la population.

```
#EstimVariance[[2]] : Accéder au deuxième élément dans la liste EstimVariance. Nous
avons déjà le fonction EstimationVar() définie au dessus.
(EstimEcarttype<-sqrt(EstimVariance[[2]]))
```

```
## [1] 2.074922
```

On calcule SE:

```
(SE<-(EstimEcarttype/sqrt(500)))
```

```
## [1] 0.09279334
```

```
(SE<-(sqrt(Variance_EstimationR)))
```

```
## [1] 0.09279334
```

```
EstimMoyenne
```

```
## [1] 2.163454
```

Intervalle de Confiance pour la moyenne:

```
limiteSup<-EstimMoyenne+1.96*SE
limiteInf<-EstimMoyenne-1.96*SE
(IntervalleConfiance<-list("limite inférieure"=limiteInf,"limite supérieure"=limiteS
up))
```

```
## $`limite inférieure`
## [1] 1.981579
##
## $`limite supérieure`
## [1] 2.345329
```

3E

On fait 1000 sondages pour trouver la distribution des estimations de la moyenne.

```
MoyenneEstimations<-function(data,n,variable,nbSond){
  MoyenneEstimation<-c()
  for(i in 1:nbSond){
  Echantillon<-data[sample(1:nrow(data),n,replace = T),]
  EstimMoyenne<-mean(Echantillon[[variable]],na.rm=TRUE) #calculer la moyenne sans t
enir compte les NA.
  MoyenneEstimation<-c(MoyenneEstimation,EstimMoyenne)
  }
  return(MoyenneEstimation)
}
Estimation250<-MoyenneEstimations(hdv2003,250,"heures.tv",1000) #taille échantillon
= 250
Estimation500<-MoyenneEstimations(hdv2003,500,"heures.tv",1000) #taille échantillon
= 500
```

On trace :

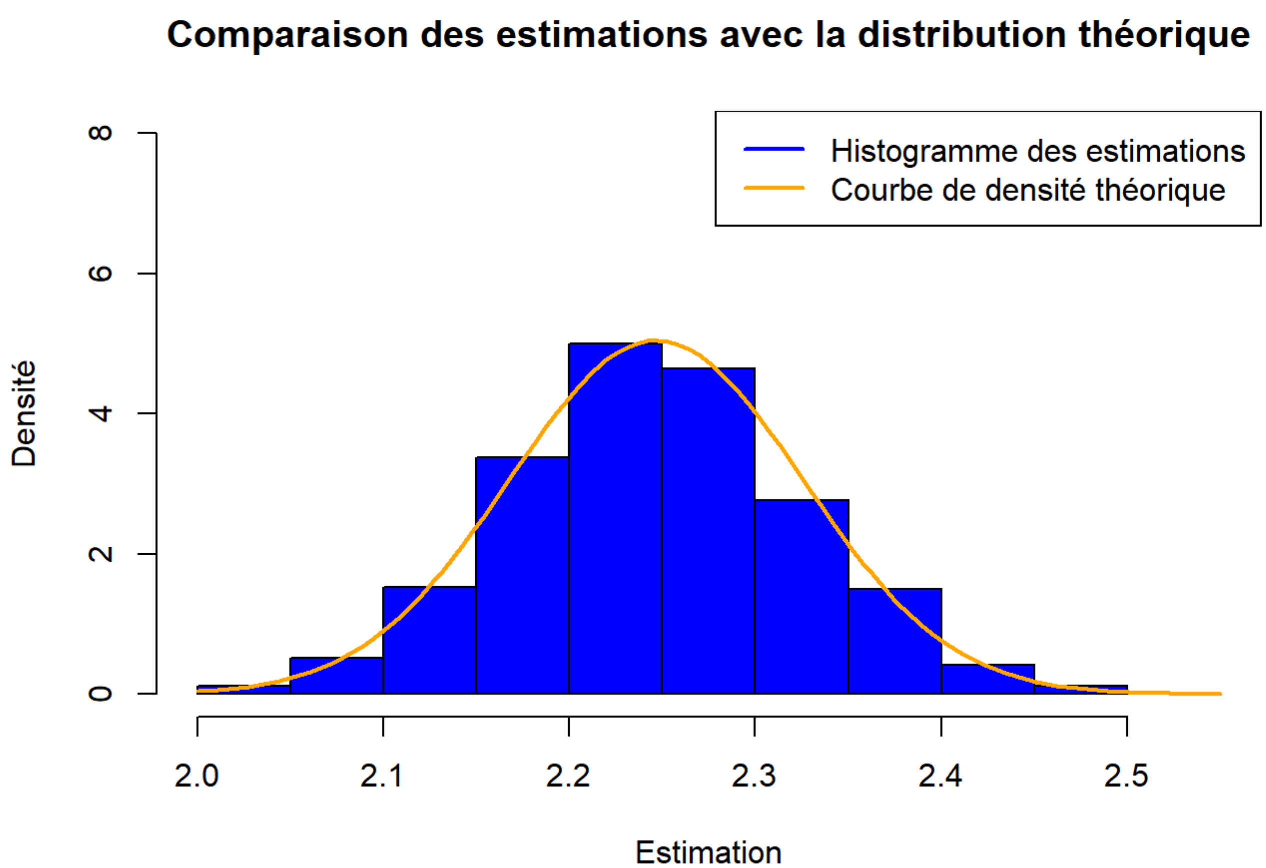
- un graphique de la comparaison des estimations avec la distribution théorique.

```
hist(Estimation500,col="blue",main="Comparaison des estimations avec la distribution
théorique",ylab="Densité",xlab="Estimation",probability = TRUE,ylim=c(0,8))

# Calculer la moyenne et l'écart-type de la distribution des estimations
mean_est <- mean(Estimation500)
sd_est <- sd(Estimation500)

# Ajouter la courbe de densité théorique
curve(dnorm(x, mean = mean_est, sd = sd_est), col = "orange", lwd = 2, add = TRUE)

# Ajouter une légende
legend("topright", legend = c("Histogramme des estimations", "Courbe de densité théo
rique"), col = c("blue", "orange"), lwd = 2, bg = "white")
```



> Nous pouvons voir que, selon le théorème central limite (TCL), la distribution des estimations se rapprochera d'une distribution normale centrée autour de la moyenne de la population lorsque la taille de l'échantillon et le nombre d'échantillons sont suffisamment grands. En revanche, la distribution de la population peut rester asymétrique

Preuve:

Lorsque le nombre de sondages est petit, la distribution des estimations ne se rapprochera pas autant d'une distribution normale centrée autour de la moyenne de la population.

On réalise 20 sondages sur des échantillons de taille 500 :

```
Estimation500<-MoyenneEstimations(hdv2003,500,"heures.tv",20)
```

On trace un graphique :

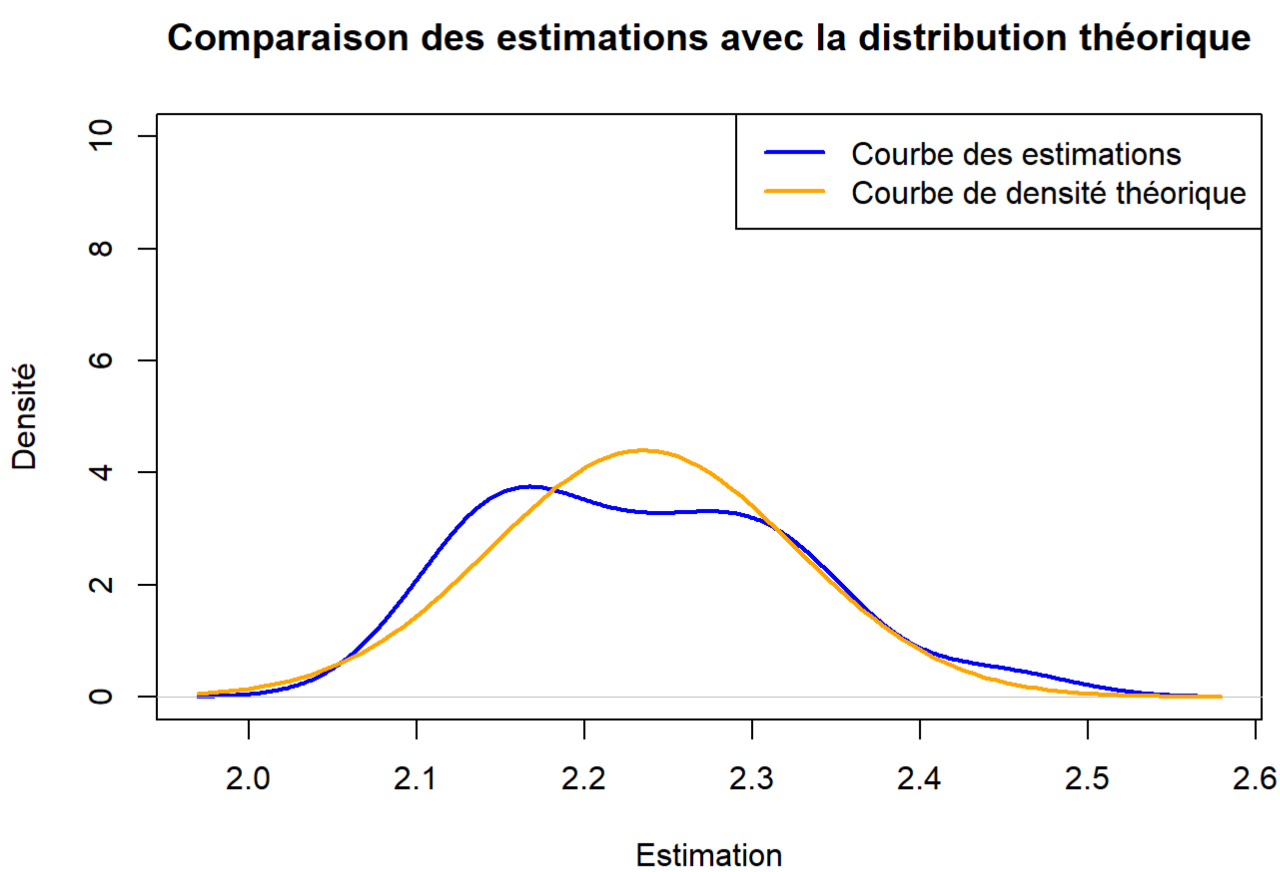
```
# Calculer la moyenne et l'écart-type de la distribution des estimations
mean_est <- mean(Estimation500)
sd_est <- sd(Estimation500)

density_estimation <- density(Estimation500)

# Tracer la courbe de densité des estimations
plot(density_estimation, col = "blue", main = "Comparaison des estimations avec la d
istribution théorique",
     xlab = "Estimation", ylab = "Densité", ylim = c(0, 10),lwd=2)

# Ajouter la courbe de densité théorique
curve(dnorm(x, mean = mean_est, sd = sd_est), col = "orange", lwd = 2, add = TRUE)

# Ajouter une légende
legend("topright", legend = c("Courbe des estimations", "Courbe de densité théoriqu
e"), col = c("blue", "orange"), lwd = 2, bg = "white")
```



Exercise 4: Amélioration

On peut encore améliorer l'estimation de l'Exercice 3 en utilisant le SAS **sans remise**. Supposons qu'on a un ensemble d'une population de taille . On retire un sous-ensemble de taille (n<N). On calcule la moyenne d'une variable X de la popula&on sur cet échantillon. (L'indice 1 correspond au

premier tirage.)

4A

Quand nous utilisons la méthode d'échantillonnage avec remplacement (SAS), cela signifie que nous prélevons un sous-ensemble sans avoir à le remettre et que le prélèvement se fait simultanément. Par conséquent, nous utilisons des combinaisons pour calculer dans ce cas. Si nous avons un sous-ensemble S_1 de taille n à partir d'un ensemble S de taille N, alors nous utiliserons la combinaison $C(N, n)$.

La combinaison $C(N, n)$ est calculée comme suit :

$$C(N, n) = \frac{N!}{n!(N-n)!}$$

Par example : en supposant N = 100 et n = 2

```
N<-100
n<-2
#Combinaison de sous-ensembles possibles
(C<-factorial(N)/(factorial(n)*factorial(N-n)))
```

```
## [1] 4950
```

On a un total de 4950 sous-ensembles possibles obtenus à partir de l'ensemble.

Puisque la probabilité de chaque sous-ensemble retiré est similaire, elle est donc égale :

$$\frac{1}{C(N, n)}$$

Alors, la probabilité totale est 1.

On reprend l'exemple :

```
(Proba<-(1/C)*100)
```

```
## [1] 0.02020202
```

```
(TotalProba<-Proba*C)
```

```
## [1] 100
```

4B

On utilise un échantillon S_1 avec la moyenne $\hat{m}_1$ afin d'estimer la moyenne de la population m, autrement dit, $\hat{m}_1$ est un estimateur pour m. Donc, $\hat{m}_1$ est la moyenne de toutes les observations dans S_1, soit

$$\hat{m}_1 = \frac{\sum_{i=1}^{n} x_i}{n}$$

On tire le sous-ensemble sans remplacement, donc l'échantillon maintenant n'est pas i.i.d. Alors, on doit tenir compte de tous les sous-ensembles tirés possibles.

L'espérance de $\hat{m}_1$ est égale à la moyenne de tous les sous-ensembles possibles

$$E(\hat{m}_1) = \frac{\sum_{i=1}^{C(N,n)} \hat{m}_i}{C(N,n)}$$

$$E(\hat{m}_1) = \frac{\sum_{i=1}^{C(N,n)} \frac{X_1 + \ldots + X_n}{n}}{C(N,n)}$$

$$E(\hat{m}_1) = \frac{\sum_{i=1}^{C(N,n)} X_1 + \ldots + X_n}{C(N,n) \cdot n}$$

Le problème est que l'on n'a pas X_1 dans tous les sous-ensembles. On compte tous les cas de X_1 qui sont contenus dans le sous-ensemble. Effectivement, quand on choisit X_1 dans un sous-ensemble, il y a $C(N-1, n-1)$ cas de sous-ensembles tirés possibles.

$$E(\hat{m}_1) = \frac{X_1 \cdot C(N-1, n-1) + \ldots + X_N \cdot C(N-1, n-1)}{C(N,n) \cdot n}$$

$$E(\hat{m}_1) = \frac{X_1 + \ldots + X_N}{N} = m$$

où:

- m est la vraie moyenne.

4D

On estime la variance de l'échantillon sans remise :

$$s^2 = \frac{\sum_{i=1}^{n} (xi - \hat{m})^2}{n}$$

```
EstimationVar<-function(data,n,variable){
  Echantillon<-data[sample(1:nrow(data),n,replace = F),]

  #utiliser le formule au-dessus:
  EstimMoyenne<-mean(Echantillon[[variable]],na.rm=TRUE)
  EcartCarree<-(Echantillon[[variable]]-EstimMoyenne)^2
  var<-sum(EcartCarree,na.rm=TRUE)/(n)

  return(list("Estimation Variance par le formule"=var))
}
EstimationVar(hdv2003,250,"heures.tv")# taille échantillon = 250
```

```
## $`Estimation Variance par le formule`
## [1] 3.416804
```

```
(EstimVariance<-EstimationVar(hdv2003,500,"heures.tv"))#taille échantillon = 500
```

```
## $`Estimation Variance par le formule`
## [1] 2.782608
```

4E

On estime la moyenne de l'échantillon sans remise :

```
EstimationM<-function(data,n,variable){
  Echantillon<-data[sample(1:nrow(data),n,replace = F),]
  EstimMoyenne<-mean(Echantillon[[variable]],na.rm=TRUE) #calculer la moyenne sans t
enir compte les NA.
  return(EstimMoyenne)
}
EstimationM(hdv2003,250,"heures.tv")#taille échantillon = 250
```

```
## [1] 2.283133
```

```
(EstimMoyenne<-EstimationM(hdv2003,500,"heures.tv"))#taille échantillon = 500
```

```
## [1] 2.258753
```

On estime la variance de l'échantillon sans remise :

```
EstimVariance#taille échantillon = 500
```

```
## $`Estimation Variance par le formule`
## [1] 2.782608
```

On suppose que la variance de l'échantillon est égale à la variance de la population.

On applique la formule suivante pour trouver **la variance de l'estimation** $\hat{m}$:

$$\mathrm{Var}(\hat{m}) = \frac{Var(m)}{n} \cdot \frac{N-n}{N-1}$$

$$\mathrm{Var}(\hat{m}) = (\frac{1-\frac{1}{N}}{1-\frac{1}{n}} \cdot E(s^2)) \cdot \frac{N-n}{N-1} \cdot \frac{1}{n}$$

$$\mathrm{Var}(\hat{m}) = (\frac{1-\frac{1}{N}}{1-\frac{1}{n}} \cdot s^2) \cdot \frac{N-n}{N-1} \cdot \frac{1}{n}$$

où :

- N est la taille de la population.
- n est la taille de l'échantillon.
- s^2 est la variance de l'échantillon.

```
#L'echantillon de taille 500
(Variance_EstimationSR<-EstimVariance[[1]]*((2000-500)/1999)*(1/500)*((1-(1/2000))/
(1-(1/500))))
```

```
## [1] 0.004182277
```

Rapelle :

L'intervalle de confiance

$$\text{Estimation} \pm z \cdot SE$$

Où :

- **Estimation** est la valeur estimée d'un paramètre statistique,
- **z** est la valeur critique associée au niveau de confiance choisi.
- **SE** est l'erreur standard, qui mesure l'incertitude de l'estimation de la valeur du paramètre. SE = $\frac{\sigma}{\sqrt{n}} = \sqrt{Var(\hat{m})}$.

Pour un niveau de confiance de 95%, z est la valeur critique associée à la probabilité de 0.975 dans une distribution normale standard. Ainsi, z équivaut à **1.96** en se référant à une table de probabilités de la loi normale standard (https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf).

On calcule SE:

```
(SE<-(sqrt(Variance_EstimationSR)))
```

```
## [1] 0.06467053
```

```
EstimMoyenne
```

```
## [1] 2.258753
```

Intervalle de Confiance pour la moyenne:

```
limiteSup<-EstimMoyenne+1.96*SE
limiteInf<-EstimMoyenne-1.96*SE
(IntervalleConfiance<-list("limite inférieure"=limiteInf,"limite supérieure"=limiteS
up))
```

```
## $`limite inférieure`
## [1] 2.131998
##
## $`limite supérieure`
## [1] 2.385507
```

On fait 1000 sondages pour trouver la distribution des estimations de la moyenne.

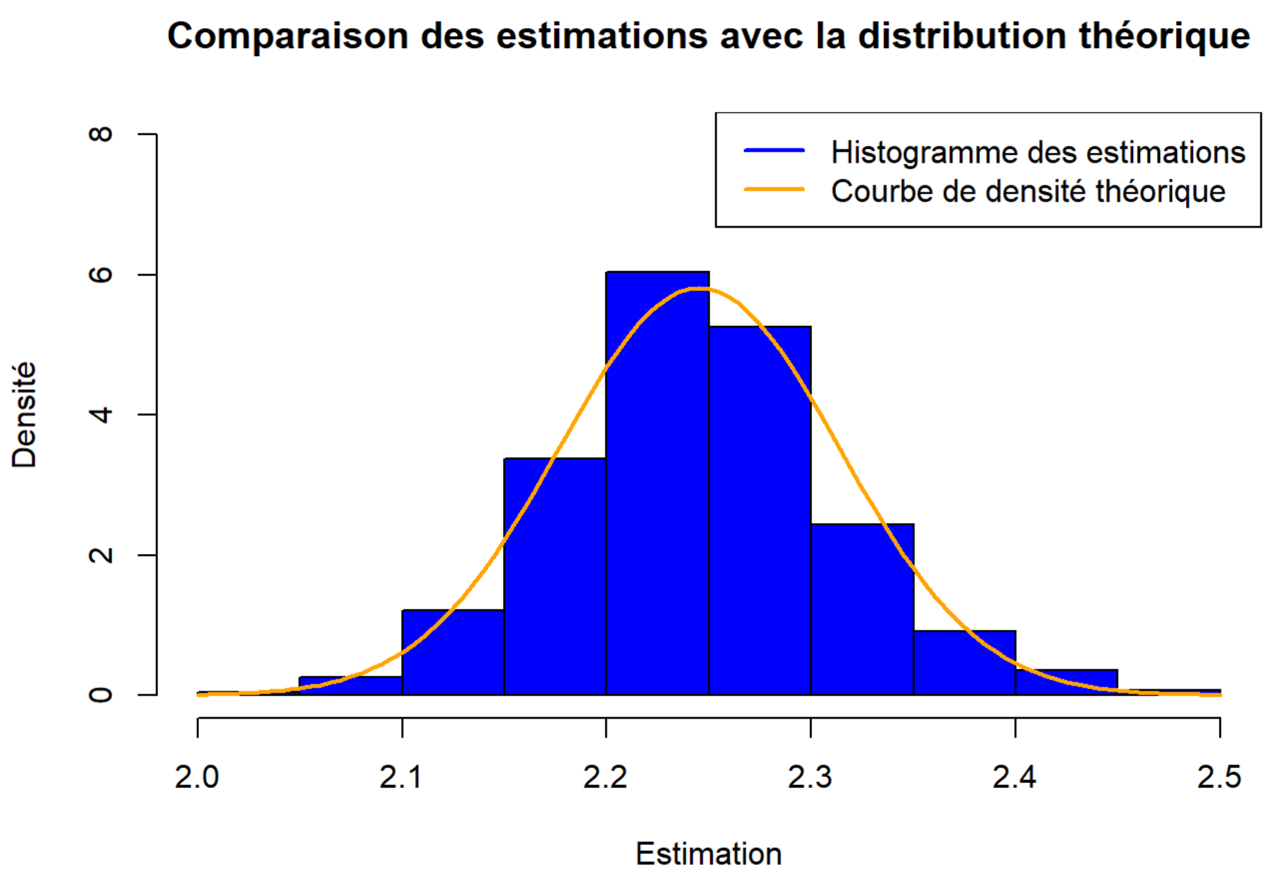
```
MoyenneEstimations<-function(data,n,variable,nbSond){
  MoyenneEstimation<-c()
  for(i in 1:nbSond){
  Echantillon<-data[sample(1:nrow(data),n,replace = F),]
  EstimMoyenne<-mean(Echantillon[[variable]],na.rm=TRUE) #calculer la moyenne sans t
enir compte les NA.
  MoyenneEstimation<-c(MoyenneEstimation,EstimMoyenne)
  }
  return(MoyenneEstimation)
}
Estimation250<-MoyenneEstimations(hdv2003,250,"heures.tv",1000) #taille échantillon
= 250
Estimation500<-MoyenneEstimations(hdv2003,500,"heures.tv",1000) #taille échantillon
= 500
```

```
hist(Estimation500,col="blue",main="Comparaison des estimations avec la distribution
théorique",ylab="Densité",xlab="Estimation",probability = TRUE,ylim=c(0,8))

# Calculer la moyenne et l'écart-type de la distribution des estimations
mean_est <- mean(Estimation500)
sd_est <- sd(Estimation500)

# Ajouter la courbe de densité théorique
curve(dnorm(x, mean = mean_est, sd = sd_est), col = "orange", lwd = 2, add = TRUE)

# Ajouter une légende
legend("topright", legend = c("Histogramme des estimations", "Courbe de densité théo
rique"), col = c("blue", "orange"), lwd = 2, bg = "white")
```



La comparaison entre les méthodes SAS:

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.3
```

```
Variance_EstimationR <-Variance_EstimationR
Variance_EstimationSR <-Variance_EstimationSR

data <- data.frame(
  "VarianceEstimationSASavecRemise" = Variance_EstimationR,
  "VarianceEstimationSASsansRemise" = Variance_EstimationSR
)

kable(data, format = "markdown")
```

VarianceEstimationSASavecRemise	VarianceEstimationSASsansRemise
0.0086106	0.0041823

> **En conclusion** : Le SAS sans remise offre généralement des estimations plus précises et une variance plus faible, mais peut être plus complexe à mettre en œuvre. Le SAS avec remise, en revanche, est plus simple à gérer, mais peut produire des estimations moins précises en raison de la possibilité de sélectionner plusieurs fois la même unité. Ainsi, le choix doit être guidé par un équilibre entre la précision nécessaire et les contraintes opérationnelles.