SAE Estimation par échantillonnage

Binh Minh TRAN

2024-09-01

Contents

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.

Notation

Symbol	Description
σ	L'écart-type de la population
σ^2	La variance de la population
m	La moyenne de la population
n	La taille de l'échantillon
e	Le nombre de sondages
s	L'écart-type de l'échantillon
s^2	La variance de l'échantillon
$\hat{\sigma}^2$	L'estimation de la variance de la population σ^2
$\hat{\sigma}$	L'estimation de l'écart-type de la population σ
$\hat{m}$	L'estimation de la moyenne de la population m

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

```
MoyenneEchantillon <- function(data, n, variable) {
    Echantillon <- data[sample(1:nrow(data), n, replace = T),
        ]
    # calculer la moyenne sans tenir compte les NA.
    EstimMoyenne <- mean(Echantillon[[variable]], na.rm = TRUE)
    return(list(Moyenne = EstimMoyenne, Echantillon = Echantillon))
}
Moyenne_taille250 <- MoyenneEchantillon(hdv2003, 250, "heures.tv")  #taille échantillon = 250
cat("La moyenne de l'échantillon de taille 250:  ", Moyenne_taille250$Moyenne)
```

```
## La moyenne de l'échantillon de taille 250:   2.4476
```

```
Moyenne_Echantillon <- MoyenneEchantillon(hdv2003, 500, "heures.tv")  #taille échantillon = 500
cat("La moyenne de l'échantillon de taille 500:  ", Moyenne_Echantillon$Moyenne)
```

```
## La moyenne de l'échantillon de taille 500:   2.2028
```

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
(moyenne_population <- 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 `Moyenne_Echantillon` (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 :

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

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.
- $\hat{m}$ est la moyenne de 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`.

```
VarianceEchantillon <- function(data, n, variable) {
    MoyenneEchantillon <- MoyenneEchantillon(data, n, variable)
    Moyenne <- MoyenneEchantillon$Moyenne
    EcartCarree <- (MoyenneEchantillon$Echantillon[[variable]] -
        Moyenne)^2
    var <- sum(EcartCarree, na.rm = TRUE)/(n - 1)

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

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

```
## $`Variance Échantillon par var()`
## [1] 4.073757
##
## $`Variance Échantillon par le formule`
## [1] 4.041036
```

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

```
## $`Variance Échantillon par var()`
## [1] 2.947411
##
## $`Variance Échantillon par le formule`
## [1] 2.947411
```

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 $\hat{m}$:

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

```
# On utilise la variance de l'échantillon de taille 500.
(Variance_Estimateur <- VariancePopulationEstim[[2]]/500)
```

```
## [1] 0.005894822
```

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
(Var_Estimateur_test <- var(hdv2003$heures.tv, na.rm = TRUE)/500)
```

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

Nombreux Échantillons

Dans ce cas, avec 1000 sondages, nous avons deux approches pour calculer la variance de l'estimation $\hat{m}$: l'utilisation de la variance estimée de la population ou celle des moyennes des échantillons.

La formule que l'on utilise pour calculer la variance de l'estimation à partir de la variance des moyennes des échantillons est la suivante :

$$\widehat{\text{Var}}(\hat{m}) = \frac{1}{e-1}\sum_{i=1}^{e}(\hat{m}_i - \bar{\hat{m}})^2$$

où :

- e est le nombre de sondages.
- $\hat{m}$ est l'estimation de la moyenne de la population.
- $\bar{\hat{m}}$ est la moyenne des moyennes des échantillons.

```
MoyennesDesEchantillons <- function(data, n, variable, nbSond) {
    MoyennesEchantillons <- c()
    for (i in 1:nbSond) {
        MoyenneEchantillon <- MoyenneEchantillon(data, n, variable)
        Moyenne <- MoyenneEchantillon$Moyenne
        MoyennesEchantillons <- c(MoyennesEchantillons, Moyenne)
```

```
    }
    return(MoyennesEchantillons)
}
Moyennes_taille500 <- MoyennesDesEchantillons(hdv2003, 500, "heures.tv",
    1000)  #taille échantillon = 500
```

```
MoyenneDesEstimations <- mean(Moyennes_taille500)
VarEstimateur <- sum((Moyennes_taille500 - MoyenneDesEstimations)^2)/(500 -
    1)
cat("Variance de l'estimation à partir de la variance des moyennes des échantillons: ",
    VarEstimateur)
```

```
## Variance de l'estimation à partir de la variance des moyennes des échantillons:  0.01244489
```

La formule utilisée pour calculer la variance de l'estimation à partir de la variance estimée de la population est la suivante :

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

où :

- $\hat{\sigma}^2$ est estimée par $\frac{1}{e}\sum_{i=1}^{e} s_i^2$, où s_i^2 est la variance du i-ème échantillon, et e est le nombre de sondages.
- n est la taille d'un échantillon.

```
VariancesDesEchantillons <- function(data, n, variable, nbSond) {
    VariancesEchantillons <- c()
    for (i in 1:nbSond) {
        MoyenneEchantillon <- MoyenneEchantillon(data, n, variable)
        Moyenne <- MoyenneEchantillon$Moyenne
        EcartCarree <- (MoyenneEchantillon$Echantillon[[variable]] -
            Moyenne)^2
        var <- sum(EcartCarree, na.rm = TRUE)/(n - 1)
        # Creer un vecteur qui stocke les variances des
        # échantillons
        VariancesEchantillons <- c(VariancesEchantillons, var)
    }
    return(VariancesEchantillons)
}
Variances_taille500 <- VariancesDesEchantillons(hdv2003, 500,
    "heures.tv", 500)  #taille échantillon = 500
cat("Variance de l'estimation à partir de la variance estimée de la population: ",
    mean(Variances_taille500)/500)
```

```
## Variance de l'estimation à partir de la variance estimée de la population:  0.006289073
```

De plus, il est remarqué que la variance de l'estimateur, c'est-à-dire la variance des moyennes des échantillons, est toujours inférieure à 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.

Supposons que $\sigma_{\textbf{population}} = \sigma_{\textbf{échantillon}}$ quand on a seulement 1 échantillon. On calcule l'écart-type de la population = $\sqrt{\text{variance de l'échantillon}}$.

> 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.

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

```
## [1] 1.716803
```

On calcule SE = $\frac{\sigma}{\sqrt{n}} = \sqrt{Var(\hat{m})}$:

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

```
## [1] 0.07677774
```

```
(SE <- (sqrt(Variance_Estimateur)))
```

```
## [1] 0.07677774
```

```
Moyenne_Echantillon$Moyenne
```

```
## [1] 2.2028
```

Intervalle de Confiance pour la moyenne:

```
limiteSup <- Moyenne_Echantillon$Moyenne + 1.96 * SE
limiteInf <- Moyenne_Echantillon$Moyenne - 1.96 * SE
(IntervalleConfiance <- list(`limite inférieure` = limiteInf,
    `limite supérieure` = limiteSup))
```

```
## $`limite inférieure`
## [1] 2.052316
##
## $`limite supérieure`
## [1] 2.353284
```

On essaie d'estimer la moyenne de la population en 10 fois afin de vérifier la confiance à 95% de l'intervalle.

```
for (i in 1:10) {
    cat("Estimateur pour la moyenne de la population: ", mean(MoyennesDesEchantillons(hdv2003,
        500, "heures.tv", 1000)), "\n")
}
```

```
## Estimateur pour la moyenne de la population:  2.246836
## Estimateur pour la moyenne de la population:  2.242336
## Estimateur pour la moyenne de la population:  2.24595
## Estimateur pour la moyenne de la population:  2.249097
## Estimateur pour la moyenne de la population:  2.243487
## Estimateur pour la moyenne de la population:  2.244782
## Estimateur pour la moyenne de la population:  2.246554
## Estimateur pour la moyenne de la population:  2.249143
## Estimateur pour la moyenne de la population:  2.240648
## Estimateur pour la moyenne de la population:  2.241639
```

> Alors, aucun cas parmi ces 10 cas ne renvoie la valeur de l'estimateur en dehors de l'intervalle de confiance que nous avons calculé ci-dessus. On peut réaliser beaucoup plus d'estimations pour la moyenne afin de trouver une valeur en dehors de cet intervalle, car un intervalle de confiance de 95% est très légitime.

3E

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

```
Estimateurs_taille500 <- MoyennesDesEchantillons(hdv2003, 500,
    "heures.tv", 1000)  #taille échantillon = 500
```

On trace :

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

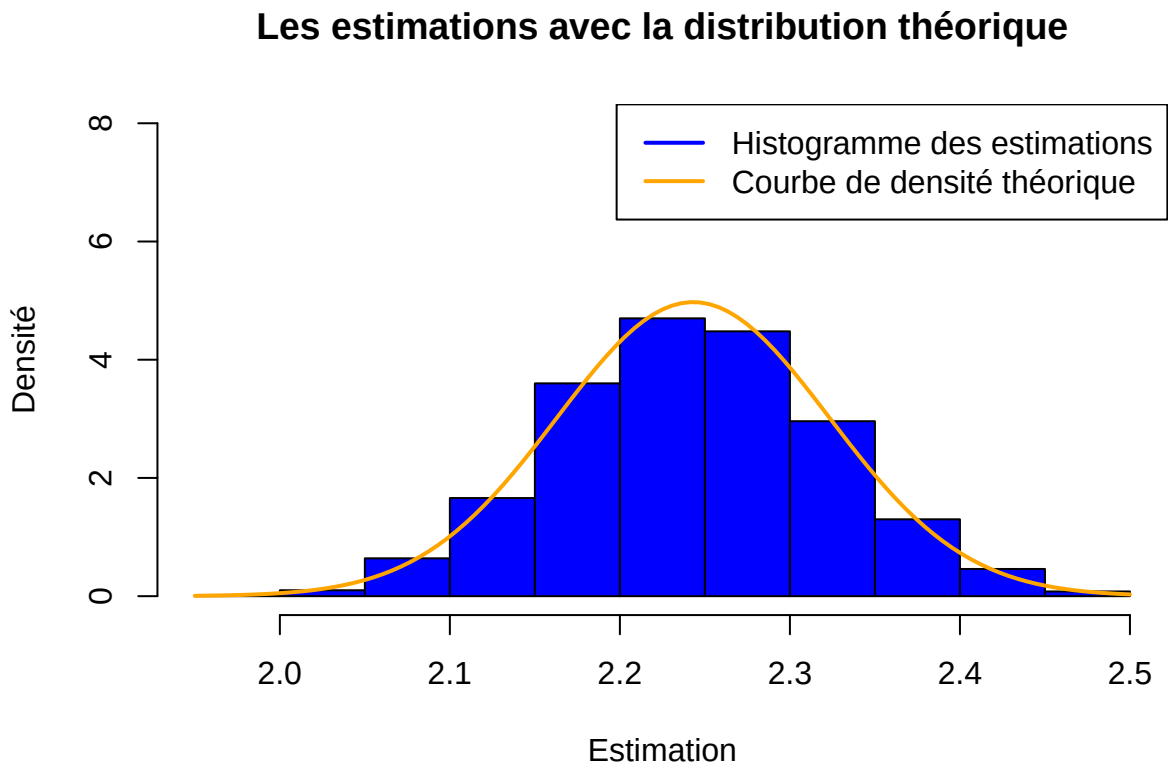
```
hist(Estimateurs_taille500, col = "blue", main = "Les estimations avec la distribution théorique",
    ylab = "Densité", xlab = "Estimation", probability = TRUE,
    ylim = c(0, 8))

# Calculer la moyenne et l'écart-type de la distribution
```

```
# des estimations
mean_est <- mean(Estimateurs_taille500)
sd_est <- sd(Estimateurs_taille500)

# 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 théorique"), col = c("blue", "orange"),
    lwd = 2, bg = "white")
```



Conclusion: 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 ou la taille de l'échantillon est petite, 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 :

```
Estimateurs_taille500 <- MoyennesDesEchantillons(hdv2003, 500,
    "heures.tv", 20)
```

On trace un graphique :

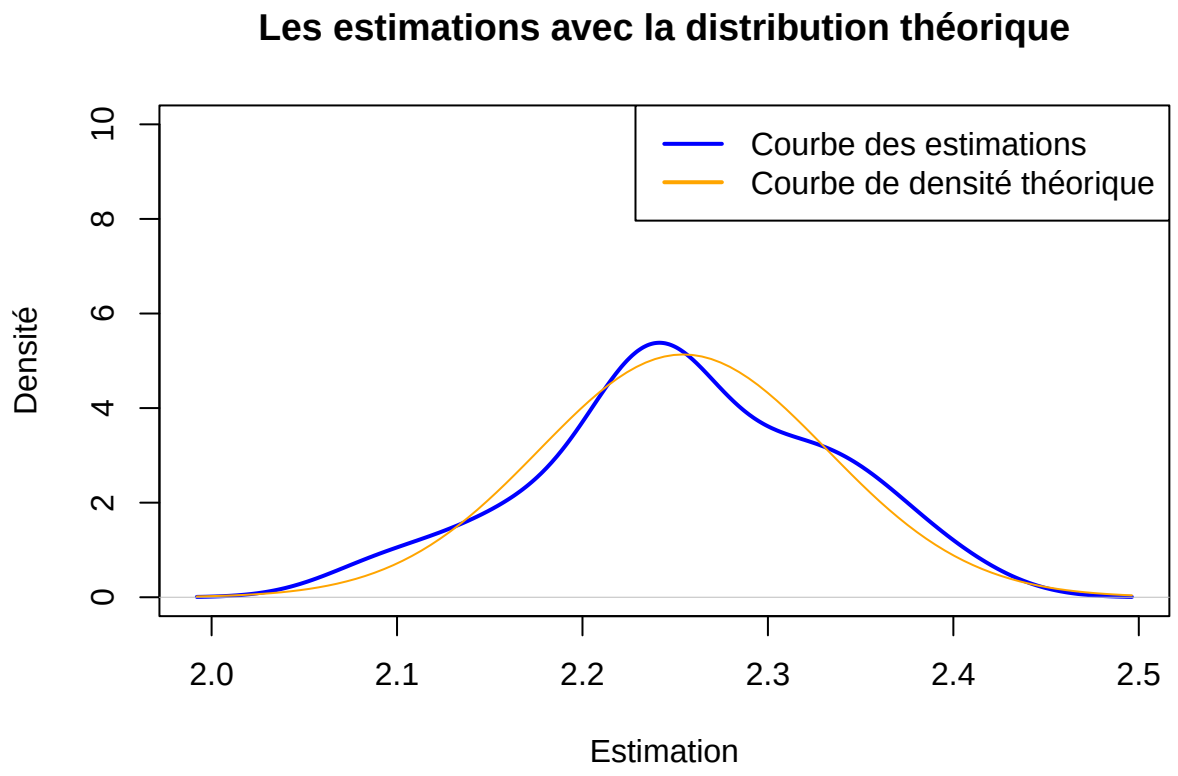
```
# Calculer la moyenne et l'écart-type de la distribution
# des estimations
mean_est <- mean(Estimateurs_taille500)
sd_est <- sd(Estimateurs_taille500)

density_estimation <- density(Estimateurs_taille500)

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

# Ajouter la courbe de densité théorique - TOUJOURS NORMAL
# DISTRIBUTION
curve(dnorm(x, mean = mean_est, sd = sd_est), col = "orange",
    lwd = 1, add = TRUE)

# Ajouter une légende
legend("topright", legend = c("Courbe des estimations", "Courbe de densité théorique"),
    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 population 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

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

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

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

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.

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

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

où:

- m est la moyenne de la population.

```
MoyenneEchantillon_SR <- function(data, n, variable) {
    Echantillon <- data[sample(1:nrow(data), n, replace = F),
        ]
    # calculer la moyenne sans tenir compte les NA.
    EstimMoyenne <- mean(Echantillon[[variable]], na.rm = TRUE)
    return(list(Moyenne = EstimMoyenne, Echantillon = Echantillon))
}
Moyenne_taille250 <- MoyenneEchantillon_SR(hdv2003, 250, "heures.tv")  #taille échantillon = 250
cat("La moyenne de l'échantillon sans remise de taille 250:  ",
    Moyenne_taille250$Moyenne)
```

```
## La moyenne de l'échantillon sans remise de taille 250:   2.3052
```

```
Moyenne_Echantillon_SR <- MoyenneEchantillon_SR(hdv2003, 500,
    "heures.tv")
cat("La moyenne de l'échantillon sans remise de taille 500:  ",
    Moyenne_Echantillon$Moyenne)  #taille échantillon = 500
```

```
## La moyenne de l'échantillon sans remise de taille 500:   2.2028
```

```
cat("La moyenne de la population: ", moyenne_population)
```

```
## La moyenne de la population:  2.246566
```

4D

On estime la variance de l'échantillon sans remise :

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

```
VarianceEchantillon_SR <- function(data, n, variable) {
    MoyenneEchantillon_SR <- MoyenneEchantillon_SR(data, n, variable)
    Moyenne <- MoyenneEchantillon_SR$Moyenne
    EcartCarree <- (MoyenneEchantillon_SR$Echantillon[[variable]] -
        Moyenne)^2
    var <- sum(EcartCarree, na.rm = TRUE)/(n)

    return(list(`Variance de l'échantillon par le formule` = var))
}
VarianceEchantillon_SR(hdv2003, 250, "heures.tv")  # taille échantillon = 250
```

```
## $`Variance de l'échantillon par le formule`
## [1] 3.063592
```

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

```
## $`Variance de l'échantillon par le formule`
## [1] 3.717722
```

```
cat("La vraie variance de la population: ", var(hdv2003$heures.tv,
    na.rm = TRUE))
```

```
## La vraie variance de la population:  3.153653
```

4E

On estime la variance de l'estimateur à partir de la variance de l'échantillon sans remise :

```
VariancePopulationEstim_SR <- Variance_Echantillon_SR  #taille échantillon = 500
```

On suppose que la variance de l'échantillon est égale à la variance de la population $\hat{\sigma}^2 = s^2$

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

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

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

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

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_EstimateurSR <- VariancePopulationEstim_SR[[1]] * ((2000 -
    500)/1999) * (1/500) * ((1 - (1/2000))/(1 - (1/500))))
```

```
## [1] 0.005587759
```

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.

On calcule SE:

```
(SE <- (sqrt(Variance_EstimateurSR)))
```

```
## [1] 0.07475131
```

```
Moyenne_Echantillon_SR$Moyenne
```

```
## [1] 2.34759
```

Intervalle de Confiance pour la moyenne:

```
limiteSup <- Moyenne_Echantillon_SR[[1]] + 1.96 * SE
limiteInf <- Moyenne_Echantillon_SR[[1]] - 1.96 * SE
(IntervalleConfiance <- list(`limite inférieure` = limiteInf,
    `limite supérieure` = limiteSup))
```

```
## $`limite inférieure`
## [1] 2.201078
##
## $`limite supérieure`
## [1] 2.494103
```

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

```
MoyenneEstimations_SR <- function(data, n, variable, nbSond) {
    MoyenneEstimation <- c()
    for (i in 1:nbSond) {
        Echantillon <- data[sample(1:nrow(data), n, replace = F),
            ]
        # calculer la moyenne sans tenir compte les NA.
        EstimMoyenne <- mean(Echantillon[[variable]], na.rm = TRUE)
        MoyenneEstimation <- c(MoyenneEstimation, EstimMoyenne)
    }
    return(MoyenneEstimation)
}
Estimation250 <- MoyenneEstimations_SR(hdv2003, 250, "heures.tv",
    1000)  #taille échantillon = 250
Estimateurs_taille500_SR <- MoyenneEstimations_SR(hdv2003, 500,
    "heures.tv", 1000)  #taille échantillon = 500
for (i in 1:10) {
    cat("L'estimateur SANS REMISE de la moyenne de la population: ",
        mean(MoyenneEstimations_SR(hdv2003, 500, "heures.tv",
            1000)), "\n")
}
```

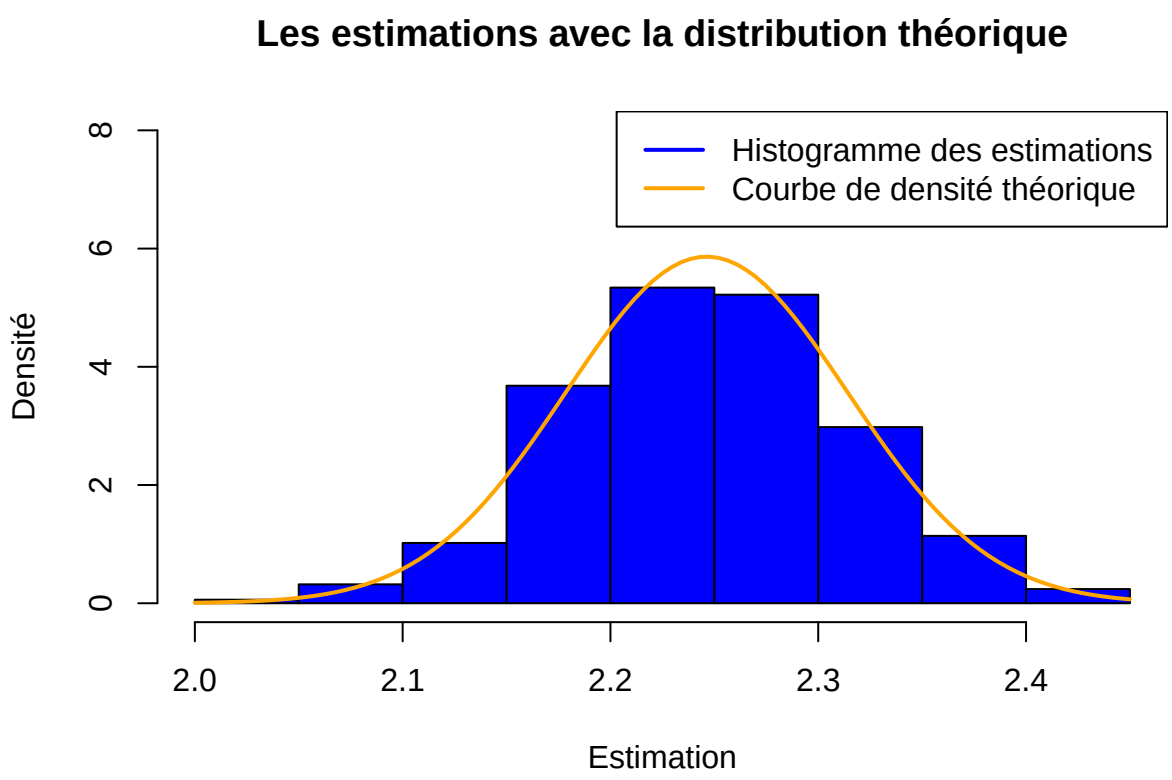
```
## L'estimateur SANS REMISE de la moyenne de la population:  2.24658
## L'estimateur SANS REMISE de la moyenne de la population:  2.2484
## L'estimateur SANS REMISE de la moyenne de la population:  2.246309
## L'estimateur SANS REMISE de la moyenne de la population:  2.246544
## L'estimateur SANS REMISE de la moyenne de la population:  2.247662
## L'estimateur SANS REMISE de la moyenne de la population:  2.243668
## L'estimateur SANS REMISE de la moyenne de la population:  2.245779
## L'estimateur SANS REMISE de la moyenne de la population:  2.248127
## L'estimateur SANS REMISE de la moyenne de la population:  2.243019
## L'estimateur SANS REMISE de la moyenne de la population:  2.245385
```

```
hist(Estimateurs_taille500_SR, col = "blue", main = "Les 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(Estimateurs_taille500_SR)
sd_est <- sd(Estimateurs_taille500_SR)

# 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 théorique"), col = c("blue", "orange"),
    lwd = 2, bg = "white")
```



La comparaison entre les méthodes SAS:

```
Variance_EstimationR <- Variance_Estimateur
Variance_EstimationSR <- Variance_EstimateurSR

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

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

VarianceEstimationSASavecRemise	VarianceEstimationSASsansRemise
0.0058948	0.0055878

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.