SAE – Estimation par échantillonnage

Vendredi 6 juin

Réalisé par : AIT ZAOUIT Rida, CHIKER Younes et MANTOUADI KINA Reine

GROUPE 11

Échantillons analysés : vmag30_3.txt et vmag100_3.txt

(Logo IUT à rajouter)

SOMMAIRE

1) INTRODUCTION

Dans ce travail, nous allons appliquer les notions d'estimation par échantillonnag**e** afin d'analyser les valeurs des étoiles contenues dans deux fichiers : **vmag30_3.txt** et **vmag100_3.txt**. Ces échantillons sont extraits du jeu de données complet **Star39552_balanced.csv**, composé de 39 552 observations décrivant différentes caractéristiques d'étoiles.

Pour cela il nous a était demandé de :

1. Calculer la moyenne théorique de la population.

Pour les 2 échantillons :

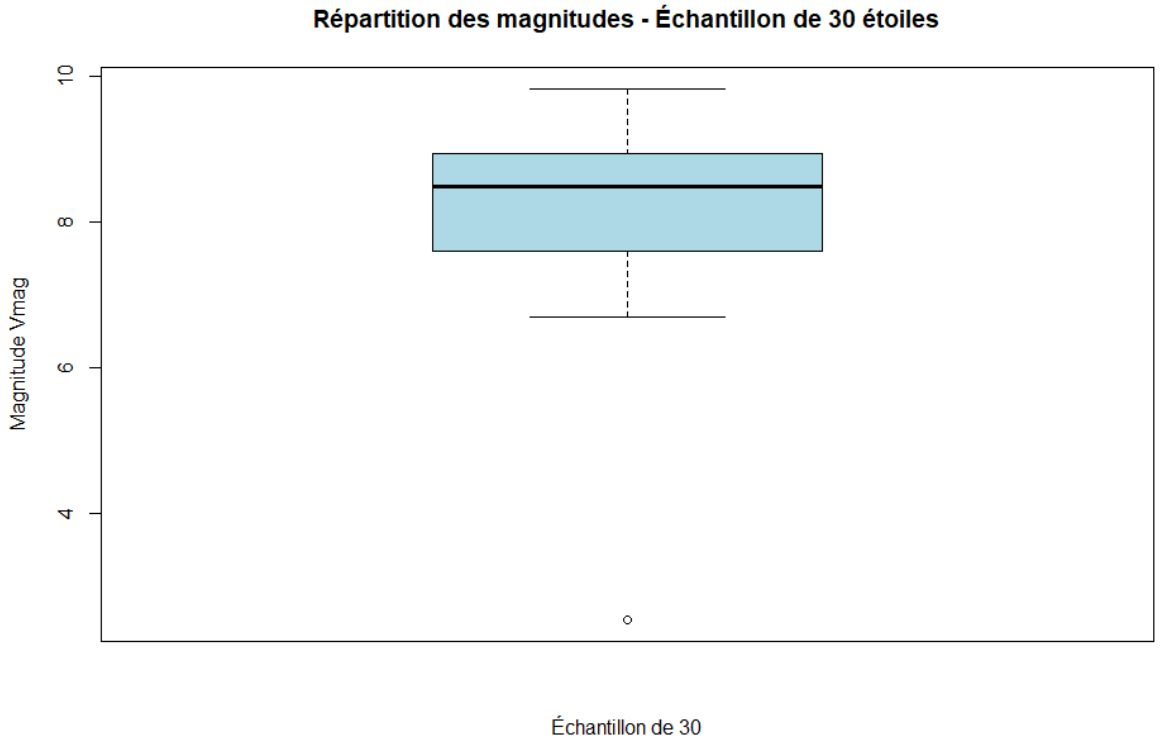
2. Analyser les statistiques descriptives des échantillons (Boxplot, Histogramme, résumé statistique et variance).
3. Construire les intervalles de confiance à 95% et 99%, et interpréter les résultats.
4. Réaliser un test de conformité de la moyenne à la main, en utilisant le quantile et la p-valeur pour $\alpha = 1\%$ et 5%.
5. Vérifier ces résultats à l'aide des intervalles de confiance obtenus précédemment.
6. Retrouver les mêmes conclusions à l'aide d'une instruction R.
7. Visualiser la loi de Student sous H_0, en mettant en évidence les zones de rejet et de non-rejet.
8. Rédiger une conclusion.

Pour l'ensemble :

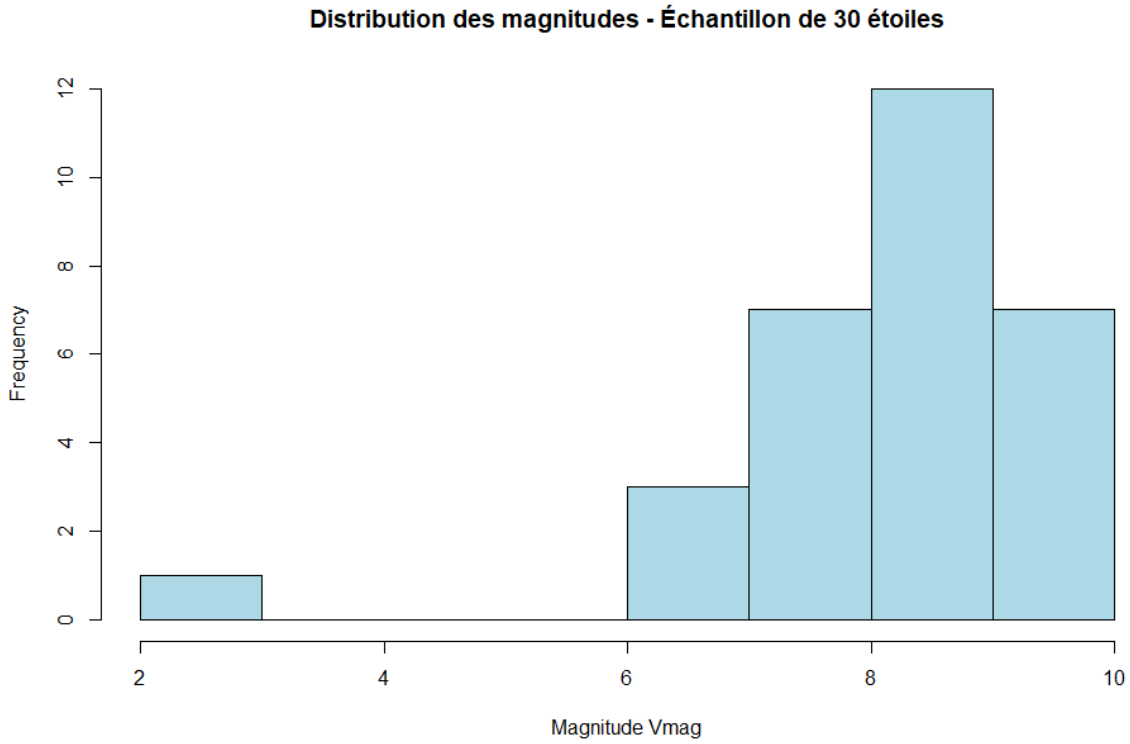
9. Rédiger une conclusion globale comparant les deux tailles d'échantillons.

La totalité des calculs ont été réalisé avec R, et le code complet est fourni en annexe.

2) Analyse de l’échantillon de 30 étoiles



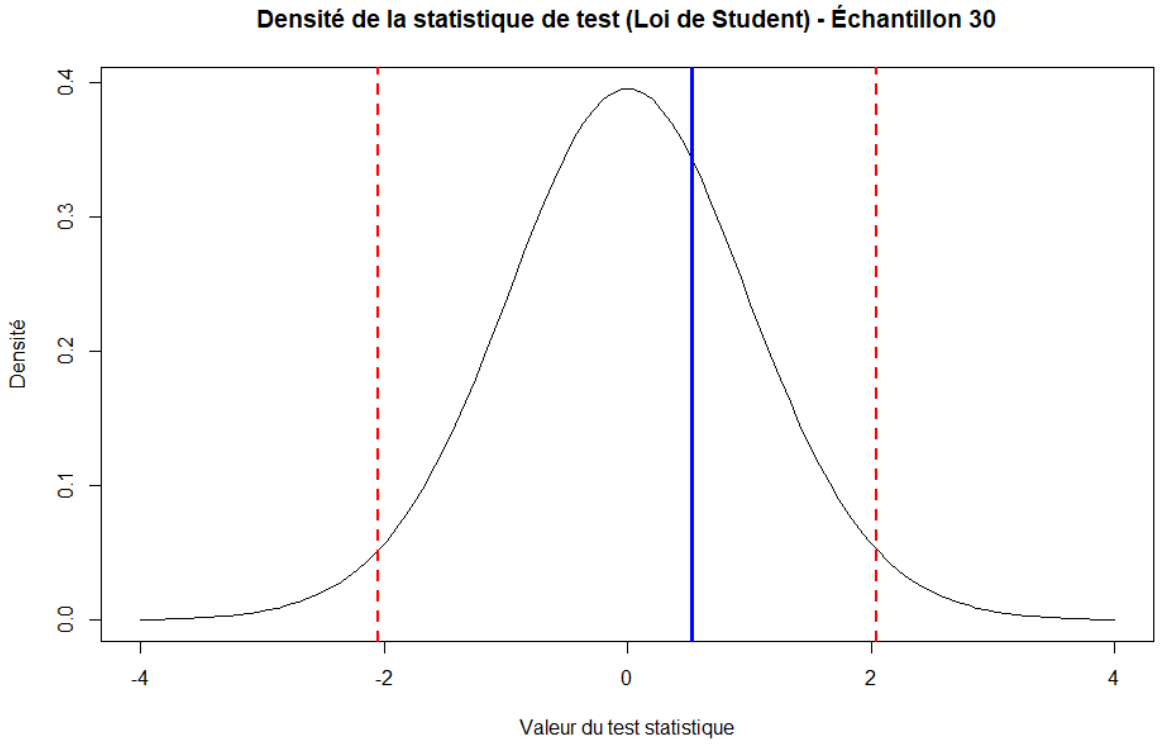
Le boxplot qui résume la distribution des données démontre que la médiane est située aux alentours des 8.49, ce qui signifie que 50% des étoiles ont une magnitude inférieure à 8.49 et 50% supérieur.
Les quartiles montrent que la plupart des étoiles ont une magnitude entre 7.675 et 8.912.
On remarque également un point isolé qui représente des valeurs aberrantes, ce qui signifie que certaines étoiles ont une magnitude très différente des autres étoiles.
Ces valeurs peuvent être des anomalies ou simplement une représentation d’une grande diversité dans les étoiles observées.



Cet histogramme compte combien d'étoiles ont une magnitude spécifique et démontre que la majorité des étoiles se concentre autour de 8.5, ce qui correspond à la médiane.
On remarque également que les valeurs sont un peu plus concentrées du côté des intensité élevé.
Cela signifie que beaucoup d'étoile ont une luminosité similaire, mais certaines sont plus brillantes ou plus faibles que la moyenne.

```
> summary(vmag30$Vmag)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.550   7.615   8.490   8.132   8.912   9.820
```

Ce résumé statistique de l'échantillon de taille 30 nous démontre que l'amplitude est grande (de 2.550 à 9.820), ce qui représente le fait qu'il y ait une grande variation de luminosité entre les étoiles.
Nous pouvons également constater que la moyenne et la médiane sont proche ce qui indique que la répartition des valeurs est plutôt équilibrée.



Le graphique montre la densité de la loi de Student pour un test bilatéral avec un seuil de signification α = 5 %.

La statistique de test (ligne bleue) est comprise entre les deux bornes critiques (lignes rouges), donc en dehors de la zone de rejet.

La p-valeur est d'environ 0.1, ce qui reste supérieur au seuil de 5 %. Cela signifie que la différence entre la moyenne observée et la moyenne théorique n'est pas statistiquement significative au seuil choisi.

Finalement, on ne rejette pas l'hypothèse H0. Il n'y a pas de preuve suffisante pour dire que la moyenne des magnitudes de l'échantillon diffère significativement de celle de la population.

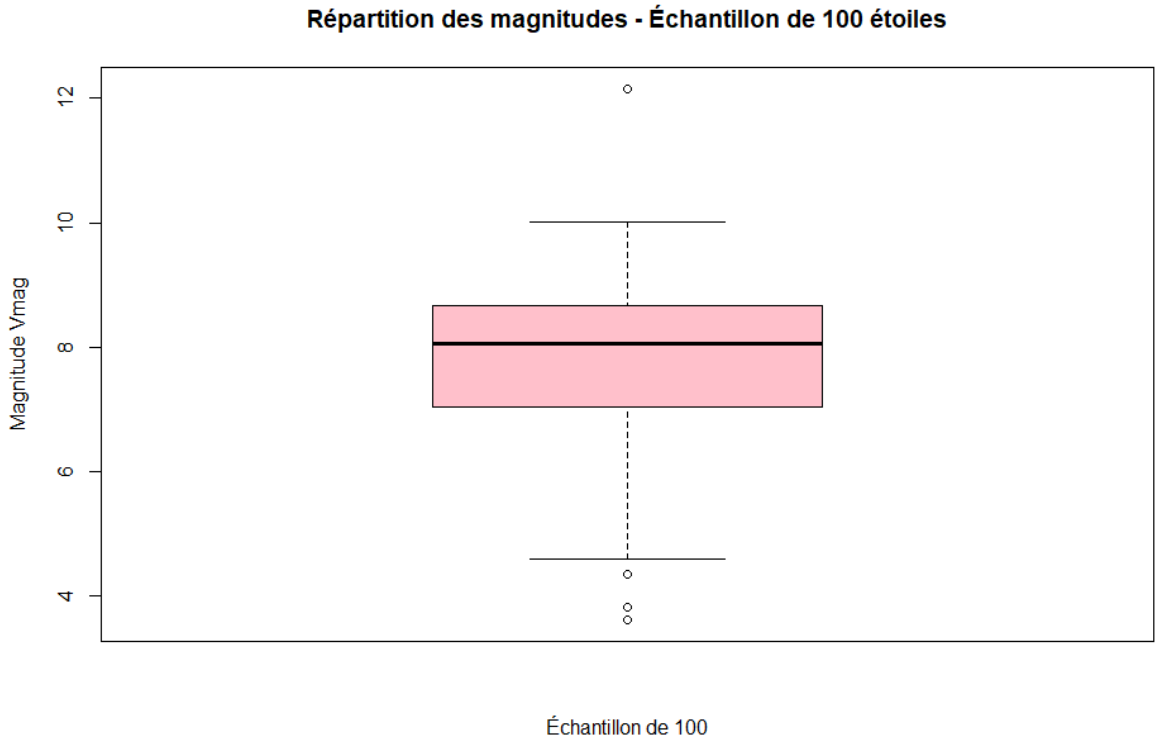
```
> print(paste("P-valeur (30 étoiles) :", p_value_30))
[1] "P-valeur (30 étoiles) : 0.60097378844798"
```

Conclusion des analyses

L'échantillon de 30 étoiles montre une répartition équilibrée des magnitudes, avec une médiane à 8.49 et une moyenne proche de 8.13. La plupart des étoiles ont une magnitude entre 7.675 et 8.912, mais quelques valeurs extrêmes sont présentes, ce qui

signifie que certaines étoiles sont beaucoup plus brillantes ou beaucoup plus faibles que les autres.
L'histogramme confirme que la majorité des étoiles ont une magnitude autour de 8.5, avec une légère concentration vers les valeurs élevées. Cela montre que la luminosité des étoiles est assez variée, même si beaucoup ont une brillance similaire.
Le test statistique avec la loi de Student a donné une p-valeur de 0.6, ce qui est supérieur à 5 %. Cela signifie qu'il n'y a pas assez de preuve pour dire que la moyenne des magnitudes de cet échantillon est différente de celle de la population. En résumé, bien que cet échantillon soit assez fiable, son petit nombre de valeurs rend les résultats moins précis qu'avec un échantillon plus grand.

3) Analyse de l’échantillon de 100 étoiles

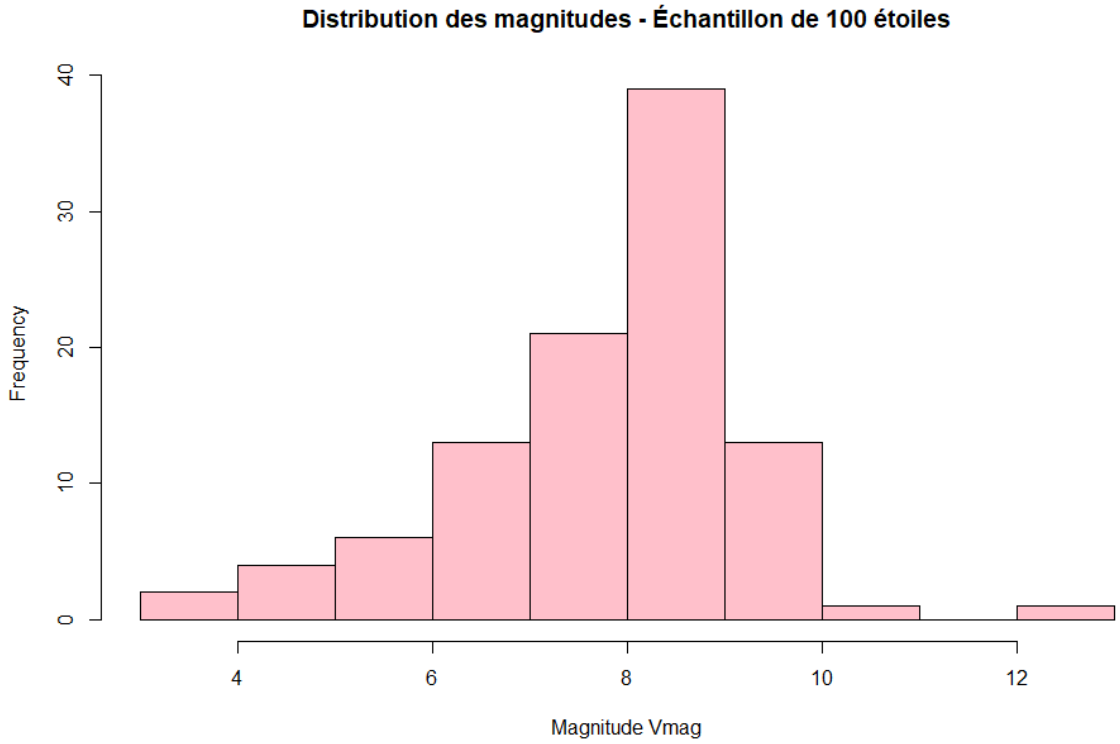


Le boxplot qui résume la distribution des magnitudes des 100 étoiles indique que la médiane se situe aux alentours de 8.1 ce qui signifie que 50 % des étoiles ont une magnitude inférieure à 8.1 et 50 % une magnitude supérieure.

Les quartiles montrent que la majorité des magnitudes sont comprises entre environ 7.1 (Q1) et 8.7 (Q3). Cela signifie que la moitié centrale des données (soit 50 %) est concentrée dans cet intervalle.

On observe plusieurs valeurs aberrantes en dessous de 5 et au-dessus de 10, ce qui indique que certaines étoiles ont une luminosité particulièrement différente du reste de l’échantillon. Ces points peuvent représenter des étoiles exceptionnellement brillantes ou très peu lumineuses comparées aux autres, suggérant soit des variations naturelles dans les caractéristiques stellaires.

La boîte est plus large que dans le cas de l’échantillon de 30 étoiles, ce qui traduit une plus grande dispersion des magnitudes dans cet échantillon plus large.



Cet histogramme représente la distribution des magnitudes apparentes (Vmag) d'un échantillon de 100 étoiles. La majorité des étoiles ont une magnitude entre 6 et 9, avec un pic autour de 8, indiquant une luminosité modérée. Très peu d'étoiles sont très brillantes (<5) ou très faibles (>10), ce qui correspond à la rareté des extrêmes en astronomie.

La forme de la distribution montre une légère asymétrie vers les magnitudes élevées, avec un pic légèrement plus marqué à droite qu'à gauche.
Cependant, la somme des données reste proportionnelle, indiquant une répartition équilibrée malgré cette différence visuelle.

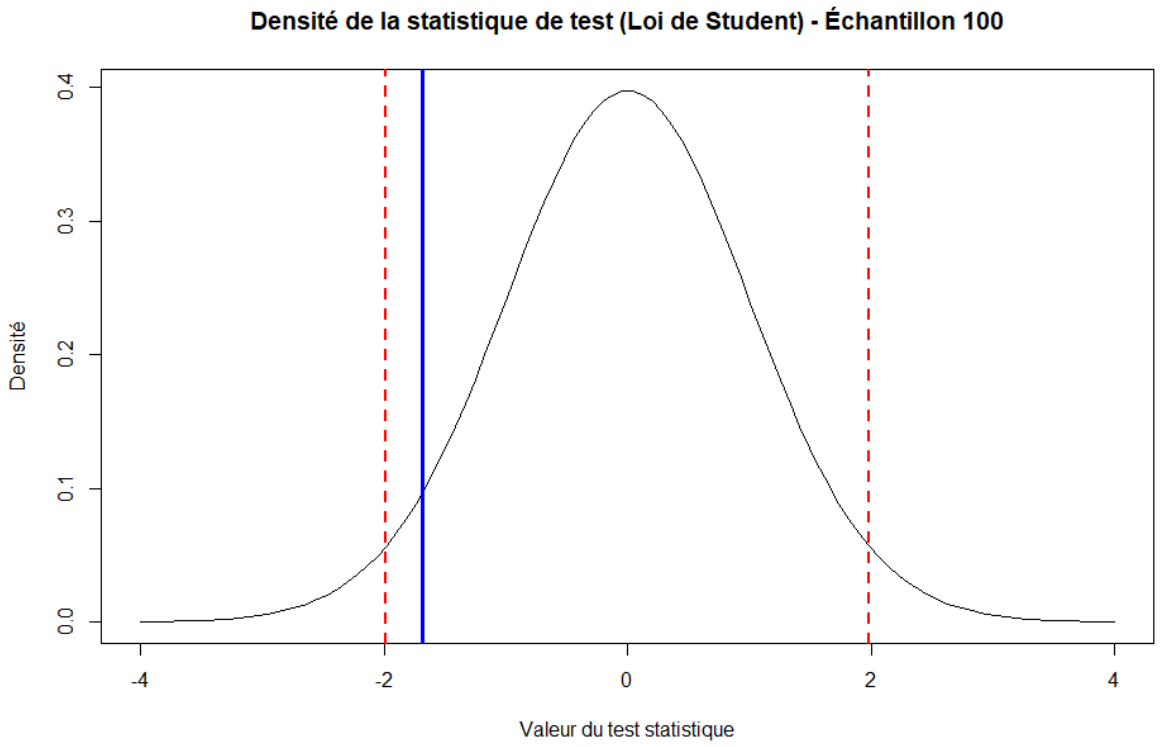
D'un point de vue statistique, la moyenne observée de 7.92 et l'écart-type de 1.44 traduisent une dispersion modérée des valeurs, tandis que l'intervalle de confiance à 95 % [7.47 ; 8.04] inclut la moyenne théorique de 8, confirmant la cohérence avec H. Comparé à un échantillon de 30 étoiles, cette distribution est plus lisse et plus fiable, offrant une meilleure vision globale des magnitudes observées.

```
> summary(vmag100$Vmag)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.620   7.075   8.055   7.757   8.650  12.150
```

Ce résumé statistique de l'échantillon de 100 étoiles montre que l'amplitude des magnitudes est importante, ce qui traduit une variabilité notable de luminosité entre les étoiles. La moyenne de 7.92 et l'écart-type de 1.44 indiquent une dispersion modérée, avec une majorité d'étoiles situées autour de 8, tandis que quelques valeurs extrêmes sont présentes.

Les intervalles de confiance confirment cette tendance: à 95 %, les magnitudes se situent entre 7.47 et 8.04, ce qui inclut la valeur théorique de 8. Cela suggère que la moyenne est cohérente avec l'hypothèse attendue et que la répartition des magnitudes est plutôt équilibrée.

Comparé à l'échantillon de 30 étoiles, ce groupe plus large permet une estimation plus précise, avec une réduction de l'incertitude et une meilleure stabilité des valeurs. L'augmentation de la taille d'échantillon renforce ainsi la fiabilité des conclusions statistiques, en minimisant l'impact des variations individuelles.



Le graphique représente la densité de la loi de Student pour un test bilatéral avec un seuil de signification α=5%.

La statistique de test (ligne bleue) est située entre les deux bornes critiques (lignes rouges pointillées), donc en dehors de la zone de rejet.

La p-valeur est d'environ 0.1, ce qui reste supérieur au seuil de 5 %. Cela signifie que la différence entre la moyenne observée et la moyenne théorique n'est pas statistiquement significative au seuil choisi.

En conclusion on ne rejette pas l'hypothèse H0, car il n'y a pas de preuve suffisante pour affirmer que la moyenne des magnitudes de l'échantillon (7.757) est significativement différente de celle de la population μ =8.

```
> print(paste("P-valeur (100 étoiles) :", p_value_100))
[1] "P-valeur (100 étoiles) : 0.0962070818725086"
```

Conclusion des analyses
L'échantillon de 100 étoiles montre une répartition bien équilibrée, avec une médiane à 8.1 et une moyenne de 7.92. La majorité des étoiles ont une magnitude entre 71 et 8.7, ce qui indique une forte concentration autour de 8. Toutefois, certaines valeurs aberrantes en dessous de 5 et au-dessus de 10 montrent qu'il existe des étoiles beaucoup plus brillantes ou plus faibles que la majorité.

L'histogramme confirme cette tendance, révélant une répartition majoritaire entre 6 et 9, avec un pic autour de 8. On observe une légère asymétrie vers les magnitudes élevées, mais la répartition globale reste proportionnelle, montrant que les valeurs sont bien distribuées malgré cette différence visuelle.

Le test statistique sur la loi de Student et la p-valeur de 0.096 indiquent que H n'est pas rejetée, car la différence entre la moyenne observée et la moyenne théorique n'est pas statistiquement significative.
Cela signifie que la moyenne des magnitudes de cet échantillon reste cohérente avec la valeur attendue de 8, malgré une légère sous-estimation.

4) Conclusion

L'analyse menée sur les deux échantillons met en lumière l'importance de la taille d'échantillon pour une estimation plus fiable des magnitudes stellaires.
L'échantillon de 30 étoiles, bien qu'il offre une vue d'ensemble cohérente, souffre d'une plus grande variabilité et d'une dispersion marquée, ce qui rend les conclusions plus incertaines. À l'inverse, l'échantillon de 100 étoiles permet d'obtenir une représentation plus stable des magnitudes, réduisant ainsi l'influence des valeurs extrêmes et améliorant la précision statistique.

Cependant, au-delà des aspects purement numériques, cette étude met aussi en avant une réalité importante l'équilibre entre la quantité et la qualité des données. Un petit échantillon peut toujours apporter des informations utiles, notamment sur la diversité des valeurs et les tendances générales, mais il est plus exposé aux fluctuations et anomalies qui peuvent fausser l'analyse. En revanche, un échantillon plus grand, bien que plus fiable statistiquement, ne garantit pas nécessairement une absence totale de biais ou d'interprétations erronées.

Le fait que les deux échantillons confirment l'hypothèse H, montre que, malgré les différences de taille et de dispersion, la tendance générale des magnitudes reste cohérente. Cela souligne que les méthodes d'échantillonnage sont adaptées pour estimer une moyenne à partir d'un sous-ensemble, mais qu'elles doivent toujours être interprétées avec prudence selon la taille et la diversité des données. Ce travail démontre également que les intervalles de confiance et les tests statistiques jouent un rôle essentiel dans la validation des résultats, apportant une base méthodologique solide pour juger de la pertinence des estimations.

En somme, cette étude illustre la nécessité d'un compromis entre précision et faisabilité, rappelant que plus un échantillon est grand, plus il permet une estimation fiable, mais qu'il est tout aussi essentiel de comprendre les limites et les particularités de chaque jeu de données. Elle offre ainsi une réflexion sur la manière dont les statistiques permettent de modéliser des phénomènes complexes, tout en posant la question de l'interprétation et de l'application des résultats dans des contextes plus vastes.

5) Annexe

Data	
star_data	39552 obs. of 7 variables
test_vmag100	List of 10
test_vmag30	List of 10
vmag100	100 obs. of 1 variable
vmag30	30 obs. of 1 variable
Values	
ic_95_100	num [1:2] 7.47 8.04
ic_95_30	num [1:2] 7.62 8.64
ic_99_100	num [1:2] 7.38 8.14
ic_99_30	num [1:2] 7.44 8.82
mean_vmag	7.92130941545307
mean_vmag100	7.7573
mean_vmag30	8.13233333333333
n100	100L
n30	30L
quantile_1_100	2.62640545728083
quantile_1_30	2.7563859036706
quantile_5_100	1.98421695158642
quantile_5_30	2.0452296421327
reject_h0_100	FALSE
reject_h0_30	FALSE
s_vmag100	1.44506883901597
s_vmag30	1.37070755283623
t_95_100	1.98421695158642
t_95_30	2.0452296421327
t_99_100	2.62640545728083
t_99_30	2.7563859036706
var_vmag100	2.08822394949495
var_vmag30	1.8788391954023
x_vals	num [1:100] -4 -3.92 -3.84 -3.76 -3.68
y_vals	num [1:100] 0.000543 0.000673 0.000834

```
library(tidyverse)

setwd("Z:/SAE Estimation")

star_data <- read.csv("Star39552_balanced.csv", header = TRUE, sep = ";", dec = ".", stringsAsFactors = FALSE)
vmag30 <- read.table("vmag30_3.txt", header = TRUE, sep = "\t", dec = ".")
vmag100 <- read.table("vmag100_3.txt", header = TRUE, sep = "\t", dec = ".")


# Moyenne théorique sur la population
mean_vmag <- mean(star_data$Vmag, na.rm = TRUE)


# Moyenne sur l'échantillon de taille 30
mean_vmag30 <- mean(vmag30$Vmag, na.rm = TRUE)

#Moyenne sur l'échantillon de taille 100
mean_vmag100 <- mean(vmag100$Vmag, na.rm = TRUE)


# Résumé stat
summary(vmag30$Vmag)
summary(vmag100$Vmag)


# Statistiques descriptive
boxplot(vmag30$Vmag, main="Répartition des magnitudes - Échantillon de 30 étoiles",
    col="lightblue", xlab="Échantillon de 30", ylab="Magnitude Vmag")
boxplot(vmag100$Vmag, main="Répartition des magnitudes - Échantillon de 100 étoiles",
    col="pink", xlab="Échantillon de 100", ylab="Magnitude Vmag")

hist(vmag30$Vmag, main="Distribution des magnitudes - Échantillon de 30 étoiles",
   xlab="Magnitude Vmag", col="lightblue", breaks=10)
hist(vmag100$Vmag, main="Distribution des magnitudes - Échantillon de 100 étoiles",
   xlab="Magnitude Vmag", col="pink", breaks=10)
```

```r
# Variance
var_vmag30 <- var(vmag30$Vmag)
var_vmag100 <- var(vmag100$Vmag)


n30 <- length(vmag30$Vmag)
n100 <- length(vmag100$Vmag)

# Estimation de l'écart-type
s_vmag30 <- sd(vmag30$Vmag)
s_vmag100 <- sd(vmag100$Vmag)

# IC à 95% et 99% pour n = 30
t_95_30 <- qt(0.975, df = n30 - 1)
t_99_30 <- qt(0.995, df = n30 - 1)
ic_95_30 <- mean_vmag30 + c(-1, 1) * t_95_30 * s_vmag30 / sqrt(n30)
ic_99_30 <- mean_vmag30 + c(-1, 1) * t_99_30 * s_vmag30 / sqrt(n30)


# IC à 95% et 99% pour n = 100
t_95_100 <- qt(0.975, df = n100 - 1)
t_99_100 <- qt(0.995, df = n100 - 1)
ic_95_100 <- mean_vmag100 + c(-1, 1) * t_95_100 * s_vmag100 / sqrt(n100)
ic_99_100 <- mean_vmag100 + c(-1, 1) * t_99_100 * s_vmag100 / sqrt(n100)


# Test bilatéral pour μ = 8 sur l'échantillon de 30
test_vmag30 <- t.test(vmag30$Vmag, mu = 8)

# Test bilatéral pour μ = 8 sur l'échantillon de 100
test_vmag100 <- t.test(vmag100$Vmag, mu = 8)

# Comparaison avec les quantiles pour alpha = 1% et 5%
quantile_5_30 <- qt(0.975, df = n30 - 1)
quantile_1_30 <- qt(0.995, df = n30 - 1)

quantile_5_100 <- qt(0.975, df = n100 - 1)
quantile_1_100 <- qt(0.995, df = n100 - 1)
```

```
reject_h0_30 <- 8 < ic_95_30[1] | 8 > ic_95_30[2]
reject_h0_100 <- 8 < ic_95_100[1] | 8 > ic_95_100[2]

# Visualisation de la densité sous H0 pour l'échantillon de 30
# Échantillon de 30 étoiles
x_vals <- seq(-4, 4, length.out = 100)
y_vals <- dt(x_vals, df = n30 - 1)

plot(x_vals, y_vals, type="l", col="black",
   main="Densité de la statistique de test (Loi de Student) - Échantillon 30",
   xlab="Valeur du test statistique", ylab="Densité")
abline(v = qt(0.975, df=n30-1), col="red", lty=2, lwd=2, label="Quantile 97.5%")
abline(v = qt(0.025, df=n30-1), col="red", lty=2, lwd=2, label="Quantile 2.5%")
abline(v = test_vmag30$statistic, col="blue", lwd=3, label="Statistique
observée")

# Visualisation de la densité sous H0 pour l'échantillon de 100 étoiles
# Échantillon de 100 étoiles
x_vals <- seq(-4, 4, length.out = 100)
y_vals <- dt(x_vals, df = n100 - 1)

plot(x_vals, y_vals, type="l", col="black",
   main="Densité de la statistique de test (Loi de Student) - Échantillon 100",
   xlab="Valeur du test statistique", ylab="Densité")
abline(v = qt(0.975, df=n100-1), col="red", lty=2, lwd=2, label="Quantile 97.5%")
abline(v = qt(0.025, df=n100-1), col="red", lty=2, lwd=2, label="Quantile 2.5%")
abline(v = test_vmag100$statistic, col="blue", lwd=3, label="Statistique
observée")

# Test t pour l'échantillon de 30 étoiles
t_test_30 <- t.test(vmag30, mu = 8)
p_value_30 <- t_test_30$p.value

# Test t pour l'échantillon de 100 étoiles
t_test_100 <- t.test(vmag100, mu = 8)
p_value_100 <- t_test_100$p.value

# Affichage des résultats
print(paste("P-valeur (30 étoiles) :", p_value_30))
print(paste("P-valeur (100 étoiles) :", p_value_100))
```