COLAB1 – Mehdi BENAYED Groupe 14

SAE : Echantillonnage

07/06/2023

Université Paris Cité

Table des matières

Introduction

La base de données « Star39552balanced.csv » est composé de 39 552 étoiles et étudie 7 variables :

- Vmag : magnitude apparente de l'étoile.
- Plx : distance entre l'étoile et la terre en années lumières,
- ePlx : écart-type de la variable Plx,
- B – V : indice de couleur de l'étoile,
- SpType : type spectral de l'étoile,
- Amag : magnitude absolue de l'étoile obtenue à l'aide de l'équation

Dans ce rapport nous allons nous baser sur la variable Plx, donc la variable mesurant les distances étoile – Terre. Nous avons 2 bases de données à disposition, une avec 30 étoiles et une autre avec 100 étoiles. L'objectif est de construire des tests et des intervalles de confiance de autour de μ et de vérifier la fiabilité d'une moyenne théorique μ_0.

1) Etude sur un échantillon de 30 observations

Dans cette partie 1ère nous allons exploiter le fichier texte « plx30_6.txt », on étudiera donc la variable plx qui contient 30 observations, soient les distances des étoiles par rapport à la terre compté en année lumière. Dans cette étude nous supposons la moyenne µ connue : la moyenne des distances en années lumières. D'après le fichier « Star39552balanced.csv » et le programme R ci-dessous :

```
df_star <- read.csv("Star39552_balanced.csv")
mean(df_star$Plx)
```

```
[1] 7.117378
```

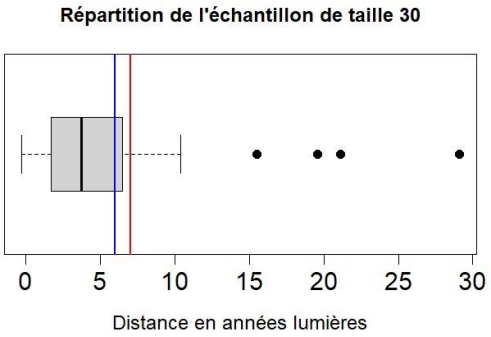
Nous obtenons une distance moyenne de µ égale à ≃7.12 années lumières.

1.a) Observation descriptive d'un échantillon de 30 observations.

Le code suivant nous permet d'obtenir une vue d'ensemble de nos données.

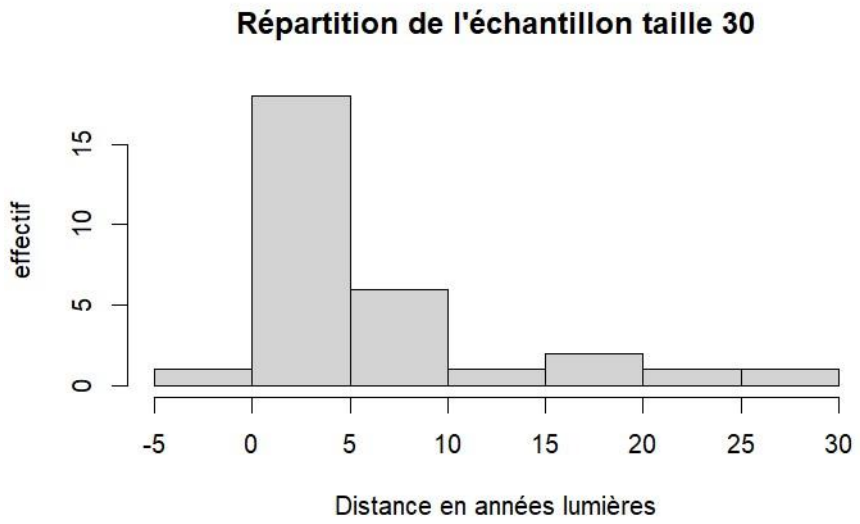
```
summary(df_star30)

boxplot(df_star30$Plx,pch=16,cex=1.2,bty="l",cex.lab=1.2,cex.axis=1.5,main = "Répartition de l'échantillon de taille 30",xlab="Distance en années lumières",horizontal=TRUE)
abline(v=moy30,lwd=2,col="blue") #estimation
abline(v=mutheo,lwd=2,col="red") #théorique

hist(df_star30$Plx, main = "Répartition de l'échantillon taille 30", xlab = "Distance en années lumières", ylab = "effectif")
```

Voici les sorties :




```
> summary(df_star30)
      Plx
Min.   :-0.280
1st Qu.: 1.785
Median : 3.720
Mean   : 5.996
3rd Qu.: 6.365
Max.   :29.080
```

Cette boîte à moustache représente la distribution de l'échantillon de taille 30, nous constatons graphiquement que le 1er quartile est égal à 1.785 années lumières, la médiane à 3.77 années lumières et le 3ème quartile équivaux à 6.365 années lumières. La droite verticale bleue représente la moyenne observée qui est de 5.996 années lumières et la rouge celle de la moyenne théorique fixé à 7 années lumières. Nous pouvons remarque que la moyenne théorique est légèrement plus élevée que la moyenne observée.



D'après cet histogramme qui illustre la distribution de notre échantillon, nous pouvons voir qu'une classe se démarque des autres, celle des étoiles avec une distance en année lumière compris entre]0 ; 5], qui est la classe modale de notre échantillon.

1.b) Construction d'intervalles de confiances de μ à 95% et 99%

Soit une étude mesurant les distances étoile – Terre compté en année lumière, nous avons comme informations à disposition :

X : la distance entre l'étoile et la Terre N : population

d'étoiles μ : la distance moyenne entre les étoiles et la

Terre σ^2 : la variance des distances entre les étoiles et

la Terre

Xn : l'estimateur sans biais de μ

S : l'estimateur sans biais de σ^2

Et notre échantillon de taille 30 (x1, …, x30) une réalisation de (X1, …,X30), obtenu à partir d'un tirage aléatoire simple.

- **Intervalle de confiance de μ à 95% (α=5%) :**

Avec nos observations nous somme dans le cas d'une loi quelconque avec un grand échantillon, c'està-dire échantillon ≥ 30. Alors d'après le Théorème central limite nous pouvons approximer Xn par une loi normale centré réduite N(0,1).

Alors on sait que $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

On sait aussi déterminer Cα d'après la fonction qnorm de tel sort que

$P(-C\alpha < Z < C\alpha) = 1 - \alpha = 0.95 \Leftrightarrow P(Z < C\alpha) = 1 - \alpha/2 = 0.975$, on applique notre fonction :

```
alpha005 <- 0.05
Calpha196 <- round(qnorm(1-alpha005/2, 0,1),2)
Calpha196
```

```
> Calpha196
[1] 1.96
```

Nous obtenons une valeur $C_{0.05} = 1.96$

Donc $P(-1.96 < \mu < 1.96) = 1 - \alpha = 0.95$

Or $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

Ce qui nous donne $P(-1.96 < \sqrt{n}*(Xn - \mu)/ \sqrt{S} < 1.96) = 0.95$

Soit $P(Xn-1.96*(\sqrt{S}/\sqrt{n}) < \mu < Xn+1.96*(\sqrt{S}/\sqrt{n})) = 0.95$

Pour trouver une estimation par intervalle on utilisera un programme R :

```
risque_005 <- paste("[",round(xn-Calpha196*(s/sqrt(n)), 3), ";", round(xn+Calpha196*(s/sqrt(n)),3), "]")
risque_005
```

Ce qui nous donne :

```
"[ 3.542 ; 8.449 ]"
```

Voici notre intervalle de confiance à 95% : [3.542 ; 8.449]. Cet intervalle nous permettra de dire si dans 95% des cas une valeur comme un μ théorique s'y trouve.

- Intervalle de confiance de μ à 99% (α=1%) :

On sait que $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

On sait aussi déterminer $C_{0.01}$ d'après la fonction qnorm de tel sort que

$P(-C\alpha < Z < C\alpha) = 1 - \alpha = 0.99 \Leftrightarrow P(Z < C\alpha) = 1 - \alpha/2 = 0.995$, on applique notre fonction :

```
> Calpha258
[1] 2.58
```

Nous obtenons une valeur $C_{0.01} = 2.58$

Donc $P(-2.58 < \mu < 2.58) = 1 - \alpha = 0.99$

Or $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

Ce qui nous donne $P(-2.58 < \sqrt{n}*(Xn - \mu)/ \sqrt{S} < 2.58) = 0.99$

Soit $P(Xn-2.58*(\sqrt{S}/\sqrt{n}) < \mu < Xn+2.58*(\sqrt{S}/\sqrt{n})) = 0.99$

Pour déterminer une estimation par intervalle de μ on exécutera un programme R utilisé plus haut:

```
> risque_001
[1] "[ 2.766 ; 9.225 ]"
```

Voici notre intervalle de confiance à 99% : [2.766 ; 9.225]. Cet intervalle nous permettra de dire si dans 99% des cas une valeur comme un μ théorique s'y trouve.

1.c) Test de comparaison de moyenne

Pour notre test, il nous faut déterminer :

- Les hypothèses : H0 et H1
- Une statistique de test Tn
- Une règle de décision avec une région de rejet

Quelles sont les hypothèses H0 et H1 ?

H0 : $\mu = 7$ H1 : $\mu \neq 7$

L'hypothèse H0, on suppose que la distance moyenne étoiles – Terre est de 7 années lumières.

L'hypothèse H1, on suppose que la distance moyenne étoiles – Terre est différente de 7 années lumières.

On souhaite déterminer une statistique de test Tn

On pose

$Tn = \sqrt{n}*(Xn - \mu)/S$

Nous sommes dans un cas avec une loi quelconque et de variance inconnue. Notre échantillon est assez grand ($n=30 \geq 30$) pour approximer une loi normale centrée réduite par le Théorème central limite $Tn = \sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

On calcule une estimation de Tn grâce à notre programme R :

```
tn <- round(abs(sqrt(n)*((xn-mutheo)/s)), 3)
tn
```

```
> tn
[1] 0.802
```

D'après la sortie R, une estimation de la statistique de test Tn est : tn = 0.802

Maintenant nous voulons établir une règle de décision

Pour cela il nous faut construire une région de rejet, soit R : $\{|Tn| > C\alpha\}$ Si

$|Tn| > C\alpha$, c'est-à-dire si $Tn < -C\alpha$ ou $Tn > C\alpha$, on rejette l'hypothèse H0.

Si $|Tn| < C\alpha$, c'est-à-dire $-C\alpha < Tn < C\alpha$, on ne rejette pas l'hypothèse H0.

Afin de déterminer une région de rejet, nous devons trouver une valeur $C\alpha$, cherchons $C_{0.05}$:

```
alpha005 <- 0.05
Calpha196 <- round(qnorm(1-alpha005/2, 0,1),2)
Calpha196
```

```
> Calpha196
[1] 1.96
```

On remarque qu'à partir de notre échantillon observé, nous avons une réalisation de Tn :

tn = 0.802

Nous avons tn = 0.802 < 1.96, nous pouvons conclure que l'hypothèse H0 n'est pas rejeté. La distance moyenne des étoiles en année lumière par rapport à la Terre n'est pas significativement différente de 7.

- Test de niveau α à 1% :

Pour un test de niveau α = 1% nous devons établir une règle de décision en construisant une région de rejet.

Afin de déterminer une région de rejet, nous devons trouver une valeur Cα, on réutilise notre programme R, elle nous renvoie :

```
> Calpha258
[1] 2.58
```

Nous avons une valeur pour $C_{0.01}$ = 2.58, ce qui nous donne :

|tn| = 0.802 < 2.58

On voit que |tn| est inférieur à 2.58, nous pouvons conclure que l'hypothèse H0 n'est pas rejeté. La distance moyenne des étoiles par rapport à la Terre n'est pas significativement différente de 7.

Après avoir déterminé une statistique de test et fait des tests de niveau α = 1% et α = 5%. On souhaite à présent déterminer une p-valeur. La p-valeur représente le plus petit risque α pour que le test de niveau α soit significatif.

p-valeur = P_{H0}([Tn| > 0.802) = 2(1-P(Tn<0.802))

Soit

```
pvaleur <- 2*(1-pnorm(tn))
```

```
> pvaleur
[1] 0.422553
>
```

Nous obtenons un arrondi de la p-valeur égale à 0.422, on sait que plus la p-valeur est proche de 0 plus le test sera significatif et :

- Si α < p-valeur, le test de niveau α ne sera pas significatif
- Si α ⩾ p-valeur, le test de niveau α sera significatif

Dans notre cas avec un test de niveau α = 5%, nous avons α = 0.05 < 0.422 = p-valeur, le test de niveau 5% n'est pas significatif.

Pour un test de niveau 1%, nous avons α = 0.01 < 0.422 = p-valeur, le test de niveau 1% n'est pas significatif.

1.d) Test avec les intervalles de confiances

Nous avons fait des estimations d'intervalles de confiance à 95% et 99%, nous avons déterminé à l'étape précédente que les tests n'étaient pas significatifs, on souhaiterait vérifier nos propos à travers les intervalles de confiances.

- Pour l'intervalle de confiance à 95% de μ :

```
ifelse(xn-Calpha196*(s/sqrt(n))<mutheo & mutheo<xn+Calpha196*(s/sqrt(n)),
       paste("Dans 95% des cas 7 se trouve dans l'intervalle",risque_005),
       "La 7 ne fait pas partie de l'intervalle")
```

```
[1] "Dans 95% des cas 7 se trouve dans l'intervalle [ 3.542 ; 8.449 ]"
>
```

On ne rejette pas l'hypothèse $\mu = 7$ car $7 \in [3.542 ; 8.449]$, cette conclusion confirme le résultat de notre test avec $\alpha = 5\%$.

- Pour l'intervalle de confiance à 99% de μ :

```
[1] "Dans 99% des cas 7 se trouve dans l'intervalle [ 2.766 ; 9.225 ]"
>
```

On ne rejette pas l'hypothèse $\mu = 7$ car $7 \in [2.766 ; 9.225]$, cette conclusion confirme le résultat de notre $\alpha = 1\%$.

1.e) Fonction R : t.test

Dans cette partie, nous cherchons une fonction R capable de créer un intervalle de confiance, réaliser un test et déterminer une p-valeur de notre échantillon. Il existe une fonction native sur R qui permet de faire cela, c'est la fonction t.test.

- Pour un risque $\alpha = 5\%$ nous retrouvons à quelques différences près les valeurs calculer à la main.

```
t.test(df_star30, mu=mutheo, conf.level=1-0.05)
```

```
data:  df_star30
t = -0.80233, df = 29, p-value = 0.4289
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 3.435503 8.555830
sample estimates:
mean of x
 5.995667
```

Nous avions trouvé une estimation de l'intervalle de confiance : [3.542 ; 8.449], une statistique de test tn = 0.802 et une p-valeur = 0.422.

De même pour un risque $\alpha = 1\%$.

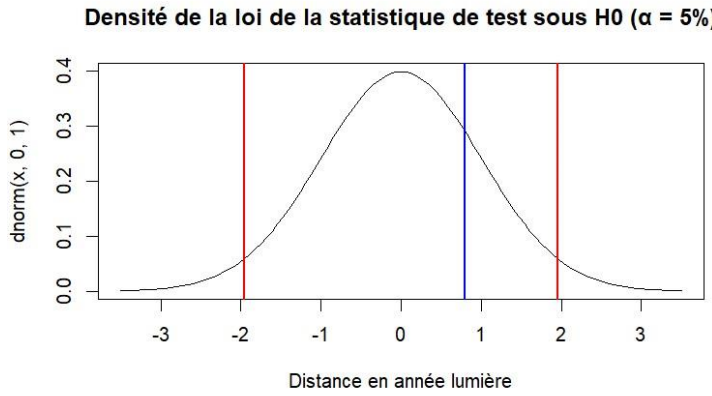
```
data:  df_star30
t = -0.80233, df = 29, p-value = 0.4289
alternative hypothesis: true mean is not equal to 7
99 percent confidence interval:
 2.545297 9.446037
sample estimates:
mean of x
 5.995667
```

Nous avions trouvé une estimation de l'intervalle de confiance : [2.766 ; 9.225], une statistique de test tn = 0.802 et une p-valeur = 0.422.

1.f) Illustration graphique des tests

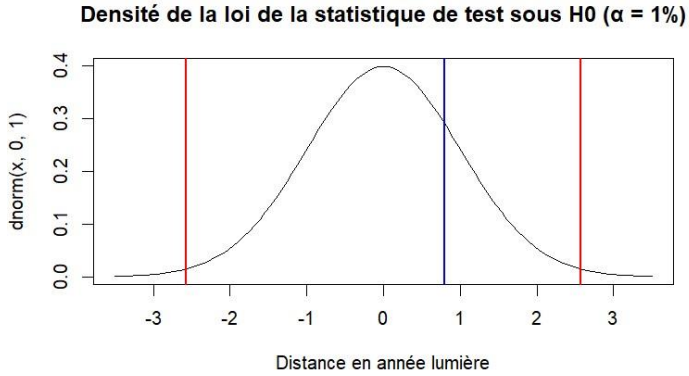
- Illustration du test sous H0 avec α = 5% :

```
curve(dnorm(x,0,1),col="black",from=-3.5,to=3.5,
      main ="Densité de la loi de la statistique de test sous H0 (α = 5%)",
      xlab = "Distance en année lumière")
abline(v=tn,lwd=2,col="blue")
abline(v=qnorm(0.025),lwd=2,col="red")
abline(v=qnorm(0.975),lwd=2,col="red")
```



La droite bleue représente la valeur de la statistique de test « tn » et en rouge les quantiles théoriques délimitant la zone de non rejet de H_0 et la zone de rejet H_0. La statistique de test se trouve dans la zone de non rejet de H_0, montrant que le test n'est pas significatif.

- Illustration du test sous H0 avec α = 1% :



Nous pouvons voir que la statistique de test se trouve dans la zone de non rejet de H_0, montrant bien que le test n'est pas significatif.

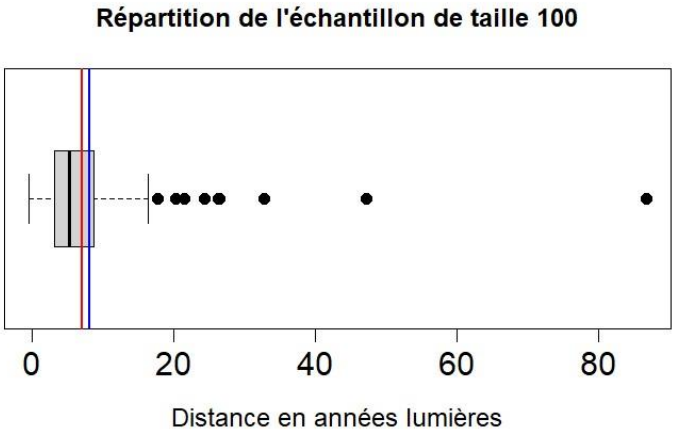
1.g) Conclusion sur l'échantillon 30 observation

Dans cette 1ère partie de notre étude, basée sur la variable plx avec 30 observations, c'est-à-dire la variable mesurant la distance entre les étoiles et la Terre. Nous avons fait des tests avec différents pourcentages de risque α = 1% et α = 5%, qui se sont avéré non significatifs, c'est-à-dire que la moyenne théorique μ_0 = 7 n'est pas significativement différente de la moyenne μ de notre échantillon pour nos 2 tests de risque α égale à 1% et 5%. La p-valeur obtenue lors de l'étude de cet échantillon étant plus élevé que les seuils de nos tests à 1% et à 5%, cela ne nous permet pas de rejeter l'hypothèse nulle.

2) Etude sur un échantillon de 100 observations

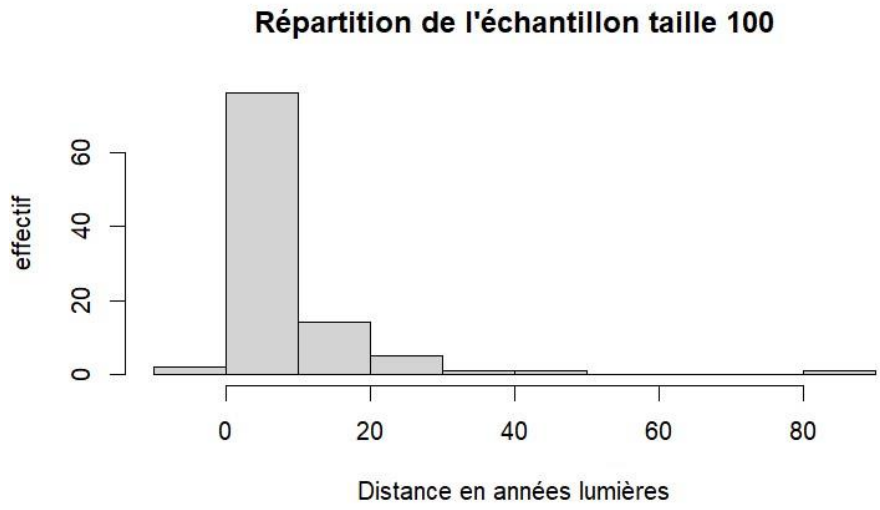
Dans cette 2nd partie nous allons exploiter le fichier texte « plx100.txt », on étudiera donc la variable plx qui contient 100 observations, soient les distances des étoiles par rapport à la terre compté en année lumière.

2.a) Observation descriptive d'un échantillon de 100 observations



	Plx
Min.	:-0.490
1st Qu.	: 3.178
Median	: 5.215
Mean	: 8.145
3rd Qu.	: 8.688
Max.	:86.760

Cette boîte à moustache représente la distribution de l'échantillon de taille 100, nous constatons graphiquement que le 1er quartile est égal à 3.178 années lumières, la médiane à 5.215 années lumières et le 3ème quartile équivaux à 8.688 années lumières. La droite verticale bleue représente la moyenne observée qui est de 8.145 années lumières et la rouge celle de la moyenne théorique fixé à 7 années lumières. On remarque que la moyenne théorique est inférieure à la moyenne observée, contrairement à la première étude sur les 30 observations.



D'après cet histogramme qui illustre la distribution de notre échantillon, nous pouvons voir que la classe modale est celle avec une distance en années lumières compris entre]0 ; 10].

2.b) Construction d'intervalles de confiances de μ à 95% et 99%

Soit une étude mesurant les distances étoile – Terre compté en année lumière, nous avons comme informations à disposition :

X : la distance entre l'étoile et la Terre N : population

d'étoile μ : la distance moyenne entre les étoiles et la

Terre σ^2 : la variance des distances entre les étoiles et

la Terre

Xn : l'estimateur sans biais de μ

S : l'estimateur sans biais de σ^2

Et notre échantillon de taille 100 (x1, …, x100) une réalisation de (X1, …,X100), obtenu à partir d'un tirage aléatoire simple.

- **Intervalle de confiance de μ à 95% (α=5%):**

Avec nos observations nous somme dans le cas d'une loi quelconque avec un grand échantillon, c'està-dire échantillon ≥ 100. Alors d'après le Théorème central limite nous pouvons approximer Xn par une loi normale centré réduite N(0,1).

Alors on sait que $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

On sait aussi déterminer $C_{0.05}$

$P(-C_{0.05} < Z < C\alpha_{0.05}) = 1 - \alpha = 0.95 \Leftrightarrow P(Z < C_{0.05}) = 1 - \alpha/2 = 0.975$, ce que notre fonction R renvoie:

```
> Calpha196
[1] 1.96
```

Nous obtenons une valeur $C_{0.05} = 1.96$

Donc $P(-1.96 < \mu < 1.96) = 1 - \alpha = 0.95$

Or $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

Ce qui nous donne $P(-1.96 < \sqrt{n}*(Xn - \mu)/ \sqrt{S} < 1.96) = 0.95$

Soit $P(Xn-1.96*(\sqrt{S}/\sqrt{n}) < \mu < Xn+1.96*(\sqrt{S}/\sqrt{n})) = 0.95$

Le programme R nous renvoie :

```
[1] "[ 6.03 ; 10.259 ]"
```

Voici notre intervalle de confiance à 95% : [6.03 ; 10.259]. Cet intervalle nous permettra de dire si dans 95% des cas une valeur comme μ théorique s'y trouve.

- **Intervalle de confiance de μ à 99% (α=1%) :**

On sait que $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

On sait aussi déterminer $C_{0.01}$ d'après la fonction qnorm de tel sort que

$P(-C\alpha < Z < C\alpha) = 1 - \alpha = 0.99 \Leftrightarrow P(Z < C\alpha) = 1 - \alpha/2 = 0.995$, notre fonction nous renvoie:

```
> Calpha258
[1] 2.58
```

Nous obtenons une valeur $C_{0.01} = 2.58$

Donc $P(-2.58 < \mu < 2.58) = 1 - \alpha = 0.99$

Or $\sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

Ce qui nous donne $P(-2.58 < \sqrt{n}*(Xn - \mu)/ \sqrt{S} < 2.58) = 0.99$

Soit $P(Xn-2.58*(\sqrt{S}/\sqrt{n}) < \mu < Xn+2.58*(\sqrt{S}/\sqrt{n})) = 0.99$

Pour déterminer une estimation par intervalle de μ on utilisera un programme R :

```
[1] "[ 5.361 ; 10.928 ]"
```

Voici notre intervalle de confiance à 99% : [2.766 ; 9.225]. Cet intervalle nous permettra de dire si dans 99% des cas une valeur comme un μ théorique s'y trouve.

2.c) Test de comparaison de moyenne

Pour notre test, il nous faut déterminer :

- Les hypothèses : H0 et H1
- Une statistique de test Tn
- Une règle de décision avec une région de rejet

Quelles sont les hypothèses H0 et H1 ?

H0 : $\mu = 7$ H1 : $\mu \neq 7$

L'hypothèse H0, on suppose que la distance moyenne étoiles – Terre est de 7 années lumières.

L'hypothèse H1, on suppose que la distance moyenne étoiles – Terre est différente de 7 années lumières.

On souhaite déterminer une statistique de test Tn

On pose

$Tn = \sqrt{n}*(Xn - \mu)/S$

Nous sommes dans un cas avec une loi quelconque et de variance inconnue. Notre échantillon est assez grand ($n=100 \geq 30$) pour approximer une loi normale centrée réduite par le Théorème central limite $Tn = \sqrt{n}*(Xn - \mu)/S \approx N(0,1)$

On calcule une estimation de Tn avec notre programme R :

```
[1] 1.061
```

D'après la sortie R, une estimation de la statistique de test Tn est : tn = 1.061

Maintenant nous voulons établir une règle de décision

Pour cela il nous faut construire une région de rejet : $R : \{|Tn| > C\alpha\}$ Si

$|Tn| > C\alpha$, c'est-à-dire si $Tn < -C\alpha$ ou $Tn > C\alpha$, on rejette l'hypothèse H0.

Si $|Tn| < C\alpha$, c'est-à-dire $-C\alpha < Tn < C\alpha$, on ne rejette pas l'hypothèse H0.

- Test de niveau α à 5% :

Nous avons déjà déterminé $C_{0.05} = 1.96$, nous savons que tn = 1.061

Nous avons $tn = 1.061 < 1.96$, nous pouvons conclure que l'hypothèse H0 n'est pas rejeté. La distance moyenne des étoiles par rapport à la Terre n'est pas significativement différente de 7.

- Test de niveau α à 1% :

Pour un test de niveau $\alpha = 1\%$ nous devons établir une règle de décision en construisant une région de rejet.

Nous connaissons la valeur $C_{0.01}$ qui est égale à 2.58, ce qui donne :

$|tn| = 1.061 < 2.58 = C_{0.01}$

On voit que $|tn|$ est inférieur à $C_{0.01}$, nous pouvons conclure que l'hypothèse H0 n'est pas rejeté. La distance moyenne des étoiles en année lumière par rapport à la Terre n'est pas significativement différente de 7.

Après avoir déterminé une statistique de test et fait des tests de niveau $\alpha = 1\%$ et $\alpha = 5\%$. On souhaite à présent déterminer une p-valeur, la p-valeur représente le plus petit risque α pour que le test de niveau α soit significatif.

$p\text{-valeur} = P_{H0}([Tn| > 1.061) = 2(1-P(Tn<1.061))$

Soit

```
[1] 0.2886899
```

Nous obtenons un arrondi de la p-valeur égale à 0.288, on sait que plus la p-valeur est proche de 0 plus le test sera significatif :

- Si $\alpha <$ p-valeur, le test de niveau α ne sera pas significatif
- Si $\alpha \geqslant$ p-valeur, le test de niveau α sera significatif

Dans notre cas avec un test de niveau $\alpha = 5\%$, nous avons $\alpha = 0.05 < 0.288$ = p-valeur, le test de niveau 5% n'est pas significatif.

Pour un test de niveau 1%, nous avons $\alpha = 0.01 < 0.288$ = p-valeur, le test de niveau 1% n'est pas significatif.

2.d) Test avec les intervalles de confiances

Nous avons fait des estimations d'intervalles de confiance à 95% et 99%, nous avons déterminé à l'étape précédente que les tests n'étaient pas significatifs, on souhaiterait vérifier nos propos à travers les intervalles de confiances.

- Pour l'intervalle de confiance à 95% de μ :

```
[1] "Dans 95% des cas 7 se trouve dans l'intervalle [ 6.03 ; 10.259 ]"
```

On ne rejette pas l'hypothèse $\mu = 7$ car $7 \in [6.03 ; 10.259]$, cette conclusion confirme le résultat de notre test avec $\alpha = 5\%$.

- Pour l'intervalle de confiance à 99% de μ :

```
[1] "Dans 99% des cas 7 se trouve dans l'intervalle [ 5.361 ; 10.928 ]"
```

On ne rejette pas l'hypothèse $\mu = 7$ car $7 \in [5.361 ; 10.928]$, cette conclusion confirme le résultat de notre test avec $\alpha = 1\%$.

2.e) Fonction t.test

Dans cette partie, nous cherchons une fonction R capable de créer un intervalle de confiance, réaliser un test et déterminer une p-valeur de notre échantillon. Il existe une fonction native sur R qui permet de faire cela, c'est la fonction t.test.

- Pour un risque $\alpha = 5\%$ nous retrouvons à quelques différences près les valeurs calculées à la main.

```
        One Sample t-test

data:  df_star100
t = 1.0609, df = 99, p-value = 0.2913
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
  6.003911 10.285289
sample estimates:
mean of x
   8.1446
```

Nous avions trouvé une estimation de l'intervalle de confiance : [6.03 ; 10.259], une statistique de test tn = 1.061 et une p-valeur = 0.288.

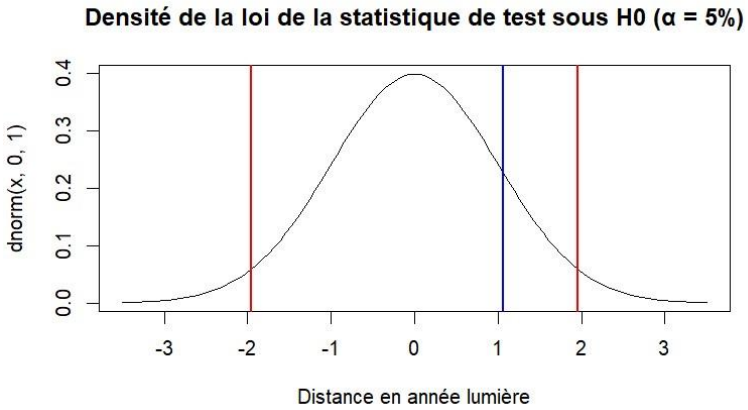
De même pour un risque α = 1%.

```
        One Sample t-test

data:  df_star100
t = 1.0609, df = 99, p-value = 0.2913
alternative hypothesis: true mean is not equal to 7
99 percent confidence interval:
  5.311081 10.978119
sample estimates:
mean of x
   8.1446
```

Nous avions trouvé une estimation de l'intervalle de confiance : [5.361 ; 10.928], une statistique de test tn = 1.061 et une p-valeur = 0.288.

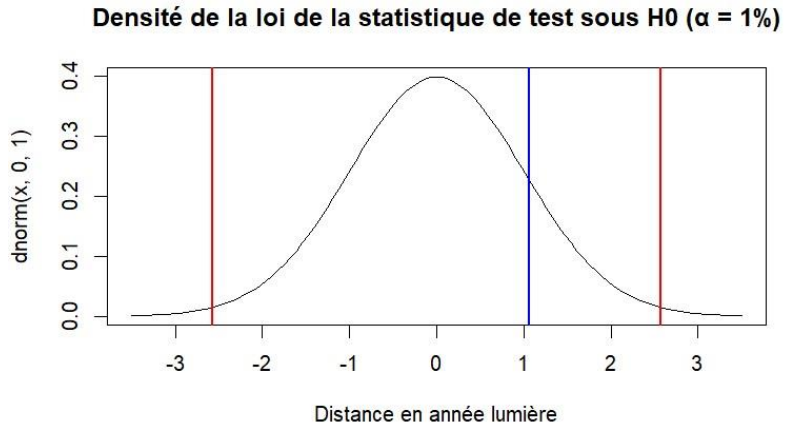
2.f) Illustration graphique des tests

- Illustration du test sous H0 avec α = 5% :



La droite bleue représente la valeur de la statistique de test « tn » et en rouge les quantiles théoriques délimitant la zone de non rejet de H_0 et la zone de rejet H_0. La statistique de test se trouve dans la zone de non rejet de H_0, montrant que le test n'est pas significatif.

- Illustration du test sous H0 avec α = 1% :



La statistique de test se trouve dans la zone de non rejet de H_0, montrant que le test n'est pas significatif.

2.g) Conclusion sur l'échantillon 100 observation

Dans cette 2nd partie de notre étude, basée sur la variable plx avec 100 observations, c'est-à-dire la variable mesurant la distance entre les étoiles et la Terre. Nous avons fait des tests avec différents pourcentages de risque α = 1% et α = 5%, qui se sont avéré également non significatifs, c'est-à-dire que la moyenne théorique $\mu_0 = 7$ n'est pas significativement différente de la moyenne μ de notre échantillon pour nos 2 tests de risque α égale à 1% et 5%. La p-valeur obtenue lors de l'étude de cet échantillon étant plus élevé que les seuils de nos tests à 1% et à 5%, cela ne nous permet pas de rejeter l'hypothèse nulle.

Conclusion

Notre étude s'est basée sur la variable Plx avec des échantillons de taille différentes, nous voulions mesurer à l'aide de méthodes statistiques des indicateurs sur la distance en année lumière entre les étoiles et la Terre. Dans le cas de notre échantillon de 30 observations, nous avons prélevé des informations telles que : 1er quartile = 1.785, la médiane = 3.720, 3ème quartile = 6.365 et la moyenne = 5.996. Nous remarquons que la moyenne théorique égale à 7 surestime la moyenne observée qui était de 5.996, mais toute fois d'après nos tests de p-valeur par rapport au seuil de significativité à 1% et 5%, nous n'avons pas pu conclure une différence significative entre la moyenne théorique et notre moyenne observée. Concernant l'échantillon de 100 observations, nous avons prélevé des informations telles que : 1er quartile = 3.178, la médiane = 5.215, 3ème quartile = 8.688 et une moyenne égale à 8.145. Nous pouvons remarquer que la moyenne théorique égale à 7 sousestime la moyenne observée qui est de 8.145, il existe une différence mais pas assez significative entre les 2 moyennes pour rejeter l'hypothèse que 7 puisse être une moyenne observée dans un échantillon, d'après nos comparaisons entre la p-valeur trouvé pour cet échantillon et les tests au niveau α = 1% et α = 5%.