Statistique : devoir, octobre 2023

Le sujet est composé de **4 exercices indépendants**. Vous devrez répondre aux questions sur un document quarto (ou éventuellement Markdown) avec une sortie au format **html**. Ce document devra afficher les codes R ainsi que les sorties qui permettent de répondre aux questions. A la fin de l'épreuve vous enverrez

- le **fichier de sortie compilé correctement au format html** ainsi que le répertoire créé par quarto (vous pouvez le zipper)
- le **fichier source au format qmd**

par email à laurent.rouviere@univ-rennes2.fr. La qualité du document quarto sera prise en compte dans le barème tout comme la structure et l'élégance des codes **R**.

Exercice 1 (Calcul de probabilités avec R)**.** Cet exercice est consacré à des calculs de probabilités et des représentations de lois classiques.

1. On considère X une variable qui suit une loi Binomiale $\mathcal{B}(50, 0.3)$.

   a) Calculer les probabilités suivantes (on donnera les résultats sans utiliser de fonctions **R** mais en justifiant brièvement).

   $$\mathbf{P}(X \leq -2), \quad \mathbf{P}(X \geq -1) \quad \text{et} \quad \mathbf{P}(X \geq 0).$$

   X prend toutes ses valeurs entre 0 et 50. Par conséquent :

   $$\mathbf{P}(X = -2) = 0, \mathbf{P}(X \geq -1) = 1 \quad \textit{et} \quad \mathbf{P}(X \geq 0) = 1.$$

   b) Calculer les probabilités (avec **R**) suivantes.

   $$\mathbf{P}(X = 13), \quad \mathbf{P}(X \leq 13) \quad \text{et} \quad \mathbf{P}(X > 13).$$

```
dbinom(13,50,0.3)
## [1] 0.1050175
pbinom(13,50,0.3)
## [1] 0.3278832
1-pbinom(13,50,0.3)
## [1] 0.6721168
```

c) Calculer les probabilités suivantes (toujours avec **R**).

$$\mathbf{P}(10 < X \leq 15), \quad \mathbf{P}(18 \leq X \leq 45) \quad \text{et} \quad \mathbf{P}(21 \leq X \leq 100).$$

```
sum(dbinom(11:15,50,0.3))
## [1] 0.4903278
sum(dbinom(18:45,50,0.3))
## [1] 0.2178069
sum(dbinom(21:50,50,0.3))
## [1] 0.04776384
```

2. On considère ici Y une variable de loi normale d'espérance 3 et de variance 1 (notée $\mathcal{N}(3,1)$).

a) Calculer les probabilités suivantes (sans le logiciel **R** et en justifiant brièvement).

$$\mathbf{P}(Y = 0) \quad \text{et} \quad \mathbf{P}(Y \geq 3).$$

La première est nulle puisque Y est une variable continue. La seconde vaut 0.5 puisque Y est centrée en 3.

b) Calculer (avec **R**) les probabilités suivantes.

$$\mathbf{P}(Y \leq 1), \ \mathbf{P}(Y < 1) \quad \text{et} \quad \mathbf{P}(Y > 1).$$

```
pnorm(1,3,1)
## [1] 0.02275013
pnorm(1,3,1)
## [1] 0.02275013
1-pnorm(1,3,1)
## [1] 0.9772499
```

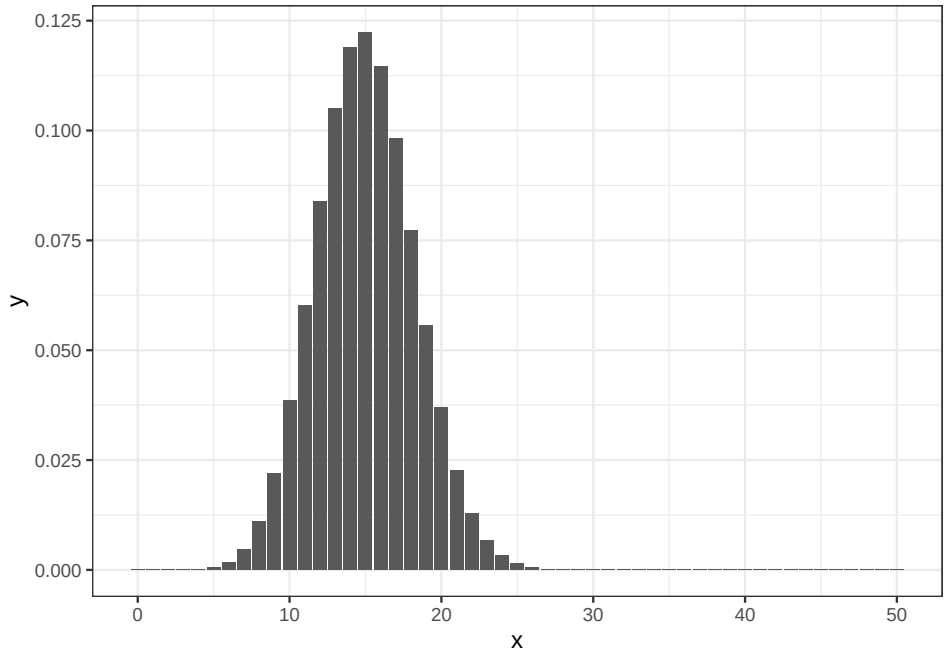
c) Calculer les probabilités (toujours avec **R**)

$$\mathbf{P}(-1 < Y \leq 3) \quad \text{et} \quad \mathbf{P}(Y \leq -1 \text{ ou } Y \geq 4).$$

```
pnorm(3,3,1)-pnorm(-1,-2,1)
## [1] -0.3413447
pnorm(-1,3,1)+(1-pnorm(4,3,1))
## [1] 0.1586869
```

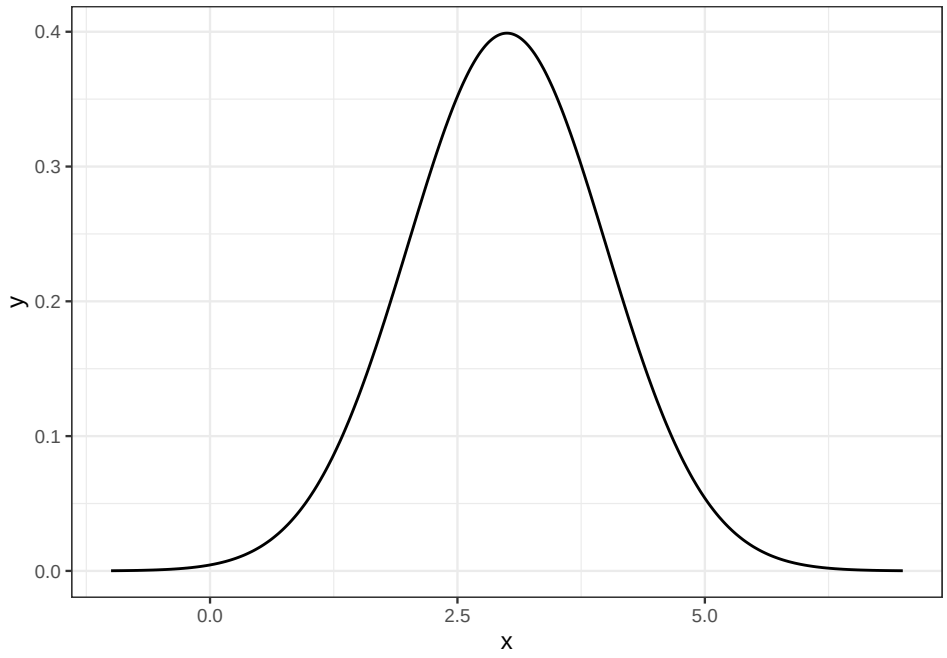
3. Représenter la loi Binomiale $\mathcal{B}(50, 0.3)$ à l'aide d'un diagramme en barre (on utilisera `ggplot`).

```
tibble(x=0:50,y=dbinom(x,50,0.3)) |>
  ggplot() +aes(x=x,y=y) + geom_bar(stat = "identity")
```



4. A l'aide de `ggplot`, représenter la densité de la loi de Y (la densité de $\mathcal{N}(3, 1)$). On prendra des valeurs entre -1 et 7 sur l'axe des abscisses.

```
tibble(x=seq(-1,7,by=0.01),y=dnorm(x,3,1)) |>
  ggplot()+aes(x=x,y=y)+geom_line()
```



Exercice 2 (Iris de Fisher)**.** On considère les données sur les iris de Fisher. On répondra aux 5 questions suivantes en utilisant les verbes du package `dplyr`.

1. Calculer la longueur de pétales moyenne de tous les iris.

```
iris |> summarize(moy=mean(Petal.Length))
##     moy
## 1 3.758
```

2. Calculer la largeur de pétales moyenne des iris de l'espèce `versicolor`.

```
iris |> filter(Species=="versicolor") |> summarize(moy=mean(Petal.Width))
##     moy
## 1 1.326
```

3. Calculer, pour chaque espèce, la moyenne et variance de la longueur de pétales.

```
iris |> group_by(Species) |>
  summarize(M=mean(Petal.Length),V=var(Petal.Length))
## # A tibble: 3 x 3
##   Species        M      V
##   <fct>      <dbl>  <dbl>
## 1 setosa      1.46 0.0302
## 2 versicolor  4.26 0.221
## 3 virginica   5.55 0.305
```

4. Pour l'espèce `setosa`, quelle est l'iris qui a la plus petite largeur de sépales ?

```
iris |> filter(Species=="setosa") |> arrange(Sepal.Width) |> slice(1)
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          4.5         2.3          1.3         0.3  setosa
#ou
iris |> filter(Species=="setosa") |> slice_min(Sepal.Width)
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          4.5         2.3          1.3         0.3  setosa
```

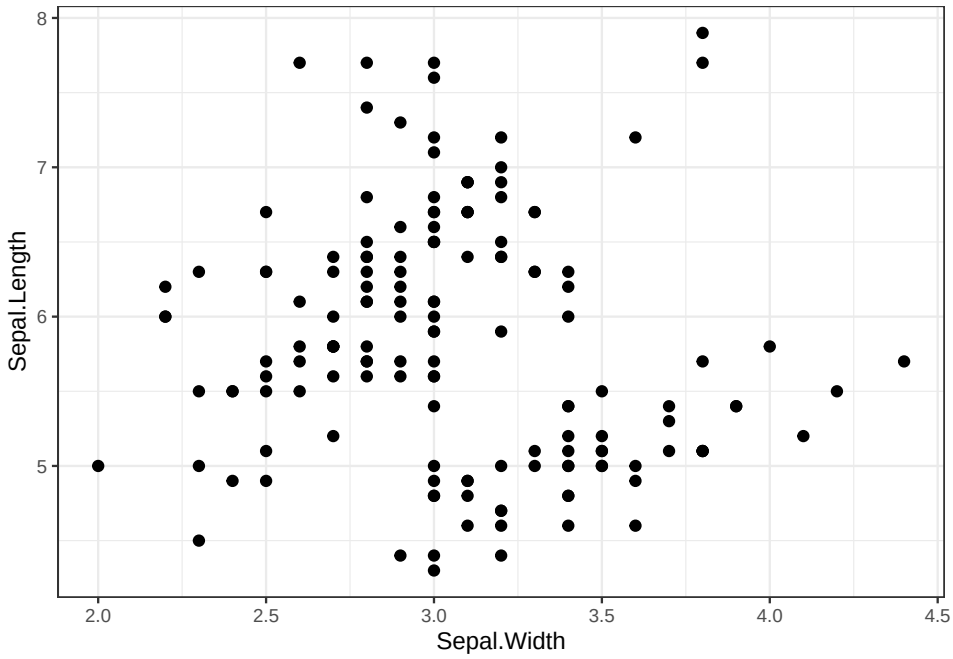
5. Combien d'iris ont une largeur de sépalse supérieure ou égale à 2.95 ?

```
iris |> filter(Sepal.Width>=2.95) |> summarize(n())
##   n()
## 1  93
```

On répondra aux 3 questions suivantes à l'aide d'un graphe **ggplot**.

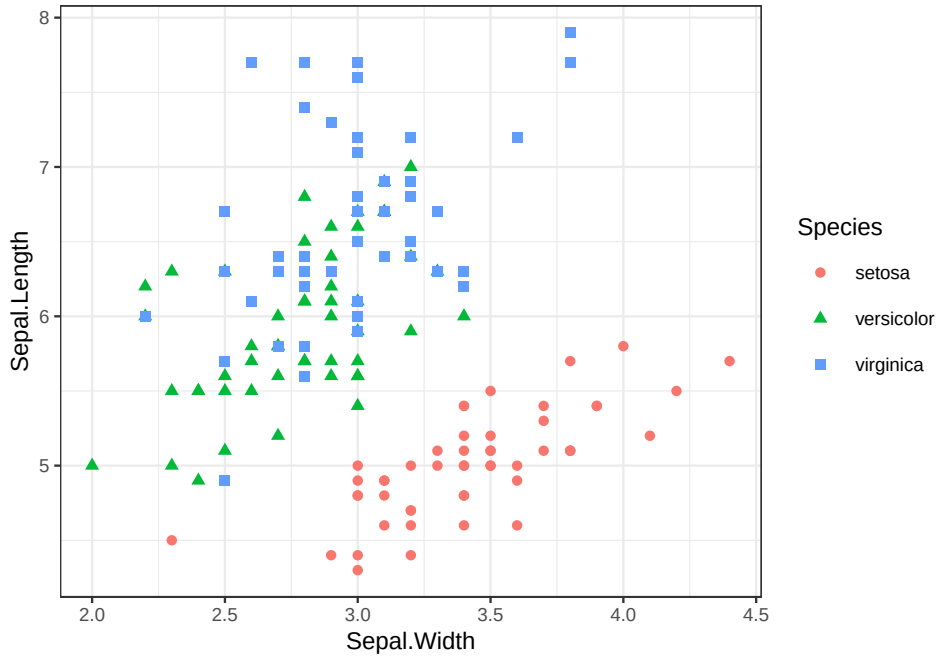
6. Construire le nuage de poins qui permet de visualiser la longueur de sépales en fonction de la largeur de sépales

```
(p1 <- ggplot(iris)+aes(x=Sepal.Width,y=Sepal.Length)+geom_point())
```



7. Même question mais avec une couleur et une forme différente pour les points en fonction de l'espèce.

```
p1+aes(color=Species,shape=Species)
```

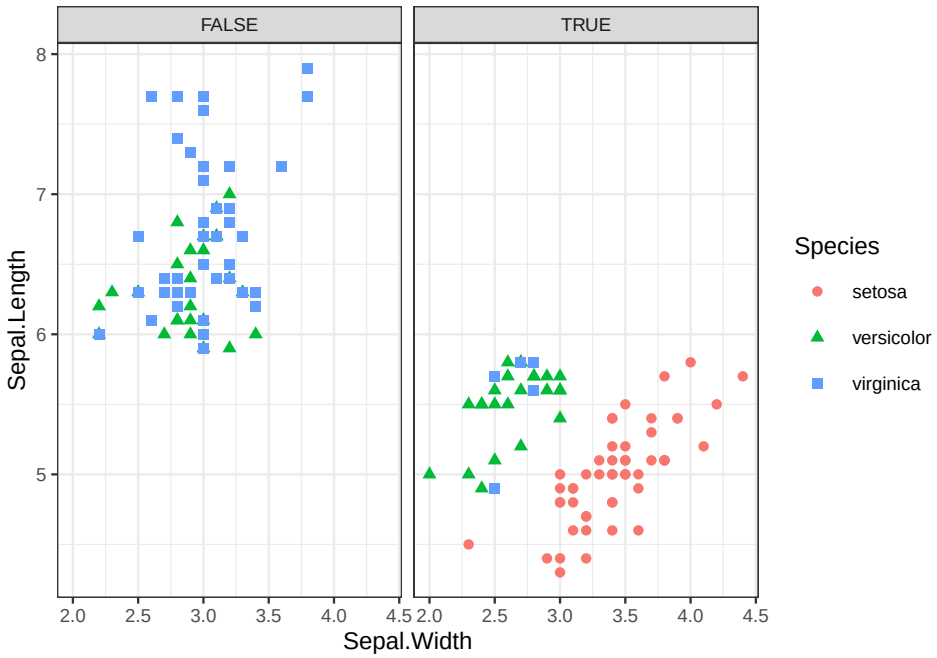


8. On souhaite maintenant conserver cette visualisation avec 2 graphes côte à côte :

   - le premier (à gauche ou à droite) pour les iris dont la longueur de sépales est inférieure ou égale à 5.8
   - le second (à gauche ou à droite) pour les iris dont la longueur de sépales est supérieure à 5.8.

   Représenter ce graphe.

```
iris |> mutate(separation=Sepal.Length<=5.8) |>
  ggplot()+aes(x=Sepal.Width,y=Sepal.Length,color=Species,shape=Species)+
  geom_point()+
  facet_wrap(~separation)
```



9. Construire un intervalle de confiance de niveau 95% pour les paramètres suivants :

   a) La largeur de pétales moyenne

```
t.test(iris$Petal.Width,conf.level=0.95)$conf.int
## [1] 1.076353 1.322313
## attr(,"conf.level")
## [1] 0.95
```

   b) La largeur de sépales moyenne de l'espèce `versicolor`

```
sep_ver <- iris |> filter(Species=="versicolor") |> select(Sepal.Width)
t.test(sep_ver,conf.level=0.95)$conf.int
## [1] 2.68082 2.85918
## attr(,"conf.level")
## [1] 0.95
#ou
iris |> filter(Species=="versicolor") |>
  summarize(bi=t.test(Sepal.Width,conf.level=0.95)$conf.int[1],
            bs=t.test(Sepal.Width,conf.level=0.95)$conf.int[2])
##        bi      bs
## 1 2.68082 2.85918
```

   c) La longueur de sépales moyenne des iris qui ont une longueur de pétales supérieure ou égale à 2

```r
lp2 <- iris |>
  filter(Petal.Length>=2) |>
  select(Sepal.Length)
t.test(lp2,conf.level=0.95)$conf.int
## [1] 6.130479 6.393521
## attr(,"conf.level")
## [1] 0.95
```

Répondre aux deux question suivantes à l'aide d'un test statistique de niveau $\alpha = 0.05$.

10. Peut-on dire que la longueur de sépales moyenne de tous les iris est égale à 6 ?

```r
t.test(iris$Sepal.Length,alternative="two.sided",mu=6)
##
##  One Sample t-test
##
## data:  iris$Sepal.Length
## t = -2.3172, df = 149, p-value = 0.02186
## alternative hypothesis: true mean is not equal to 6
## 95 percent confidence interval:
##  5.709732 5.976934
## sample estimates:
## mean of x
##  5.843333
```

11. Peut on dire que la largeur de sépales moyenne de l'espèce `versicolor` est différente de celle de l'espèce `virginica`.

```r
ver <- iris |> filter(Species=="versicolor")
vir <- iris |> filter(Species=="virginica")
t.test(ver$Sepal.Width,vir$Sepal.Width,alternative="two.sided")
##
##  Welch Two Sample t-test
##
## data:  ver$Sepal.Width and vir$Sepal.Width
## t = -3.2058, df = 97.927, p-value = 0.001819
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.33028364 -0.07771636
## sample estimates:
## mean of x mean of y
##     2.770     2.974
```

Exercice 3 (Importation et fusion)**.** On considère deux jeux de données qui se trouvent dans les fichiers :

- `donnees1_2023.csv` : 100 lignes et 2 colonnes (`ID` pour l'identifiant d'un individu et `A2020` pour un indicateur mesuré en 2020.
- `donnees2_2023.csv` : 100 lignes et 2 colonnes (`id` pour l'identifiant d'un individu et `A2021` pour un indicateur mesuré en 2021.

On précise que ces deux tables contiennent les mêmes individus identifiés dans la première colonne.

1. Importer les données et calculer la moyenne de l'indicateur en 2020 et 2021

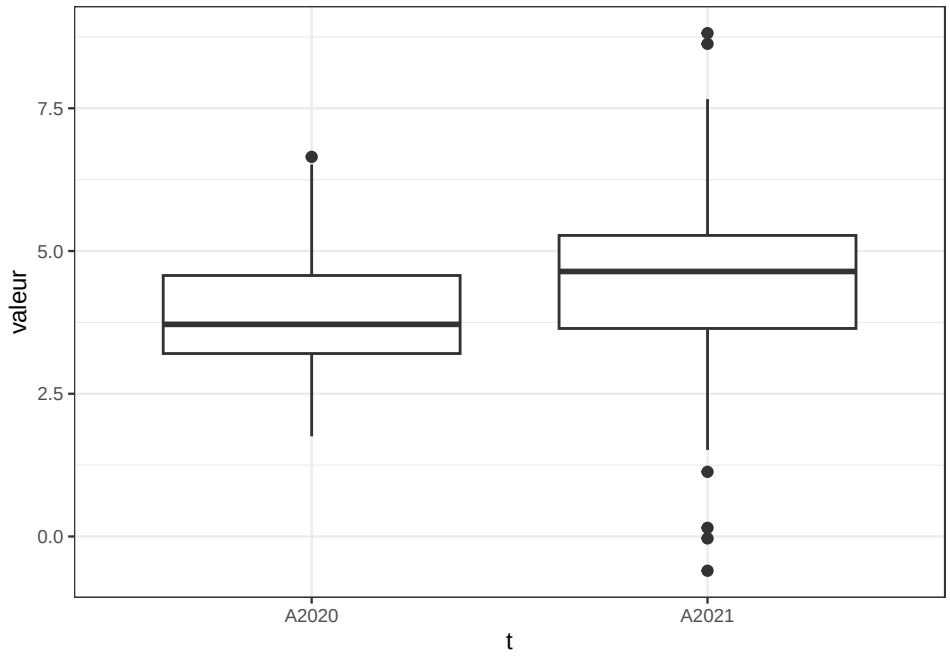
```
tbl1 <- read_csv("donnees1_2023.csv")
tbl2 <- read_csv("donnees2_2023.csv")
tbl1 |> summarise(M20=mean(A2020))
## # A tibble: 1 x 1
##     M20
##   <dbl>
## 1  3.94
tbl2 |> summarise(M21=mean(A2021))
## # A tibble: 1 x 1
##     M21
##   <dbl>
## 1  4.44
```

2. Effectuer une jointure complète : on obtiendra une nouvelle table avec 100 lignes et 3 colonnes.

```
tbl <- full_join(tbl1,tbl2,by=join_by(ID==id))
```

3. À l'aide d'un boxplot, comparer la distribution de l'indicateur en 2020 et 2021.

```
tbl_l <- tbl |> pivot_longer(-ID,names_to = "t",values_to = "valeur")
ggplot(tbl_l)+aes(x=t,y=valeur)+geom_boxplot()
```



4. Effectuer un test de comparaison de moyenne de niveau 5% pour comparer les moyennes de l'indicateur en 2020 et 2021. On n'oubliera pas de donner la conclusion du test.

```
t.test(tbl$A2020,tbl$A2021)
##
##  Welch Two Sample t-test
##
## data:  tbl$A2020 and tbl$A2021
## t = -2.577, df = 164.5, p-value = 0.01084
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8727809 -0.1155408
## sample estimates:
## mean of x mean of y
##  3.943238  4.437399
```

5. Rappeler ce que signifie le niveau de 5%.

Exercice 4 (Tests)**.** On dispose de $n = 50$ observations $x_1, \dots, x_n$ issues d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ avec μ et σ^2 inconnus. On souhaite tester les hypothèses

$$H_0 : \mu = 4 \quad \text{contre} \quad H_1 : \mu > 4.$$

avec le test vu en cours au niveau $\alpha = 5\%$.

1. Quelle est la zone de rejet de ce test

```
# c'est la zone défini après ce quantile
qt(0.95,49)
## [1] 1.676551
```

2. Après calcul, on trouve que la valeur observée de la statistique de test vaut 1.95. Quelle est la conclusion du test ?

La statistique tombe dans la zone de rejet, on rejette l'hypothèse nulle.

3. Calculer la probabilité critique de ce test. Est-ce que le résultat est cohérent avec la réponse de la question précédente ?

```
1-pt(1.95,49)
## [1] 0.02845485
```

4. Toujours pour le même test, quelle est la probabilité de rejeter H_0 à raison lorsque la vraie valeur de μ vaut 5 et que la variance des données S^2 est égale à 4.

Si $\mu = 5$ la loi de la statistique change mais le processus de décision reste identique. On a alors

$$T = \sqrt{n}\frac{\bar{x}_n - 5}{S} \sim \mathcal{T}_{n-1}.$$

On a alors

$$\begin{aligned}
\mathbf{P}_{H_1}(R_{H_0}) =& \mathbf{P}\left(\sqrt{n}\frac{\bar{x}_n - 5}{S} > t\right) \\
=& \mathbf{P}\left(\bar{x}_n > \frac{tS+4}{\sqrt{n}}\right) \\
=& \mathbf{P}\left(T > t + (4-5)\frac{\sqrt{n}}{S}\right) \\
=& 1 - F_{\mathcal{T}_{n-1}}\left(t + (4-5)\frac{\sqrt{n}}{S}\right).
\end{aligned}$$

On trouve

```
1-pt(qt(0.95,49)+(4-5)*sqrt(49)/2,df=49)
## [1] 0.9628309
```

Cette probabilité correspond à la puissance du test lorsque $\mu = 5$.