École Nationale d'Ingénieurs de Tunis

Rapport de projet statistique de Test d'hypothese

Elaboré par :
Lakhzouri Roukaya

Classe :
2AMINDS

Encadré par :
Mme. Anissa RABHI

Année universitaire : 2023/2024

Table des matières

Table des figures

Chapitre 1

Exercice 1

1.1 Énoncé de l'exercice :

On considère une expérience aléatoire consistant à lancer un dé bien équilibré. On note D la variable aléatoire égale au résultat obtenu.

1. Quelle est la loi de D ? Quelle est son espérance µ et sa variance σ^2 ?
2. Créer une fonction De_faces(N) qui renvoie un vecteur ligne de taille N dont les coefficients sont des variables aléatoires indépendantes de même loi que D.
3. Tracer l'histogramme d'un échantillon de N = 1000 valeurs suivant la loi de D.
4. Écrire une fonction moy_emp(X) qui prend en entrée un vecteur (ligne ou colonne) X et qui renvoie sa moyenne empirique.
5. Calculer la moyenne empirique d'un échantillon de N = 10000 valeurs suivant la loi de D. Pouvait-on s'attendre à ce résultat ? Justifier votre réponse.

Soit n un entier naturel non nul. On note $D_n$la variable aléatoire égale à la moyenne empirique de n copies indépendantes de D.

6. Écrire un programme De_Moyennes(n, N) qui renvoie un vecteur ligne de taille N, dont les composantes sont des variables aléatoires indépendantes de même loi que D_n.
7. Tracer l'histogramme d'un échantillon de N = 1000 valeurs indépendantes suivant la loi de D_{100}.
8. Quelle est l'allure de cet histogramme ? Quelle est la "loi approchée" de D_{100} ? Justifier.

On considère le jeu (récursif) suivant :

Le joueur commence par lancer un dé ; si le résultat d'un lancer est 6, le joueur a un nouveau lancer, sinon il doit arrêter ; le score du joueur est la somme de tous les résultats des dés qu'il a lancés.

9. Est-il possible d'avoir un score de 13 avec ce jeu ? Si oui, avec quelle probabilité ?
10. Écrire une fonction JeuDeDes(N) qui renvoie un échantillon de N valeurs indépendantes qui simulent le score d'un joueur de ce jeu de dés.
11. Tracer l'histogramme d'un échantillon pour N = 10000.
12. Calculer la fréquence empirique de la valeur 13, toujours pour N = 10000 et comparer avec la réponse donnée à la question 9.
13. Estimer, à l'aide de la simulation, l'espérance du score d'un joueur et la probabilité que le score soit supérieur à 10.

1.2 Solution :

1.2.1 Question 1 :

D est la Variable Aleatoire qui représente le résultat obtenu après le lancer d'un dé équilibré donc $D \hookrightarrow \mathcal{U}(6)$ On a dans ce cas

$$E(D) = \frac{6+1}{2} = 3.5$$

$$\sigma^2 = V(D) = \frac{6^2-1}{12} = \frac{35}{12}$$

1.2.2 Question 2 :

```
#question 2 :
De.faces<-function(N)
{
  return(sample(1:6, size = N, replace = TRUE))
}
```

FIGURE 1.1 – Code sur R de fonction De.faces(N)

1.2.3 Question 3 :

```
#question 3
echantillon <- De.faces(1000)
# Tracer l'histogramme
hist(echantillon, main = "Histogramme de l'échantillon", xlab = "Valeurs",
ylab = "Fréquence", col = "lightblue", border = "black")
```

Figure 1.2 – Code sur R d'histogramme d'un échantillon suivant la loi de D

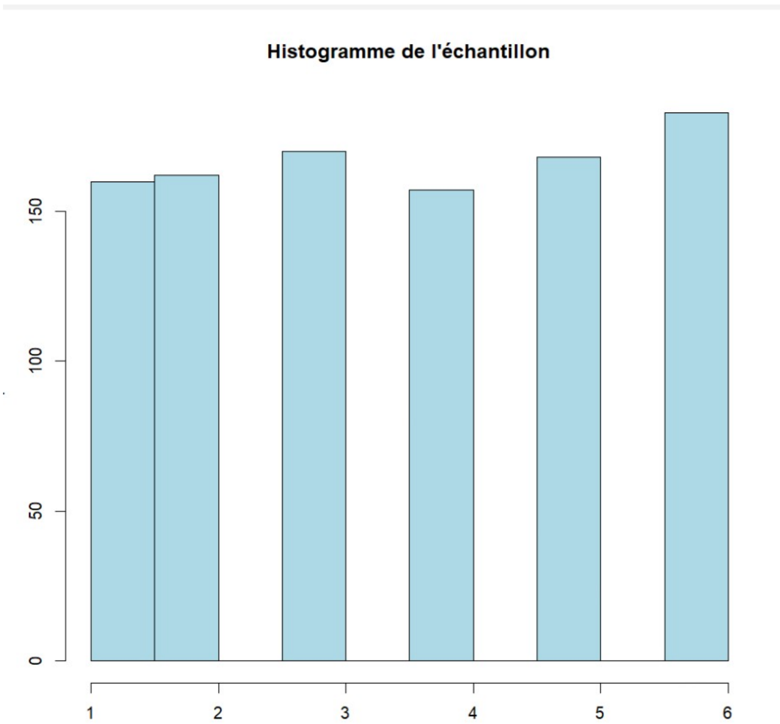


Figure 1.3 – histogramme d'un échantillon suivant la loi de D

Interpretation de l'histigramme : Pour une distribution uniforme discrète, chaque issue a la même probabilité d'occurrence, ce qui signifie que chaque barre de l'histogramme devrait être d'une hauteur similaire, représentant approximativement $\frac{\text{taille totale de l'échantillon}}{6} \approx 166$ ce qui justifie la configuration de l'histogramme.

1.2.4 Question 4 et 5 :

on a par definition : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ou n est la taille de l'échantillon, $(X_1,.........,X_n)$ un n-échantillon d'une variable aléatoire X dans l'échantillon

```
#question4:
moy.emp<-function(X)
{
  return((1/length(X))*sum(X))
}
#question 5:
echantillon_2<-De.faces(10000)
s=moy.emp(echantillon_2)
print(s)#s=3.4999 est tres proches de E(X)
```

FIGURE 1.4 – Code sur R de fonction moy.emp(X)

Il est observé que la moyenne empirique de la variable aléatoire D, calculée à l'aide de la fonction `moy.emp(D)`, se rapproche considérablement de son espérance E(D). Cette proximité est conforme à la prévisibilité du théorème des grands nombres, d'autant plus que la taille de l'échantillon N est très grande.

$$\lim_{n\to\infty} \bar{D}_n = E(D)$$

1.2.5 Question 6 et 7 et 8 :

D_n est la moyenne empirique de n copies indépendantes de D, alors D_n est définie comme suit :

$$D_n = \frac{1}{n}\sum_{i=1}^{n} D_i$$

où D_i représente chaque observation individuelle de la variable aléatoire D.

```
#question 6:
N=1000
De.Moyennes<-function(n, N)
{
  moyennes <- numeric(N)
  for(i in 1:N )
  {
    D_i<-De.faces(n)
    D_n=moy.emp((D_i))
    moyennes[i]=D_n
  }
  return(moyennes)
}
```

FIGURE 1.5 – Code sur R de fonction De.moyenne(n,N)

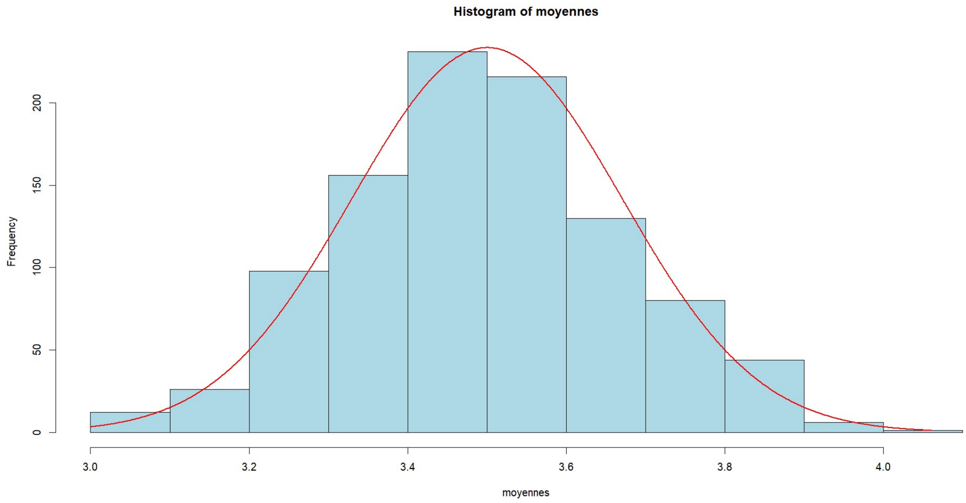


FIGURE 1.6 – l'histogramme d'un échantillon suivant la loi de D_{100}

Dans notre cas, nous avons un échantillon de taille $N = 1000$ en prenant la moyenne de $n = 100$. L'allure de l'histogramme de D_{100} suggère une distribution qui se rapproche de la loi normale $\mathcal{N}(E(D_n), V(D_n))$ avec

$$E(D_n) = E(D) = 3.5 \text{ et } V(D_n) = \frac{V(D)}{n} = \frac{35}{1200}$$

Ce qui est prévisible d'après le Théorème Central Limite. Cependant, l'analyse graphique seule ne suffit pas à démontrer que la "loi approximative" de D_{100} est une loi normale. C'est pourquoi nous faisons appel aux tests d'hypothèses, plus précisément en nous appuyant sur une approche numérique telle que le test de normalité prédéfini :

`shapiro.test`(D_{100}) p-value

Ce test oppose, par construction, les deux hypothèses :
H_0 : D_{100} est une loi normale
H_1 : D_{100} n'est PAS une loi normale
L'application du test sur D_{100} :
p-value $= 0.2787 > 0.05$

Comme la p-valeur est supérieure à 0.05, on ne rejette pas H_0. Ainsi, on peut considérer, au seuil $\alpha = 5\%$, que D_{100} suit une loi normale $\mathcal{N}(E(D_n), V(D_n))$.

```r
#question 8 :
# Superposer la courbe de la loi normale
mu <- 3.5
sigma <- sqrt(35/1200)
x <- seq(min(moyennes), max(moyennes), length = 1000)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y * N * diff(hist(moyennes, col = "lightblue",border = "black")
                      $breaks[1:2]), col = "red", lwd = 2)
# Test statistique
shapiro.test(moyennes)
```

Figure 1.7 – code de l'histogramme d'un échantillon suivant la loi de D_{100} et test d'hypothese

1.2.6 Question 9 :

Soit S :la variable aléatoire qui représente le score du joueur

R_i :le résultat obtenu à la i-ème lancée. il est claire que $R_i \hookrightarrow \mathcal{U}(6)$

Si le jeu s'arrête dès qu'un nombre autre que 6 est obtenu, alors la seule possibilité d'obtenir un score de 13 est d'obtenir successivement un 6, puis un autre 6, puis un 1. Dans ce cas, la probabilité serait effectivement :

$$\begin{aligned} P(S=13) &= P((R_1=6)\cap(R_2=6)\cap(R_3=1)) \\ &= P(R_1=6)\cdot P(R_2=6)\cdot P(R_3=1) \\ &= \frac{1}{6}\times\frac{1}{6}\times\frac{1}{6} = \left(\frac{1}{6}\right)^3 \approx 0.00462. \end{aligned}$$

1.2.7 Question 10 :

```r
#question 10 :
JeuDeDes<-function(N)
{
  scores=numeric(N)
  for(i in 1:N)
  {
    score<-0
    resultat=De.faces(1)# Lancer un dé
    while(resultat==6)
    {
      score=score+resultat #Ajouter le résultat au score
      resultat=De.faces(1) #Répéter le lancement de Dé
    }
    score=score+resultat
    scores[i]=score

  }
  return(scores)
}
```

Figure 1.8 – Code sur R de fonction JeuDedes(N)

1.2.8 Question 11 :

```
#question 11
N=10000
scores<-JeuDeDes(N)
hist(scores, main = "Histogramme de scores", xlab = "scores",
     ylab = "Fréquence", col = "lightblue", border = "black")
```

FIGURE 1.9 – Code sur R de fonction JeuDedes(N)

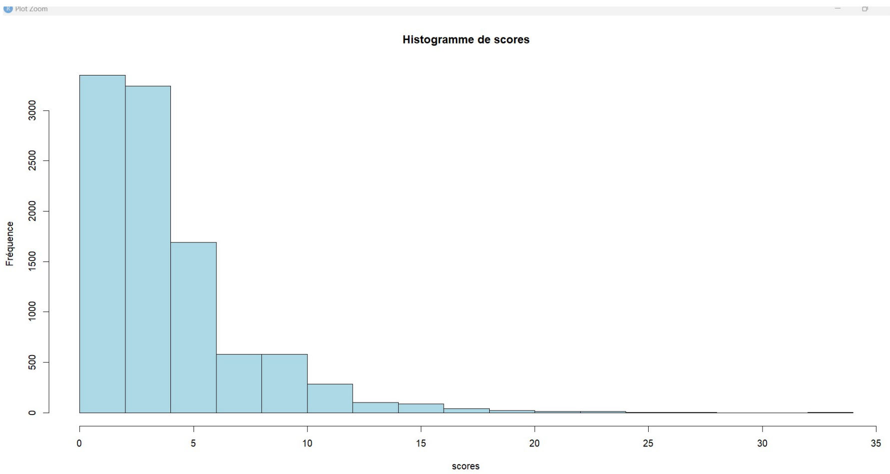


FIGURE 1.10 – histogramme d'un échantillon des scores

La probabilité d'obtenir un score inférieur ou égal à 5 ($P(\mathrm{S} \leq 5)$) est de $\frac{5}{6}$, tandis que la probabilité d'obtenir un score de 6 ($P(\mathrm{S} = 6)$) est de $\frac{1}{6}$. Cela explique l'augmentation de la fréquence des scores inférieurs ou égaux à 5 et la diminution des scores supérieurs à 6.

1.2.9 Question 12 :

```
#question 12
freq=0
for (i in 1:N)
{
  if(scores[i]==13)
  {
    freq=freq+1
  }
}
freq_emp=freq/N
#freq_emp=0.0041
```

FIGURE 1.11 – methode de calcule de freq$_{e}mp$

le resultat obtenue est 0.0041 qui est tres proche de la probabibilte P(S=6) ce qui est explique par le theoreme de Grand nombre : Sur un grand nombre de lancers d'un dé équilibré à six faces numérotées de 1 à 6, on peut approcher le moyenneme empirique $\S = D_N = \frac{1}{N}\sum_{i=1}^{N} D_i$ par E(D) or on a

$$\bar{(}D) = \frac{1}{N}\sum_{i=1}^{N} D_i = \frac{1}{N}(n_1 + n_2 \times 2 + n_3 \times 3 + n_4 \times 4 + n_5 \times 5 + n_6 \times 6) = \sum_{k=1}^{6} k \times f_k$$

avec n_k = nombre de fois d'obtenir l'entier k

$$\sum_{k=1}^{6} n_k = N$$

et

$$f_k = \frac{n_k}{N}$$

donc f_k represente la fréquence empirique de k .

or ona $E(D) = \sum_{k=1}^{6} k \times p_k$ qui correspond à la définition de son espérance avec p_k probabilités associées pour chaque valeur k on a les fréquences d'apparition fk se rapprochent des probabilités associées pour chaque valeur quand $N \to +\infty$

1.2.10 Question 13 :

```
#question 13
esp=sum(scores)/N
f=0
for (i in 1:N)
{
  if(scores[i]>=10)
  {
    f=f+1
  }
}
freq_emp2=f/N
print(freq_emp2)
```

FIGURE 1.12 – calcule de l'esperance du score d'un joueur et la probabilite que le score soit superieur à 10

ona l'esperance du score est égale à 4.2307 et la probabilite que le score soit superieur à 10 est égale à 0.0845

Chapitre 2

Exercice 2 :

2.1 Énonce de l'exercices

Un groupe de 10 sujets entreprend sur une période de six mois un programme d'enrichissement cognitif destiné à améliorer leurs processus de traitement de l'information. Pour évaluer l'effet du programme, on fait passer aux sujets deux tests de niveaux comparables, l'un avant, l'autre après la période d'apprentissage, dont voici les scores :

$$\text{Score avant} : 5, 2, 8, 9, 5, 3, 25, 8, 8, 4, 1, 2$$

$$\text{Score après} : 6, 2, 9, 10, 6, 1, 3, 10, 6, 7, 2, 10$$

À l'aide d'un test paramétrique de comparaison, décider au niveau $\alpha = 5\%$ si les sujets ont en moyenne amélioré leur processus de traitement de l'information après le programme d'enrichissement cognitif.

2.2 Solution :

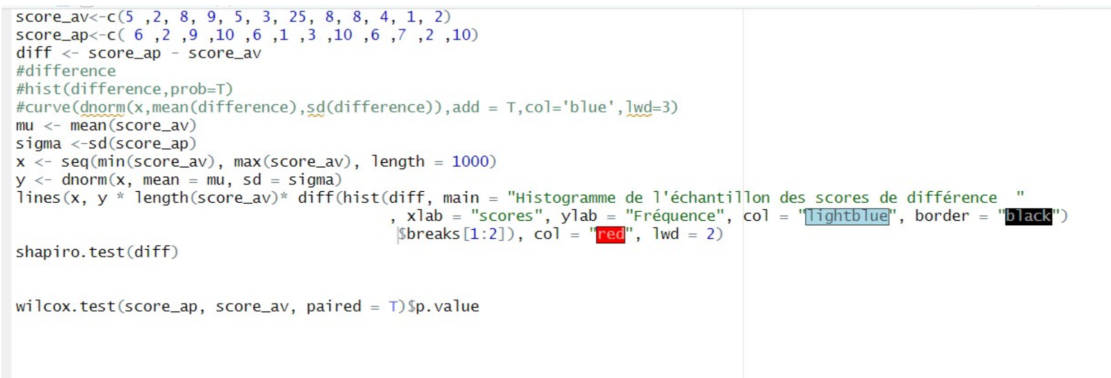
2.2.1 Code :

Modele score$_a vant : lascoredelapopoulationavantleprogramed'enrichissementscore_a pres :$ $lascoredelapopoulationapresleprogramed'enrichissement$

Test (Hypothèse nulle (H_0) : Il n'y a pas de différence significative entre les scores avant et après le programme d'enrichissement cognitif, c'est-à-dire $\mu_{\text{avant}} = \mu_{\text{après}}$. Hypothèse alternative (H_1) : Il y a une différence significative entre les scores avant et après le programme d'enrichissement cognitif, soit $\mu_{\text{avant}} \neq \mu_{\text{après}}$.

Il est évident que les deux échantillons sont appariés. Pour aborder cette problématique, nous commençons par tester la normalité de la différence entre les deux scores avant et après le traitement en utilisant la commande `shapiro.test`. La p-value dans ce cas est de 0.000273, inférieure au niveau de signification $\alpha = 0.05$. Par conséquent, la différence ne suit pas une distribution normale.

Dans ce cas, nous ne pouvons pas utiliser le test de Student. La solution consiste à utiliser un test de Wilcoxon, également appelé test des rangs signés. Ce test est employé pour comparer deux échantillons appariés ou deux échantillons indépendants lorsque les données ne suivent pas une distribution normale ou lorsque l'hypothèse d'homoscédasticité n'est pas respectée. La p-value dans ce cas est de 0.4998, supérieure au niveau de signification $\alpha = 0.05$, ce qui nous conduit à accepter l'hypothèse nulle.)



```r
score_av<-c(5 ,2, 8, 9, 5, 3, 25, 8, 8, 4, 1, 2)
score_ap<-c( 6 ,2 ,9 ,10 ,6 ,1 ,3 ,10 ,6 ,7 ,2 ,10)
diff <- score_ap - score_av
#difference
#hist(difference,prob=T)
#curve(dnorm(x,mean(difference),sd(difference)),add = T,col='blue',lwd=3)
mu <- mean(score_av)
sigma <-sd(score_ap)
x <- seq(min(score_av), max(score_av), length = 1000)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y * length(score_av)* diff(hist(diff, main = "Histogramme de l'échantillon des scores de différence  "
                                          , xlab = "scores", ylab = "Fréquence", col = "lightblue", border = "black")
                                          $breaks[1:2]), col = "red", lwd = 2)
shapiro.test(diff)


wilcox.test(score_ap, score_av, paired = T)$p.value
```

Figure 2.1 – code ex2

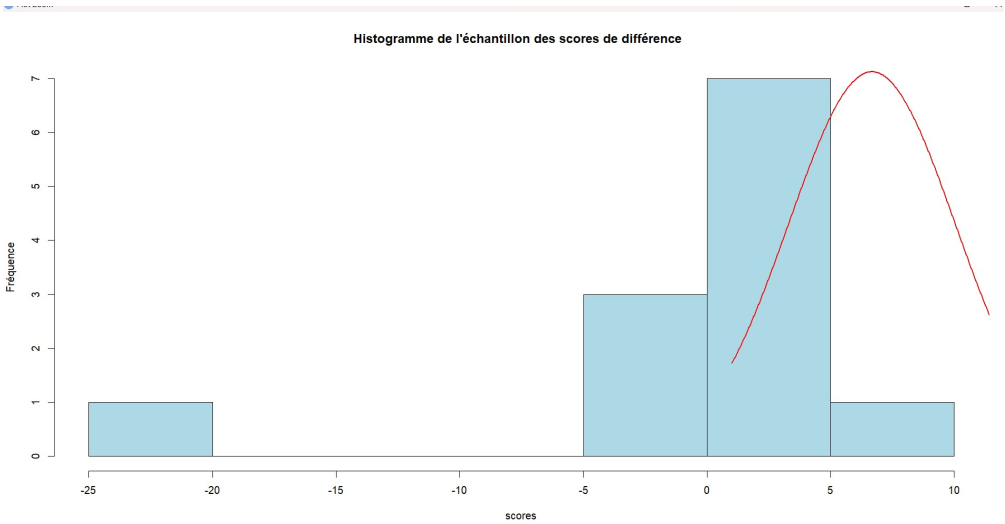


Figure 2.2 – histogramme de la différence

Chapitre 3

Exercice 3

3.1 Énoncé de l'exercice :

Dans un article scientifique, un biologiste donne la longueur L en mm des œufs trouvés dans les nids de deux espèces d'oiseaux :
Dans des nids de petite taille :

$$19,5; 22,1; 21,5; 21,9; 22,0; 21,0; 22,3; 21,0; 20,3; 20,9; 22,0; 22,0; 23,8; 21,2; 21,0.$$

Dans des nids de taille plus grande :

$$22,35; 23,9; 20,8; 25,8; 25,0; 24,0; 23,8; 21,7; 22,8; 23,1; 23,5; 23,0; 23,0; 23,1.$$

1. Vérifier que dans chacune des deux espèces, la longueur suit une loi normale.
2. Pouvez-vous dire, au seuil $\alpha = 5\%$, que les deux espèces ont la même variance ?
3. Si oui, tester l'hypothèse que l'oiseau adapte la taille de ses œufs à la taille du nid dans lequel il pond.

3.2 Solution :

3.2.1 Le code :

Le code complet dans R Studio se divise principalement en trois parties : la première partie vise à vérifier si la longueur suit une distribution normale, la deuxième présente un test d'hypothèse bilatéral visant à déterminer si les deux espèces ont la même variance, et la dernière partie évalue l'hypothèse selon laquelle l'oiseau ajuste la taille de ses œufs en fonction de la dimension du nid dans lequel il les dépose

```
#question 1 :
Lp=c(19.5,22.1,21.5,21.9,22.0,21.0,22.3,21.0,20.3,20.9,22.0,22.0,23.8,21.2,21.0)
Lg=c(22.35,23.9,20.8,25.8,25.0,24.0,23.8,21.7,22.8,23.1,23.5,23.0,23.0,23.1)
# Test de normalité pour Lp
mu <- mean(Lp)
sigma <-sd(Lp)
x <- seq(min(Lp), max(Lp), length = 1000)
y <- dnorm(x, mean = mu, sd = sigma)
lines(x, y * length(Lp)* diff(hist(Lp,
                                   main = "Histogramme de l'échantillon Dont des nids de petite taille ",
                                   xlab = "tailles des œufs", ylab = "Fréquence", col = "lightblue",
                                   border = "black")$breaks[1:2]), col = "red", lwd = 2)

# Tests statistiques
shapiro.test(Lp)
#W = 0.95149, p-value = 0.5841
# Test de normalité pour Lg graphiquement
mu2 <- mean(Lg)
sigma2 <-sd(Lg)
x <- seq(min(Lg), max(Lg), length = 1000)
y <- dnorm(x, mean = mu2, sd = sigma2)
lines(x, y * length(Lg)* diff(hist(Lg, main = "Histogramme de
                                   l'échantillon Dont des nids de grand taille ",
                                   xlab = "tailles des œufs", ylab = "Fréquence",
                                   col = "lightblue", border = "black")$breaks[1:2]), col = "red", lwd = 2)
# Test statistique
shapiro.test(Lg)
#W = 0.97064, p-value = 0.3973

#question 2
var.test(Lp,Lg)
#p_value=0.4676642

#question3
t.test(Lp,Lg,var.equal = T)
```

Figure 3.1 – code de l'exercice 3

3.2.2 Question 1 :

Lp V.A modélisant la longueur en mm d'un oeuf de la petite taille .

Lg V.A modélisant la longueur en mm d'un oeuf de grande taille . En adoptant une approche graphique et en superposant les histogrammes de fréquences des échantillons observés avec les courbes des distributions normales qui sont, par hypothèse, associées à ces échantillons, nous pouvons formuler l'hypothèse que les variables aléatoires L_1 et L_2 sont, dans une première approximation avec prudence, susceptibles d'être modélisées par des lois normales.

Pour résoudre cette ambiguïté, une méthode numérique a été employée en utilisant le test de normalité prédéfini

`shapiro.test(L)$p value`.

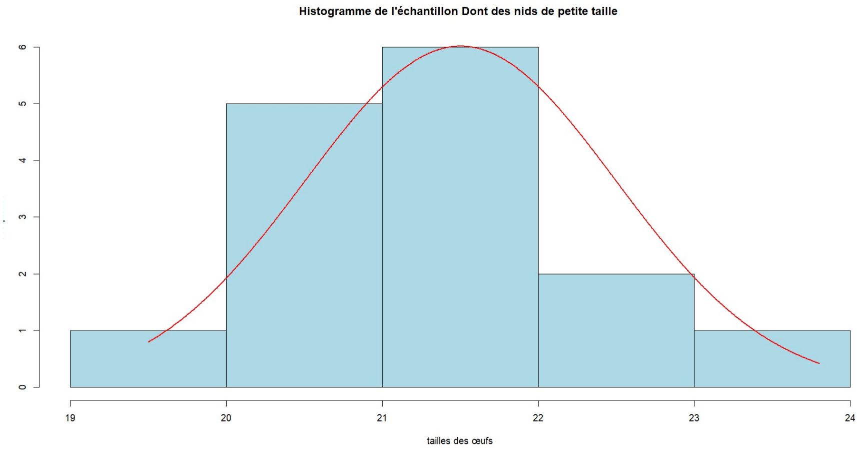


FIGURE 3.2 – histogramme de Lp

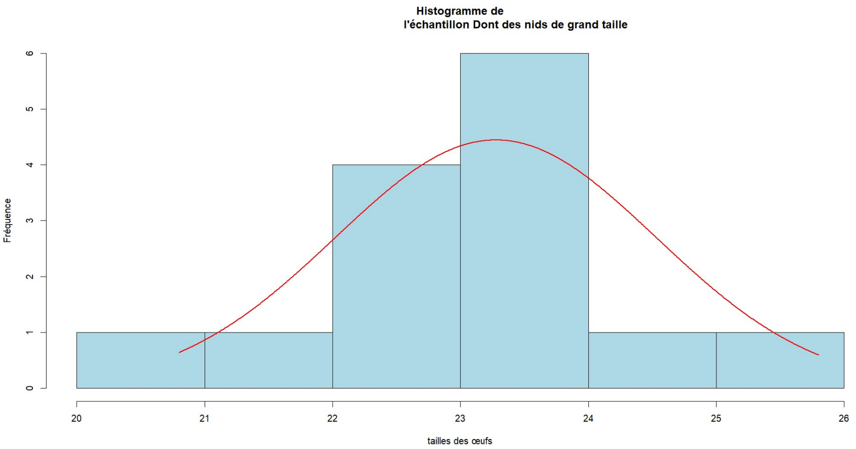


FIGURE 3.3 – histogramme de Lg

Ce test confronte, par sa conception, deux hypothèses :

$$H_0 : L \text{ suit une loi normale} \quad \text{contre} \quad H_1 : L \text{ NE suit PAS une loi normale}$$

en fixant le seuil de signification alpha établi à 0,05. L'application de ce test sur les variables L_p et L_g a produit respectivement les résultats suivants :

$$\text{p\_valeur pour } L_p : 0.4529(> 0.05)\text{p\_valeur pour } L_g : 0.8851(> 0.05)$$

nous constatons que les deux valeurs sont supérieures à alpha. Par conséquent, nous ne disposons pas de suffisamment de preuves pour rejeter l'hypothèse nulle, indiquant que les données des échantillons semblent suivre une distribution normale.

3.2.3 Question 2 :

Modeles : Lp : Lp la longueur en mm des oeufs de première espèce de petite taille . Lg la longueur en mm des oeufs de deuxièmes espèce de grande taille. soit (X1p,..,xnp)=(19.5,22.1,21.5,21.9,22.0,21.0,22.3,21.0,20.3,20.9,22.0,22.0,23.8,21.2,21.0) sont les echantillons indepandant de Lp et soit

(x1g,..,xng)=(22.35,23.9,20.8,25.8,25.0,24.0,23.8,21.7,22.8,23.1,23.5,23.0,23.0,23.1) sont les echantillons indepandant de Lg $x_1 = \frac{1}{np}\sum_{i=1}^{n_p} x_i p$ $x_2 = \frac{1}{ng}\sum_{i=1}^{n_g} x_i g$

$$s1 = \sqrt{\frac{1}{n_p - 1}\sum_{i=1}^{n_p}(x_{p,i} - \bar{x}_1)^2}$$

$$s2 = \sqrt{\frac{1}{n_g - 1}\sum_{i=1}^{n_h}(x_{g,i} - \bar{x}_1)^2}$$

on consiedre donc : F $\sim F(\nu_1, \nu_2),\ (\nu_1, \nu_2) = \begin{cases} (n_p - 1, n_g - 1) & \text{si } s_1 > s_2, \\ (n_g - 1, n_p - 1) & \text{si } s_2 > s_1. \end{cases}$

$f_{\text{obs}} = \left(\max(s_1, s_2) \cdot \frac{\max(s_1,s_2)}{\min(s_1,s_2)}\right)^2$ et $p_{\text{value}} = 2P(F \geq f_{\text{obs}})$

les Tests : pour tester si les deux variances sont egaux. On considère le test bilaterale suivant :

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

On recourt à la commande : `var.test(Lp, Lg)$p.value` L'exécution renvoie :

p-value = 0.5213536 > $\alpha = 0.05$: Il y a non rejet de l'hypothèse nulle pour le seuil $\alpha = 0.05$.

On acceptera donc que :

$$\sigma_1^2 = \sigma_2^2 := \sigma^2$$

3.2.4 Question 3 :

Dans cette question, nous allons évaluer l'hypothèse selon laquelle l'oiseau ajuste

Modéle : soit (X1p,..,xnp)=(19.5,22.1,21.5,21.9,22.0,21.0,22.3,21.0,20.3,20.9,22.0,22.0, 23.8,21.2,21.0) sont les echantillons indepandant de Lp et soit

(x1g,..,xng)=(22.35,23.9,20.8,25.8,25.0,24.0,23.8,21.7,22.8,23.1,23.5,23.0,23.0,23.1) sont les echantillons indepandant de Lg

$x_1 = \frac{1}{np} \sum_{i=1}^{n_p} x_i p$ $x_2 = \frac{1}{ng} \sum_{i=1}^{n_g} x_i g$

$$s1 = \sqrt{\frac{1}{n_p - 1} \sum_{i=1}^{n_p} (x_{p,i} - \bar{x}_1)^2}$$

$$s2 = \sqrt{\frac{1}{n_g - 1} \sum_{i=1}^{n_h} (x_{g,i} - \bar{x}_1)^2}$$

on considere : $t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_p} + \frac{1}{n_g}}}$ et p_value= P(T tobs) avec T T (), = np + ng 2,

Les tests Pour ce faire, nous réalisons un test d'hypothèse bilatéral :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2$$

Et étant donné qu'il n'y a pas de différence significative entre les variances des deux groupes d'échantillons, la commande `t.test` est utilisée. Cette commande prend en arguments deux échantillons indépendants et renvoie la p-value, nous informant ainsi sur le degré de significativité du test.

La p-value de ce test est égale à 0.0002366, ce qui est nettement inférieur à 0.001. Par conséquent, le rejet de l'hypothèse nulle (H_0) est très significatif, avec une probabilité d'erreur de 5%. Ainsi, on peut conclure que la taille moyenne des œufs diffère en fonction de la taille du nid, indiquant que l'oiseau ajuste effectivement la taille de ses œufs en fonction de la dimension du nid dans lequel il pond.