

Fiche de Travaux Pratiques

Analyse et traitement des données Probabilités et statistiques

Master 1 Informatique Parcours « Données et Systèmes Connectés » (DSC) Saint-Étienne, France

Fabrice Muhlenbach

Laboratoire Hubert Curien, UMR CNRS 5516
Université Jean Monnet de Saint-Étienne
18 rue du Professeur Benoît Luras
42000 SAINT-ÉTIENNE, FRANCE
<http://perso.univ-st-etienne.fr/muhlfabr/>

Résultats attendus

L'objectif de cette séance de TP est de se familiariser avec les fonctions existant dans \mathbb{R} pour travailler avec les probabilités, les variables aléatoires et les distributions de probabilités.

1 Probabilités

Approche de la probabilité par la fréquence relative

Pile ou face ?

La probabilité qu'un événement E (éventuellement inconnu), noté par $P(E)$, est défini par la valeur approchée par la fréquence relative de l'occurrence de E dans une longue série d'essais d'une expérience aléatoire.

Avec \mathbb{R} , nous représenterons la fréquence relative des fois où l'on tombe sur « face » comme une fonction du nombre de lancers d'une pièce de monnaie non truquée. Avec \mathbb{R} , la fonction `runif` génère des valeurs aléatoires (cette fonction nécessite le chargement préalable du package `stats`), par exemple `runif(1)` donnera une valeur comprise entre 0 et 1. Avec une pièce non truquée, il y a 2 événements possibles : « pile » ou « face », ayant chacun la même chance de se produire $= \frac{1}{2}$.

Tout d'abord, afficher les résultats de 10 lancers aléatoires 0 ou 1 (nous conviendrons arbitrairement de considérer la valeur 0 pour l'événement « pile » et la valeur 1 pour l'événement « face ») en utilisant la fonction `runif` et la fonction `round` (qui permet d'arrondir une valeur décimale à l'entier le plus proche).

```
require(stats)
for (i in 1:10) {
```


```
x <- round(runif(1))
print(x)
}
```

Ensuite, représenter la fréquence relative du nombre de « face » comme une fonction du nombre de lancers de la pièce (allant de 1 à 1000 lancers). Pour cela, il faut utiliser un vecteur, ici appelé *proba*, et initialiser une variable de somme, appelée *sum*, à une valeur de 0. La valeur de *proba* prendra la valeur de la fréquence relative du nombre de fois où la pièce est tombée sur « face », et cette fréquence sera simplement calculée par le nombre de « face » (donné par *sum*) divisé par le nombre de lancers.

Enfin, représenter graphiquement la fréquence relative calculée en tant que graphe linéaire et ajouter une ligne avec la fonction **segments** à un niveau de 0,5 sur la fréquence relative des valeurs de « face ».

```
proba <- as.vector(1:1000)
sum <- 0
for (i in 1:1000) {
  x <- round(runif(1))
  sum <- sum + x
  proba[i] <- sum / i
}

plot(1:1000, proba[1:1000], "l", xlim=c(1,1000), ylim=c(0,1))
segments(0,0.5,1000,0.5, col="red")
```

Vous pouvez comparer vos résultats avec ceux de vos voisins ou relancer le traitement qui, avec ses valeurs aléatoires, produira un autre résultat. Modifiez votre code  pour avoir la représentation de la fréquence relative du nombre de « face » comme une fonction du nombre de lancers allant de 1 à 10 000 lancers de pièce (au lieu de 1000 lancers). Quelle est la différence entre le nouveau résultat et le précédent ?

2 Variables aléatoires et distributions de probabilité

2.1 Distribution de probabilité pour des variables aléatoires discrètes

Illustration avec un dé

Représenter par un histogramme la probabilité des résultats obtenu par un dé (non pipé).

La valeur $max = 6$ est le nombre maximal de points que l'on peut obtenir avec un dé ordinaire à 6 faces. Le nombre total d'événements est égal à 6 (les 6 faces du dé). Comme le dé n'est pas pipé, la probabilité d'avoir la valeur 1, 2, 3... ou 6 est la même, c'est-à-dire qu'elle est égale à $\frac{1}{6}$.

```
max <- 6
nb_events <- max
proba <- as.vector(1:max)
for (i in 1:max) {
  proba[i] <- 1/nb_events
}
```

Nous pouvons maintenant représenter par un graphique en barres la probabilité d'avoir chaque valeur d'un dé (allant de 1 à 6). Ces valeurs sont considérées comme des facteurs. Pour simplifier les instructions d'affichage, nous utiliserons la fonction `qplot` du package `ggplot2` afin de tracer ce graphique.

```
require( ggplot2 )
qplot( factor( 1:max ) , proba [ 1:max ] ,
      xlab=" Valeur_obtenue_avec_un_seul_de" ,
      ylab=" Probabilite " ) + geom_bar( stat=" identity " )
```


Remarque : dans ce qui suit, les codes seront donnés en anglais car il n'est pas possible de reprendre les caractères accentués du français (les copies de codes écrits en  comportant des caractères accentués sont mal interprétés par \LaTeX)

Illustration avec deux dés

Représenter par un histogramme de probabilité les résultats obtenus par la somme de deux dés (non pipés).

Avec deux dés, combien de possibilités avons-nous d'obtenir les valeurs suivantes :

- 1?
- 2?
- 3?...

Quel est le nombre total d'événements ?

Quel est le score maximal que l'on peut obtenir avec deux dés ?

```
max_dice <- 6
score_max <- max_dice * 2
nb_events <- max_dice ^ 2

proba <- rep.int( 0 , score_max )
proba <- as.vector( proba )

for ( i in 1:max_dice ) {
  for ( j in 1:max_dice ) {
    proba[ i+j ] <- proba[ i+j ] + 1/nb_events
  }
}

qplot( factor( 1: score_max ) , proba [ 1: score_max ] ,
      xlab=" Value_obtained_with_the_sum_of_two_dice " ,
      ylab=" Probability " ) + geom_bar( stat=" identity " )
```

2.1.1 Espérance d'une variable aléatoire discrète

L'espérance (valeur moyenne) d'une variable aléatoire discrète X , notée $E(X)$, est définie ainsi :

$$E(X) = \sum_{\text{toutes les valeurs possibles de } x} x \times P(x)$$

Calculer la valeur de l'espérance (en anglais : *expectation*) de la somme de deux dés (non pipés).

```
expec <- 0
for (i in 1:score_max) {
  expec <- expec + proba[i] * i
}
print(expec)
```

Quelle est cette espérance ? Est-ce que cette valeur est cohérente avec le résultat de l'histogramme de probabilité tracé auparavant ?

2.1.2 Variance et écart-type d'une variable discrète

La variance d'une variable aléatoire discrète X , notée par σ_X^2 ou $V(X)$, est définie ainsi :

$$\sigma_X^2 = V(X) = \sum_{\text{toutes les valeurs possibles de } x} [x - E(X)]^2 \times P(x)$$

L'écart-type (en anglais : *standard deviation*) de X est $\sigma_X = \sqrt{\sigma_X^2}$.

Calculer la variance et l'écart-type de la somme de deux dés (non pipés).

```
variance <- 0
for (i in 1:score_max) {
  variance <- variance + (expec - i)^2 * proba[i]
}
print(variance)
st_deviation <- sqrt(variance)
print(st_deviation)
```


Il existe une autre manière de calculer la variance : $V(X) = E(X^2) - E^2(X)$

```
expec_square <- 0
for (i in 1:score_max) {
  expec_square <- expec_square + proba[i] * i^2
}
variance_bis <- expec_square - (expec^2)
print(variance_bis)
```

Le plus souvent, le dénominateur $n - 1$ est utilisé au lieu de n pour donner un estimateur non biaisé de la (co-)variance d'observations de variables indépendantes et identiquement distribuées (i.i.d.).

```
u_variance <- variance * (nb_events / (nb_events - 1))
```

2.1.3 Fonctions statistiques de R


 est un langage spécifiquement conçu pour l'analyse de données et les statistiques. Nous pouvons travailler avec un vecteur *sum* enregistrant les différentes valeurs de somme de deux dés et nous pouvons appliquer directement les fonctions statistiques *mean*, *var* et *sd* pour obtenir les valeurs respectives de l'espérance (la moyenne), la variance et l'écart-type. En outre, nous pouvons représenter l'histogramme avec la fonction *hist*. Notons que l'écart-type et la variance calculés sont par défaut des estimateurs non biaisés.

```
sum <- rep.int(0,nb_events)
sum <- as.vector(sum)
event <- 0

for (i in 1:max_dice) {
  for (j in 1:max_dice) {
    event <- event + 1
    sum[event] <- i+j
  }
}

hist(sum, breaks = c(1:12), col = "blue1")
mean(sum)
var(sum)
sd(sum)
```

2.2 Distribution de probabilité pour des variables aléatoires continues

Durant cette séance de TP, nous avons vu qu'il est important de pouvoir comparer les différentes variables d'un jeu de données. Ceci peut être fait au moyen d'une matrice de graphiques avec  avec la fonction *pairs* qui effectue des comparaisons par paire.

Charger le jeu de données *iris* (mesures en centimètre des variables correspondant aux longueurs et largeurs des sépales et pétales de 3 espèces d'iris) et afficher le résumé (*summary*) des données.

```
data("iris")
summary(iris)
```

Afficher ensuite la représentation des comparaisons par paire des 4 premières variables continues, ainsi que le coefficient de corrélation entre toutes les variables à part la 5^{ème} qui représente la classe :

```
pairs(iris[-5], bg=iris$Species, pch=21)
cor(iris[-5])
```

Charger le package *PerformanceAnalytics*. Si ce package n'est pas encore présent dans votre liste de package, il faut l'installer au préalable.

```
install.packages("PerformanceAnalytics",
```

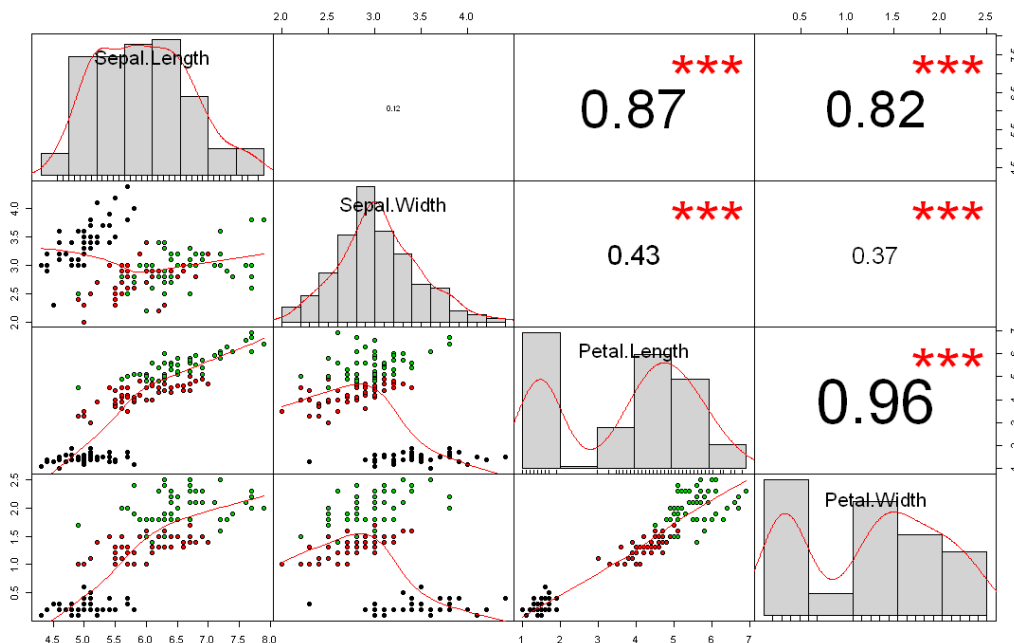
```
lib="U:/My_R_personal_library")
library(PerformanceAnalytics)
```

Ensuite, afficher le diagramme de corrélation du jeu de données des iris :

```
chart.Correlation(iris[-5], bg=iris$Species, pch=21)
```


Ce diagramme comporte de nombreuses informations :

- en diagonale, il y a les distributions unimodales, affichées comme des histogrammes et des graphiques de densité à noyau ;
- à la droite de la diagonale, il y a les valeurs des corrélations par paire entre les variables, avec le niveau de significativité de la corrélation représenté par un nombre d'étoiles en rouge ;
- plus le coefficient de corrélation est élevé, plus grande est la taille de la police de caractères dans laquelle est écrit le coefficient de corrélation ;
- à gauche de la diagonale, il y a la matrice de nuages de points, avec une ligne rouge destinée à faciliter l'interprétation de la manière dont les deux variables représentées sont reliées entre elles.



Ce graphique combine un grand nombre d'informations en une seule représentation graphique et est simple à afficher.

2.3 Loi normale $\mathcal{N}(\mu, \sigma)$

Il existe différentes lois de probabilité ayant leurs traductions avec des fonctions spécifiques dans  :

- la loi **binomiale** d'une variable X , qui est définie par X = nombre de succès observés sur n essais. Ainsi, si X suit une variable binomiale (notée $X \equiv \mathcal{B}(n, p)$), les fonctions pertinentes dans \mathbb{R} pour une variable aléatoire X binomialement distribuée sur n essais avec une probabilité de succès p sont :
 - `dbinom(x, n, p)`, pour trouver $P(X = x)$
 - `pbinom(x, n, p)`, pour trouver $P(X \leq x)$
 - `qbinom(q, n, p)`, pour trouver c tel que $P(X \leq c) = q$
 - `rbinom(n, x, p)`, pour générer n valeurs indépendantes de X
- la loi **géométrique** $\mathcal{G}(p)$ d'une variable X , qui est définie comme le nombre d'essais jusqu'à ce que le premier succès soit rencontré (incluant l'essai réussi). Dans \mathbb{R} , `dgeom(x, p)` donne la densité, `pgeom(q, p)` donne la fonction de distribution et `rgeom(n, p)` génère de manière aléatoire n valeurs suivant cette distribution géométrique ;
- la distribution de Poisson $\mathcal{P}(\lambda)$ est une probabilité de distribution discrète qui exprime la probabilité qu'un nombre donné d'événements se produise dans un intervalle donné de temps ou sur un espace donné avec la connaissance d'un taux moyen et indépendant du temps ou du dernier événement. Avec \mathbb{R} , `dpois($x, lambda$)` donne la densité, `ppois($q, lambda$)` donne la fonction de distribution, `qpois($p, lambda$)` donne la fonction quantile et `rpois($n, lambda$)` produit la génération aléatoire d'une distribution de Poisson de paramètre $lambda = \lambda$.

Dans la suite, nous verrons moins l'utilisation de ces trois lois que l'emploi de la loi normale. Une loi normale $\mathcal{N}(\mu, \sigma)$ est une courbe en cloche symétrique. Elle est caractérisée par la moyenne μ et l'écart-type σ . La valeur μ décrit la valeur où la courbe est centrée et σ décrit l'ouverture de la courbe autour du centre.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Dans \mathbb{R} , il existe un grand nombre de fonctions permettant de travailler avec une loi normale. La densité, la fonction de distribution, la fonction quantile et la fonction de génération aléatoire d'une loi normale sont facilement calculées avec la moyenne (μ) égale au paramètre *mean* et l'écart-type (σ) égal au paramètre *sd*.

La fonction `dnorm` donne la densité, `pnorm` donne la fonction de distribution, `qnorm` donne la fonction quantile et `rnorm` génère des valeurs aléatoires suivant cette loi normale.

Si nous considérons une loi normale avec une moyenne $\mu = 0$ et un écart-type $\sigma = 1$, nous avons une fonction de densité de probabilité appelée « distribution normale standardisée ». Cette distribution normale peut être tracée dans \mathbb{R} ainsi (en gros trait de couleur bleue) :

```
x=seq(-4,4,length=200)
y=1/sqrt(2*pi)*exp(-x^2/2)
plot(x,y,type="l",lwd=5,col="blue")
```

Au lieu de la fonction mathématique exacte classique de la loi normale, nous pouvons tracer plus simplement cette fonction grâce à l'instruction `dnorm` (en courbe plus fine, en jaune). Nous attribuons à l'instruction `plot` la valeur `TRUE` de manière à ce que le nouveau tracé n'efface pas le précédent.

```
par(new = TRUE)
y=dnorm(x,mean=0,sd=1)
plot(x,y,type="l",lwd=1,col="yellow")
```

Afficher sur le même graphique trois différentes fonctions de densité de probabilité, la première pour $\mu = 10$ et $\sigma = 5$ (en rouge), la seconde pour $\mu = 40$ et $\sigma = 2,5$ (en vert) and la troisième pour $\mu = 70$ et $\sigma = 10$ (en bleu).

```
x <- seq(-10,100,.1)
normdensity1 <- dnorm(x,mean=10,sd=5)

normdensity2 <- dnorm(x,mean=40,sd=2.5)

normdensity3 <- dnorm(x,mean=70,sd=10)

plot(x, normdensity1, type="l", col="red",
      ylim=range(c(normdensity1, normdensity2, normdensity3)))
par(new = TRUE)
plot(x, normdensity2, type="l", col="green",
      ylim=range(c(normdensity1, normdensity2, normdensity3)),
      axes = FALSE, xlab = "", ylab = "")
par(new = TRUE)
plot(x, normdensity3, type="l", col="blue",
      ylim=range(c(normdensity1, normdensity2, normdensity3)),
      axes = FALSE, xlab = "", ylab = "")
```

68%, 95,7% and 99,7%

Examinons la probabilité qu'un nombre sélectionné aléatoirement d'une distribution normale standardisée ait l'occasion de se produire avec un écart-type autour de la moyenne. Cette probabilité est représentée par l'aire sous la courbe de la loi normale entre les valeurs en abscisse $x = -1$ et $x = 1$. Cette aire sera colorée en gris avec la fonction `polygon` :

```
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=2, col="blue")
x=seq(-1,1,length=100)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```

Pour trouver la densité de probabilité représentée par l'aire comprise entre $x = -1$ et $x = 1$, nous pouvons soustraire à l'aire située à gauche de $x = -1$ de l'aire située à gauche de $x = 1$:

```
pnorm(1,mean=0,sd=1)-pnorm(-1,mean=0,sd=1)
```

Nous pouvons agir de même avec une aire comprise entre deux écarts-types et trois écarts-types :

```
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=2, col="blue")
```



```
x=seq(-2,2,length=200)
y=dnorm(x)
polygon(c(-2,x,2),c(0,y,0),col="gray")
pnorm(2,mean=0,sd=1)-pnorm(-2,mean=0,sd=1)
pnorm(3,mean=0,sd=1)-pnorm(-3,mean=0,sd=1)
```

Problème

Une distribution suivant une loi normale avec une moyenne de 3500 grammes et un écart-type de 600 grammes est un bon modèle de probabilité de distribution de la variable continue X représentant le poids de naissance d'un nouveau-né humain né à terme [Exercice du cours].

Question 1

Quelle est la proportion des poids de naissance qui sont compris entre 2900 and 4700 grammes ?

Solution

```
pnorm(4700,mean=3500,sd=600)-pnorm(2900,mean=3500,sd=600)
```

Question 2

Quel est le poids de naissance w qui ne se produit que dans 2,5% des cas ?


Solution

```
qnorm(1 - 2.5 / 100, mean = 3500, sd = 600)
```

Une autre solution peut être obtenue en indiquant que l'on ne s'intéresse qu'à la queue de distribution supérieure (*upper tail*), ou l'aile supérieure, de la distribution de la loi normale :

```
qnorm(2.5 / 100, mean = 3500, sd = 600, lower.tail = FALSE)
```

Réponse

Les réponses sont les résultats calculés par . Vous pouvez comparer vos résultats aux valeurs présentées dans le cours. De plus, vous trouverez ci-dessous une solution graphique à la question 2.

```
x <- seq(0, 7000, length=200)
y <- dnorm(x, mean= 3500, sd = 600)
# same as y <- dnorm((x - 3500) / 600)

plot(x, y, type="l", lwd=2, col="gray")

z <- qnorm(2.5 / 100, mean = 3500, sd = 600, lower.tail = FALSE)
```

```
x <- seq(z, 7000, length=100)
y <- dnorm(x, mean= 3500, sd = 600)

polygon(c(z, x, 7000), c(0, y, 0), col="navy")

mtext("What birth weight is exceeded only in 2.5% of the cases?",
      side=1, adj=1, col="navy")
```

Exercices sur les tests d'hypothèse

Test unilatéral sur la moyenne de la population avec variance connue

Supposons qu'un fabricant d'ampoules électriques affirme que la durée de vie moyenne d'une ampoule est supérieure à 10 000 heures. Dans un échantillon de 30 ampoules, il a été constaté qu'elles ne durent en moyenne que 9 900 heures. Supposons que l'écart-type de la population est de 120 heures. Au niveau de signification de 0,05, pouvons-nous rejeter l'affirmation du fabricant ?

Réponse

Hypothèse nulle : $H_0 : \mu \geq 10000$

On calcule ensuite la statistique de test $z = \sqrt{n} \left(\frac{\bar{X} - \mu_0}{\sigma} \right)$:

```
xbar = 9900          # moyenne simple sur l'échantillon
mu0 = 10000          # valeur de la moyenne
sigma = 120          # écart-type sur la population
n = 30               # taille de l'échantillon
z = (xbar - mu0)/(sigma/sqrt(n))
z                    # statistique de test
```

La valeur de z est -4.564355 .

On calcule la valeur critique à 0,05 :

```
alpha = .05
z.alpha = qnorm(1 - alpha)
-z.alpha          # valeur critique
```

La statistique de test $-4,5644$ est inférieure à la valeur critique de $-1,6449$. Par conséquent, au niveau de signification de 0,05, nous rejetons l'affirmation H_0 selon laquelle la durée de vie moyenne d'une ampoule est supérieure à 10 000 heures.

Test bilatéral sur la moyenne de la population avec variance connue

Supposons que le poids moyen des manchots royaux trouvé dans une colonie en Antarctique l'année dernière était de 15,4 kg. Dans un échantillon de 35 manchots à la même époque cette année dans la même colonie, le poids moyen des pingouins est de 14,6 kg. Supposons que l'écart-type de la population est de 2,5 kg. Au niveau de signification de 0,05, pouvons-nous rejeter l'hypothèse nulle selon laquelle le poids moyen des manchots ne diffère pas de l'année dernière ?

Réponse

Hypothèse nulle : $H_0 : \mu = 15.4$

Hypothèse alternative : $H_1 : \mu \neq \mu_0 = 15.4$

On calcule ensuite la statistique de test $z = \sqrt{n} \left(\frac{\bar{X} - \mu_0}{\sigma} \right)$:

```
xbar = 14.6          # moyenne simple sur l'échantillon
mu0 = 15.4           # valeur de la moyenne
sigma = 2.5          # écart-type sur la population
n = 35               # taille de l'échantillon
z = (xbar - mu0)/(sigma/sqrt(n))
z                    # statistique de test
```

La valeur de z est -1.893146 .

On calcule les valeurs critiques à 0,05 :

```
alpha = .05
z.half.alpha = qnorm(1 - alpha/2)
c(-z.half.alpha, z.half.alpha)
```

Les résultats donnés sont -1.9600 et 1.9600 .

La statistique de test -1.8931 se situe entre les valeurs critiques -1.9600 et 1.9600 . Par conséquent, au niveau de signification de 0,05, nous ne rejetons pas l'hypothèse nulle H_0 selon laquelle le poids moyen des manchots de cette année ne diffère pas de l'année dernière.

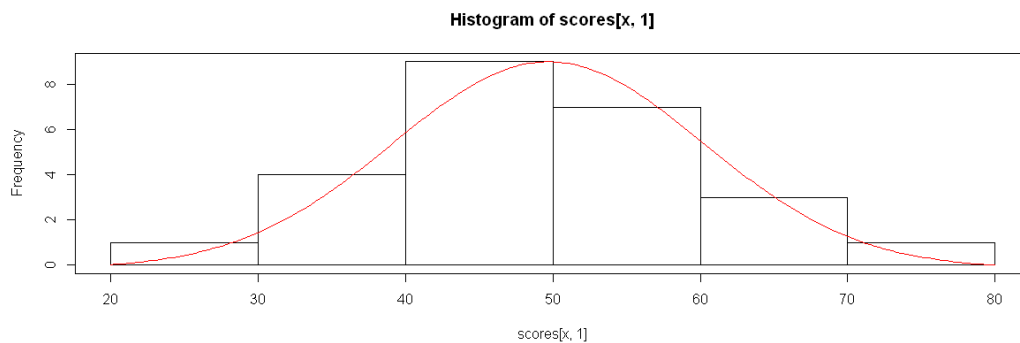
Exemple

Importer dans RStudio les résultats (*scores*) de 25 étudiants à un examen (avec une note sur une échelle de 0 à 100 points). Les données sont : 39 41 47 58 65 37 37 49 56 59 62 36 48 52 64 29 44 43 41 52 57 54 72 50 et 50. Vous pouvez télécharger le fichier *scores.txt* depuis la plate-forme de cours *Claroline*. À partir du moment où les données sont importées, celles-ci seront stockées dans un data frame appelé *scores* comportant 25 observations et 1 variable.

Tracer les résultats de manière linéaire (un nuage de point en fonction de l'ordre des élèves), puis un histogramme de ces notes. Calculer la moyenne et l'écart-type, puis, sur le même graphique, tracer la distribution suivant la loi normale qui approxime les données réelles des résultats (sans avoir de nom pour les axes des abscisses et des ordonnées afin de ne pas écraser les noms des axes du graphique précédent).

```
x <- seq(1:25)
plot(x, scores[x,1])
hist(scores[x,1])
mu <- mean(scores[,1])
sigma <- sd(scores[,1])
x2=seq(20,80,length=200)
par(new = TRUE)
y=dnorm(x2,mean=mu,sd=sigma)

plot(x2,y,type="l",lwd=1,col="red", xlab="", ylab="", yaxt="n")
```



Suivant cette approximation normale, combien y a-t-il d'étudiants que vous vous attendez à trouver ayant une note supérieure ou égale à 55, 60, 65 ou 70? Combien y en a-t-il en réalité?

```
# How many students do you expect to find with a score
# equal or greater than 55?

# Answer by using the normal distribution:
# 25 is the number of students

round(25 * (pnorm(55, mean=mu, sd=sigma, lower.tail=FALSE)))

# By counting the real values:

nb_scores <- 0
for (i in 1:25)
{
  if (scores[i,1] >= 55)
    nb_scores <- nb_scores + 1
}
print(nb_scores)
```

Exercice

En Travaux Dirigés, nous nous sommes intéressés à l'événement « nombre de fois où la pièce apparaît sur “pile” dans un ensemble de 10 lancers ». Les pièces de monnaie utilisées lors des lancers étant supposées non truquées, la probabilité de tomber sur « pile » est de 1/2.

En TD, sur nos 102 essais de 10 lancers de pièces, nous avons obtenu les résultats suivants :

Nb « pile »	0	1	2	3	4	5	6	7	8	9	10
Cas rencontrés	0	3	8	12	18	27	17	9	6	2	0

Question initiale

Combien y a-t-il de modalités différentes pour cet événement « nombre de fois où la pièce apparaît sur “pile” dans un ensemble de 10 lancers »?

Questions complémentaires et ébauches de solutions

Combien y a-t-il de possibilités de n'avoir aucune fois « pile » sur 10 lancers ?

Ici, la seule situation possible est lorsque tous les lancers donnent « face » : F-F-F-F-F-F-F-F-F, ce qui donne une probabilité de n'avoir aucun « pile » égale à $\left(\frac{1}{2}\right)^{10} \approx 0,1 \%$.

Combien y a-t-il de possibilités de faire une seule fois « pile » sur 10 lancers ?

Par rapport à la situation précédente, il y a 10 combinaisons possibles : P-F-F-F-F-F-F-F-F (« pile » dès le premier lancer et ensuite que des lancers avec « face »), F-P-F-F-F-F-F-F-F (« pile » au deuxième lancer), ..., F-F-F-F-F-F-F-P-F (« pile » à l'avant-dernier lancer) et F-F-F-F-F-F-F-F-P (« pile » au dernier lancer).

La probabilité d'avoir une fois « pile » est donc $10 \times \left(\frac{1}{2}\right)^{10} \approx 0,98 \%$.

Pour tomber 2 fois sur « pile », il y a 10 positions possibles pour le premier lancer « pile », mais 9 positions possibles pour le second lancer « pile », à diviser par deux : P-P-F-F-F-F-F-F-F, P-F-P-F-F-F-F-F-F, ..., F-F-F-F-F-F-F-P-P, soit $10 \times 9/2 = 45$.

La probabilité d'avoir deux fois « pile » est donc de $4,39 \%$.

Rappel

Soit E un ensemble fini de cardinal n et k un entier naturel.

Une k -combinaison de E se calcule ainsi :

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Travail demandé :

- Écrire une formule ou une fonction qui calcule les combinaisons dans \mathbb{R} en utilisant la fonction `factorial` pour le calcul de la factorielle.
- Créer un data frame comportant deux variables : le nombre de « pile » et la probabilité associée à cet événement (vous aurez besoin d'utiliser votre fonction de calcul des combinaisons).
- Vérifier que la somme des probabilités associées vaut bien 1.
- Faire une représentation graphique de cette distribution.
- Calculer la moyenne des valeurs en utilisant les valeurs du data frame. Deux manières sont possibles :
 1. faire la somme du nombre de « pile » multiplié par la probabilité associée à chaque événement ;
 2. diviser la somme du nombre de « pile » par le nombre de modalités différentes de l'événement « nombre de fois où la pièce apparaît sur «pile» dans un ensemble de 10 lancers ».
- Calculer l'écart-type :
 1. initialiser une variable faisant la somme des carrés à zéro ;
 2. ajouter à cette variable «somme des carrés» le produit de la probabilité associée à chaque événement par le carré de la différence entre le nombre de « pile » de chaque événement et la moyenne du nombre de « pile » ;
 3. extraire la racine carrée de la variance (la somme des carrés des différences pondérées) pour obtenir l'écart-type.
- Représenter sur le même graphique la fonction théorique d'une loi normale avec la moyenne et l'écart-type calculés. Que peut-on en conclure ?
- Comparer avec les valeurs obtenus en TD. Que peut-on en conclure à nouveau ?

3 Lectures conseillées

Quelques propositions de lectures en rapport avec ce TP :

- Adler (2010), “R in a Nutshell – a Desktop Quick Reference,” Chapter 17 “Probability Distributions.”
- Baayen (2008), “Analyzing Linguistic Data : A Practical Introduction to Statistics using R,” Chapter 3 “Probability distributions.”
- Chihara and Hesterberg (2011), “Mathematical Statistics with Resampling and R.”
- Cohen and Cohen (2008), “Statistics and Data with R : An Applied Approach Through Examples,” Part II “Probability, densities and distributions.”
- Crawley (2005), “Statistics : An Introduction using R,” Chapter 5 “Single Samples.”
- Dalgaard (2008), “Introductory Statistics with R,” Chapter 3 “Probability and distributions.”
- Peck et al. (2012), “Introduction to Statistics and Data Analysis,” Chapter 6 “Probability” and Chapter 7 “Random Variables and Probability Distributions.”
- Teetor (2011), “R Cookbook”, Chapter 8 “Probability,” Chapter 9 “General Statistics.”
- Venables et al. (2013), “An Introduction to R,” Chapter 8 “Probability distributions.”

Références

- Adler, J. (2010). *R in a Nutshell – a Desktop Quick Reference*. O’Reilly.
- Baayen, R. H. (2008). *Analyzing Linguistic Data : A Practical Introduction to Statistics using R*. Cambridge University Press.
- Chihara, L. M. and T. C. Hesterberg (2011). *Mathematical Statistics with Resampling and R*. John Wiley & Sons, Ltd.
- Cohen, Y. and J. Y. Cohen (2008). *Statistics and Data with R : An Applied Approach Through Examples*. John Wiley & Sons, Ltd.
- Crawley, M. J. (2005). *Statistics : An Introduction using R*. John Wiley & Sons, Ltd.
- Dalgaard, P. (2008). *Introductory Statistics with R* (2nd ed.). Springer.
- Peck, R., C. Olsen, and J. L. Devore (2012). *Introduction to Statistics and Data Analysis* (4th ed.). Brooks / Cole, Cengage Learning.
- Teetor, P. (2011). *R Cookbook*. O’Reilly.
- Venables, W. N., D. M. Smith, and the R Core Team (2013). An introduction to R –notes on R : A programming environment for data analysis and graphics.
URL <http://cran.r-project.org/doc/manuals/R-intro.html>.