TP01- DEMARRER AVEC R

Objectifs : installer les logiciels, démarrer avec R

PRELIMINAIRES

1. Commencez par installer les logiciels R et Rstudio, en suivant les liens sur https://www.r-project.org et https://www.rstudio.com.

2. Installer le package *survival*

La **fenêtre de script** affiche les commandes que vous exécutez au moyen du menu dans le langage de programmation de R : elles peuvent être modifiées à la main et exécutées à nouveau pour en personnaliser certains aspects. L'intégralité des commandes exécutées (le script) peut être enregistré sous la forme d'un fichier texte.

La **fenêtre de sortie** permet de visualiser les résultats des commandes demandées. Les sorties peuvent aussi être enregistrées sous la forme d'un fichier texte.

La **fenêtre de messages** gère l'affichage des messages du programme : elle est notamment utile en cas d'erreur car elle donne des indications sur la cause de celle-ci.

Le bouton "Données" précise quel est le jeu de données courant. Pour visualiser l'ensemble des jeux de données disponibles dans l'espace de travail et pour en changer, il suffit d'appuyer sur ce bouton. L'enregistrement de l'espace de travail enregistre tous les jeux de données et toutes les variables créées durant la session.

Sauvegarder le script au format R

La commande "Fichier->Sauver le script ..." permet de sauvegarder le script au format texte. Sauvegarder le script sous le nom TP01_NomPrenom.r

Ouvrir et exécuter un script

Le menu "Fichier->Ouvrir un script" permet d'ouvrir un script enregistre préalablement dans la fenêtre de script. Pour l'exécuter, il suffit de sélectionner les lignes `à exécuter et d'appuyer sur le bouton "Soumettre" (à droite).

COMMANDES USUELLES SOUS R

Calculs simples

```
10*9*8*7*6*5*4*3*2*1
log(2)
log(2)/log(10)
```

```
?log
log10(2)
log(2,base=10)
```

Vecteurs

```
x<-c(1,4,5)
x
y<-seq(1,9,2) # ou bien :
y<-2*(1:5)-1
y
y[2]
```

```
y[-2]
xy<-y[c(1,4,5)]# ou puisque x=c(1,4,5)
xy<-y[x]
xy
y[2:4]
```

```
y+1
y[1:4]+1
# ou une écriture équivalente
```

```
(y+1)[1:4]
x<-2:4
y*x
```

```
vec<-rnorm(10)
vec
length(vec[vec>0])
```

```
yvec<-log(vec)
vec2<-yvec[!is.na(yvec)]
length(vec2)
```

Tableaux

```
mat<-matrix(1:15,ncol=5,byrow=T)
mat
mat[2:3,c(2,4)]
```

```
mat[mat[,1]<3,1]
mat[mat[,1]<3,]
```

1.1.Exercice (Prise en main)

1. Ouvrez un nouveau script. Enregistrez votre fichier sous TP01_NomPrenom.r.

2. Créez le vecteur (1, 2, 3, 4, 5).

Assigner le vecteur précédent à X.

Vérifiez le contenu de X.

3. Créez le vecteur Y avec les valeurs (1, 4, 9, 16, 25).

4. Vérifiez que X et Y ont la même longueur.

5. Tracer les points définis par les deux vecteurs X et Y par *plot(X,Y)*. Modifier le symbole : *pch=2*, puis *pch=3*, etc. Changez le type : *type="b"*, puis *type="l"*.

Que font les option *pch=* et *type= ?*

Changez la couleur : *col="red"*, puis *col="blue"*, etc. Ajouter un titre, ajouter des étiquettes sur les deux axes (options : *main* = "title",*xlab* = "xlabel",*ylab*= "ylabel").

6. Additionner la courbe $y = x^2$ par *curve(x^2,add=TRUE)*.

1.2.Exercice (Prise en main)

1. Créez le vecteur X contenant tous les entiers de 0 à 7. 2. Multipliez X par 5, divisez-le par 5, ajoutez-y 5.

3. Calculez la somme de X, ses sommes cumulées. 4. Calculez la racine carrée de X, sa troisième puissance.

1.3.Exercice (Prise en main)

1. Créez le vecteur X contenant (0, 1, 4, 9, 16).

Extraire de X le sous-vecteur avec les indices 3 et 5.

Extraire toutes les valeurs supérieures à 2.

Extraire toutes les valeurs supérieures à 2 et inférieures à 10.

2. Créez le vecteur Y contenant 5 colonnes de 1 (*rep(1,5)*).

Créez le vecteur Z contenant la séquence de 3 à 11 avec un pas de 2 (*seq(3,11,by=2)*).

Concaténez X, Y, Z.

Reliez-les en rangées. (commande *rbind()*)

Reliez-les en colonnes et affectez le résultat à XYZ. (commande *cbind()*)

3. Calculez les sommes des lignes et des colonnes de XYZ.

4. Extraire de XYZ :

a) la ligne numéro 4 b) les numéros de la colonne 3

c) lignes avec indices 3,5 et les colonnes avec indices 2,3 d) les lignes de telle sorte que X soit supérieur à 2.

e) les colonnes intitulées "Y" et "Z".

1.4.Exercice (table de la loi normale centrée-réduite)

Nous allons utiliser la fonction *integrate()* pour construire la table statistique des probabilités de la loi N(0,1).

Notons $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ la fonction de répartition de la loi N(0, 1).

Il est bien connu que $\Phi(-x)=1-\Phi(x)$. Par conséquent, on peut se contenter de construire la table pour les valeurs positives de x.

1. Programmez la fonction *phi()* qui prend un vecteur x de longueur n comme paramètre d'entrée et renvoie le vecteur des valeurs $\frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$, i = 1, ..., n.

2. Utilisez la fonction *integrate()* pour calculer $\Phi(x)$ pour tous les x du vecteur suivant :

quantiles <- seq(0,5.5,by=0.1). Vous les stockerez dans un vecteur nommé *probs*.

3. Comparez, au moyen de la fonction *all.equal()*, les résultats que vous avez obtenus avec ceux donnés par la fonction *pnorm()*.

4. Tracez la courbe des valeurs de $\Phi(x)$ pour tous les x et tous les $-x$ dans *quantiles* (fonction *rev()*).

5. Ajoutez à ce graphique, en bleu, la courbe de la fonction *pnorm()*. Vérifiez que ces deux courbes sont parfaitement superposées.

TP02- STATISTIQUES UNIVARIEES AVEC R

Objectifs : Maitriser les commandes basiques en relation avec une variable aléatoire ou un échantillon

Exercice

0. Pre-charger les données

```
install.packages("questionr")
library(questionr)
data("hdv2003")
d <- hdv2003

#sinon
d <- read.csv("C:/Users/Spinel/Desktop/Dataset_TPStatUnivariees.csv")
```

Variable quantitative

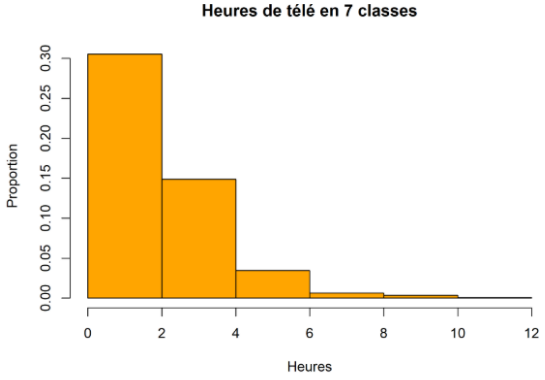
1. Déterminer les caractéristiques suivantes de la variable *d$heures.tv* : moyenne, ecart-type, minimum, maximum, étendue, médiane, quartiles, 1er décile, et 9ème decile : en utilisant les fonctions *mean (), sd (), min (), max () range (), median(), quantile(), summary()*. Comme il y a des données manquantes, utilisé l'option *na.rm = TRUE* si nécessaire.

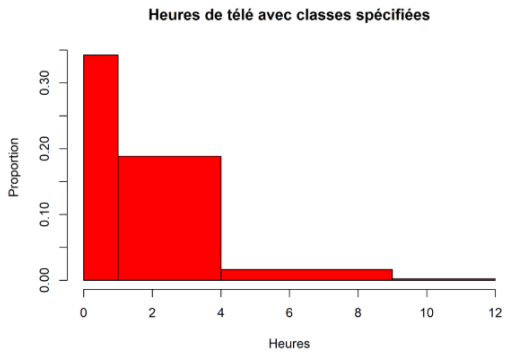
2. Tout cela est bien pratique, mais pour pouvoir observer la distribution des valeurs d'une variable quantitative, il n'y a quand même rien de mieux qu'un graphique. Générer un histogramme de la répartition des valeurs. Celui-ci peut être généré très facilement avec la fonction *hist* auquel vous appliquerez les options Ici, les options *main, xlab et ylab* permettent de personnaliser le titre du graphique, ainsi que les étiquettes des axes.

De nombreuses autres options existent pour personnaliser l'histogramme, parmi celles-ci on notera :

- *probability* si elle vaut TRUE, l'histogramme indique la proportion des classes de valeurs au lieu des effectifs.
- *breaks* permet de contrôler les classes de valeurs. On peut lui passer un chiffre, qui indiquera alors le nombre de classes, un vecteur, qui indique alors les limites des différentes classes, ou encore une chaîne de caractère ou une fonction indiquant comment les classes doivent être calculées.
- *col* la couleur de l'histogramme

Réaliser les histogrammes suivants :





2. Densité et répartition cumulée

Représenter la fonction de repartition de cet échantillon au moyen de la fonction *density* et *plot*. Pour la fonction densité, utiliser la fonction *na.rm = TRUE* indique que l'on souhaite retirer les valeurs manquantes avant de calculer cette courbe de densité.

Calculer la fonction de répartition empirique ou *empirical cumulative distribution function* en anglais avec la fonction *ecdf*. Représenter le résultat via la fonction *plot*.

3. Les boites à moustaches

Les boîtes à moustaches, ou boxplots en anglais, sont une autre représentation graphique de la répartition des valeurs d'une variable quantitative. Elles sont particulièrement utiles pour comparer les distributions de plusieurs variables ou d'une même variable entre différents groupes, mais peuvent aussi être utilisées pour représenter la dispersion d'une unique variable.

Représenter la boite à moustache de la série étudiée (avec titre et labels).

Variable qualitative

4. Tris à plat

Parmi nos enquêtés combien trouve-t-on de femmes et d'homme ? Utiliser la fonction *table* appliquée aux données de sexe *d$sexe*

Quand le nombre de modalités est élevé, on peut ordonner le tri à plat selon les effectifs à l'aide de la fonction sort. Combien y a-t-il d'occupation différentes ?

Classer par du plus petit effectif au plus grand effectifs les différentes occupations.

Classer par du plus grand effectif au plus petit effectifs les différentes occupations. (option : *decreasing = TRUE*))

À noter que la fonction table exclut par défaut les non-réponses du tableau résultat. L'argument *useNA* de cette fonction permet de modifier ce comportement :

- useNA="no" (valeur par défaut), les valeurs manquantes ne sont jamais incluses dans le tri à plat
- useNA="ifany", une colonne NA est ajoutée si des valeurs manquantes sont présentes dans les données
- useNA="always", une colonne NA est toujours ajoutée, même s'il n'y a pas de valeurs manquantes dans les données.

Pour la colonne satisfait *d$trav.satisf* avec leur travail, combien de personnes ne se déclarent pas satisfaites ?

Que fait la commande *summary* appliquée *d$trav.satisf*

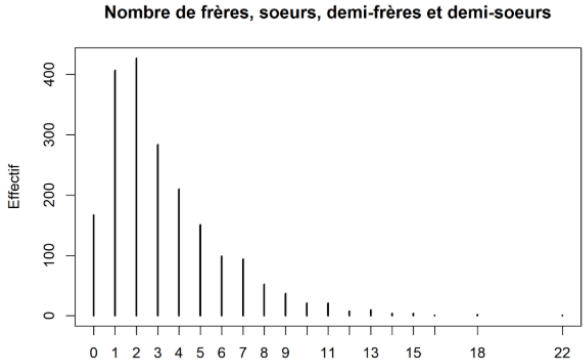
Fréquence

5. Obtenir un tableau des fréquences du type de travail *d$qualif* avec la fonction *freq.* Interpréter les colonnes % et *val%* ?

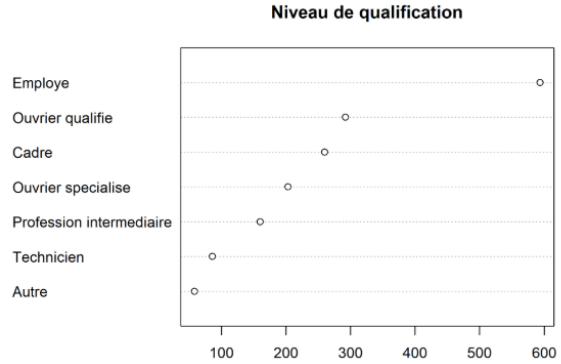
Il est possible d'ajouter des options à la fonction *freq*. Taper la commande suivante et interpréter son résultat et chaque option introduite.

```
freq(d$qualif, cum = TRUE, total = TRUE, sort = "inc", digits = 2, exclude = NA)
```

6. Pour les variables qualitatives, les diagrammes en bâtons sont utilisés automatiquement par R lorsqu'on applique la fonction générique *plot* à un tri à plat obtenu avec *table*. En choisissant les données représentant le nombre de frères et de sœur, reproduire le graphique suivant :



Pour les autres types de variables qualitatives, on privilégiera les diagrammes de Cleveland, obtenus avec la fonction *dotchart*. On doit appliquer cette fonction au tri à plat de la variable, obtenu avec *table*. Pour des raisons interne a R, il est conseiller d'appliquer *as.matrix* à la table obtenue via *table*. Celle-ci converti la table en une matrice. Reproduire le graphique suivant pour les données *d$qualif*, utiliser également la fonction *sort*



Changer le marqueur o (« rond ») pour un triangle avec l'option *pch*

TP03- LOI DISCRETES (BINOMIALE ET POISSON)

Objectifs : notion de probabilité d'un événement; calculer les probabilités et les quantiles à partir d'une distribution binomiale ou de Poisson;

PRELIMINAIRES

Soient $(X_1, X_2, \cdots, X_n)_{n \in \mathbb{N}}$ variables aléatoires indépendantes de Bernoulli, chacune de ces variables ayant pour paramètre p, la variable aléatoire Sn définie par $S_n = \sum_{i=1}^{i=n} X_i$ suit **une loi binomiale** de paramètres n et p, on note Sn ~ $\mathcal{B}(n,p)$ d'espérance et variances connues : $E(X) = np$ et $V(X) = np(1-p)$. On a également pour tout entier k, 0⩽k⩽n, $P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$

La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé, si ces événements se produisent avec une fréquence moyenne ou espérance connue et indépendamment du temps écoulé depuis l'événement précédent. Par exemple, si un certain type d'évènements se produit en moyenne 4 fois par minute , pour étudier le nombre d'évènements se produisant dans un laps de temps de 10 minutes, on choisit comme modèle une loi de Poisson de paramètre λ.

X ~ P(λ) ⇔ *pour tout* $k \in \mathbb{N}$, on a $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$. Dans ce cas, on a $E(X) = \lambda$ et $V(X) = \lambda$

AIDE SUR LES FONCTIONS

```
help(sample)
sample(1:100, 1)        # choisir un nombre entre 1 et 100
sample(1:6, 10, replace = T) # lancer un dé 10 fois
sample(1:6, 10, T, c(.6,.2,.1,.05,.03,.02))  # un dé non équilibré

help(dbinom)
dbinom(k, n, P) #probabilité d'avoir k réussite parmi n tirages
help(qbinom)
help(rbinom)
help(rpois)
dpois(0:2, lambda=4)    # probabilités de 0,1,2 pour la loi Poisson(4)
```

1.1.Exercice (Modélisation du jeu de Yam)

Le statisticien considère que dans le monde qui nous entoure, les phénomènes qui s'y produisent constituent un vaste enchevêtrement d'événements aléatoires, qui peuvent être modélisés de façon simplifiée par des variables aléatoires.

1. Nous pouvons commencer par évoquer l'exemple simple, et classique, du lancer d'une pièce de monnaie. Cette expérience peut être assimilée à l'observation du côté PILE ou FACE à chaque lancer. Et nous pouvons modéliser cela par une variable aléatoire X de loi Bernoulli de paramètre ½ (fonction *rbinom()*). Cette expérience peut être recréée à l'intérieur de l'ordinateur. Créez une fonction nommée X permettant de simuler le lancer d'une pièce. Vous pouvez maintenant effectuer quelques lancers de votre pièce virtuelle.

2. Nous pouvons également proposer une modélisation de l'expérience consistant à lancer un dé. Celle-ci peut être assimilée à l'observation du nombre de points que l'on voit sur la face supérieure du dé à chaque lancé. Et nous pouvons modéliser tout cela (si le dé n'est pas truqué) par une variable aléatoire X de loi (de fonctionnement) uniforme discrète sur {1,2, 3,4, 5, 6}. Cette expérience peut ainsi être recréée à l'intérieur de l'ordinateur. Créez une fonction nommée *lancer.le.dé()* en utilisant la fonction *sample()*. Vous pouvez maintenant effectuer quelques lancers de votre dé virtuel.

3. Pour simuler le jeu du Yams, nous allons créer une fonction nommée *yams()* qui permet de lancer cinq dés virtuels. Créez cette fonction en utilisant le paramètre size et le paramètre replace de la fonction *sample()*.

4. Estimez la probabilité d'obtenir un yams, c'est-à-dire cinq dés identiques sur un même lancer de cinq dés (indice : utilisez les fonctions *apply(), replicate() et unique()*). Vous devriez trouver une valeur proche de $\frac{1}{6^4}$.

1.2.Exercice (Loi Binomiale)

D'après l'expérience passée, on sait qu'une intervention chirurgicale donnée à 90% de chances de réussir. Cette opération va être réalisée sur 5 patients. Soit X la variable aléatoire égale au nombre de succès sur les 5 tentatives.

0. Faire connaissance avec les commandes de R permettant de générer tout ce qui a trait à la loi de binomiale *help(dbinom), help(qbinom), help(rbinom)), help(qbinom).*

1. Quelle distribution de probabilité proposez-vous comme modèle pour X ? Quelles sont les valeurs ? Quelles sont les probabilités des différentes valeurs ? Quelle est leur somme ?

2. Calculez l'espérance, la variance et l'écart-type théoriques (par les formules). Calculer la médiane, le premier et le troisième quartile de cette distribution. Pourquoi la médiane et le troisième quartile sont-ils tous les deux égales à 5 ?

3. Quelle est la probabilité que l'intervention chirurgicale réussisse les 5 fois ? exactement 3 fois ? au plus 3 fois ? au moins 3 fois ? de 2 à 4 fois ?

4. Affecter à X un échantillon simulé de taille *N=100* de la distribution binomiale avec les paramètres 5 et 0,9. Calculez les fréquences relatives des différentes valeurs.

1.3.Exercice (Simulation de lois discrètes)

1. Que font ces lignes de commande ?

```
sample(0:1,10,replace = TRUE,prob=c(0.25,0.75))
rbinom(20,1,0.5)
rnbinom(15,10,0.45)
```

2. Générer un échantillon aléatoire de taille 10 suivant les lois suivantes:

• une loi uniforme discrète sur l'ensemble {1, 2, 3, 4, 5, 6} (exemple du lancer d'un dé). Utiliser la fonction *sample().* • B(5, 0.5); • P(3); • G(0.1).

3. Une urne contient 5 boules blanches et 10 boules noires. On tire successivement et sans remise 3 boules dans l'urne. Soit X la v.a. représentant le nombre de boules blanches tirés. Faire un arbre de la situation, en déduire la loi de probabilité de X (un tableau), Générer 10 réalisations indépendantes de X avec la fonction *sample().*

4. Soit X une var dont la loi est donnée par P(X = 0) = 0.2, P(X = 2) = 0.5, P(X = 5) = 0.3. Simuler 1000 réalisations de X (fonction *sample()*) et préciser les effectifs associés aux valeurs de X (fonction *table()*)

1.4.Exercice (Loi binomiale)

Lors d'une séance d'identification, on demande à 6 témoins d'identifier un meurtrier parmi 4 suspects, dont vous-même.

1. Si chacun des 6 témoins choisit au hasard, quelles sont vos chances :

a) de ne pas être signalé ? b) d'être signalé exactement une fois ? c) d'être signalé deux fois ou plus ?

2. Il s'avère que 2 des 6 témoins vous ont identifié comme le meurtrier. En ce qui concerne l'alinéa 1c), vous attendez-vous à ce que le juge pense que cela pourrait être attribuable à ce qui suit une chance ?

3. Et si 4 des 6 témoins vous ont identifié ?

1.5.Exercice (Loi de poisson)

1. Représenter le graphe de la densité d'une v.a.r. X∼P(1) pour $x\epsilon\{0,1,2,\dots,8\}$

2. Construire un vecteur simul2 qui est composé de 1000 éléments dont les valeurs sont des réalisations d'une var X ∼ P(1).

3. Faire l'histogramme des fréquences de simul1

Dans un autre graphe, afficher le diagramme en bateon de la loi de X. Comparer, les deux graphes parraissent ils similaires ?

4. Faire le boxplot de simul2

1.6.Exercice (Loi de poisson)

1. Faire connaissance avec les commandes de R permettant de générer tout ce qui a trait à la loi de Poisson *help(rpois)*

2. Générer un échantillon de taille 10 pour une loi de Poisson P(5). Générer un échantillon de taille 25 pour une loi de Poisson P(15). Que remarquez-vous ?

Donner les valeurs de la densité et de la fonction de répartition des lois de Poisson P(5) et P(15) en x=-1, x=4 et x=13.

Donner le 1er décile et le 3ème quartile des lois P(5) et P(15)

En construisant une grille entre 0 et 20 (commande *seq*), tracer sur un graphique les densités des lois P(1), P(3), et P(12).

3. Tracer les diagrammes en bâtons pour des tailles d'échantillons valant 5, 50, 500 et 5000 d'un loi de Poisson P(10). Que remarquez-vous ?

1.7.Exercice (Loi de poisson)

Une entreprise fabrique des supports d'auvents utilisés notamment dans la construction des stades. Ces pièces sont réalisées en béton. Soit X la variable aléatoire qui à chaque production de 50 pièce associe le nombre de supports défectueux qu'elle contient. La production est suffisamment importante pour qu'on puisse assimiler

tout tirage de 50 supports à 50 tirages aléatoires et indépendants. On admet que X~P(1). Donner à 0,001 près :

a) la probabilité de n'avoir aucune pièce défectueuse sur un tirage de 50 supports

b) la probabilité d'en avoir au moins 4 c) la probabilité d'en avoir au plus 2

1.8.Exercice (Loi de poisson)

Considérons la production d'une entreprise réalisant des résistors pour des fours électriques. Ces résistors sont fabriqués à partir de fil métallique livré en bobine. Le fil utilisé présente des défauts soit de diamètre soit d'homogénéité qui rendent le résistor inutilisable. On considère un très grand nombre de bobines et on constate que sur l'ensemble des résistors produits avec le fil d'une bobine, il y a en moyenne 6 résistors défectueux. On admet que la variable aléatoire Z, qui à tout fabrication utilisant une bobine prélevée au hasard associe le nombre de résistors défectueux, suit une loi de Poisson P(λ)

a) Déterminer la valeur de λ b) Calculer la probabilité qu'il n'ait aucun résistor défectueux

c) La probabilité qu'il n'y ait pas plus de 6 résistors défectueux

1.9.Exercice (Loi de poisson)

Application à l'EuroMillions : un joueur du loto a 1 chance sur 76 millions de gagner le gros lot. Il y a en moyenne 40 millions de joueurs par tirage.

a) quelle loi (avec paramètres) permet ici de déterminer la probabilité d'avoir 0, 1 2 ou 10 gagnants ? calculer ses probabilités.

b) A l'évidence, cette loi décrit ici un nombre petit devant le nombre d'expérience. Pour ce genre d'évènement rares, une bonne approximation est en général apportée par la loi de Poisson de même espérance. Recalculez ces probabilités.

c) Comparer de la même manière les deux lois pour la probabilité d'avoir 5 piles sur 10 lancers de pièce

TP04- LOI A DENSITE (UNIFORME ET EXPONENTIELLE)

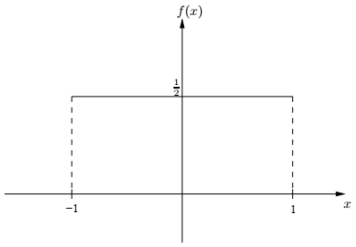
Objectifs : calculer les probabilités et les quantiles à partir d'une distribution uniforme ou d'exponentielle ;

PRELIMINAIRES

X suit une **loi uniforme** ⇔ il existe a et b deux réels tels que a < b

et $f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a,b] \\ 0 & \text{si } x \notin [a,b] \end{cases}$

Si $X \sim U([a,b])$, alors $E(X) = (b+a)/2$ et $Var(X) = (b-a)^2/12$



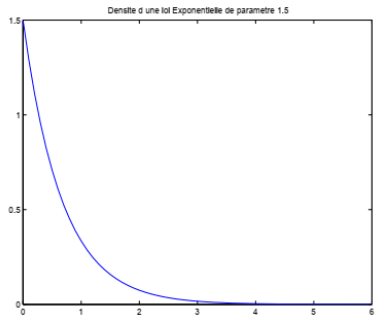
Densité d'une loi uniforme, a = −1 et b = 1

X suit une **loi exponentielle de paramètre α**, notée $E(\lambda)$

⇔ pour tout $x \in \mathbb{R}$ sa densité $f(x) = \alpha e^{-\alpha x} \, \|_{[0,+\infty[}(t)$

Soit $X \sim E(\lambda)$, alors $E(X) = 1/\lambda$ et $Var(X) = 1/\lambda^2$

Proposition : La loi exponentielle vérifie la propriété d'absence de mémoire. Soit X un v.a. de loi exponentielle $E(\lambda)$. Alors pour tous s, t > 0, P(X > t + s/X > t) = P(X > s)



Densité d'une loi Exponentielle de paramètre 1.5

AIDE SUR LES FONCTIONS

```
#Pour effectuer un calcul R avec la loi de probabilité, il suffit d'utiliser une des
#lettres d, p, q, r suivi de l'appellation unif,exp,chisq,t,…
runif(m,min=a,max=b)            # m tirages de Loi uniforme
w <- rexp(1000, rate=.1)        # 1000 tirages, loi exponentielle
rchisq(m,df=r)          # m tirages de Loi du Chi-deux à r degrés de liberté χ 2 (r) :
rt(m,df=r)                      # m tirages de Loi de Student à r degrés de liberté :
```

1.1.Exercice (Loi à densité)

A- Etude de la loi $f(x) = \frac{3}{2}\sqrt{x}$ sur [0, 1]

1. Vérifiez que f(x) est une densité au moyen de la fonction *integrate()*.

2. Simulez un échantillon de taille 1 000 selon la loi définie par la densité $f(x) = \frac{3}{2}\sqrt{x}$ sur [0, 1].

3. Calculez les moyennes et variances empiriques.

4. Comparez avec les valeurs théoriques.

5. Calculez et comparez les probabilités théoriques et empiriques des classes suivantes :

[0, 0.30],]0.30, 0.50],]0.50, 0.70],]0.70, 0.85],]0.85, 1].

1.2.Exercice (Loi uniforme)

1. Que faut-il pour générer une loi Uniforme sur [0,1] ?

A l'aide du logiciel R, nous utilisons la fonction *runif()*. Vous pouvez voir ses caractéristiques avec h*elp(runif).* Cette fonction tiendra le rôle de notre générateur de nombres aléatoires.

2. Générer à l'aide de la fonction *runif()* 5 réalisation entre 0 et 1. Tracer un graphique permettant de les visualiser

A l'aide de ce graphique, est-il possible de retrouver la loi de départ.

3. Retourner au 2/ et refaire les questions précédentes avec n=50, n500 et n=5000.

Quelle conclusion pouvez-vous en tirer ?

4. Les commentaires ci-dessus sont bien difficiles à faire dans bien des cas. De quel outil disposez-vous pour vous donner une idée plus précise de la loi de probabilité de cette variable ?

Cet outil est disponible dans R avec la fonction *hist*(). Regardez ces caractéristiques avec *help(hist)*. Refaire les tests précédents avec les différentes tailles d'échantillons Que remarquez-vous ?

1.3.Exercice (Loi uniforme sur un carré)

1. Simulez 1 000 observations de (X1, X2) suivant la loi uniforme sur le carré [0, 1]×[0, 1].

2. Obtenez une approximation de la probabilité que la distance de (X1, X2) au côté le plus proche soit inférieure à 0.25.

3. Même question pour la distance au sommet le plus proche. (se servir des lignes de code ci-dessous en expliquant ce qu'elle font)

```
dist.som <- function(coord){
d <- min(coord[1]^2+coord[2]^2,coord[1]^2+coord[4]^2,
coord[3]^2+coord[2]^2,coord[3]^2+coord[4]^2)
d <- sqrt(d)
}
```

4. Essayez d'identifier la loi théorique de la variable distance au côté le plus proche : espérance, variance, densité

1.4.Exercice (Loi uniforme: fonctions dunif punif qunif)

1. Utilisation de la fonction *dunif()*. Quelle est la valeur de la densité d'une Uniforme U([0;5]) en 0.2 ? d'une U([-17;-25])en -19 ? d'une U([6;12]) en 10 ?

2. Utilisation de *punif()*.

Quelle est la valeur de la fonction de répartition d'une Uniforme U([0;5]) en 0.2 ? d'une U([-25,-17]) en -19?

Donnez la valeur, pour une U([1;10]) de P(u<1,9), P(u>7), P(4<u<10), P(4<u<12), P(3<u<8) et P(u<0,9).

3. Utilisation de *qunif()*. Le pième quartile p de la distribution de la variable aléatoire X peut être défini comme une valeur x telle que $P(X \leq x) \geq p$.

Exemple de quantiles : la médiane (p=1 /2), les quartile (p=1/4, V2 ou 3/4), les déciles (p=1110 etc ...). A l'aide de R, donner la médiane d'une U([2;4]). Commenter le résultat.

Donner

- Le premier décile d'une U([8;19]),
- La médiane, le 2e quartile et le 5ème décile d'une U([1;15])
- Le 7ème décile d'une U([-5;7]).

1.5.Exercice (Loi exponentielle/uniforme)

Propriété : Si U est de loi Uniforme sur [0,1] alors X—F^{-1}(U) a comme fonction de répartition F.

1. A l'aide de cette propriété, construire un générateur pour la loi Exponentielle E(α), dont la densité est $f(t) = \alpha e^{-\alpha t} \|_{[0,+\infty[}(t)$ pour tout $t \epsilon \mathbb{R}$. Cette méthode s'appelle la méthode d'inversion de la fonction de répartition.

2. Générer un échantillon *a* de taille 5 de loi Exponentielle E(1). Le visualiser à l'aide de la commande *plot()*. Que remarquez-vous ?

Faites de même avec la loi uniforme et visualiser le sur le même graphique. Est-il possible de distinguer les 2 lois ?

Pour une meilleure visualisation on peut également utiliser l'histogramme. Refaire les questions précédentes avec n=50, n=500 puis n=5000. Conclure.

3. Ce générateur est déjà implémenté dans R avec la commande *rexp()*. Vous pouvez regarder ses caractéristiques à l'aide de la commande *help(rexp)*. Générer un échantillon de taille 5 à l'aide de cette commande. Comparez-le graphiquement à un échantillon généré à l'aide de la méthode d'inversion de la fonction de répartition. Pouvez-vous conclure que ces 2 commandes génèrent bien la même loi ?

Refaire cette question 3. avec des échantillons de taille 50, 500 et 5000.

4. Toujours en faisant varier les tailles d'échantillon, comparez les E(1), E(10), E(100) et E(1000). Que contrôle le paramètre X, de la fonction exponentielle ?

1.6.Exercice (Loi exponentielle : fonctions dexp pexp qexp)

1. Utilisation de la fonction *dexp()*.

Quelle est la valeur de la densité d'une Exponentielle E(0,9) en 0.2 ?d'une E(9) en -19? d'une E(2,9) en 10?

2. Utilisation de *pexp()*.

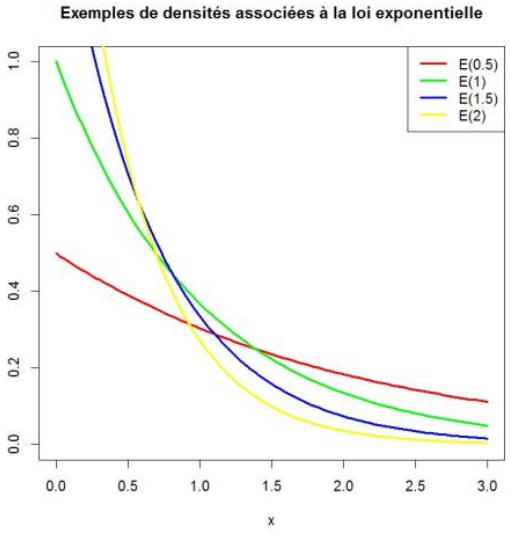
Quelle est la valeur de la fonction de répartition d'une E(0,9) en 0.2 ? d'une E(4,5) en -19 ou en 19? Donnez la valeur, pour une E(3) de P(u<1.›9). P(u>7), P(4<u<10), P(4<u<l2), P(3<u<8) et P(u<0).

3. Utilisation de *qexp()*.

Donnez la médiane d'une E(1), le 1er décile d'une E(4), la médiane et le 5ème décile d'une E(5) et le 7e décile d'une E(2).

1.7.Exercice (Loi exponentielle)

Proposer des commandes R permettant d'obtenir le graphique suivant :



TP05- LOI NORMALE

Objectifs : calculer les probabilités et les quantiles à partir des distributions normales.

PRELIMINAIRES

X suit une **loi Normale** $\mathcal{N}(m,\sigma)$
⇔ pour tout x ∈ ℝ, sa densité f est donnée par

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Si X ~ $\mathcal{N}(m,\sigma)$, alors E(X) = m et Var(X) = σ^2

Remarque : on parle de loi normale centrée réduite lorsque m=0 et σ=1

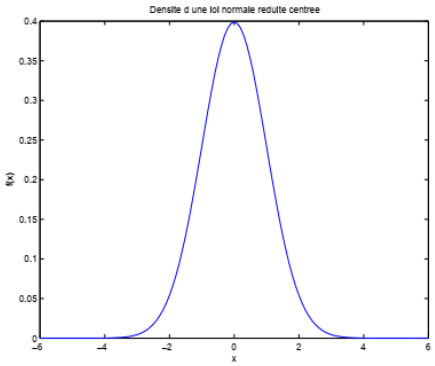


Figure 1. Densité d'une loi normale réduite centrée

Proposition : Si X ~ $\mathcal{N}(m,\sigma^2)$, alors $Z = \frac{X-m}{\sigma}$ suit une loi normale centrée réduite $\mathcal{N}(0,1)$

AIDE SUR LES FONCTIONS

```
#Pour effectuer un calcul R avec la loi normale, il suffit d'utiliser une des lettres
#d, p, q, r suivi de l'appellation norm. En voici quelques exemples:
qnorm(0.975)    # quantile d'ordre 0.975 de la loi N(0,1)
dnorm(0)        # valeur de la fonction de densité en 0 de la loi N(0,1)
pnorm(1.96)     # valeur de la fonction de répartition en 1.96 de la loi N(0,1)
rnorm(20)       # génération de 20 réalisations indépendantes suivant la loi N(0, 1)
pnorm(12,mean=10,sd=2)  # P(X < 12) pour la loi N(10,4)
qnorm(.75,mean=10,sd=2) # 3ème quartile de la loi N(10,4)
x <- rnorm(100)      # 100 tirages, loi N(0,1)
rnorm(10,mean=5,sd=0.5) # génération de 10 réalisations indépendantes suivant la loi
N(5, 0.25)
```

1.1.Exercice (Loi normale)

Pour tracer la densité de la loi $\mathcal{N}(0,1)$, on utilise la fonction plot:

```
x=seq(from=-3,to=3,by=0.1)
plot(x,dnorm(x),type="l",ylab="fonction de densité")
```

On pourra aussi utiliser la fonction *curve()*:

```
curve(dnorm(x),from=-3,to=3,ylab="fonction de densité")
```

Tracer la densité d'une loi $\mathcal{N}(10,4)$ puis sa fonction de répartition sur l'intervalle [4; 16].

1.2.Exercice (Loi normale)

Séparer l'écran graphique en 3 (1 ligne, 3 colonnes). Utiliser les commandes *par(mfrow = c(3, 1))* et *par(mfrow = c(1,1))* en début et fin d'instruction.

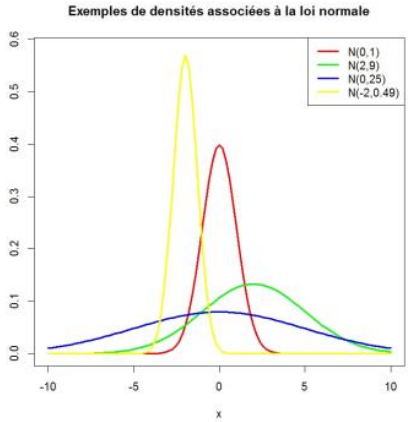
Dans la première fenêtre, représenter le graphe de la densité d'une v.a.r. X $\sim \mathcal{N}(4,1)$, puis ajouter dans la même fenêtre, avec une autre couleur, celui de la densité d'une v.a.r. Y$\sim\mathcal{N}(5,1)$.

Dans la deuxième fenêtre, représenter le graphe de la densité d'une v.a.r. X $\sim \mathcal{N}(4,1)$ puis ajouter dans la même fenêtre, avec une autre couleur, celui de la densité d'une v.a.r. Y$\sim \mathcal{N}(4,4)$.

Dans la troisième fenêtre, représenter le graphe de la densité d'une v.a.r. X$\sim\mathcal{N}(4,1)$ puis ajouter dans la même fenêtre, avec une autre couleur, celui de la densité d'une v.a.r. Y$\sim\mathcal{N}(5,4)$.

1.1.Exercice (Loi normale)

Proposer des commandes R permettant d'obtenir le graphique ci-contre



1.2.Exercice (Loi normale)

On suppose que le poids d'un foie gras peut être modélisé par une v.a.r. X suivant la loi normale N (550, 100^2), l'unité étant le gramme. Quelle est la probabilité qu'un foie gras pèse

a) moins de 650 grammes ? b) plus de 746 grammes ? c) entre 550 grammes et 600 grammes ?

1.3.Exercice (Loi normale)

Pour cette exercice, nous utiliserons la proposition rappelée en préliminaires.

1. Simuler un échantillon de taille 500 suivant une loi normale $\mathcal{N}(50,9)$. Stocker cet échantillon dans un vecteur x.

2. Créer ensuite le vecteur y = (x−50)/3. Il s'agit de l'échantillon centré réduit obtenu à partir de l'échantillon initial.

3. Exécuter les commandes suivantes en les expliquant

```
hist(x)
hist(x,breaks=8,probability = T)
lines(density(x),col='red')

plot(density(y),col="red",type="l",xlab="y",ylab="densité",main="")
z=seq(-3,3,0.1)
lines(z,dnorm(z),col="blue")
```

On pourra ajouter une légende avec la fonction *legend.*

4. Tracer la fonction de répartition empirique de y (utiliser la fonction *plot.ecdf*). Superposer sur le même graphique la fonction de répartition théorique de la loi $\mathcal{N}(0,1)$.

1.4. Exercice (Loi normale)

Soit X une v.a.r. suivant la loi normale $\mathcal{N}(0,1)$dont la densité f_X est connue (cf. préliminaires),

1. Écrire une nouvelle fonction densnorm R équivalente à la fonction *dnorm()*. Commande *densnorm = function(x) {…}*

2. Vérifier numériquement que $\int_{-100}^{100} f_X(t)dt \cong 1$

3. Séparer l'écran graphique en 2 en introduisant avant la commande *par(mfrow = c(1, 2))* : dans la fenêtre 1, représenter le graphe de la densité f_X et, dans la fenêtre 2, celui de la fonction de répartition de X.

4. Calculer $P(X \leq 2.2)$, $P(X \geq 1.7)$, $P(0.2 \leq X < 1.4)$ et $P(|X| \leq 1.96)$.

5. Déterminer le réel x vérifiant P(X ≤ x) = 0.98.

1.5. Exercice (Loi normale)

L'apport nutritionnel total (en Kcal par jour) dans la population témoin est en moyenne de 2970 et l'écart type de 251. Chez les coureurs, la moyenne est de 3350 et l'écart-type de 223.

1. Prenons l'exemple d'une personne prise au hasard dans la population témoin. Diriez-vous que les chances que l'apport soit inférieur à 3000 sont supérieures à 1/2 ? qu'il est supérieur à 3000 sont inférieurs à 1/2 ? Répétez l'opération pour les coureurs.

2. Quelle proportion des apports dans la population témoin est inférieure à 2 600 ? supérieur à 3400 ? entre 2600 et 3400 ? Répétez l'opération pour les coureurs.

3. Quelle est la limite inférieure de sorte que 1 % de la population témoin se situe sous la limite inférieure ? Qui est telle que 1% des coureurs sont au-dessus de ? Utilisez ces bornes pour représenter les deux densités sur la même parcelle. Ajoutez des lignes verticales à 2600 et 3400. A qui la courbe la plus à droite correspond-elle ? À quelle population s'adresse le correspond à une courbe plus étroite ? Identifier sur le graphique les zones correspondant à les probabilités calculées à la question 2.

4. Quel apport est tel que 5 % de la population témoin est inférieur ? supérieur ? Répéter pour 0,5 %, 50 %. Répétez l'opération pour les coureurs.

5. Toujours en utilisant les mêmes limites inférieure et supérieure, représentent les deux cdf sur le même tracé. A quelle population correspond la courbe inférieure ? A quelle population correspond la courbe la plus raide ? Où, sur le graphique, pourriez-vous lire les réponses aux questions 2 et 4 ?

6. Si un millier de personnes étaient choisies au hasard dans la population témoin et classées en fonction de l'augmentation de l'apport, quelle quantité le 400e mangerait-il ? le 800e ? Répétez l'opération pour les coureurs.

7. Considérons que deux personnes ont été choisies au hasard, l'une parmi la population témoin et l'autre parmi les coureurs. Quelle est la distribution de probabilité de la différence des apports ? Calculez la limite inférieure et la limite supérieure, de sorte que 1 % des différences soient inférieures à la limite inférieure et 1 % supérieures à la limite supérieure. Utilisez ces limites pour tracer la densité de la différence ; ajoutez des lignes verticales vertes à 0 et 400. Quelles seraient les chances que le coureur mange plus que l'autre ? manger plus de 400 Kcal de plus par jour ? Identifiez les zones correspondantes sur le graphique.

8. Répétez l'exercice en remplaçant les coureurs par des cyclistes, dont la moyenne est de 3880 et l'écart-type de 450.

9. Répétez l'opération en remplaçant les coureurs par des skieurs alpins, pour lesquels la moyenne est de 3524 et l'indice écart type 352.

TP06- LOI DES GRANDS NOMBRES

Objectifs : calculer les probabilités et les quantiles à partir d'une distribution binomiale ; comprendre la loi des grands nombres : sur un grand nombre d'expériences, la fréquence relative d'un événement s'approche de sa probabilité théorique.

PRELIMINAIRES

Théorème (Loi des grands nombres) : Soit $(Xi)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes de même loi telles que Var(X_1) < ∞. on définit i) la somme de ces variables aléatoires $S_n = \sum_{i=1}^{i=n} X_i$, ii) la moyenne empirique est donnée par $\overline{S_n} = \frac{S_n}{n}$. Pour tout $\varepsilon > 0$, on a $\lim_{n \to \infty} P(|\overline{S_n} - E(X_1)| \leq \varepsilon) = 1$

1.1. Exercice (Lois des grands nombres)

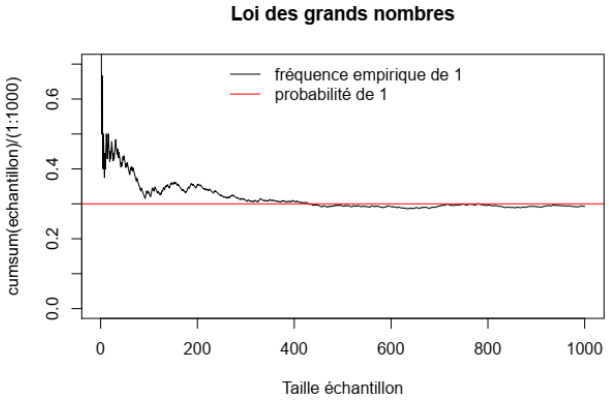
On tire n fois une pièce de monnaie biaisée telle que la probabilité d'obtenir face est 0.3 et on calcule ensuite la proportion des faces obtenues.

1. Caractériser la loi X du lancer de monnaie biaisée.

Générer pour n=1000 un telle échantillon que vous stockerez dans une variable que vous appellerez échantillon

2. Reproduire le graphique ci-contre en utilisant la fonction *cumsum()* pour obtenir le nombre des faces en fonction de *n*.

3. Générer le même graphique pour n=4000. Que conclure sur la fluctuation de la moyenne empirique en fonction de la taille de l'échantillon ?



1.2. Exercice (loi des grands nombres)

1. Charger TP06_tauber.csv.

```
TA <- read.table("CheminJusquAuFichier/TP06_tauber.csv",header=TRUE,sep=";")
```

Expliquer de quoi est compose le fichier TP06_tauber (nombre d'observations, variables)

Attribuer la hauteur de colonne à la variable H. via la commande H <-*TA[,"height"]*

Combien d'enfants de l'échantillon ont une taille supérieure à 110 cm ?

Combien ont une taille de 110 ou moins ?

2. Tracer les fréquences cumulées pour l'événement H>110, avec les ordonnées dans l'intervalle (0,1) : option de tracé *ylim=c(0,1)*.

3. Assigner au vecteur *rH* une permutation aléatoire de H via la fonction *sample*. Superposer sur le même graphique Fréquences cumulées pour *rH>110*, en bleu.

4. Assigner au vecteur *iH* les valeurs de H, triées par ordre croissant via la fonction *sort*. Superposer sur l'écran mêmes graphiques fréquences cumulées pour *iH>110*, en vert.

Dans quel intervalle est la constante de la courbe verte ?

5. Assigner au vecteur *dH* les valeurs de H, triées par ordre décroissant. Superposer sur l'écran mêmes graphiques fréquences cumulées pour *dH>110*, en rouge. Dans quel intervalle est l'intervalle constante de la courbe rouge ?

1.3. Exercice (Loi Binomiale et loi des grands nombres)

Par l'expérience passée, on sait qu'une intervention chirurgicale donnée à 90% de chances de réussir. Cette opération va être réalisée sur 5 patients. Soit X la variable aléatoire égale au nombre de succès sur les 5 tentatives.

1. Quelle distribution de probabilité proposez-vous comme modèle pour X ? Quelles sont les valeurs ? Quelles sont les probabilités des différentes valeurs ? Quelle est leur somme ?

2. Calculez la moyenne théorique, la variance, l'écart-type, la médiane, le premier et le troisième quartile de cette distribution. Pourquoi la médiane et le troisième quartile sont-ils tous les deux égales à 5 ?

3. Quelle est la probabilité que l'intervention chirurgicale réussisse les 5 fois ? exactement 3 fois ? au plus 3 fois ? au moins 3 fois ? de 2 à 4 fois ?

4. Affecter à X un échantillon simulé de taille *N=100* de la distribution binomiale avec les paramètres 5 et 0,9. Calculez les fréquences relatives des différentes valeurs. Comparez avec les probabilités théoriques. Répéter (plusieurs fois) pour *N=1e4*, *N=1e6*.

5. Affecter à X un échantillon simulé de taille *N=1e4*. Tracez la fonction de distribution cumulative empirique de X en bleu (*ecdf(x)*). Superposer les probabilités cumulatives théoriques de la distribution binomiale avec les paramètres 5 et 0.9 comme points rouges.

6. Tracez en points bleus les moyennes cumulées de X (sommes cumulées divisées par (1:N)) contre (1:N). Ajouter une ligne rouge horizontale marquant la valeur théorique de la moyenne de X.

7. Tracez la moyenne cumulative de X<=3 contre (1:N), en points bleus. Ajouter une ligne rouge horizontale marquant la probabilité théorique.