**L3 EBO**

**TP2 Exploration et analyse des données environnementales**

**Manipulation et présentation des données avec R**

1. **Explorer les données sur une variable :**

*Ouvrir le fichier trees.txt.*

*- Essayez les 2 fonctions suivantes avec la variable « Height » et déterminer à quoi elles servent :*

> range(VA)

> c(min(VA),max(VA))

*- Continuez avec les fonctions :*

> sum(VA)

> cumsum(VA)

*- Pour avoir le résumé :*

> summary (VA)

ou

> fivenum(VA)

- On peut créer ses propres fonctions :

*Calculez comme vu dans le TP1 l’écart-type de la variable « Height ».*

*Maintenant tapez :*

> std<-function(x) sqrt (var(x))

> std(Height)

*Que vous donne la fonction que vous venez de créer ? Réfléchissez à sa syntaxe et trouver à quoi correspondent les différents termes utilisés.*

- *Que font les commandes ?*

> sort(VA)

> rank(VA)

> sample(VA, 20)

- *Pour aller un peu plus loin, essayez la macro suivante, que fait-elle ?*

> res <- c(); nn = 1000;

for (i in 1:nn) {

x = sample(Girth, 10);

res[i] = mean(x,na.rm=T);

}

*Tapez :*

> hist(res, prob=T)

> hist(res, freq=F)

*Observez-vous une différence entre les 2 histogrammes ?*

*La distribution est-elle unimodale? Si non, comment est-elle et pourquoi ? Refaire le vecteur « res » en vous arrangeant pour (1) ne pas modifier le jeu de données trees.txt et (2) que la distribution soit unimodale de façon certaine.*

1. **Manipuler des data frames**

On demande aux étudiants d'un amphi d'indiquer leur sexe (h/f), leur poids (en kg) et leur taille (en cm). Les données sont contenues dans le fichier HF.txt.

- *Ouvrez le fichier HF.txt.*

- *Que font les macros :*

> names(nom\_tableau)[3]= "taille"

> nom\_tableau [,(dim(nom\_tableau)[2]+1)]=ifelse(nom\_tableau $sexe=="h",0,1)

- *Renommez la colonne nouvellement créée.*

- *Retirez la ligne 32 de ce fichier de façon temporaire puis de façon définitive.*

*- Supprimez la colonne que vous avez créé juste avant.*

1. **Les lois de probabilité**

- *Tapez :*

> help(Normal)

> help(Binomial)

> help(Poisson)…

On remarque dans ces "help" que l'on a accès pour chacune des lois (LOI), à la densité de probabilité (d\_LOI), à la probabilité (probabilité cumulée)(p\_LOI), aux quantiles (q\_LOI) et on peut générer des nombres pseudo-aléatoires (r\_LOI).

*- Préparez une fenêtre graphique pouvant accueillir 6 graphiques en même temps.*

*Essayez les macros suivantes et de comprendre à quoi elles servent :*

> s1=rnorm(10,mean=2)

> summary(s1)

> s2=rnorm(100,mean=2)

> summary(s2)

> s3=rnorm(10000,mean=2)

> summary(s3)

> hist(s1, prob=T)

> hist(s2, prob=T)

> hist(s3, prob=T)

> plot(density(s1))

> x=seq(-5 , 5, by=.01)

> lines(x,dnorm(x,mean=2),col=2)

> plot(density(s2))

> lines(x,dnorm(x,mean=2),col=2)

> plot(density(s3))

> lines(x,dnorm(x,mean=2),col=2)

*Que remarquez-vous entre s1, s2 et s3 ?*

- *Essayez :*

> pnorm(1.96)

> qnorm(0.975)

*Où avez-vous déjà vu ces valeurs ?*

*- Qu’est ce qu’une variable binomiale ?*

*Essayez la macro suivante, que fait-elle ?*

> n=20; p=0.3; mu=n\*p; s2=n\*p\*(1-p);

> x = rbinom(100, n, p)

> hist(x, freq=FALSE)

> mean(x)

> var(x)

- Essayez les macros suivantes, que font-elles ?

> n = length(x)

> plot(sort(x),(1:n)/n, type= "s")

> hist(x, prob=T, col=grey(0.4))

> curve(dnorm(x,mu,s2),add=T)

> qqnorm(x);qqline(x)

> plot(qpois(ppoints(x),mu),sort(x))

> plot(qbinom(ppoints(x),n,p),sort(x))

> n=50; p=0.03; mu=n\*p; s2=n\*p\*(1-p);

> x = rbinom(100, n, p)

> plot(qpois(ppoints(x),mu),sort(x))

> x = rnorm(100)

> hist(x,freq=F)

> curve(dnorm(x),add=T)

- *Ouvrez à nouveau le fichier HF.txt (tel qu’il était avant les modifications des questions précédentes)*.

*Créez un objet avec les données de taille des hommes uniquement et un avec celles des femmes uniquement.*

*Vous allez créer la fonction de densité de probabilité des données concernant la taille des hommes à l’aide des macros suivantes :*

> hist(table\_data\_hommes, proba = TRUE, col = grey(0.8), main = "Taille des hommes")

> provi <- seq(160, 200, length = 50)

> lines(provi, dnorm(provi, mean(table\_data\_hommes), sd(table\_data\_hommes)), lwd = 2)

*Faites la fonction de densité de probabilité de la taille des femmes. La fonction concernant la population devra être tracée en bleu.*

- Pour calculer un intervalle de confiance à 95% de la moyenne d’une variable numérique :

> t.test(VA)$conf.int[1:2]

Il est possible de préciser le taux de confiance (ici 90%) :

> t.test(VA, conf.level=0.90)$conf.int[1:2]

**Exercice1 :**

Pour déterminer la concentration en glucose d’un échantillon sanguin, on effectue des dosages à l’aide d’une technique expérimentale donnée. On considère que le résultat de chaque dosage est une variable aléatoire normale. On effectue 10 dosages indépendants, qui donnent les résultats suivants (en g/l) : 0.96, 1.04, 1.08, 0.92, 1.04, 1.18, 0.99, 0.99, 1.25, 1.08

1. Calculer une estimation de la concentration moyenne en glucose de cet échantillon.

2. Calculer un intervalle de confiance de cette concentration à 95%.

**Exercice 2 :**

2% des micro-ordinateurs d’une marque donnée tombent une fois en panne par mois d’utilisation. Aucun ordinateur ne tombe deux fois en panne dans le même mois.

Une entreprise décide d’acquérir 150 ordinateurs de ce type.

1. Calculez la probabilité des évènements suivants :

- le nombre mensuel de pannes est 5,

- le nombre mensuel de pannes est au plus égal à 3.

2. Déterminez le nombre minimum *n* tel que la probabilité de l’événement « le nombre de pannes est au plus de *n* » soit supérieure à 0,99.