INITIATION AU LOGICIEL R - TP 2

Distributions de probabilités, représentations graphiques

Exercice 1 : Comparaison de distributions

1. Soumettre le code suivant et commenter la sortie obtenue :

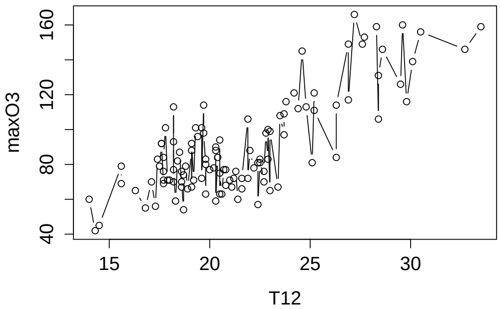
```
curve(dnorm(x),from=-4, to=4)
title("Comparaison de distributions")
curve(dt(x,5), add=TRUE,col=2)
curve(dt(x,30), add=TRUE,col=3)
```

2. Soumettre et commenter le code suivant :

```
dev.new()
x=seq(-4,4,0.01)
plot(x,dnorm(x), type="l")
lines(x,dt(x,5),col="2")
lines(x, dt(x,30), col="3")
title("Comparaison de distributions")
legend(x=-4,y=0.4,legend=c("loi normale","loi de Student à 5 ddl", "loi de Student à 30
ddl"),col=1:3,lwd=1,cex=0.7)
```

Exercice 2 : Représentation d'un nuage de points

1. Importer les données du fichier `ozone.txt` dans un data-frame nommé `ozone` (utiliser la fonction **read.table**) puis tracer le nuage de points du maximum d'ozone (`maxO3`) en fonction de la température (`T12`) avec la commande `plot(maxO3 ~ T12,data=ozone)`.
2. Relier les points du nuage avec des lignes en utilisant l'instruction **type** de la fonction **plot** (on pourra consulter l'aide en ligne de la fonction **plot** avec la commande `?plot` ou `help(plot)`).
3. En utilisant **order**, tracer le graphique suivant :



4. Sur le même graphique, tracer, en rouge et en pointillés, une droite horizontale représentant la valeur moyenne de la variable `max03` (utiliser la fonction **abline**).
5. Sur le même graphique, tracer en bleu la droite de régression d'équation $y = -27,42 + 5,469x$.

Exercice 3 : Loi des grands nombres

1. Après avoir fixé la graine du générateur aléatoire à 123 (**set.seed**), simuler un échantillon $(x_1, \dots, x_{1000})$ de longueur 1000 provenant d'une loi de Bernoulli de paramètre $p = 0,6$.
2. Calculer les moyennes successives $M_\ell = \frac{1}{\ell}\sum_{i=1}^{\ell} X_i$ pour $\ell = 1, \dots, 1000$ (fonction **cumsum**). Tracer la courbe joignant les points de coordonnées (ℓ, M_ℓ) puis ajouter la droite d'équation $y = 0,6$.

Exercice 4 : Théorème central limite pour la loi de Bernoulli

1. Utiliser la fonction **rbinom** pour simuler $n = 10000$ réalisations $S_1, \dots, S_{10000}$ d'une loi binomiale de paramètres N et p (fixer $N = 10$ et $p = 0,5$). Ranger dans un vecteur $U10$ les quantités $\dfrac{S_i - Np}{\sqrt{Np(1-p)}}$. Faire de même avec $N = 100$ et $N = 1000$ pour obtenir deux nouveaux vecteurs $U100$ et $U1000$.
2. Représenter dans une même fenêtre graphique les histogrammes de $U10$, $U100$ et $U1000$ en superposant à chaque fois la densité de la loi normale centrée réduite (**dnorm**). Expliquer en quoi ce graphique illustre le théorème central limite.

Exercice 5 : Intervalle de confiance pour une proportion

Soit $X_1, \dots, X_n$ n réalisations indépendantes de la loi de Bernoulli de paramètre p. On rappelle qu'un intervalle de confiance de niveau asymptotique $1-\alpha$ (avec $0 < \alpha < 1$) pour p est donné par

$$\left[\bar{X}_n \pm q_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\right], \qquad (1)$$

où $\bar{X}_n$ et S_n^2 désignent respectivement la moyenne empirique et la variance empirique de $\{X_1, \dots, X_n\}$ et $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale standard.

1. Simuler $N = 10000$ échantillons de taille n de la loi de Bernoulli de paramètre p (fixer p à 0.3 et n à 30). On rangera les $n \times N$ valeurs simulées dans une matrice de taille $n \times N$, nommée `mat`.
2. Calculer les N réalisations de $\bar{X}_n$ et S_n (utiliser la fonction **apply** sur la matrice `mat`).
3. Calculer N intervalles (1) de niveau de confiance asymptotique 0.95 (i.e. $\alpha = 0,05$) et afficher à l'écran deux colonnes contenant respectivement les N bornes inférieures et les N bornes supérieures de ces intervalles.
4. Calculer la proportion des N intervalles qui contiennent effectivement la valeur $p = 0.3$. Que constate-t-on ? Répéter l'expérience en prenant $n = 100$. Répéter l'expérience en prenant $\alpha = 0,01$ puis $\alpha = 0,1$.

Exercice 6 : Test sur la moyenne d'un échantillon gaussien de variance inconnue

Soit $X_1, \dots, X_n$ n réalisations indépendantes de la loi normale $\mathcal{N}(m, \sigma^2)$. On rappelle qu'une règle de décision au niveau α (avec $0 < \alpha < 1$) pour le problème de test de $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ consiste à rejeter H_0 lorsque $|\bar{X}_n - m_0| > c$, où le seuil critique c est donné par

$$c = \mathcal{T}_{1-\alpha/2}(n-1) \cdot \frac{S_n}{\sqrt{n}}$$

($\mathcal{T}_{1-\alpha/2}(n-1)$ désigne le quantile d'ordre $1-\alpha/2$ de la loi de Student à $(n-1)$ degrés de liberté et S_n^2 désigne la variance empirique de $\{X_1, \dots, X_n\}$).

1. Simuler $N = 10000$ échantillons de taille $n = 10$ de la loi $\mathcal{N}(1,1)$. On rangera les $n \times N$ valeurs simulées dans une matrice de taille $n \times N$, nommée `mat`.
2. Calculer les N réalisations de $\bar{X}_n$ et S_n (utiliser la fonction **apply** sur la matrice `mat`).
3. On souhaite tester $H_0 : m = 1$ contre $H_1 : m \neq 1$. Calculer les N valeurs du seuil critique pour les N échantillons simulés (on prendra $\alpha = 0,05$).
4. Pour chacun des N échantillons simulés, comparer $|\bar{X}_n - m_0|$ au seuil critique correspondant et calculer la proportion de tests (parmi les N tests réalisés) qui amène à rejeter H_0 à tort. Commenter la valeur obtenue.
5. Répéter l'expérience avec $n = 30$, avec $\alpha = 0,01$ puis $\alpha = 0,1$.