Feuille d'Exercices

Vincent Guillemot

31 janvier 2009

Installation de R, installation de packages

Le logiciel se trouve sur le site du *Comprehensive R Archive Network* : http ://cran.r-project.org/. Nous nous intéresserons plus particulièrement à l'utilisation de R sous Windows, aussi il faut aller chercher l'application sur la page http ://cran.r-project.org/bin/windows/base/. L'installation se fait classiquement, il faut juste faire attention, selon que l'on a les droits administrateur ou pas, à bien cocher la case "Enregistrer le numéro de version dans le registre".

Fichier .R

Le code à compléter se trouver sur arel, il s'agit du fichier `codeacompleter.R`. Le télécharger et le mettre dans le dossier de travail de votre choix. Il faudra bien sûr changer le dossier de travail courant, c'est ce à quoi sert la toute première ligne de code de ce fichier !

Obtenir de l'aide sur une fonction

Pour obtenir de l'aide sur les fonctions de R, essayez la commande `help.start()`.

Début du TP

Pour commencer, il faut installer R, bien sûr, mais aussi profiter du début du TP pour installer les packages du projet Bioconductor en exécutant les commandes du point 0.0. Pendant que ça s'installe, en profiter pour passer à la suite !

Exercice 1 : Variables aléatoires continues

Exemple

Une variable aléatoire X admet pour densité de probabilité : $f(x) = ax(2-x)$ pour $x \in [0;2]$ et 0 ailleurs.

1. Calculer a.
2. Déterminer la moyenne, la variance et l'écart type de cette variable.
3. Tracer cette densité avec le R. Indication : utiliser la fonction `plot` pour tracer et la commande `seq(from=0,to=2,length=100)` .

Densités connues

Pour chaque type de variable aléatoire à densité que nous avons vue, tracer sa densité (chercher dans l'aide en ligne de R) et lui superposer l'histogramme d'un échantillon de taille n. Indications :

- variable uniforme, trouver l'aide de la commande `runif`,
- variable du Khi-deux, trouver l'aide de la commande `rchisq`,
- variable de Fisher, trouver l'aide de la commande `rf`,
- variable de Student, trouver l'aide de la commande `rt`.

Un exemple pour une variable aléatoire normale vous est proposé.

Exercice 2 : Estimation

Propriétés des estimateurs de la moyenne et de la variance

Dans une population de taille $N = 4$, une variable Y peut prendre uniquement les valeurs suivantes de façon équiprobable

$$-4, -2, 2, 4$$

1. Calculez la moyenne μ et la variance σ^2 de Y dans cette population.
2. On effectue des prélèvements d'échantillons de taille $n = 2$ sans remise dans cette population. Énumérez tous les échantillons possibles. Pour chacun d'entre eux, calculez leur moyenne empirique et leur variance empirique.
3. Vérifiez que la moyenne empirique est un estimateur sans biais de la moyenne μ.
4. Calculez la variance de la moyenne empirique.

Étude de la moyenne empirique sur des données simulées

Générez, avec le logiciel R, 1000 échantillons de taille 47 et stockez les dans une matrice (matrix) à 1000 lignes et 47 colonnes, échantillons qui seront gaussiens de moyenne nulle et de variance 1.

1. On veut calculer la moyenne empirique d'échantillons de taille 2, 5, 10 et 30. Utilisez les 2 premières colonnes pour calculer 1000 moyennes sur 2 individus, les colonnes 3 à 7 pour 5 individus, les colonnes 8 à 17 pour 10 individus et 18 à 47 pour 30 individus.
2. Construisez un histogramme de ces 1000 moyennes pour chaque taille d'échantillon.
3. Reprendre les mêmes questions pour une loi uniforme sur $[-1, 1]$.

Calcul de la taille d'un échantillon pour une précision donnée - *déjà vu* -

On a pesé 15 poulpes mâles adultes pêchés au large des côtes Mauritaniennes. On suppose que, pour cette espèce de poulpes, les poids sont répartis selon une loi normale d'espérance μ et de variance σ^2. Le tableau ci-dessous donne l'échantillon des 15 valeurs obtenues :

1150	1500	1700	1800	1800	1850	2200	2700
2900	3000	3100	3500	3900	4000	5400	

1. Donnez une estimation de μ et σ^2 à partir des données.

2. Construisez un intervalle de confiance pour μ au niveau $\alpha = 5\%$. Donnez l'amplitude de cet intervalle.
3. Si n désigne la taille de l'échantillon, donnez l'amplitude de l'intervalle de confiance pour μ au niveau $\alpha = 5\%$ en fonction de n.
4. Tracer l'histogramme de cet échantillon.

Exercice 3 : Tests d'hypothèses

Codez en R les tests de Fisher de comparaison des variances, de Student pour la comparaison des moyennes, et de Wilcoxon pour la comparaison des médianes. On supposera dans cet exercice que les données en entrée de tous ces tests ne sont pas appariées.

Testez vos fonctions sur des petits jeux de données simulés.

Test sur de vraies données

Chargez les données `mais_MSR.xls`. Chaque colonne représente des mesures effectuées sur des plantes indépendantes de la masse sèche de leurs racines. Certaines ont reçu un traitement, d'autres non ! Déterminez avec un test de Student s'il y a une différence significative entre les deux échantillons.

Comparaison de

Exercice 4 : Analyse différentielle d'un jeu de données transcriptomiques

Allez sur le site GEO pour récupérer le jeu de données GSE6467 :
http ://www.ncbi.nlm.nih.gov/geo/query/acc.cgi ?acc=GSE6467 Récupérez un fichier compressé contenant les données brutes (sans normalisation) qui se présentent sous la forme de fichier ayant une extension `.CEL`. Stockez ces fichiers (après décompression) dans un dossier nommé CEL dans votre dossier de travail.

Suivez les indications du code pour effectuer la normalisation des données brutes. Une fois cela fait, vous devrez vous même :

(1) enlever les sondes contrôles commençant par la chaîne de caractères `AFFX`. (indications : utiliser la commande `grep` pour localiser ces sondes, puis la syntaxe `Matrice[- indices , ]` pour enlever les lignes spécifiées dans le vecteur `indices` de la matrice `Matrice`),

(2) faire une boucle `for` pour identifier les gènes différentiellement exprimés (d'une part avec un test de Student et d'autre par avec un test de Wilcoxon),

(3) utiliser la fonction `p.adjust` pour calculer les p-values ajustées (stocker ces p-values ajustées dans le vecteur `padj`).

Comparez avec le jeu de données Golub.

Que concluez-vous sur la première étude ?