

Devoir Maison 1

Aude Sportisse

Ce devoir maison est à envoyer par mail (aude.sportisse@upmc.fr (mailto:aude.sportisse@upmc.fr)) pour le lundi 28 septembre (soir) en format .Rmd ou .ipynb et .html.

Il comporte deux exercices, on pourra utiliser la fiche 1.

Exercice 1

Nous allons nous intéresser au jeu de données decathlon (<https://audesportisse.github.io/files/decathlon.csv>). Ce jeu de données comprend les performances des athlètes pour les dix épreuves du décathlon (10 premières colonnes), le classement des athlètes (colonne 11), les points obtenus (colonne 12) et la compétition où cela s'est déroulé (colonne 13).

1. Charger le jeu de données avec le nom des colonnes et le nom des lignes (qui correspond à la première colonne du csv) en utilisant la fonction **read.table** et en spécifiant correctement les arguments `sep`, `row.names` et `header`.
2. Faire une analyse rapide des données (résumé des données, dimension, type des variables).
3. Nous allons nous intéresser à la variable **Longueur**. A l'aide de la fonction **sum**, calculer sa moyenne et son écart-type empiriques. A l'aide de la fonction **sort** et **ceiling**, déterminer la médiane empirique de cette variable.

On rappelle que le quantile empirique d'ordre p d'un échantillon de taille n est défini par

$$x_p(n) = \inf\{x \in \mathbb{R}, F_n(x) \geq p\} = X_{(\lceil np \rceil)},$$

où F_n est la fonction de répartition empirique et $X_{(1)} \leq \dots \leq X_{(n)}$ est l'échantillon ordonné. La médiane empirique correspondant au quantile empirique d'ordre $1/2$.

4. Retrouver les indicateurs des questions précédentes à l'aide des fonctions **mean**, **sd** et **median**. A quoi peuvent être dues les différences éventuelles avec les réponses des questions précédentes selon vous ?
5. Donner la moyenne et les écarts-types des 10 premières colonnes en utilisant la fonction **apply**.
6. Donner la moyenne des points (colonne 12) des joueurs qui ont participé au Decastar et celle des joueurs qui ont participé au JO pour en déduire quelle compétition semble la plus relevée. Laquelle réunissait le plus d'athlètes ?

Exercice 2

Cet exercice est un extrait de "Probability: Theory and Examples", Rick Durrett, page 316 deuxième édition.

Le professeur demande à l'étudiant 1 d'écrire au tableau une séquence de 100 nombres, au hasard, compris entre 0 à 9. L'étudiant 2 choisit un des dix premiers nombres de la séquence écrite au tableau et ne le dit pas au professeur. Il note bien la position qu'a ce nombre dans la séquence. L'étudiant va parcourir la séquence de nombres de la manière suivante: si le nombre choisi est 7 et que c'est le deuxième nombre de la liste, la nouvelle position sera égale à $2+7=9$.

Par exemple, supposons que les nombres écrits au tableau sont 1 7 8 9 0 2 1 4 3 6 4 9 ... Si 7 est choisi à la deuxième position, c'est ensuite le nombre à la $2+7=9$ ème position qui est choisi, c-à-d 3 et c'est ensuite le nombre à la $9+3=12$ ème position, c-à-d 9...

Si le nombre est 0, la nouvelle position sera calculée en ajoutant 10 à la position initiale. L'étudiant 2 parcourt la séquence tant que la nouvelle position ne dépasse pas 100. Quant à lui, le professeur choisit le premier nombre de la séquence de nombres et parcourt ensuite la séquence de nombres en utilisant la même méthode que l'étudiant. La probabilité que le professeur n'obtienne pas la même position finale que l'étudiant 2 est approximativement égale à 0.026.

Le but de cet exercice est d'écrire un programme permettant de retrouver cette probabilité de 0.026.

Pour estimer la probabilité d'un évènement (ici l'évènement est "le professeur se trompe"), nous pouvons répéter l'expérience un grand nombre de fois et compter combien de fois l'évènement se produit. On estime donc la probabilité p par C/n , où C est la variable aléatoire de comptage et n le nombre d'expériences.

Nous allons d'abord simuler une seule expérience (questions 1 à 5).

1. Simuler la séquence de nombre que l'étudiant 1 écrit au tableau. (Pour simplifier la suite, vous pouvez supposer que l'étudiant 1 écrit des nombres de 1 à 10).
2. Le professeur choisit le premier nombre de cette séquence (le premier nombre a l'indice (= la position) 1 dans la séquence). Ecrire le code qui donne l'indice du nombre que le professeur obtient à la fin. Attention cet indice ne doit pas être supérieur à 100 !
3. L'étudiant 2 choisit (au hasard) un nombre parmi les 10 premiers, simuler l'indice du nombre qu'il choisit.
4. Ecrire le code qui donne l'indice du nombre que l'étudiant 2 obtient à la fin.
5. On rappelle que l'évènement "le professeur se trompe" se produit si l'indice du nombre du professeur obtenu à la fin est différent de celui de l'étudiant 2. Tester si l'évènement s'est produit dans cette expérience.
6. En faisant un grand nombre d'expériences, retrouver la probabilité que le professeur a de se tromper.