Devoir maison 2

Aude Sportisse

14/09/2020

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 et .html.

Il comporte deux exercices, on pourra utiliser les fiches 2, 3 et 4.

Exercice 1

Nous allons nous intéresser au jeux 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 premières colonne du csv) en utilisant la fonction **read.table** et en specifiant correctement les arguments sep, row.names et header.

Dans la suite, donner un titre à vos graphiques et vérifier que les titres des axes sont indicatifs.

2. Représenter un nuage de points de la variable **Longueur** on fonction de la variable **Points**. Que remarque-t-on ?
3. Représenter le même nuage de point avec une couleur différente par compétition (la compétition est donnée par la variable Competition). Dans cette question, tracer 3 graphiques. Utiliser les couleurs par défault dans un premier temps (premier graphique) puis choisir ses propres couleurs (deuxième graphique). Utiliser ensuite des formes différentes au choix en fonction de la compétition (troisième graphique).
4. Sur le nuage de points obtenu avec différentes couleurs par compétition, tracer un lissage avec la méthode "loess". En spécifiant l'argument group dans la fonction aes, tracer deux courbes de lissage, une pour chaque compétition (idéalement, les courbes d'une compétition ont la même couleur que les points de cette compétition).
5. Tracer le boxplot de la variable Longueur en fonction de la compétition. Tracer un autre boxplot de la variable X100m en fonction de la compétition. Interpréter.
6. Représenter par un graphique le nombre d'individus présents dans chacune des compétitions.
7. Représenter l'histogramme de la variable Longueur en réglant le nombre de classes, superposer l'estimateur de la densité. Les données sont-elles gaussiennes ?
8. Dans l'exercice 3 de la Fiche 3, nous avons déterminer un intervalle de confiance asymptotique pour la moyenne μ d'un échantillon $X = (X_1, \dots, X_n)$ qui n'est pas forcément gaussien. Cet intervalle asymptotique de niveau $1 - \alpha$ est le suivant:

$$\left[\bar{X}_n - \frac{q\hat{\sigma}}{\sqrt{n}} \; ; \; \bar{X}_n + \frac{q\hat{\sigma}}{\sqrt{n}}\right]$$

avec q le quantile d'ordre $(1-\alpha/2)$ d'une gaussienne centrée réduite, $\hat{\sigma}$ l'écart-type empirique et $\bar{X}$ la moyenne empirique de X. Ecrire une fonction qui prend en entrée X et α et qui renvoie l'intervalle de confiance.

9. A l'aide de cette fonction, calculer un intervalle de confiance de niveau 95% et 90% de la Longueur. Retrouver le résultat avec la fonction **t.test**. Expliquer "à la main" lequel et pourquoi des intervalles de confiances est le plus petit.

Exercice 2

Au Tour de France, plusieurs classements importent dont le classement général pour le maillot jaune (qui récompense le meilleur coureur) et le classement par équipes (qui récompense la meilleure équipe).

Le classement individuel a largement été dominé par l'équipe Ineos Grenadiers (alias Sky) de 2012 à 2019 (victoire tous les ans d'un coureur de l'équipe sauf en 2014). Pourtant, Ineos n'a remporté qu'une seule fois le classement par équipes, alors que la Movistar l'a remporté 4 fois de 2012 à 2019.

Pour être bien classé au Tour de France (sans rêver du maillot jaune, rêvons d'abord du top 10…), vaut-il mieux avoir un contrat avec Ineos ou avec la Movistar, ou n'y a-t-il pas de différence ?

Le but de l'exercice est de tester l'indépendance entre deux variables qualitatives, le classement individuel d'un joueur (top 10, 11ième à 50ième, + de 50ième) et son appartenance à l'équipe Ineos/Movistar.

1. Les données sont contenues dans la matrice suivante, appelée tableau de contingence $(N_{ij})_{1\leq i\leq I, 1\leq j\leq J}$. Créer la matrice sur R.

	Top10	**11-50**	**+50**
Ineos	11	22	28
Movistar	12	17	31

Pour tester l'indépendance de deux variables qualitatives, on teste l'hypothèse nulle H_0 : "les deux variables sont indépendantes" contre l'hypothèse alternative H_1 : "les deux variables ne sont pas indépendantes". Pour cela, on construit la statistique de test suivante :

$$T_n = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{ij} - \frac{N_{i\bullet}N_{\bullet j}}{n})^2}{\frac{N_{i\bullet}N_{\bullet j}}{n}},$$

où

- N_{ij} est l'effectif pour la modalité i de la première variable et la modalité j de la seconde,
- $N_{i\bullet}$ correspond à l'effectif pour la modalité i de la première variable
- $N_{\bullet j}$ correspond à l'effectif pour la modalité j de la deuxième variable, I et J les nombres de modalités de chacune des variables.
- n représentant l'effectif total.

On peut montrer que, sous l'hypothèse H_0,

$$T_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{(I-1)\times(J-1)},$$

tandis que, sous H_1,

$$T_n \xrightarrow[n\to\infty]{p.s.} +\infty.$$

On rejette donc H_0 si l'observation $t_n = T_n(\omega)$ de la statistique de test prend une grande valeur.

2. Effectuer le test au niveau 5% "à la main" en expliquant bien les lignes du code et en ayant le code le plus concis possible.
3. A l'aide de la fonction **chisq.test**, faire le test.
4. Retrouver la p-value à la main et interpréter.