Programmation Statistique sous R
EXAMEN
Documents autorisés
Durée : 1H30

Antoine ROLLAND et Anthony SARDELLITTI

24 Mars 2023

CONSIGNES

Vous avez 1h30 pour réaliser l'ensemble du devoir. Vous devez vous connecter sur les sessions EXAM. Il est conseillé de copier-coller le fichier .csv sur votre Bureau. Le cours en ligne a été téléchargé et est disponible sous format HTML (à ouvrir avec EDGE et pas Chrome). Si lors de la navigation quelques liens ne fonctionnent pas, il suffit d'ouvrir directement le fichier HTML souhaité avec EDGE.

Voici les recommandations pour que votre travail soit pris en compte et pour ne pas avoir de pénalités :

- Le script R est à rendre sur la session EXAM
- Votre fichier doit se nommer **NOM_PRENOM** et doit être un script R
- Le rendu est individuel
- Pour chaque question n'oubliez pas :
  - d'écrire le code R permettant de répondre à la question
  - de préciser le numéro de la question et de l'exercice en commentaire.

Exemple :

```
# Ex 1-b
Data<-read.csv()
```

Présentation

Pour cet examen, on utilise le fichier `games` qui présente les statistiques de parties d'échecs sur le site Lichess.org. Aux échecs, le joueur avec les pions blancs joue le premier coup de la partie.

Voici une description des données :

- `id` : L'id de la partie
- `rated` : Si la partie est classée
- `turns` : nombre de coups dans la partie
- `victory_status` : Le type de victoire (Draw = Match nul , Resign = Abandon , mate = échec et mat , outoftime = temps écoulé)
- `winner` : le vainqueur
- `white_id` : id du joueur blanc
- `white_rating` : classement du joueur blanc
- `black_id` : id du joueur noir
- `black_rating` : classement du joueur noir
- `moves` : l'ensemble des mouvements de la partie
- `opening_eco` : l'id du type d'ouverture
- `opening_name` : nom de l'ouverture
- `opening_ply` : nombre de coups dans la phase d'ouverture

Conseils

- Pensez à copier le dataset sur votre bureau
- N'ouvrez pas le fichier avec excel mais privilégiez le bloc note ou Notepad+
- Pensez à vous autocorriger avant de vous lancer dans la question suivante
- Pensez à sauvegarder votre script régulièrement

1 Exercice 1 : Importer les données

a. Importez le jeu de données `games` avec la fonction `read.csv( )` **uniquement** dans un objet appelé `df`.

b. Combien de parties ont été jouées ?

c. Affichez un résumé des données avec la fonction adaptée.

d. Dans le dataframe, transformez les variables `victory_status`, `winner` , `rated`, et `opening_name` en type factor

e. Supprimez la colonne `moves` dans le dataframe.

f. Créez une variable `rating_difference` qui est correspond à la différence de classement entre les deux joueurs. On souhaite que cette différence soit en valeur absolue. Aidez-vous de la fonction `abs()` pour cela.

2 Exercice 2 : Statistiques descriptives

a. Quel est le nombre de tours moyen d'une partie ?

b. Calculez les **centiles** du nombre de tours. D'après le résultat, quelle est le pourcentage de parties jouées en 9 tours ou moins ?

c. Proposez un représentation graphique pertinente pour étudier la distribution du nombre de tours des parties.

d. Quel est l'écart-type de la variable `rating_difference` ?

e. Combien de joueurs sont représentés dans le jeu de données

f. Calculez la répartition en pourcentage de la variable `winner`. Représenter cette répartition dans un diagramme circulaire.

g. Calculer la répartition des `victory_status`. Représentez cette répartition dans un diagramme en barre.

3 Exercice 3 : Requêtes

Pour chaque question, il est recommandé de sauvegarder le résultat de la requête dans un objet puis de le visualiser dans une vue pour vérifier.

a. Construire une requête pour extraire les 10 parties avec le moins de coups joués.

b. Construire une requête pour extraire les 10 parties avec le plus d'écart de niveau au classement.

c. Construire une requête pour avec uniquement les parties classées où un joueur moins bien classé que l'adversaire a gagné.

d. Construire une requête avec les 10 joueurs qui ont gagné le plus de parties.

4 Exercice 4 : Statistique

Les joueurs d'échecs et les théoriciens s'accordent à dire que les Blancs commencent le jeu avec un certain avantage. Les statistiques compilées depuis 1851 militent en ce sens, montrant que les Blancs gagnent toujours un peu plus souvent que les Noirs, marquant généralement entre 52 % et 56 % des points. Au vu de notre échantillon, peut-on confirmer cette théorie ? On prendra un risque de 5% ? Pour cela, vous aurez besoin de la formule pour calculer un intervalle de fluctuation d'une proportion. Attention, il ne faut pas tenir compte des parties où il y a eu match nul.

$$p \pm z\sqrt{\frac{p(1-p)}{N}}$$