Analyse statistique des parties d'échecs en ligne

Emmanuel GUEISSAZ Quentin FOURNIER

2023-05-05

Introduction

Origine

Le jeu de données utilisé dans cette étude statistique provient du site www.kaggle.com. Vous pouvez le trouver en suivant le lien suivant : https://www.kaggle.com/datasets/ulrikthygepedersen/online-chess-games.

Description

Ce jeu de données contient des informations sur une vaste gamme de parties d'échecs en ligne. Le fichier contient les informations de 20058 parties dont l'individu est la partie.

Les échecs sont un jeu de plateau stratégique pour deux joueurs. Il se joue sur un plateau divisé en cases carrées et composé de 64 cases au total, alternant entre les couleurs claires (généralement blanc) et foncées (généralement noir). Chaque joueur dispose d'un ensemble de 16 pièces, comprenant un roi, une dame, deux tours, deux fous, deux cavaliers et huit pions.

L'objectif du jeu est de mettre le roi de l'adversaire en échec et mat, ce qui signifie qu'il est attaqué sans possibilité d'échapper à la capture. Les joueurs déplacent leurs pièces de manière alternée, en utilisant des stratégies tactiques et des mouvements spécifiques pour attaquer les pièces adverses, défendre les leurs et contrôler le plateau.

Chaque pièce a des règles de mouvement spécifiques. Par exemple, la tour se déplace horizontalement ou verticalement sur le plateau, le fou se déplace en diagonale, le cavalier effectue des mouvements en forme de L, tandis que la reine peut se déplacer dans toutes les directions sur le plateau.

Le jeu d'échecs est apprécié à la fois comme un loisir et comme un sport intellectuel. Il exige une réflexion stratégique, la prévision des coups de l'adversaire et la capacité de planifier des séquences de coups. Les échecs sont joués à différents niveaux de compétence, du niveau amateur au niveau professionnel, avec des tournois organisés à l'échelle mondiale.

Dans le cadre de cette analyse statistique des parties d'échecs en ligne, nous allons examiner les données pour comprendre les schémas de jeu, les performances des joueurs et les facteurs qui peuvent influencer les résultats des parties.

Questions intéressantes :

- Y a-t-il une association entre le classement du vainqueur et le nombre de tours joués dans une partie d'échecs ?
- Est-ce que le classement du vainqueur influence le nombre de tours joués dans une partie ?
- Est-ce que la couleur impacte directement le résultat de la partie ?

Nettoyage des données

Nous n'avons pas rencontré de problème avec les données, il ne nous a pas semblé pertinent de faire un nettoyage.

Variables quantitatives

Nombre de tours d'une partie

- Cette variable dicrète représente le nombre de tours joués lors d'une partie d'échecs en ligne.
- Elle est mesurée en termes de la quantité d'échanges de coups entre les joueurs.
- Cette variable permet d'évaluer la durée et la complexité d'une partie.

Classement du vainqueur

- Cette variable continue indique le classement du vainqueur d'une partie d'échecs en ligne.
- Plus elle est haute, plus le joueur est censé être bon.

- Le classement permet de déterminer la performance des joueurs.

Variables qualitatives

Couleur du vainqueur

- Cette variable représente la couleur (withe (blanc) ou black (noir)) du joueur qui remporte la partie.

Partie classée ou non

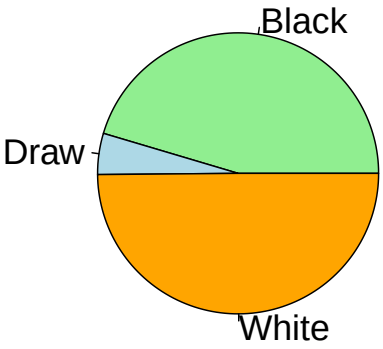
- Cette variable indique si une partie est classée ou non.
- Une partie classée est généralement enregistrée dans un système de classement officiel, tandis qu'une partie non classée est jouée à des fins de pratique ou de divertissement.

Analyses

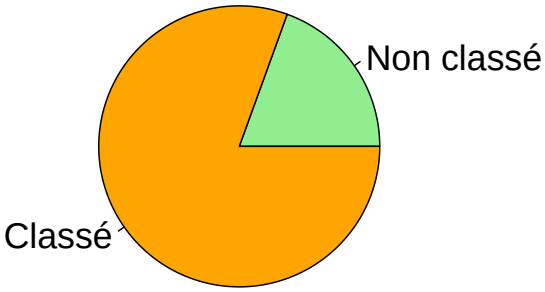
Analyse univariée

Exploration statistique

Répartition des gagnants



Répartition des parties



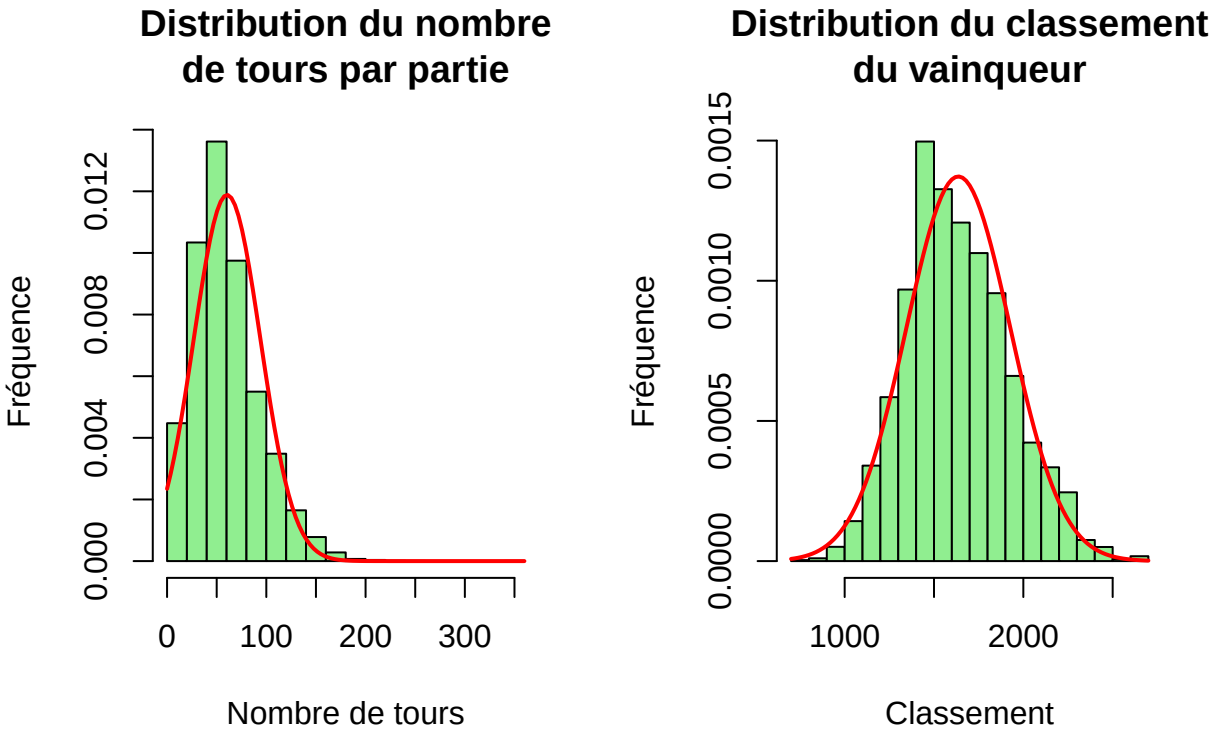
Qualitatives

- On remarque les blancs gagnent un peu plus souvent,
- On remarque aussi que la plupart des parties sont en mode classé.

Quantitatives Les analyses statistiques effectuées sur les parties d'échecs en ligne révèlent les résultats suivants :

- La moyenne du nombre de tours réalisés dans chaque partie est de 60.47 tours.
- La variance du nombre de tours réalisés dans chaque partie est de 1126.98.
- La moyenne des classements des partie est de 1636.35.
- La variance des rangs est de 8.454669×10^4.

Distribution normale



En examinant l'histogramme du nombre de tours par partie et la courbe rouge représentant la distribution normale, on peut observer une certaine similarité entre les deux. La courbe rouge semble suivre approximativement la forme de l'histogramme, suggérant une possible adéquation à une distribution normale.

Tests de confiance

Les tests de confiance ont été réalisés pour évaluer les intervalles de confiance des variables quantitatives étudiées. Voici les résultats obtenus :

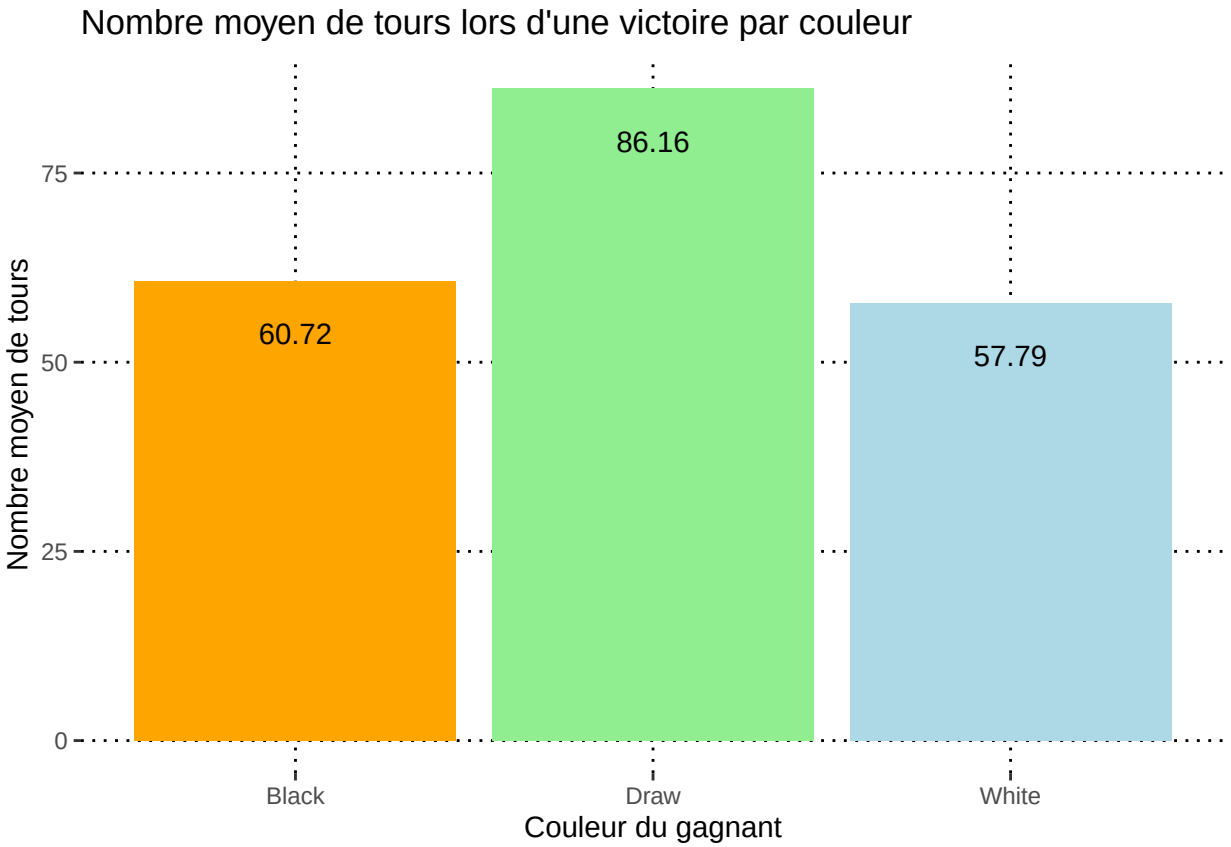
- Nombre de tours : L'intervalle de confiance à 95 % pour le nombre de tours réalisés dans chaque partie est de 60 à 60.93. Cela signifie qu'avec un niveau de confiance de 95 %, on peut estimer que la moyenne réelle du nombre de tours dans la population se situe entre 60 et 60.93.
- Classement du vainqueur : L'intervalle de confiance à 95 % pour le classement du vainqueur est de 1632.33 à 1640.38. Cela signifie qu'avec un niveau de confiance de 95 %, on peut estimer que la moyenne réelle du classement du vainqueur dans la population se situe entre 1632.33 et 1640.38.

Analyse de la proportion de vainqueur par couleur

Les résultats de l'analyse indiquent que l'intervalle de confiance à 95% pour la proportion de vainqueurs blancs est de `0.49` à `0.51`, tandis que l'intervalle de confiance à 95% pour la proportion de vainqueurs noirs est de `0.45` à `0.46`. On en déduit donc que les blancs gagnent plus souvent les parties d'échec en ligne.

Analyse multivarié

Quantitative x Qualitative



Nous observons que la moyenne du nombre de tours pour les parties remportées par les blancs est de 60.72 tours, tandis que pour les parties remportées par les noirs, la moyenne est de 57.79 tours. En ce qui concerne les parties déclarées comme une égalité, la moyenne du nombre de tours est de 86.16 tours.

Le test t-test est généralement utilisé pour comparer les moyennes de deux groupes. Cependant, dans notre cas, la variable “winner” a trois niveaux distincts : “Black”, “Draw” et “White”. Le t-test ne peut pas être appliqué directement car il ne permet de comparer que deux groupes à la fois. Par conséquent, nous utilisons le test de Kruskal-Wallis, qui est un test non paramétrique adapté à notre situation. Le test de Kruskal-Wallis nous permet de comparer les distributions des variables “turns” entre les différents niveaux de la variable “winner” et de déterminer s’il y a des différences significatives entre ces groupes.

Le test de Kruskal-Wallis a produit une statistique de test (chi-carré) de 397.68 avec 2 degrés de liberté. La valeur de p associée est inférieure à 2.2e-16, ce qui indique une différence significative entre les médianes des “turns” pour les différentes catégories de la variable “winner”.

Cela suggère qu’il existe une corrélation significative entre la couleur du vainqueur et le nombre de tours joués dans la partie. Cependant, il ne permet pas d’identifier quelles paires de groupes diffèrent significativement les unes des autres.

Qualitative x Qualitative

Table 1: Tableau des fréquences relatives de la couleur du vainqueur selon la classification de la partie

	Black	Draw	White
Non classé	0.0859009	0.0115166	0.0971682
Classé	0.3681324	0.0358460	0.4014358

Parmi les parties non classées, on observe que 8.59% ont été remportées par les joueurs noirs, 1.15% se sont terminées par une égalité et 9.72% ont été remportées par les joueurs blancs.

Parmi les parties classées, 36.81% ont été remportées par les joueurs noirs, 3.58% se sont terminées par une égalité et 40.14% ont été remportées par les joueurs blancs.

On peut remarquer que les proportions de victoires diffèrent entre les parties classées et non classées. Les parties classées semblent présenter un pourcentage plus élevé de victoires pour les joueurs blancs, tandis que les parties non classées montrent une distribution plus équilibrée.

Le test du chi carré a été utilisé pour évaluer l'association entre la variable "Couleur du vainqueur" (Black, Draw, White) et la variable "Partie classée" (Non classé, Classé) dans notre jeu de données.

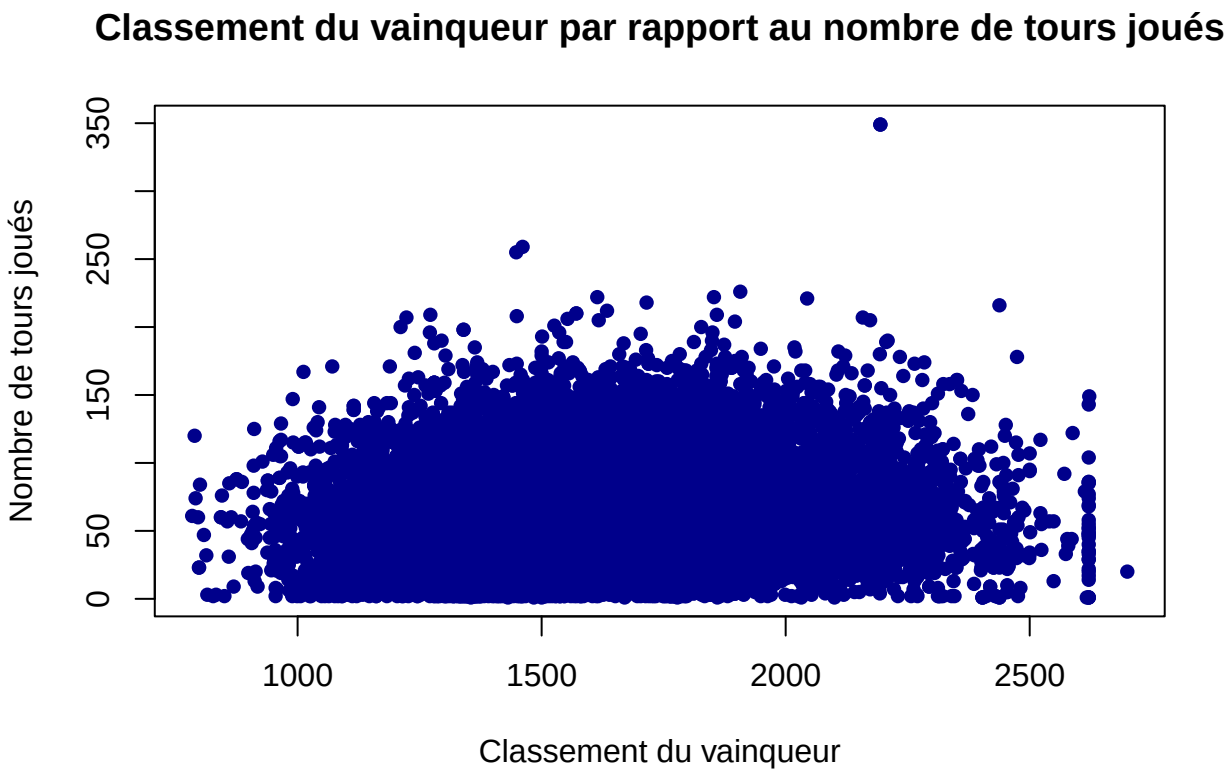
```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 15.995, df = 2, p-value = 0.0003363
```

Les résultats du test indiquent une statistique de test (X-squared) de 15.995 avec 2 degrés de liberté, ce qui correspond à un niveau de significativité de p = 0.0003363.

Cette analyse met en évidence une association statistiquement significative entre la couleur du vainqueur et le fait que la partie soit classée ou non. Autrement dit, la couleur du vainqueur et le statut de la partie semblent être liés d'une manière non aléatoire.

Ces résultats suggèrent que la classification des parties et la couleur du vainqueur sont des variables importantes à prendre en compte lors de l'analyse des jeux d'échecs. Cela pourrait avoir des implications pour la stratégie de jeu et les performances des joueurs.

Quantitative x Quantitative



Lors de l'analyse du nuage de points entre le classement du vainqueur et le nombre de tours joués dans la partie, nous constatons une dispersion importante des points et une corrélation faible de `0.101138`. Cela suggère qu'il n'y a pas de relation linéaire forte entre ces deux variables.

La faible corrélation indique que le classement du vainqueur ne semble pas être fortement influencé par le nombre de tours joués. En d'autres termes, le nombre de tours ne semble pas être un facteur déterminant dans le résultat du classement du vainqueur.

Il est important de prendre en compte que d'autres facteurs ou variables pourraient influencer le classement du vainqueur dans les parties de jeu d'échecs.

Nous avons effectué un test de normalité de Shapiro-Wilk sur un échantillon aléatoire de 5000 valeurs de la variable winClassement dans notre jeu de données. Les résultats du test indiquent une statistique de test (W) de 0.9885 et une valeur de p très faible (< 2.2e-16).

Ces résultats suggèrent fortement que l'échantillon ne suit pas une distribution normale. La faible valeur de p indique un rejet de l'hypothèse nulle selon laquelle l'échantillon provient d'une distribution normale.

Les résultats sont identiques pour le nombre de tours. Malgré l'apparence graphique suggérant une distribution gaussienne des variables, les tests statistiques ont révélé que celles-ci ne suivent pas une distribution normale.

```
##
## Call:
## lm(formula = turns ~ winClassement, data = chess_games_datas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.964 -23.796  -5.096  18.236 282.022
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.136e+01  1.348e+00   30.68   <2e-16 ***
## winClassement 1.168e-02  8.111e-04   14.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.4 on 20056 degrees of freedom
## Multiple R-squared:  0.01023,    Adjusted R-squared:  0.01018
## F-statistic: 207.3 on 1 and 20056 DF,  p-value: < 2.2e-16
```

Les coefficients du modèle indiquent que l'intercept (constante) est de 41.36 et le coefficient de la variable winClassement est de 0.01168. Les valeurs p associées aux coefficients sont très faibles (< 2e-16), ce qui indique une forte significativité statistique.

La statistique t associée au coefficient de winClassement est de 14.40, ce qui suggère une relation significative entre le classement du vainqueur et le nombre de tours joués. La valeur de R carré multiple est de 0.01023, ce qui indique que seulement 1.02% de la variance dans le nombre de tours peut être expliquée par le classement du vainqueur.