Analiza partii szachowych

Martyna Bielec, Maciej Karczewski

2022-12-15

Wprowadzenie

Strona internetowa Lichess jest drugą co do wielkości platformą na świecie, która umożliwia grę w szachy online. Użytkownicy mogą grać z komputerem lub między sobą. Każdy gracz ma przypisany ranking (początkowo ma on wartość 1500), który jest ustalany na podstawie jego umiejętności. Jeśli użytkownik rozgrywa partię "rankingową", w jej wyniku jego ranking rośnie lub maleje.

Dane, które zostaną poddane analizie, zawierają informacje o ponad 20 000 rozgrywkach szachowych ze strony Lichess Games. Dostępne są na witrynie kaggle. Składają się z następujących kolumn:

- id (zawiera elementy typu "character") - numer ID rozgrywki
- rated ("character") - wartość "True" lub "False" w zależności od tego, czy partia jest "rankingowa"
- created_at ("numeric") - czas rozpoczęcia gry
- last_move_at ("numeric") - czas zakończenia gry
- turns ("numeric") - liczba posunięć
- victory_status ("character") - status zakończenia gry: "mate" dla zakończenia matem, "resign" dla zakończenia przez poddanie się jednej ze stron, "draw" dla remisu lub "outoftime" dla rozgrywki zakończonej z powodu upływu czasu
- winner ("character") - zwycięzca gry: "white", jeśli wygrał gracz biały, "black", jeśli czarny lub "draw" w przypadku remisu
- increment_code ("character") - składa się z dwóch liczb rozdzielonych znakiem "+"; pierwsza oznacza liczbę minut przypadających na gracza na wszystkie jego ruchy, druga liczbę sekund, o którą zwiększa się czas gracza po wykonaniu ruchu
- white_id ("character") - ID gracza białego
- white_rating ("numeric") - ranking gracza białego
- black_id ("character") - ID gracza czarnego
- black_rating ("numeric") - ranking gracza czarnego
- moves ("character") - wszystkie ruchy zapisane w standardowej notacji szachowej
- opening_eco ("character") - standaryzowany kod dla rozgranego otwarcia
- opening_name ("character") - nazwa rozgranego otwarcia
- opening_ply ("character") - liczba ruchów w fazie otwarcia

Celem analizy będzie odpowiedzenie na pytanie - jakie jest prawdopodobieństwo różnych zakończeń gry: wygranej białego gracza, czarnego lub remisu? Będziemy rozważać ten problem pod kątem rozegranego otwarcia, rankingu graczy, różnicy w rankingu między graczem białym i czarnym oraz liczby posunięć w partii.

Użyjemy metody bootstrap do stworzenia przedziałów ufności estymowanego prawdopodobieństwa. Będą to przedziały percentylowe, zaznaczone na wykresach czerwonymi kreskami.

Obróbka danych

Usunięcie niepotrzebnych kolumn

Nasze dane zawierają dużo kolumn, z których nie korzystamy. Usuniemy niepotrzebne kolumny, w wyniku czego nasza tabela z danymi wygląda następująco:

```
##   turns winner white_rating black_rating                          opening_name
## 1    13  white         1500         1191       Slav Defense: Exchange Variation
## 2    16  black         1322         1261 Nimzowitsch Defense: Kennedy Variation
## 3    61  white         1496         1500  King's Pawn Game: Leonardis Variation
## 4    61  white         1439         1454 Queen's Pawn Game: Zukertort Variation
## 5    95  white         1523         1469                        Philidor Defense
## 6     5   draw         1250         1002   Sicilian Defense: Mongoose Variation
```

Usunięcie duplikatów i wierszy z brakiem danych

W następnym kroku usuwamy duplikujące się wiersze oraz wiersze, w których nie ma danych.

Zastosowanie One Hot Encoding dla kolumny "winner"

Wynik naszej partii znajduje się w kolumnie "winner". Ta kolumna zawiera 3 wartości: "white" , "black" i "draw". Dla ułatwienia analizy stworzymy 3 dodatkowe kolumny o odpowiednich nazwach "white_win", "black_win" oraz "draw". Każda z tych kolumn ma wartość 1 lub 0. Gdy w kolumnie "winner" znajduje się "white", czyli zwyciężył biały gracz, to 1 przypisujemy do "white_win", a do "black_win" oraz "draw" przypisujemy 0. Analogicznie postępujemy dla pozostałych wyników.

```
df <-  df %>% mutate(
    white_win = ifelse(winner =='white',1,0),
    black_win = ifelse(winner =='black',1,0),
    draw = ifelse(winner =='draw',1,0)
)
```

Wybór najczęstszych otwarć

W raporcie przeanalizujemy prawdopodobieństwo różnych zakończeń rozgrywki także ze względu na wybrane otwarcie. Aby uzyskać rzetelne wyniki, nie możemy uwzględnić otwarć, które zostały rozegrane za mało razy. Okazuje się jednak, że kolumna "opening_name" zawiera 1477 unikalnych wartości. Dzieje się tak, ponieważ dane rozróżniają wiele wariantów poszczególnych otwarć. Przykładowo "Obrona Sycylijska" została rozegrana w 181 różnych wariantach, takich jak:

```
## [1] "Sicilian Defense: Mongoose Variation"
## [2] "Sicilian Defense: Bowdler Attack"
## [3] "Sicilian Defense: Smith-Morra Gambit #2"
## [4] "Sicilian Defense: Canal Attack |  Main Line"
## [5] "Sicilian Defense: Dragon Variation |  Yugoslav Attack |  Main Line"
```

W celu umożliwienia analizy otwarć stworzymy nową kolumnę, zawierającą jedynie nazwę otwarcia (bez uwzględnienia wariantu). Nazywamy ją "openings_general". Zawiera ona 149 wartości unikalnych. Przeanalizujemy te otwarcia, których zostało rozegranych więcej niż 500, czyli:

```
## # A tibble: 10 x 2
##    openings_general     total_count
##    <chr>                      <int>
##  1 Sicilian Defense            2581
##  2 French Defense              1377
##  3 Queen's Pawn Game           1203
##  4 Italian Game                 952
##  5 King's Pawn Game             897
##  6 Ruy Lopez                    832
##  7 English Opening              715
##  8 Scandinavian Defense         707
##  9 Philidor Defense             670
## 10 Caro-Kann Defense            584
```

Kolumna "total_count" oznacza liczbę rozegranych gier z wykorzystaniem danego otwarcia.

Analiza danych

Prawdopodobieństwo różnych zakończeń rozgrywki bez warunkowania.

W celu rozważenia od czego zależy prawdopodobieństwo różnych zakończeń rozgrywki, najpierw oszacujemy to prawdopodobieństwo w ogólności. Liczba rozegranych partii wynosi 19618, z czego gracz biały wygrał 9782 razy, a czarny 8916 razy.

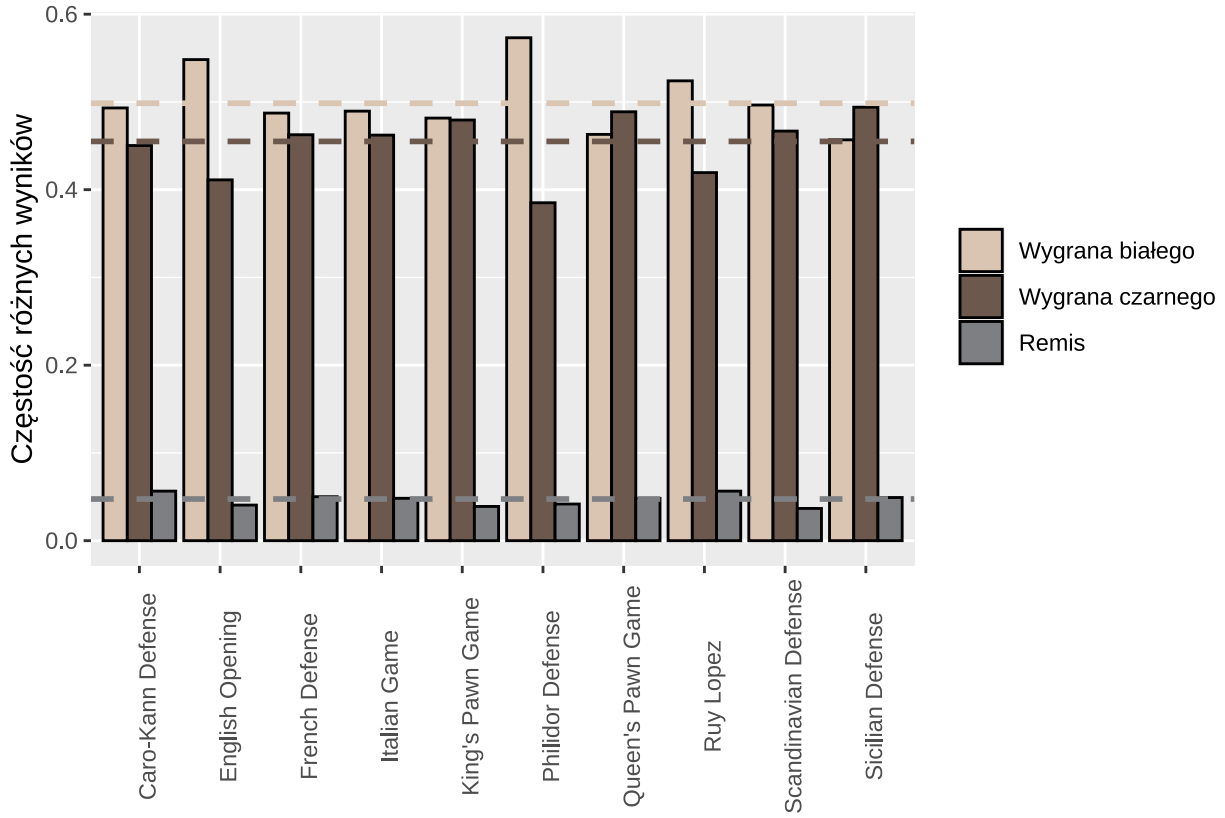
Wynika z tego, że szacowane prawdopodobieństwo wygrania gracza białego wynosi w przybliżeniu 49,85%, gracza czarnego 45,40%, a remisu 4,75%.

W dalszej części analizy sprawdzimy, jak to prawdopodobieństwo się zmienia, jeśli odpowiednio uwarunkujemy rozgrywkę.

Warunkowanie otwarciem

Przeanalizujemy jak zmienia się prawdopodobieństwo różnych wyników partii w zależności od otwarcia. Pod uwagę bierzemy dziesięć najpopularniejszych, gdzie dla każdego mamy co najmniej 500 obserwacji. Obliczamy częstość wygranych oraz remisu dla poszczególnych otwarć.

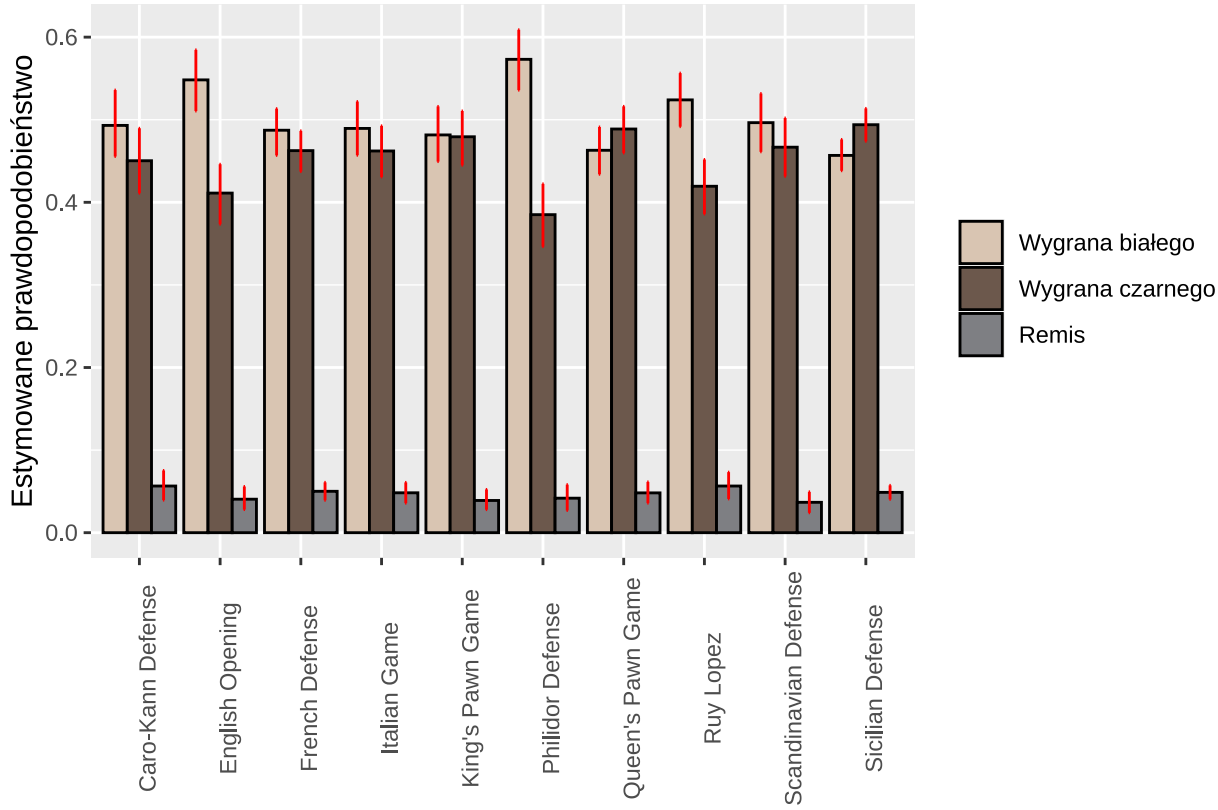
Przerywanymi liniami oznaczamy częstość poszczególnych wyników bez warunkowania: beżowa linia to częstość wygrywania gracza białego, brązowa czarnego, a szara remisu.



Z powyższego wykresu wynika, że biały gracz wygrywa statystycznie częściej niż dla średniej rozgrywki dla obrony Philidora ("Philidor Defense"), Otwarcia Angielskiego ("English Opening") oraz partii hiszpańskiej ("Ruy Lopez"). Szanse czarnego gracza na wygraną zwiększają się natomiast przy rozegraniu pozostałych z analizowanych otwarć oprócz obrony Caro-Kann ("Caro-Kann Defense"), gdzie zmalała jednocześnie częstość wygrywania gracza białego, natomiast zwiększyła się częstość remisu.

Przy dwóch z analizowanych otwarć gracz czarny nie wygrywał częściej niż gracz biały: przy obronie sycylijskiej ("Sicilian Defense") oraz otwarcia pionkiem hetmańskim ("Queen's Pawn Game").

Aby sprawdzić dokładność naszych obliczeń oraz wysunąć wnioski z danych, dołożymy przedziały ufności utworzone za pomocą metody bootstrap na poziomie ufności równym $0,05$.

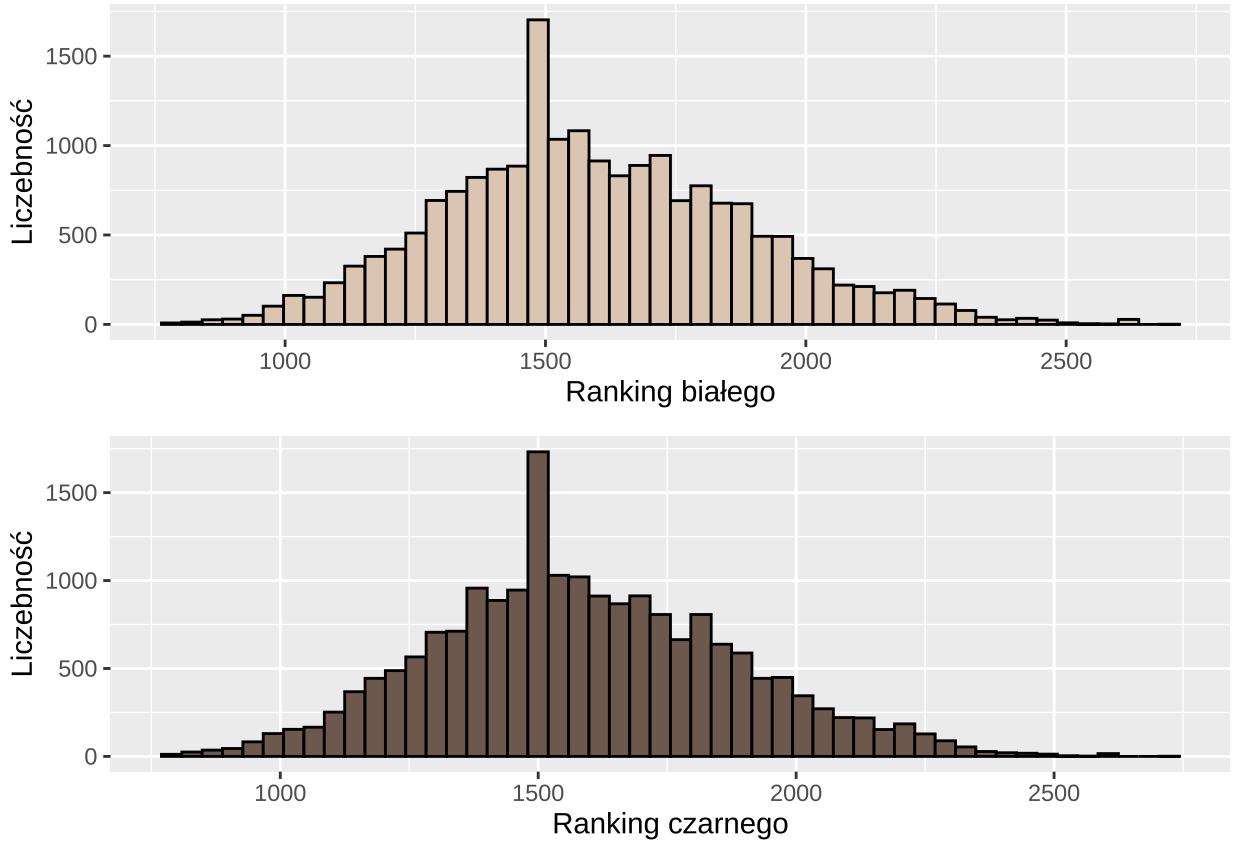


Okazuje się, że na poziomie ufności $0,05$ nie możemy stwierdzić dla wszystkich otwarć, który gracz ma większą szansę na wygraną. Konkretne wnioski możemy wysunąć jednak dla Otwarcia Angielskiego, Obrony Philidora oraz partii hiszpańskiej. Dla nich stwierdzamy na poziomie ufności $0,05$, iż prawdopodobieństwo wygrania białego gracza jest wyższe niż prawdopodobieństwo wygrania czarnego. Prawdopodobieństwo remisu nie różni się natomiast znacząco dla żadnego otwarcia.

Warunkowanie rankingiem białego

Gracze mają różne poziomy rankingu, więc sprawdzimy jak wyglądają prawdopodobieństwa różnych zakończeń rozgrywki w zależności od rankingu białego gracza.

Na początku zobaczmy, jak wygląda rozkład rankingu białego gracza, a jak czarnego, aby wykluczyć statystyczną przewagę jednego z nich.



Wykresy sugerują, że rozkład rankingu białego gracza i czarnego są bardzo podobne. Rodzi się pytanie, jak wyglądają ich średnie, mediany, wariancje, skośności czy kurtozy.

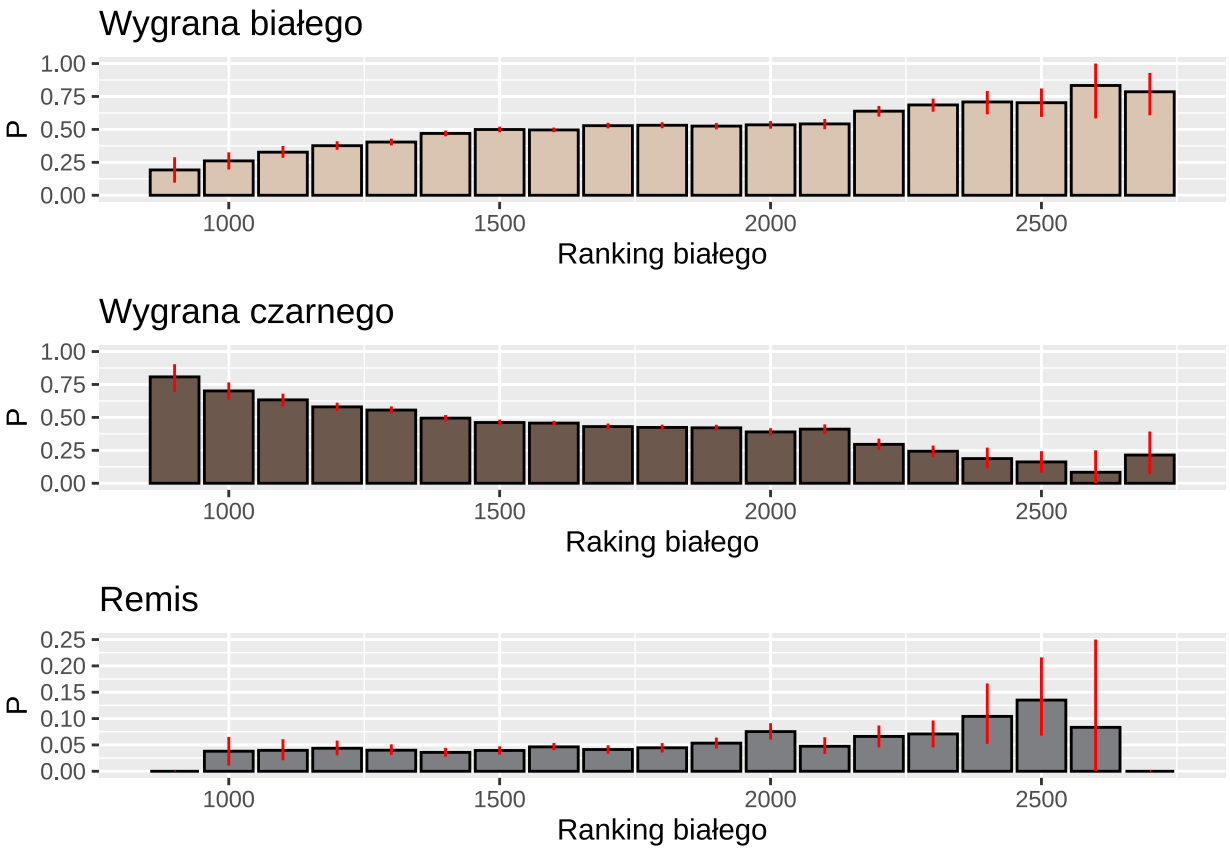
	biały	czarny
średnia	1596,15	1588,39
mediana	1567	1562
wariancja	84221,18	84214,67
skośność	0,30	0,25
kurtoza	0,029	-0,052

Analizując statystyki opisowe zawarte w tabelce, możemy dojść do wniosku, że biały ma średnio większy ranking. Ta różnica wynosi jedynie około 8 punktów, co przy średnim rankingu nieco mniejszym niż 1600 jest znikomą przewagą. Dodatkowo możemy odczytać, że mediana białego jest o 5 punktów większa od mediany czarnego co znowu jest małą różnicą w stosunku do mediany powyżej 1560. Możemy zobaczyć, że biały ma minimalnie większą wariancję, ale różni się tylko o 7, co przy wariancji białego równej $84221,18$ nie jest znaczącą różnicą. Następnie możemy zauważyć, że rozkład rankingu czarnego ma mniejszą skośność, ale obydwa rozkłady mają na tyle małe skośności w stosunku do odchylenia standardowego, że możemy przyjąć, iż ich rozkłady są symetryczne. Estymowane kurtozy są na tyle blisko 0, że możemy przyjąć, że rozkłady mają tyle samo danych odstających co rozkład normalny.

Powyższa analiza statystyk opisowych pozwala stwierdzić, że biały ma średnio przewagę, która nie jest jednak znacząca w naszej analizie.

By móc analizować wpływ rankingu białego na prawdopodobieństwo różnych zakończeń rozgrywki podzielimy ranking na przedziały. Każdy przedział ma szerokość 100. Przykładowo do przedziału, który nazwiemy 1400, trafiają obserwacje, dla których gracz biały ma ranking większy lub równy 1300 i mniejszy niż 1400.

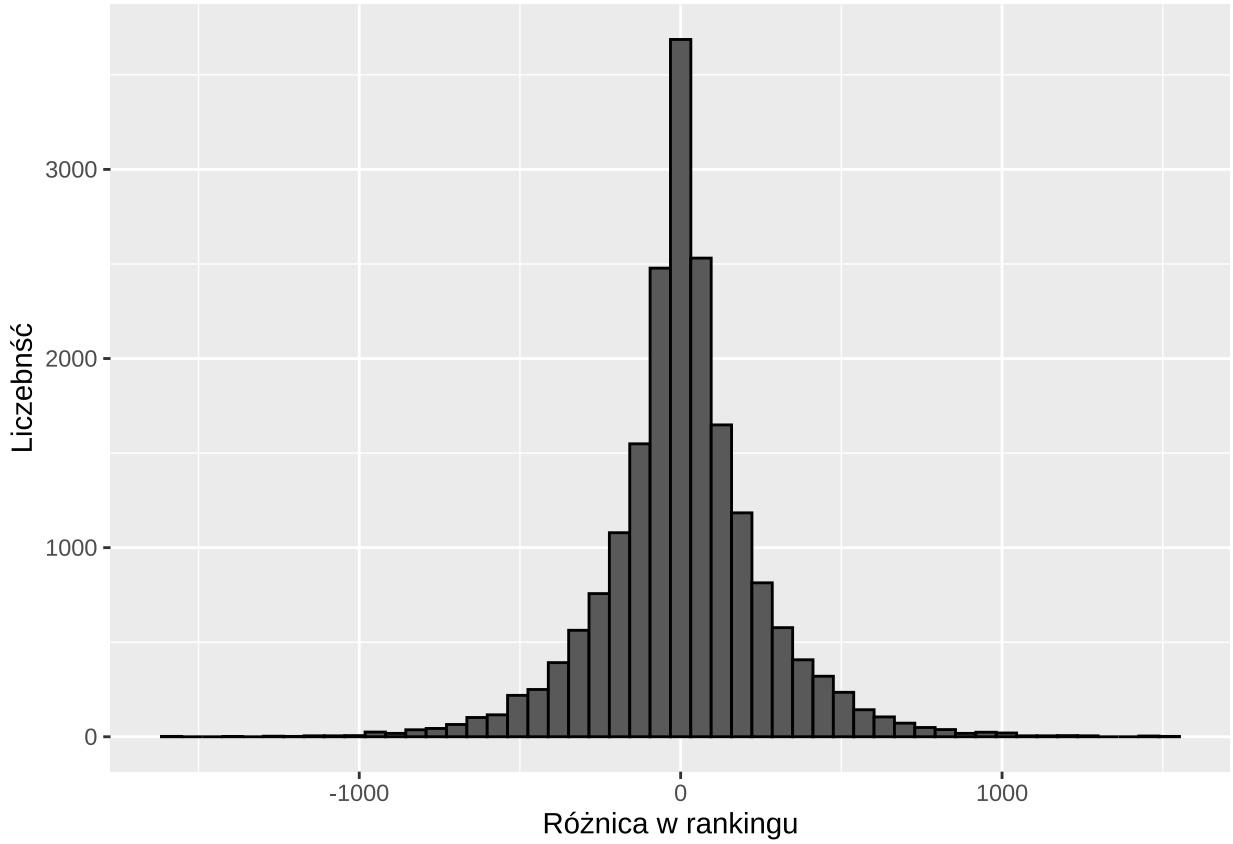
Sprawdzimy teraz jak wyglądają prawdopodobieństwa końcowych wyników, w zależności od przedziału rankingu białego gracza.



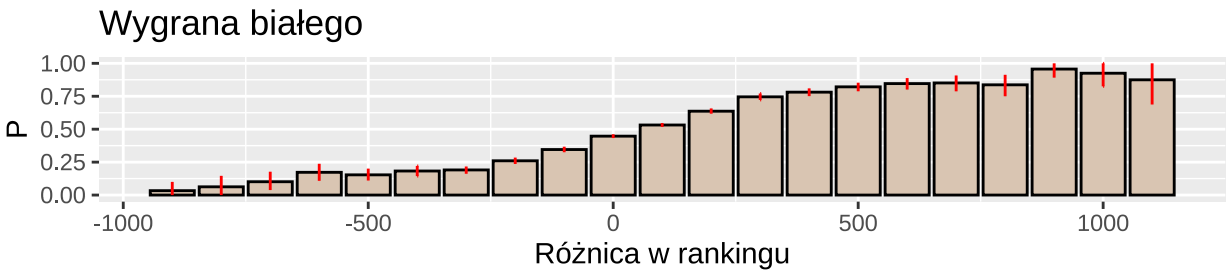
Analizując wykres możemy zauważyć, że wraz ze wzrostem rankingu białego gracza rośnie także prawdopodobieństwo jego wygranej, natomiast maleje prawdopodobieństwo wygrania czarnego. Estymowane wartości prawdopodobieństwa remisu sugerują, że jego prawdopodobieństwo rośnie. Natomiast przedziały ufności dla wysokich rankingów są na tyle szerokie, że nie możemy wysunąć konkretnych wniosków na poziomie ufności $0,05$.

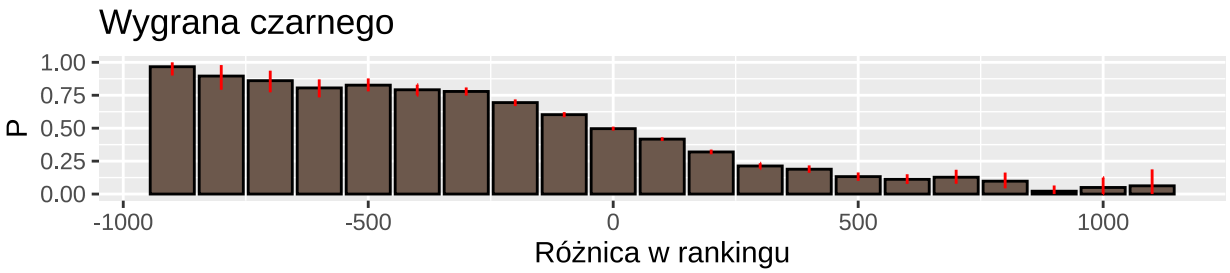
Warunkowanie różnicą rankingu białego i czarnego gracza

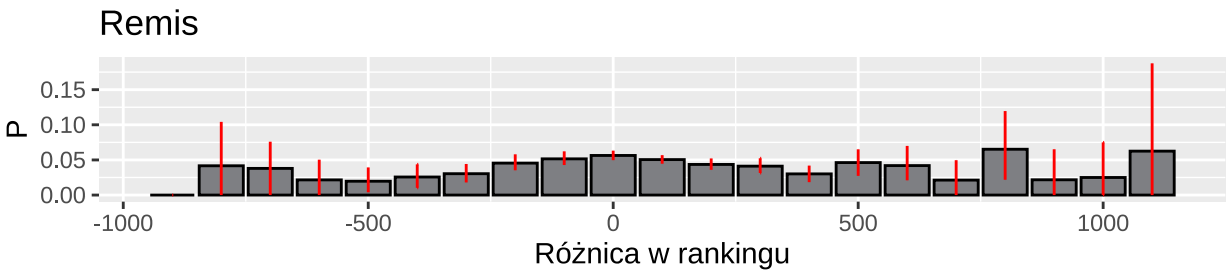
Analizowaliśmy jak wyglądają prawdopodobieństwa zakończenia partii od rankingu białego gracza, ale ranking czarnego też powinien mieć na nie wpływ. Dodamy nową kolumnę, która będzie oznaczać różnice między rankingiem białego a czarnego gracza. Rozkład różnic w rankingu wygląda następująco:



Jak możemy zobaczyć rozkład różnic w rankingu jest niemal symetryczny, a średnia wartość jest w okolicach 0. Dokładnie wynosi $7,56$. Podzielimy różnice na kategorię o długości 100 tak jak w przypadku rankingu gracza białego. Zobaczymy teraz jak wyglądają prawdopodobieństwa zakończenia partii w zależności od przedziału różnic. Do analizy bierzemy przedziały, które zawierają ponad 10 gier.



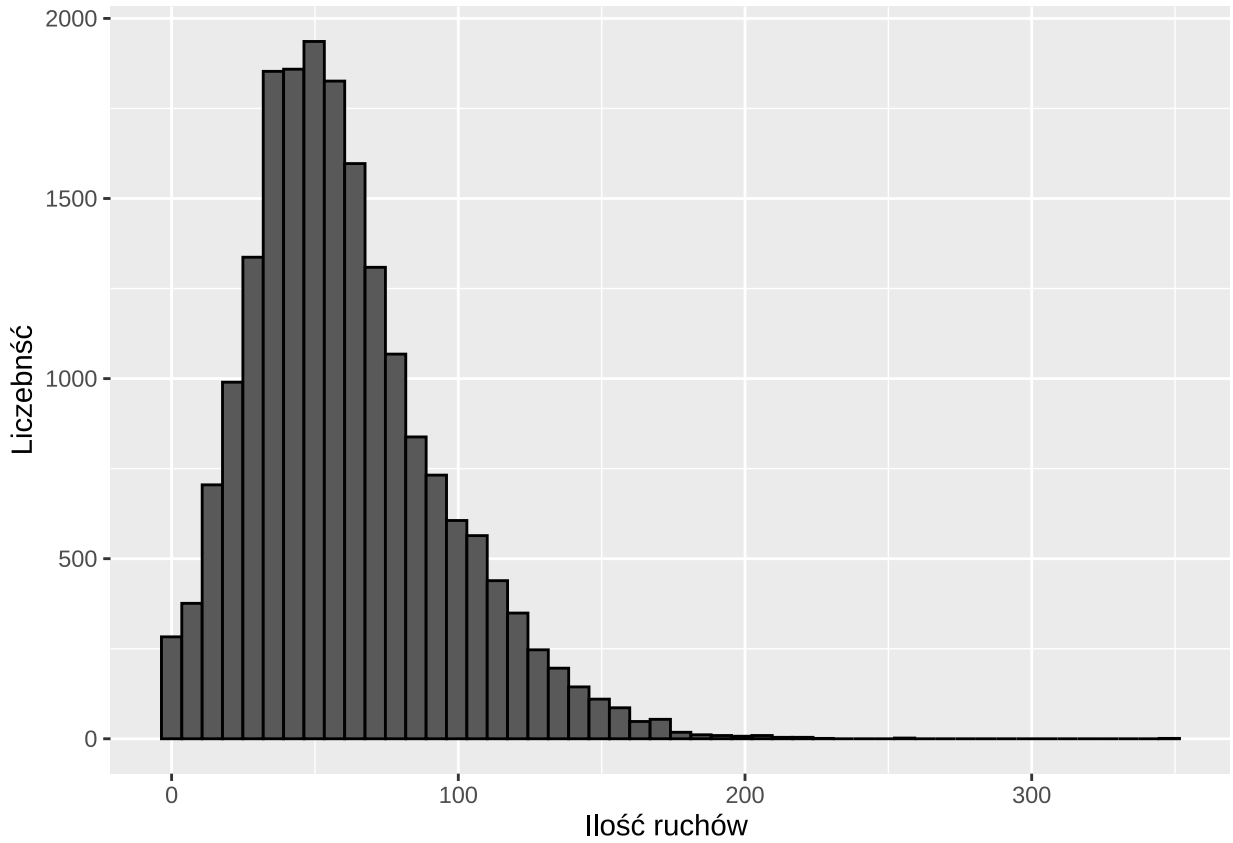




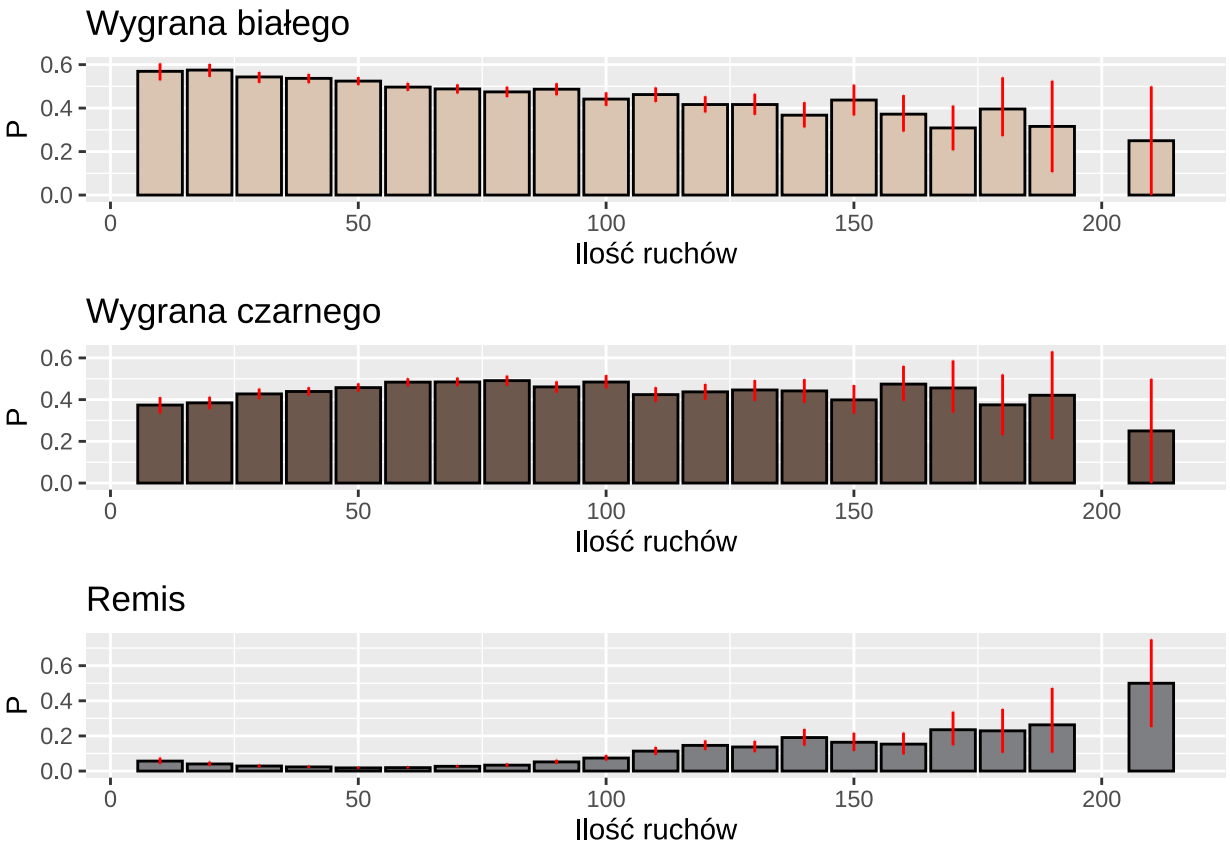
Z powyższych wykresów wynika, że wraz ze wzrostem różnicy w rankingu rośnie prawdopodobieństwo wygranej białego gracza, natomiast maleje szansa na wygranie czarnego. W przypadku remisu ciężko wywnioskować zależność.

Warunkowanie ilością ruchów

Partie w szachach mają różne długości. Sprawdzimy najpierw ile trwają partie w naszych grach.



Jak możemy zauważyć, większość gier trwa mniej niż 100 ruchów. Nasze dane podzielimy na przedziały o długości 10 gdzie pierwszy przedział to $[0, 10)$. Sprawdzimy jak wyglądają prawdopodobieństwa różnych zakończeń partii w zależności od przedziałów ilości posunięć. Bierzemy pod uwagę przedziały, które zawierają ponad 10 gier.



z wykresów możemy odczytać że prawdopodobieństwo wygrania białych maleje wraz z kolejnymi ruchami, natomiast prawdopodobieństwo wygrania czarnych rośnie do około 100 ruchu, by potem maleć. Natomiast prawdopodobieństwo remisu rośnie wraz z kolejnymi posunięciami.

Wnioski

Z analizy wynika, że na pozytywny wpływ na prawdopodobieństwo wygranej białego gracza mają: - rozegranie obrony Philidora, Otwarcia Angielskiego oraz partii hiszpańskiej - jego wysoki ranking - przewaga rankingowa nad czarnym graczem - szybkie zakończenie rozgrywki.

Pozytywnie na prawdopodobieństwo wygrania czarnego gracza (i jednocześnie negatywnie na wygranie białego) wpływają: - rozegranie obrony francuskiej, partii włoskiej, otwarcia pionkiem królewskim, otwarcia pionkiem hetmańskim, obrony skandynawskiej oraz obrony sycylijskiej - niski ranking białego gracza - przewaga rankingowa czarnego gracza nad białym - średni czas trwania rozgrywki, czyli około 50-100 ruchów.

Prawdopodobieństwo remisu zwiększa się znacząco jedynie dla długich rozgrywek (powyżej 100 ruchów).

Analizę można rozszerzyć o dodatkowe warunkowania.