STAT 516 Final Project Report

Trung Dang[1]

[1]tmdang@umass.edu - SPIRE ID 33858723

Contents

1 Introduction

In this report, we explore the transfer trends of chess players worldwide. We use the data from the FiveThirtyEight's Article: American Chess Is Great Again, and the dataset of all FIDE-officiated transfers from 2001 to 2017. More insights into the dataset will be provided in the following section of the report. We will delve into the article's central claim, which attributes the recent USA Team victories to importing top players worldwide, by comparing US-imported players' quality (ELO/Ranking) versus the rest of the world. Finally, we will discuss the result's implications, and provide additional reference resources to back our claims. To reproduce the result, please find the code here

2 Dataset and Preprocessing

We are provided with 439 transfers (after removing duplicates, filtered by player's ID). At each data point, we are provided with the ID of the player, the "from" and "to" Federation, and the transfer date. This is quite uninformative; after all, how can we tell the player's strength based on their ID and Federation?

2.1 Data preprocessing

Fortunately, the official governing body of chess - the Fédération Internationale des échecs (FIDE) - provides us with API to query this information. One can try and look up this information himself at: https://fide-api.vercel.app. Using Python 3.12, I managed to fetch the following data for each and every one of the player's ID listed in the original dataset:

- **Name** (First, Last)
- **FIDE Title** (Grandmaster (GM), International Master (IM), FIDE Master (FM), Candidate Master (CM))
- **Rating** (ELO): The higher the rating is, the stronger the player is.

We also remove duplicates and drop empty fields (Python's Pandas library provides a lightning-fast and ergonomic implementation of this step).

The final processed dataset can be found here

2.2 Data Analysis

A closer look at the average of the top 5 players in the team reveals:

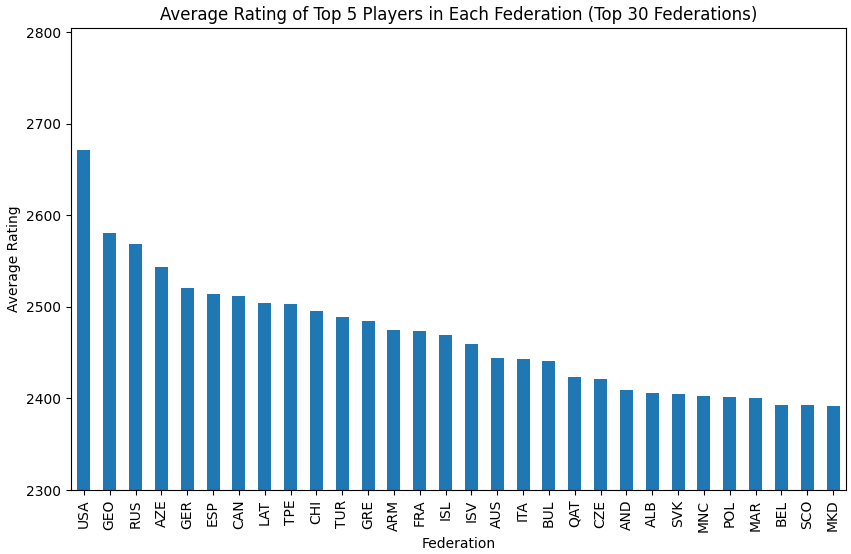

Figure 1: Average Rating of the Top 5 players in each country

We chose 5 because this is as many as a team can register on a Chess Olympiad. We can see that the USA beats the second country (Georgia) by a wide margin (almost 100 ELO ratings). In total, the average rating of USA transfers also exceeds the average of the overall average (including the USA):

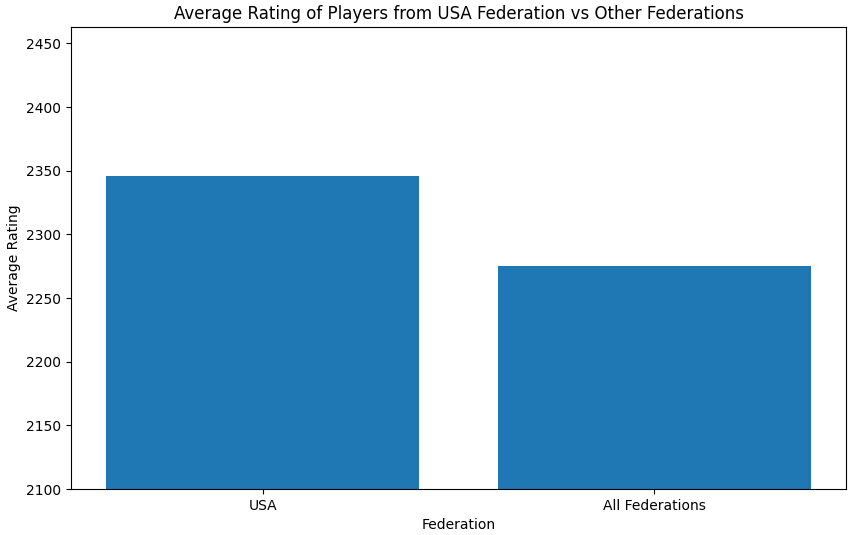

Figure 2: Average Rating of US Transfers versus Average of Global Transfers

On an individual basis, we analyze the Federation of the top 20 transferred players:

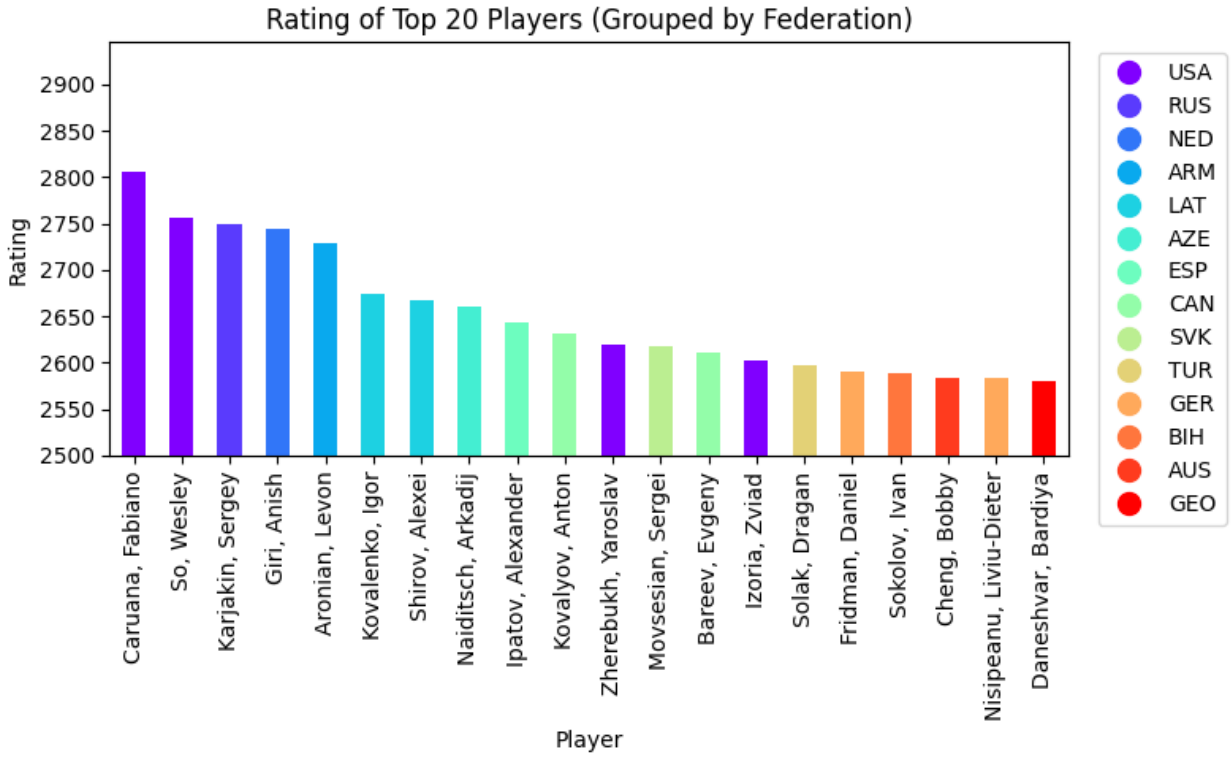

Figure 3: Ratings of the Top 20 players on the transfers list. The color bars are grouped by the Federation of the players

Note that, among the top 20 transferred players, 4 of them are from the USA. Moreover, the top 2 players (Fabiano Caruana and Wesley So) are the only ones currently in the world's top 10.

This evidence implies a difference in the rating of US-transfers and others-transfers, and suggests that the USA is better than other countries at Chess.

2.3 Assumptions and Notations

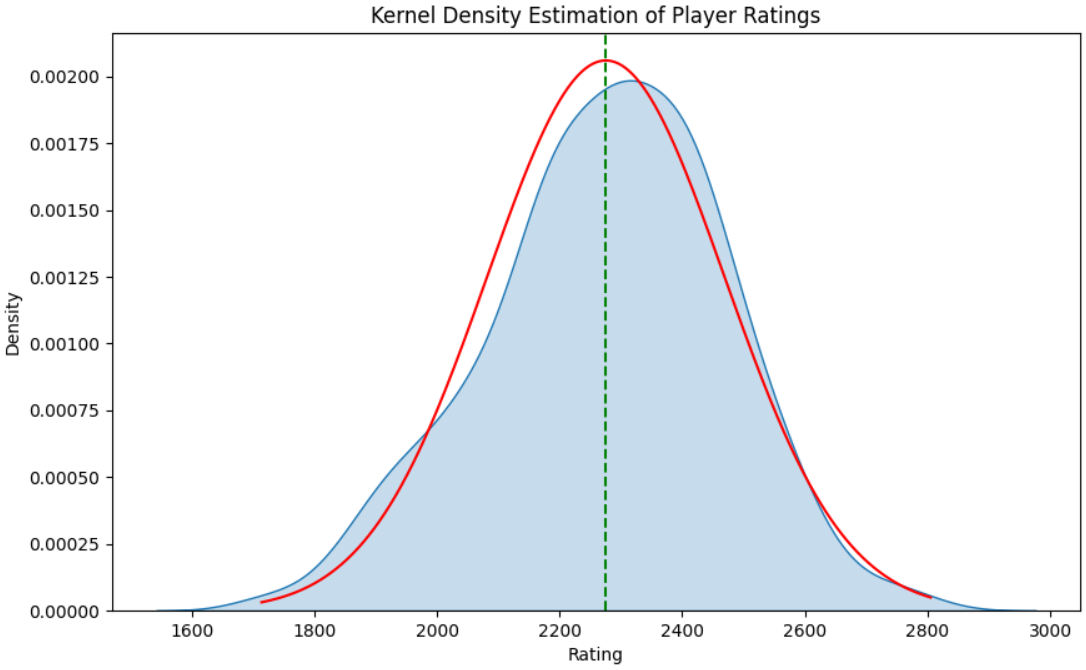

Figure 4: Kernel Density Estimation of the dataset

Using Python's Kernel Density Estimation to estimate the density of the player's rating, we can see that the distribution of players' rating is very close to the normal distribution $\mathcal{N}(\mu, \sigma^2)$. This is predictable by assuming the players' ratings are I.I.D random variables and applying the Central Limit Theorem. Therefore, we make the following assumptions:

- The player's ratings are i.i.d random variables
- These transferred players are equivalent to a random sample of a country's chess players so that any of the following statistical measurements on the dataset provide a good estimation of the overall population
- The sampled population is large enough ($n_W = 439 - 50 = 389, n_U = 50$)
- The mean and variance is finite.

Additionally, as we will be dealing with the populations of USA players and the worldwide players, we will denote any statistics of the USA players with a subscript "U" ($\mu_U, \theta_U, \sigma_U$,...), and any statistics of the world players with a subscript "W" ($\mu_W, \theta_W, \sigma_W$, ...).

Since the dataset's size is relatively large and beyond human's capacity to provide detailed calculations, simple statistics will be calculated with Python.

3 Statistical Analysis

3.1 Estimating a population of interest

Because we are interested in the difference between the US players and the rest of the world, we will split the dataset into 2 parts: $X_1, X_2, \cdots, X_{50}$ are the ratings of the US players, and $Y_1, Y_2, \cdots, Y_{389}$ are the ratings of the other players. Some statistics are listed below:

$$\begin{aligned} n_U &= 50, n_W = 439 - 50 = 389 \\ \mu_U &= 2345.6, \mu_W = 2266.30 \\ \sigma_U &= 198.50, \sigma_W = 191.42 \end{aligned}$$

The parameter of interest would be:

$$\theta = \mu_U - \mu_W$$

We can choose the estimator $\hat{\theta}$ to be:

$$\hat{\theta} = \bar{X}_n - \bar{Y}_n = 2345.6 - 2266.30 = 79.3$$

This means that we expect the mean rating of US players to be higher than the mean of the rest of the world by 79.3.

We know from class that $\bar{X}_n, \bar{Y}_n$ are unbiased expectations of μ_X, μ_Y. By the linearity of expectation, $\hat{\theta}$ is also unbiased. So:

$$Bias(\hat{\theta}) = 0$$

Therefore, the Variance of the estimator is:

$$\begin{aligned} Var(\hat{\theta}) &= Var(\bar{X}_n - \bar{Y}_n) \\ &= \frac{1}{n_U^2} \sum Var(X_i) + \frac{1}{n_W^2} \sum Var(Y_i) \\ &= \frac{1}{n_U} \sigma(X)^2 + \frac{1}{n_W} \sigma(Y)^2 \\ &= \frac{1}{50} (198.5)^2 + \frac{1}{389} (191.42)^2 \\ &= 882.23 \end{aligned}$$

The MSE of the estimator is:

$$MSE(\hat{\theta}_W) = Var(\hat{\theta}) + Bias(\hat{\theta}) = 882.23 + 0 = 882.23$$

Since the variance of the estimator is finite, and the estimator is unbiased:

$$\lim_{\substack{n_U \to \infty \\ n_W \to \infty}} MSE(\hat{\theta}) = \lim_{\substack{n_U \to \infty \\ n_W \to \infty}} Var(\hat{\theta}) = \lim_{\substack{n_U \to \infty \\ n_W \to \infty}} \sigma(X)/n_U + \sigma(Y)/n_W = 0$$

Therefore, the estimator is consistent.

3.2 Building a confidence interval

We will provide a confidence interval for the same population parameter $\mu_U - \mu_W$. We again note the following criteria of the dataset:

- $X_1, X_2, \cdots, X_{n_U}$ are IID
- $Y_1, Y_2, \cdots, Y_{n_W}$ are IID
- $X's$ and $Y's$ are independent (which follows immediately from the fact that all players' ratings are independent)
- Distributions are unknown
- n_U, n_W are large
- coverage is approximate

Then, a $(1-\alpha) - level$ two-sided CI is given by:

$$\bar{X}_{n_U} - \bar{Y}_{n_W} \pm z_{\alpha/2}\sqrt{\hat{\sigma}_X^2/n_U + \hat{\sigma}_Y^2/n_W}$$

A 99%, 95%, and 90% confidence intervals are:

$$CI(0.99) = 79.3 \pm 2.58\sqrt{882.23} = [2.67, 155.93]$$
$$CI(0.95) = 79.3 \pm 1.96\sqrt{882.23} = [21.08, 137, 52]$$
$$CI(0.90) = 79.3 \pm 1.64\sqrt{882.23} = [30.59, 128.01]$$

Again, the coverage of these intervals is only approximate.

However, we note that none of these intervals intersect with $(-\infty, 0]$, which again suggests that the mean of USA players is higher than the mean rating of the rest of the world.

3.3 Hypothesis testing

We want to test the hypothesis that the mean rating of USA players is higher than the mean rating of the rest of the world. So the null hypothesis is:

$$H_0 : \mu_X - \mu_Y = 0$$

and the alternative hypothesis is:

$$H_A : \mu_X - \mu_Y > 0$$

Because we don't know the distributions of the populations, we rely on large-sample hypothesis testing, which provides asymptotical type I error control. The p-value of the estimation is:

$$T_n = \frac{(\bar{X}_{n_U} - \bar{Y}_{n_W}) - 0}{\sqrt{\hat{\sigma}_X^2/n_U + \hat{\sigma}_Y^2/n_W}} = \frac{79.3}{\sqrt{882.23}} = 2.668$$

Note that since we are testing for the one-sided alternative, the rejection region is: $\{T_n > z_\alpha\}$. We have $z_{0.004} = 2.66 < T_n = 2.668 < 2.67 = z_{0.0038}$, so the hypothesis will reject the null hypothesis with any $\alpha > 0.004$, and fail to reject the null hypothesis with any $\alpha < 0.0038$.

This hypothesis test is very strong, and we are very likely to reject the null hypothesis with very small significance level α. This implies that the mean of the USA players are very likely to be higher than the mean of the rest of the world.

4 Conclusions and further discussions

By treating the dataset of chess player transfers as a representative body of the entire professional chess players, we found that the mean strength of the US chess players is higher than that of the rest of the world. With an unbiased estimator value of 79.3, and a 99% confidence interval of [2.67, 155.93], the null hypothesis (mean ratings are the same) is rejected with significance level as low as $\alpha = 0.004$. For reference, a widely accepted convention proposed by Sir Ronald Fisher uses $\alpha = 0.05$, so the strength of our hypothesis is somewhat significantly higher than required.

These findings lend support to the claim made in the FiveThirtyEight article that the United States has been actively recruiting top international talent, contributing to the success of the US chess team in recent years. By attracting a disproportionate number of highly rated grandmasters and international masters from around the world, the US has raised the overall strength of its chess pool.

Note that our analysis is based heavily on the assumption that the sample of transfers reflects the entire country's professional players population. This is acceptable, given that the transfers happen on a wide range of players (i.e. countries don't just import top-ranking talents but also rising prodigies and mid-level players). However, this analysis does not prove causation - importing talents does not necessarily lead to better results, nor does not doing so lead to bad results (for example, Indian players are performing phenomenally in the world championship, and India barely imports any talent). Other factors like financial incentives, training facilities, coaching, etc. likely also play a role in the US team's performance. Additionally, this dataset only covers transfers up to 2017, so more recent years are not reflected.

Future analysis could examine transfers in other sports or fields to see if the US exhibits a similar aggressive recruitment of top global talent. It would also be interesting to track

the long-term performance of these transferred players representing their new federations. Do they maintain their pre-transfer rating? How have they impacted youth development and coaching? Investigating these questions could yield additional insights into the dynamic world of chess migrations.

References

1. https://fivethirtyeight.com/features/american-chess-is-great-again/ FiveThirtyEight, American Chess Is Great Again, Aug. 8, 2017

2. https://github.com/fivethirtyeight/data/tree/master/chess-transfers FiveThirtyEight, Dataset of "American Chess Is Great Again", Last Updated Dec. 6, 2022

3. https://github.com/dmtAtUMass/umass-stat516-project Trung Dang, GitHub Repository of this analysis

4. https://github.com/dmtAtUMass/umass-stat516-project/blob/main/data_rated.csv Trung Dang, Processed Data of this analysis

5. https://en.wikipedia.org/wiki/Chess_Olympiad Wikipedia, Chess Olympiad

6. https://ratings.fide.com/top.phtml FIDE, Standard Top 100 players May 2024

7. https://fide-api.vercel.app/docs FIDE, Public API Endpoints

8. https://pandas.pydata.org/ Python Pandas documentation

9. https://seaborn.pydata.org/generated/seaborn.kdeplot.html Python Seaborn Kernel Density Estimation (KDE)

Acknowledgment

Thank you for being an amazing teacher ^^!