Analysis of Different Parameter in ODI Cricket and Studying Different Hypothesis using Statistical Methods

Statistical Methods and Application I (STAT 500)

University of Colorado, Boulder

Fall 2023

Sasi Jyothirmai Bonu
ID: 110987886

Jeet Choksi
ID: 110988624

Divya Nallawar
ID: 110916558

Tanay Shukla
ID: 110857542

TABLE OF CONTENTS

Abstract 3

Introduction 3

Data 4

Methods 5

4.1 Exploratory Data Analysis 5

4.2 ANOVA test 10

4.3 Correlation Analysis 11

4.3.1 Spearman Correlation Coefficient 12

4.3.2 Pearson Correlation Coefficient 12

4.3.3 Chi square test 13

Results 15

Conclusions 16

References 17

Abstract

This study explores the complicated world of One Day International (ODI) cricket, a popular sport around the world that is known for its statistical intricacy and unpredictability. In order to identify the variables affecting a team's performance, the study examines patterns and trends in ODI cricket data, concentrating on important metrics including strike rates, batting and bowling averages, and team performance. Through analyzing the correlation between statistical metrics and countries, the study aims to pinpoint critical components that lead to success. This research not only clarifies the complex dynamics of one-day international cricket (ODI) but also has the ability to influence strategic choices and provide insightful information to teams that want to succeed internationally.

Introduction

Cricket is a popular sport all around the world, but it's also frequently a game of statistics and uncertainty. One Day International (ODI) cricket is a dynamic format that blends skill, strategy, and unpredictability into every match within this vast arena. The excitement of the game and the abundance of statistical information it produces are not the only things that make ODI cricket so special. To shed light on the nuances of the game, this initiative examines patterns, trends, and insights found in ODI cricket data. Each team in a One Day International has a certain amount of overs to bat and bowl, usually fifty overs each side. The format necessitates striking a balance between disciplined bowling to limit the opposition's scoring and aggressive batting to collect runs. Metrics like strike rates (number of runs scored per 100 balls faced), bowling

average (average runs conceded per wicket taken), batting average (average runs scored per dismissal), and team totals (total runs scored by a team in an innings) are among those used in the analysis of ODI cricket data.

Previous research has examined cricket statistics and their relationship to national success as well as the variables that influence victory. Prior studies have examined the optimal way to arrange the players in a lineup.[1] Studies have also been conducted regarding the significance of fielding. According to research, strike rates, bowling and batting averages, and team performance are crucial factors in One-Day International cricket. Our research goal has been to identify the elements that are influenced by countries in order to determine which features are crucial in predicting a team's victory. This project's main question is: What statistical patterns and trends show up when One Day International cricket data is analyzed, and how are these measures related to a team's performance? This topic attempts to identify which metrics are most influenced by cricket teams, or more specifically, the countries they represent.

This analysis may actually have an impact on strategy and aid in the success of the team.

Data

The data was extracted in CSV formats from ESPN cricketInfo Website. Both bowling and batting make up the two fundamental components. There were three pieces each for the two sections, ODI (One Day International), T20 (Twenty twenty), and test (test cricket). Just the One Day International portion of the data, which included around 2500 player records from around 46 cricket-playing nations worldwide, was taken into consideration. The data was quite thorough and had information of the number of years a player has been active, number of games played by the player, number of innings played, total runs, number of not outs, highest score of the player, average, balls faced of the player, strike rate (number of runs scored per 100 balls/average number of balls

bowled per wicket), total number of 50s, total number of 100s, and total number of ducks.

The absence of analysis on test cricket or T20 data may have resulted in a significant bias against countries and players with a large number of ODI matches played, or none at all. It is also a well-known fact that many cricket players are skilled batters or bowlers; however, it is challenging to identify a single nation that dominates the game. As a result, it was decided to analyze the batting and bowling data separately and determine which elements may be employed for the study.

The only accessible data for analysis is till the year 2019. This indicates that the dataset does not include any events, advancements, or modifications related to the game of cricket that took place after 2019. The regulations, individual performances, team tactics, and other aspects of cricket are always changing, making it a dynamic sport. In light of the constantly shifting nature of cricket, a lot may have changed in the past four years. When what is given is not representative of the actual underlying conditions, bias may develop.

Methods

Exploratory Data Analysis

There was only one column in the original dataset for the player and the nation. The player name and country columns in this particular column were split into two different columns in order to do the analysis by nation. Similar to this, the player's career was represented by a single column that was divided into two columns to indicate the beginning and conclusion of their playing years. Since the majority of the columns contained quantitative data, they were all constrained to be formatted in numeric form. The batting data contained numerous null values that were all dropped during the

data-cleaning procedure, whereas the bowling data was nearly entirely clean and devoid of null values. In both the bowling and batting datasets, a large portion of the data was player-oriented. The data was changed to make it country-wise, but in order to be effective, it was necessary to identify which aspects need alteration.

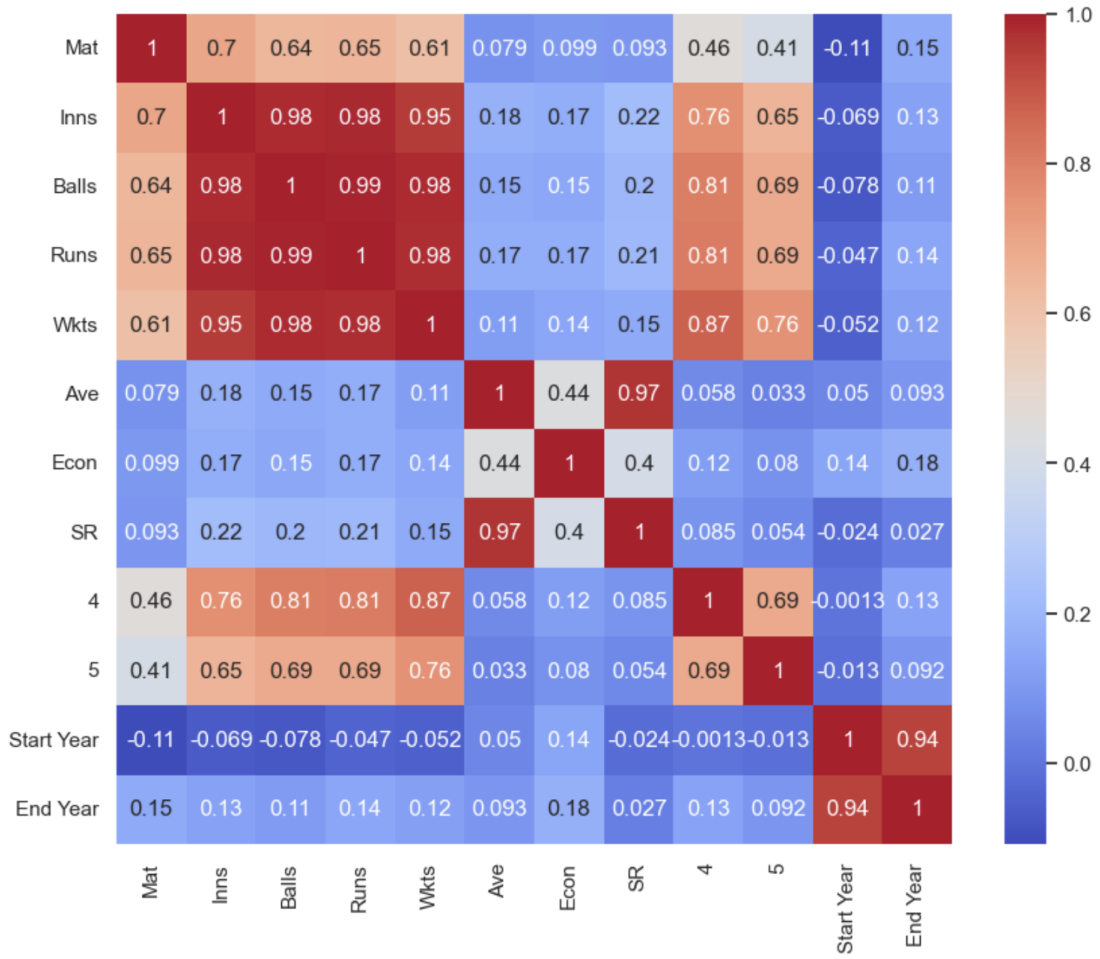

Figure 1. Heatmap of bowling data

In order to discover which factors were associated and to view the correlation values among all the quantitative variables, two preliminary heatmaps were created. This was

done to select which correlated features to form hypothesis with and which ones to avoid.

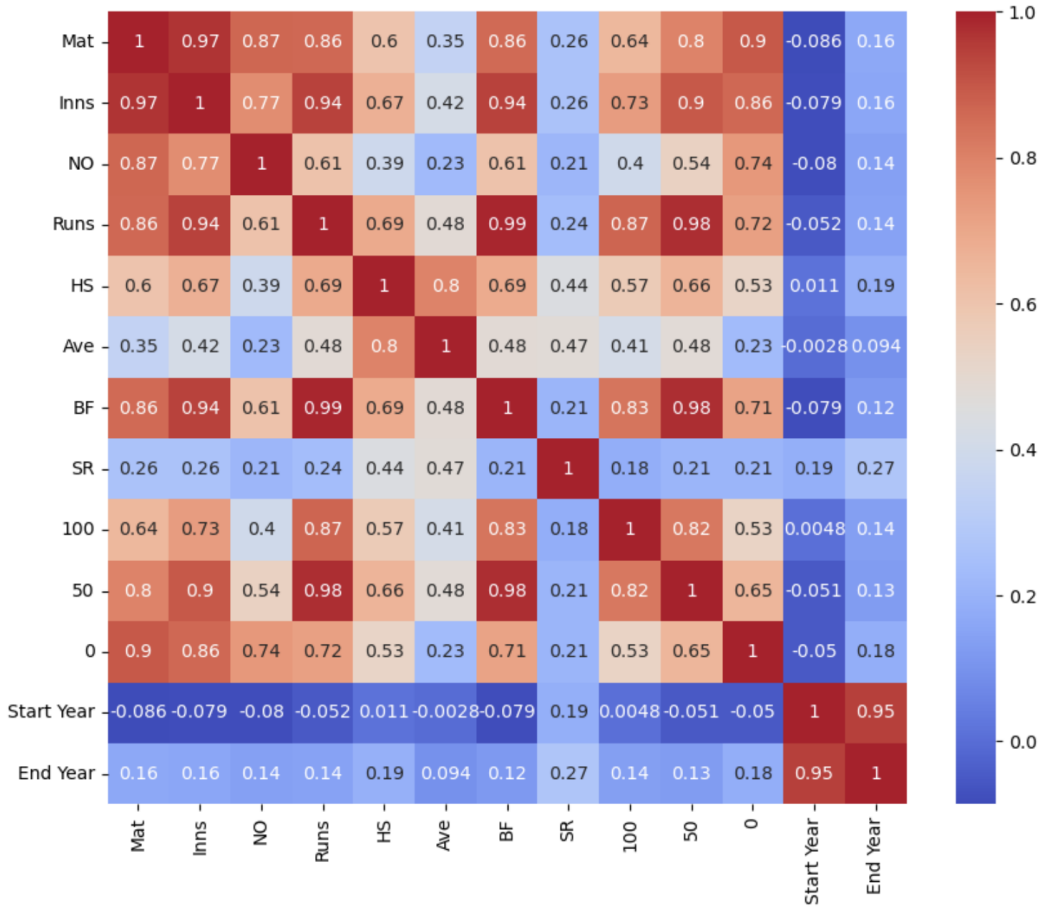

Figure 2. Heatmap of batting data

A preliminary analysis was conducted to search for correlations between the columns, and data distribution was examined by creating histograms. Following some investigation, each nation's mean number of runs and wickets taken were computed and plotted in order to go closer to the goal.

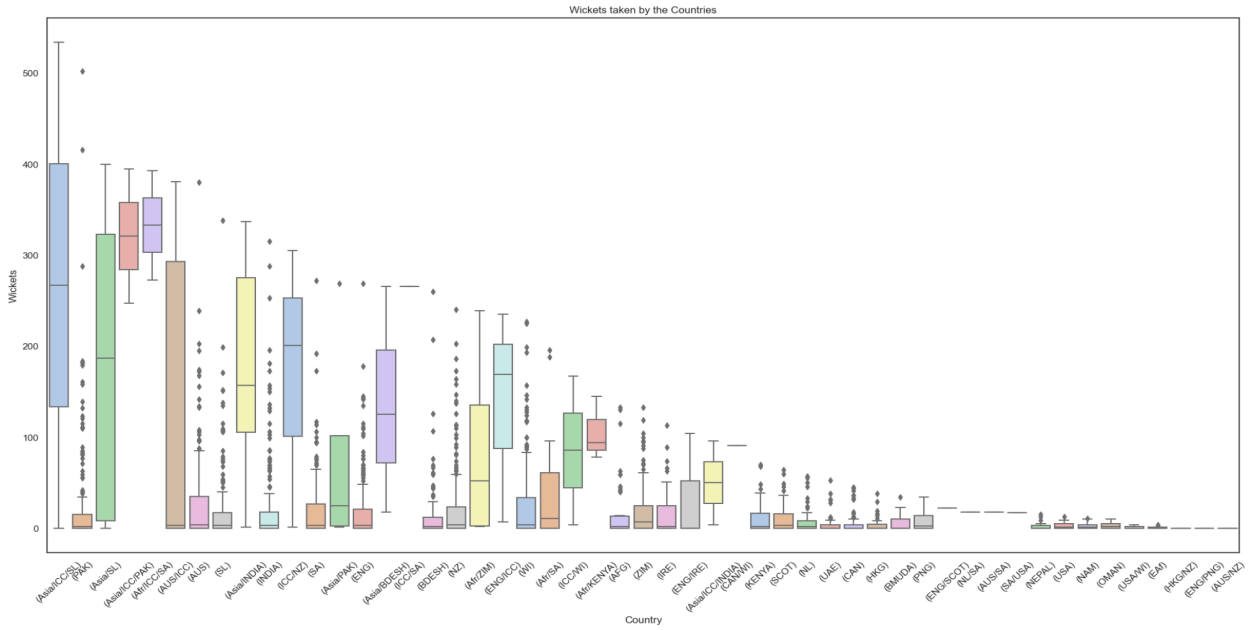

Figure 3. No. of wickets taken by the countries

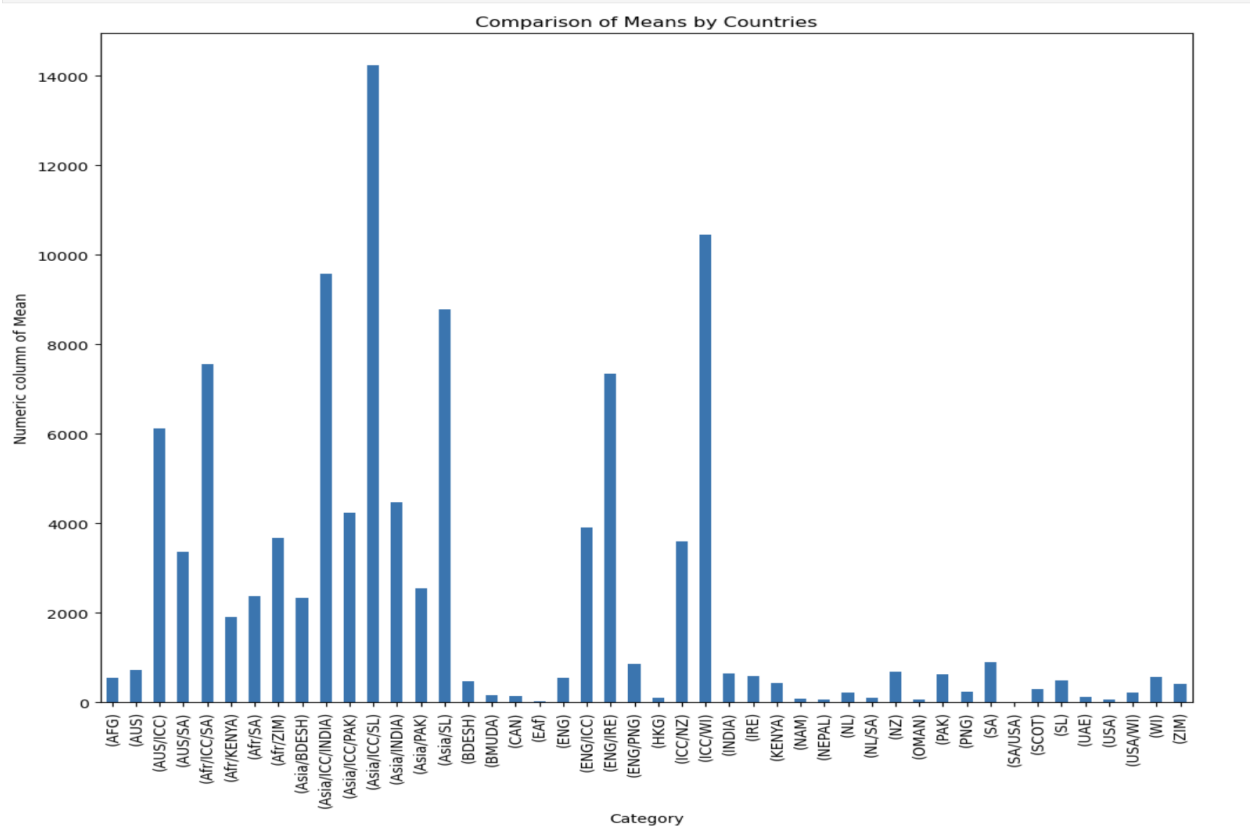

Figure 4. Average number of runs by the countries

Although this provides some direction for the investigation by indicating which country is strong at batting (figure 4) and which is at bowling (figure 3), our goal was to identify the characteristics that are most affected by the cricket-playing nations. The next goal was to determine how each nation distributed the overall number of centuries.

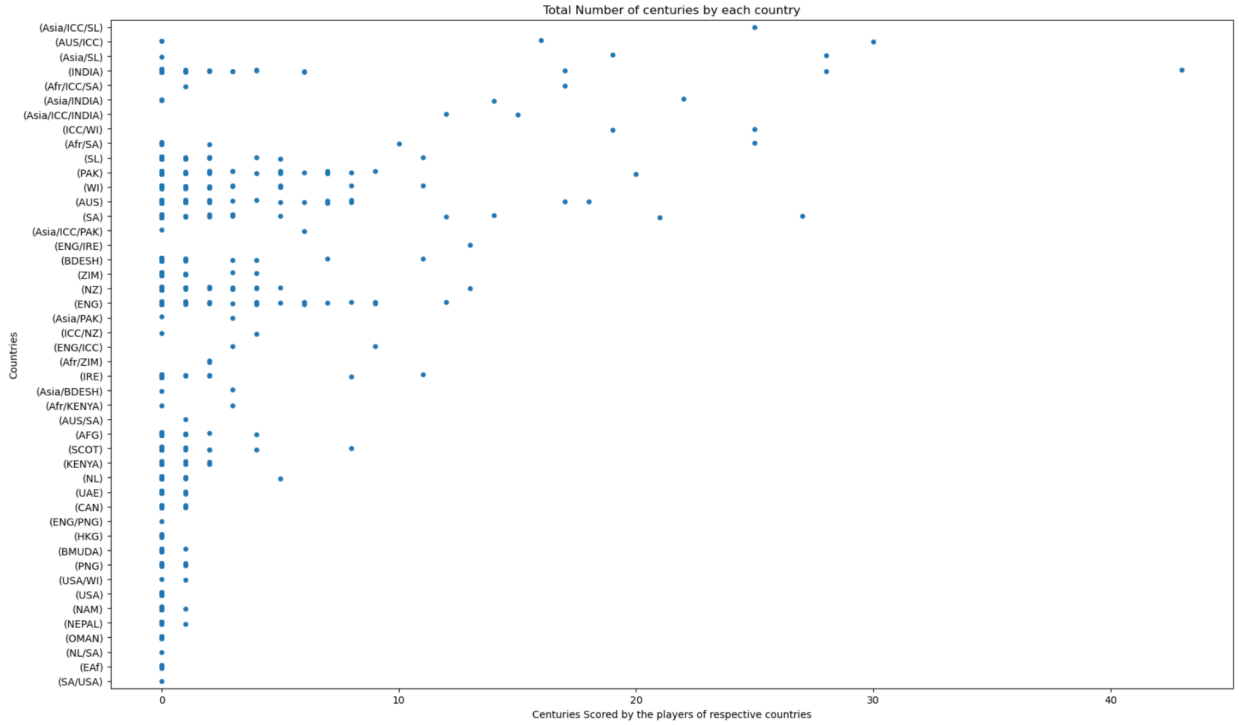

Figure 5. No. of centuries by the countries

This visualization demonstrated that the nations with the highest number of centuries—such as Sri Lanka and India—also have the highest number of century-scoring players. This is one example of how the number of centuries could vary depending on the nation to which they belong. If this is the case for centuries, it may also be the case for the number of ducks scored by each nation, a hypothesis that will be investigated in the course of hypothesis testing.

ANOVA Test

We developed a few theories to determine what aspect of bowling a country influences. Our initial idea suggested that the economic rates of various countries would differ. Considering that our goal was to ascertain how one quantitative dependent variable and one independent variable related to one another, a one-way analysis of variance (ANOVA) was performed on the economy rates (number of runs conceded in an over) to see if there were any noteworthy variations across the nations. Analysis of Variance, or ANOVA, is a statistical method for examining how group means in a sample differ from one another.[2] Comparing the variability within groups to the variability between group means is the fundamental concept of ANOVA. There may be variations in the group means if the between-group variance is noticeably greater than the within-group variability.[3] A p-value and an F-statistic are generated by the test. The alternative hypothesis is accepted and the null hypothesis is rejected if the p-value is less than the selected significance level, which is typically 0.05. Likewise, the null hypothesis is rejected when the F-statistic is high. The P value was clearly high (0.278) and the F statistic was fairly low (1.109), indicating that there is no discernible difference in the economy rates of various nations.

A similar method has been used to look into potential variations in bowling averages. However, because the F statistic was low (1.061) and the P value was high (0.358), the null hypothesis could not be rejected, much like in the case of economy rates.

Hypothesis	F statistic	P value
Economy rate	1.1092098359851377	0.2789434255900259
Bowling average	1.0615778034454826	0.35825004618253226

Table 1. F statistic and P values of bowling Hypotheses

The ANOVA test has been used to test the hypothesis that there is no significant difference between the mean strike rates (the average number of runs scored per 100 balls faced) of the countries in terms of batting ability. The test resulted in a low P value and a high F statistic, which were sufficient to reject the null hypothesis and demonstrate that mean strike rate varies by nation. Another hypothesis was to compare batting average of different players/countries and determine that there is a significant difference between them. This hypothesis was rejected because of the low P-value and the F-statistic's bigger than 1, which suggest that batting averages for various players and nations are very similar.

Hypothesis	F statistic	P value
Mean strike rate	2.6123222155644483	4.7397555314339173e-08
Batting average	2.18490831019087	1.1958839242468402e-05

Table 1. F statistic and P values of batting Hypotheses

Correlation Analysis

A statistical technique for determining the degree of connection between two quantitative variables is correlation analysis. A weak correlation indicates that there is little to no association between the variables, whereas a high correlation indicates that two or more variables have a strong relationship. The degree of relationship is quantified by the correlation coefficient, which is the expression of the correlation analysis result. This numeric value generally ranges from -1 to 1. A perfect positive correlation is shown by a correlation coefficient of +1, which means that as one variable rises, the other variable rises in proportion.[4] A perfect negative correlation, on the other

hand, denotes a correlation value of -1, which means that when one variable rises, the other falls proportionately. Furthermore, no linear relationship between the two variables is shown by a correlation coefficient of 0.

Spearman Correlation Coefficient

A third hypothesis was tested to look into the fact that good bowlers also maintain a good bowling average. The Spearman Correlation Coefficient was used to test if the two features are correlated.

It is an indicator of statistical dependency between two variables that is non-parametric in nature which is denoted by ρ. The data's ranks serve as the foundation for the Spearman correlation. It was used since the linearity of the relationship between the variables is not assumed by Spearman correlation.[5] In this case, the Spearman correlation coefficient shows a significant link, and the P-value is zero, indicating that the bowling average shows a continuous trend as the number of wickets taken rises.

Pearson Correlation Coefficient

Since most of the data were composed of quantifiable variables, the Pearson correlation coefficient was applied multiple times in the study to determine the correlation between different data elements.[6] The degree and direction of a linear relationship between two continuous variables are measured statistically by the Pearson correlation coefficient, or r.

With n data points, the following formula can be used to get the Pearson correlation coefficient between two variables, X and Y.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

When attempting to determine whether two quantitative variables have a linear relationship with one another, Pearson's correlation is the method widely used.[7] Pearson correlation coefficient was used test the hypothesis that both a player's strike rate and batting average are relative; if the former is higher than the latter, so is the strike rate. It is evident that if one of them rises, the other will follow suit, as indicated by the positive Pearson correlation coefficient of 0.474 and low P-value (4.214605328785315e-96), which point to a positive linear correlation between the batting average and strike rate. It was proposed that batting and bowling averages across nations have a significant linear correlation, that there is a relationship between batting and bowling performance. It was concluded that the hypothesis can be rejected based on the p-value we obtained from the Pearson correlation analysis, since it shows that there is only a very weak linear relationship between batting and bowling averages. Similar correlation analysis was applied to theorize impact of time on batting and bowling performance. In the dataset, there is a significant positive linear correlation between the average runs scored and the average number of wickets taken across the years. Based on the statistical evidence supplied by the Pearson correlation test, the hypothesis of a strong positive linear relationship between mean runs and mean wickets across different years can be accepted given the strong positive correlation and the very low p-value.

Chi - Square Test

The frequency of centuries and ducks scored by players and whether or not they are independent of the nation they belong to was yet another hypothesis to be tested. The frequency of centuries and ducks against nominal data was to be tested for the

hypothesis, which is a perfect occassion to use Chi square test.[8] Initially, a contingency table was made to calculate the joint distribution. A contingency table is a table used in statistics to show the joint distribution of two or more categorical variables. It is often referred to as a cross-tabulation or crosstab. The frequency of occurrences for a specific combination of categories from the variables under study is represented by each cell in the table. To examine correlations between categorical variables, contingency tables are often used along with chi-square test of independence. If the observed frequencies deviate from what would be predicted if the variables were independent, the chi-square test can be used to assess if there is a significant connection between the variables.[9] With the hypothesis proposed, the higher CHi-square test value and 0.0 P-value suggested a strong association between the ducks and centuries scored with the countries.

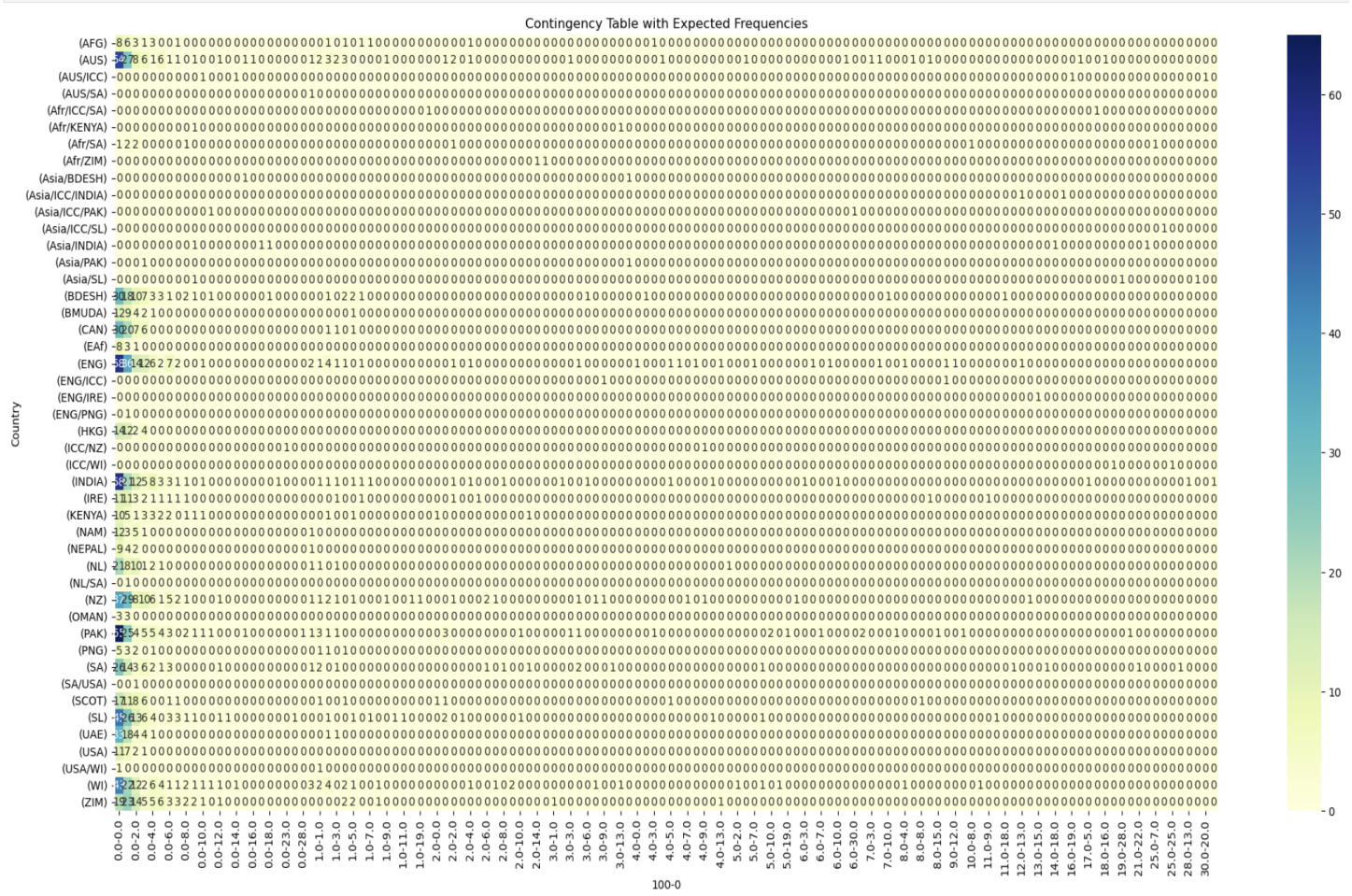

Figure 6. Contingency table

Results

In an effort to identify the characteristics that are impacted by the nations that the players are from. The theory that withstood testing was that bowling average and economy rates are unaffected by players and the nations they are a part of. Nonetheless, a trend was noted showing an increase in both the number of wickets taken and the bowling rate. Similarly, in terms of batting, a player's mean strike rates, centuries and ducks are all determined by the nation to which they belong. The results showed that strike rates and batting average were positively correlated.

A combination of the two datasets led us to hypothesize that there is a significant linear correlation between batting and bowling averages among countries which proved to be untrue.

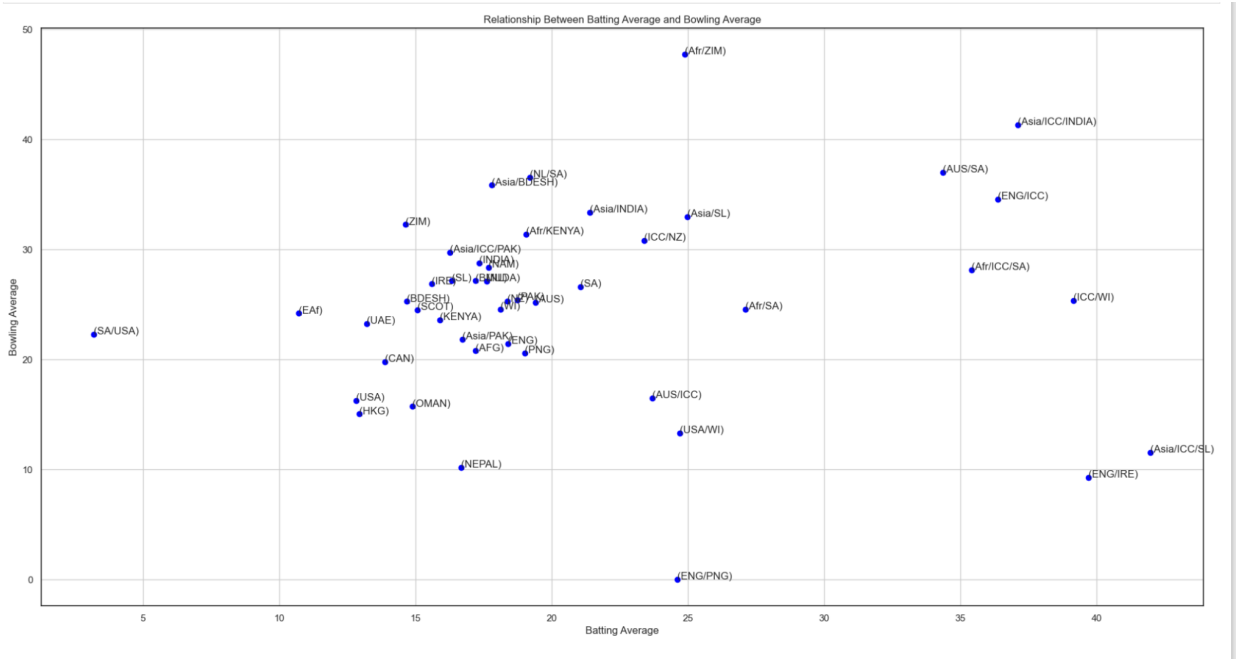

Figure 7. Relationship between batting and bowling averages

Additionally, a significant association between the average runs scored and the average number of wickets taken over the years was discovered.

Conclusions

Our research disproves some of our assumptions on the impact of nationality on the attributes of cricket players. While there was a clear trend showing an increase in wickets taken and bowling rates, bowling average and economy rates proved to be resistant to national affiliations. On the other hand, there was a noticeable correlation between a player's nationality and batting characteristics like as strike rates, centuries, and ducks scored. Notably, it was found that strike rates and batting averages positively correlated. Our data disproves a significant linear association between bowling and batting averages across nations, which is contrary to many predictions.

Presentation Video Link

References

Perera, G. H. P. (2015, December 16). *Cricket Analytics*. https://summit.sfu.ca/item/16247

Cuevas, A., Febrero, M., & Fraiman, R. (2004, August 1). *An anova test for functional data*. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2003.10.021

St»hle, L., & Wold, S. (1989, November 1). *Analysis of variance (ANOVA)*. Chemometrics and Intelligent Laboratory Systems. https://doi.org/10.1016/0169-7439(89)80095-4

Franzese, M., & Iuliano, A. (2019, January 1). *Correlation Analysis*. Elsevier eBooks. https://doi.org/10.1016/b978-0-12-809633-8.20358-0

Artusi, R., Verderio, P., & Marubini, E. (2002, April). Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval. *The International Journal of Biological Markers*, *17*(2), 148–151. https://doi.org/10.1177/172460080201700213

Johnson RA, Wichern DW (2015) Applied multivariate statistical analysis, 6th edn. Pearson, Upper Saddle River

Schober, P., Boer, C., & Schwarte, L. A. (2018, May 1). *Correlation Coefficients: Appropriate Use and Interpretation*. Anesthesia & Analgesia. https://doi.org/10.1213/ane.0000000000000286

McHugh, M. L. (2013, January 1). *The Chi-square test of independence*. Biochemia Medica. https://doi.org/10.11613/bm.2013.018

Franke, T. M., Ho, T., & Christie, C. A. (2011, November 8). The Chi-Square Test. *American Journal of Evaluation*, *33*(3), 448–458. https://doi.org/10.1177/1098214011426594