Comparing yearly batting average for Rahul Dravid using Bayesian Hierarchical Model

Aditya Ranade

April 23, 2022

1 Introduction

Cricket is the most popular game in India and is played and watched in every corner of the country. Consistency is an important and desired quality for any sportsperson. If we look closely at the most well known and best players in any sports, more often then not, they are known for their consistent performances. In this project we analyse the cricketing career of Rahul Dravid, considered to be one of the best test match player to have played the game of Cricket. He was nicknamed 'The Wall' in the sense he would be the player standing between win and loss for the opponent. He was also called 'Mr. Dependable' for his consistent performances.He played cricket at the international level representing India for 17 years which is remarkable for any sportsperson at the International level.

We will compare the average number of runs scored per innings in each of the 17 years of the said player's career. During the career of the said player, Cricket was majorly played in 2 formats. First is Test matches which are played over 5 days and each team gets to bat and bowl for 2 innings each. Second is One Day International matches (ODI) where each teams gets to bat and bowl for one inning each of 50 overs. For this project, we consider the data for test matches only.

2 Data

The data has been scraped from crickinfo website. For this report, we have considered the number of runs scored in a match, the year/season of the match, the format of the match and the location of the match (if the match is played in home country or away from home). Since some games get abandoned due to various reasons most likely due to bad weather, those were not considered. Similarly the games where the player did not get a chance to bat have also not been considered.

A note on the interpretation of average according to official cricket record keeping is the number of runs scored by the player between two times the player gets out. However, the average we are considering in this project is in the usual sense where we divide the total runs scored by the number of innings taken to score the runs. To illustrate, if a player bats and scores 300 runs cumulatively in 10 matches in a season and gets out only once then the average will be 300 according to the official records but for the project it will be 30.

Here, season is the calender year since international cricket is played throughout the year. The way the game is played, we can expect the variation to be extremely high. For example, runs scored by any player who gets an opportunity to bat is a non-negative integer. So in a match a player can get out on 0 and in the next match the player can score 100 runs. Similarly if a player plays 20 matches in a season, high variation can be expected. In fact the way the formats are designed, variation in One Day Internationals (ODI) can be expected to be much higher than Test matches.

3 Exploratory data analysis

We take a look at the boxplot of the runs scored by Rahul Dravid in each season throughout the career in figure 1.

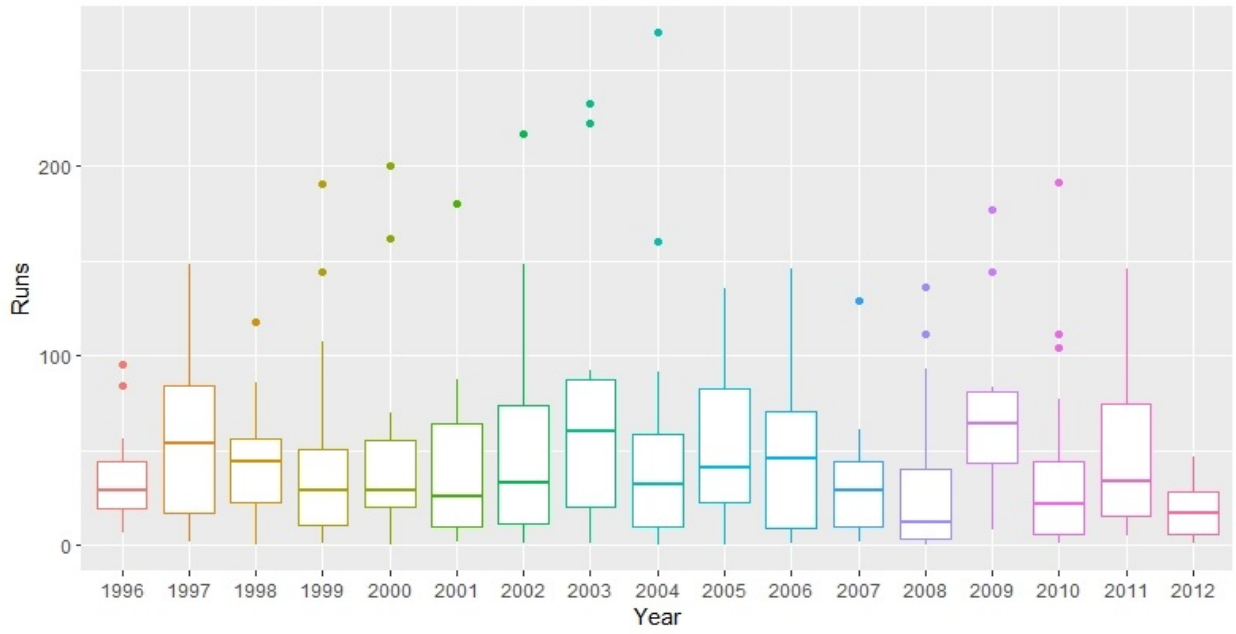

Figure 1: Boxplots for runs scored each year

A boxplot shorter in height indicates a lower variation for the particular year. Higher location of the boxplot indicates higher average for the particular year. A boxplot higher in location and compact in size indicates a good season for the player. Since a decent variation in the size of boxes is visible over the years, it would be a good idea to consider a model which allows separate variation parameter for every year.

4 Methods

4.1 Model

We will consider a Hierarchical Bayesian model to compare the batting average for the said player through seasons using .

The model can be described as follows.

$$Y_i \overset{i.i.d}{\sim} N(\mu_{year[i]}, \sigma^2_{year[i]})$$
$$\mu_{year} \overset{ind}{\sim} N(\theta, \tau^2)$$
$$\theta \sim N(\theta_0 = 50, \sigma_0 = 16)$$
$$\sigma_{year} \sim Ca^+(0, 20)$$
$$\tau^2 \sim Uniform(10, 50)$$

Y is runs scored in a particular inning. μ_{year} and σ^2_{year} indicates the average and variance respectively for a particular calender year / season.

4.2 Tools

The data was analysed using JAGS through the rjags package in R to generate the samples from the posterior distribution of the parameters. We run 3 chains in model for each format with 10,000 iterations.

5 Results

5.1 Test Match Career

Test match is played over 5 days and each team bats alternately twice and the team that scores the more number of runs cumulatively wins the test match. The match is considered a draw if all the innings do not get complete in 5 days where an innings is complete when 10 out of 11 batsmen get out. A batsmen can bat till the innings is complete or till the particular batsmen gets out. Test cricket is considered the toughest format of cricket.

Rahul Dravid was widely considered to be one of the most consistent player. In figure 2, we look at the posterior 95% credible intervals (values can be found in table 1 in Appendix A) for the average in each seasons (years 1996 to 2012) of test career. 95% credible interval indicates the batting average for a particular year lies in the interval with 95% probability. Since the number of matches played in a season is different, the variance is slightly on the higher side in general.

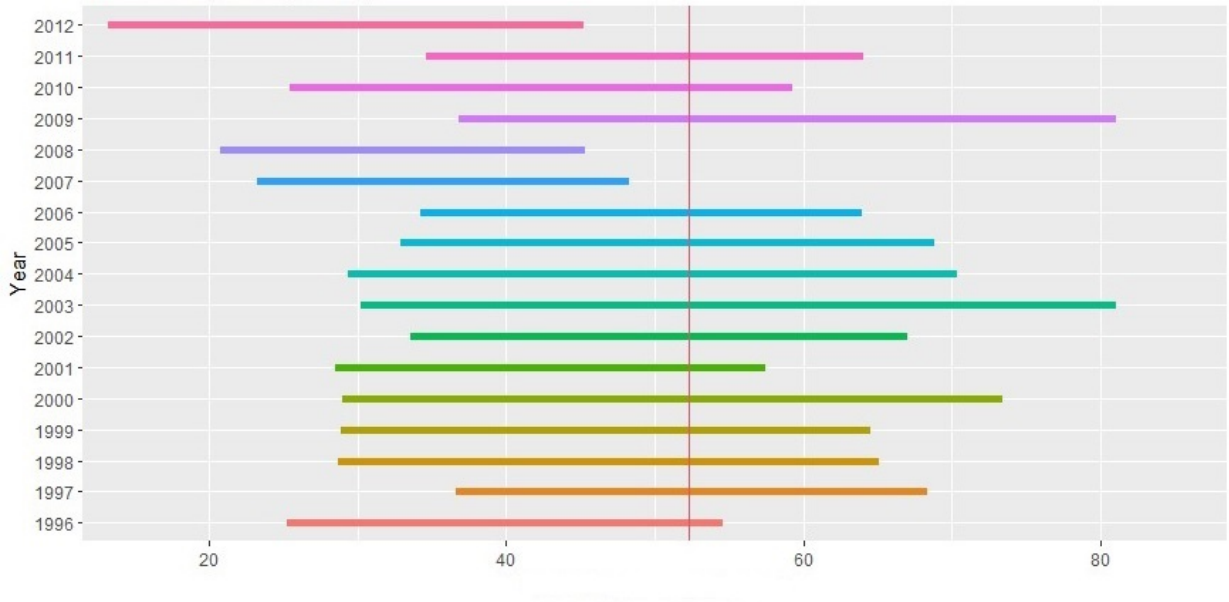

Figure 2: Credible Intervals of averages by year in test career

Since we have a model which allows seperate variance parameter for every year, the intervals are of different lengths. It can be seen that the player was having a lean patch in year 2007 and 2008 where the interval does not contain the career average. Similarly the interval for the year 2012 is considerably lowe than all the other years. Bad form is a part of every sportsperson's career and Rahul Dravid is no exception. But overall the player has had a good career. We can see from the intervals that years 2002-2006 were the best years in the player's career.

The posterior plot for the model parameters θ and τ can be seen in figure 3 below (Posterior 95% credible interval values can be found in table 1 in Appendix A) and they seem to be on expected lines. Trace plots

indicates proper mixing and convergence for each of the parameters listed. For the reasons discussed earlier, the variance is considerably high in the data which is evident in the plot for τ.

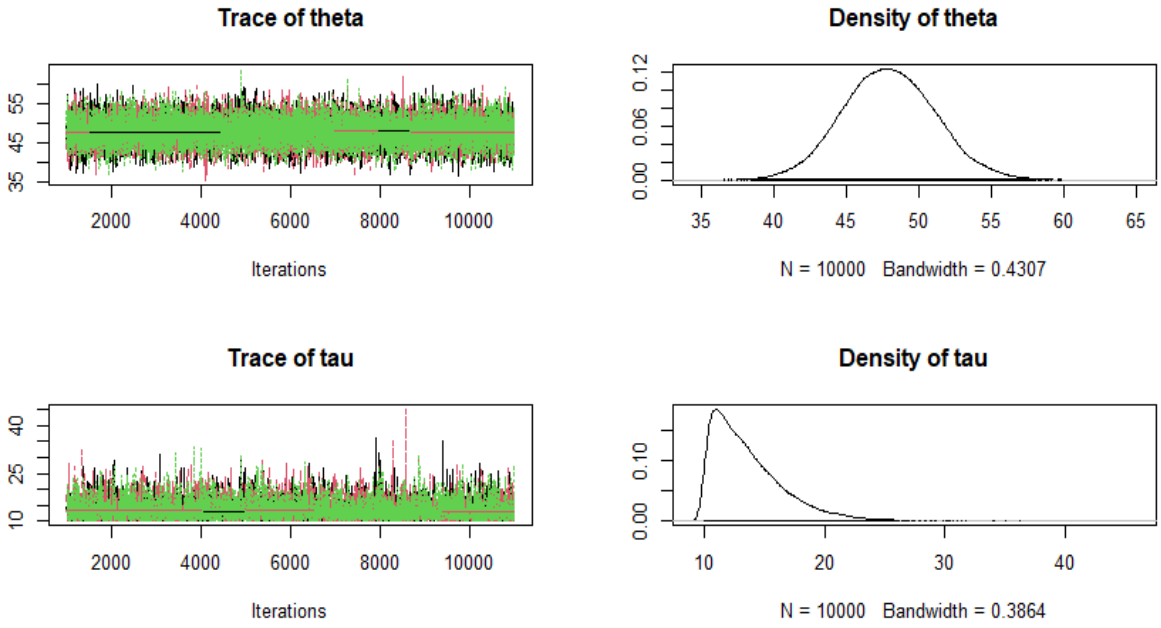

Figure 3: Posterior plots for model parameters

Since there are mean parameter for 17 years, their posterior density plots can be found in Appendix A in figure 5. We do not include trace plots for the mean of each year. However trace plot for 3 randomly chosen years (1999, 2003 and 2010) can be found in appendix A in figure 6 which indicates proper mixing and convergence.

6 Discussion

As far as the analysis is concerned, the posterior for μ and σ remain similar even if we change the prior or hyperprior parameters. Whereas the posterior for τ (the standard deviation of the mean in each of the season) is sensitive to the prior choice the number of observations per season. Hence a high variation is expected and more so since the sample points (in each year) are on the lower side as compared to career sample points in our data.

To check if the model used behaves well, a posterior predictive p-value is calculated using data replicates from the posterior distribution. We calculate the difference in runs scored on home soil (matches played in India) and away soil (matches played outside India) using a standard t-test using our observed data and replicates from posterior $p(y^{rep}|y)$. Comparison is made between the test statistic for observed data and predicted data. The plots for the test match career can be seen in figure 4. Since the observed test statistic is around the middle of the replicated statistic, it is safe to say the model works well.

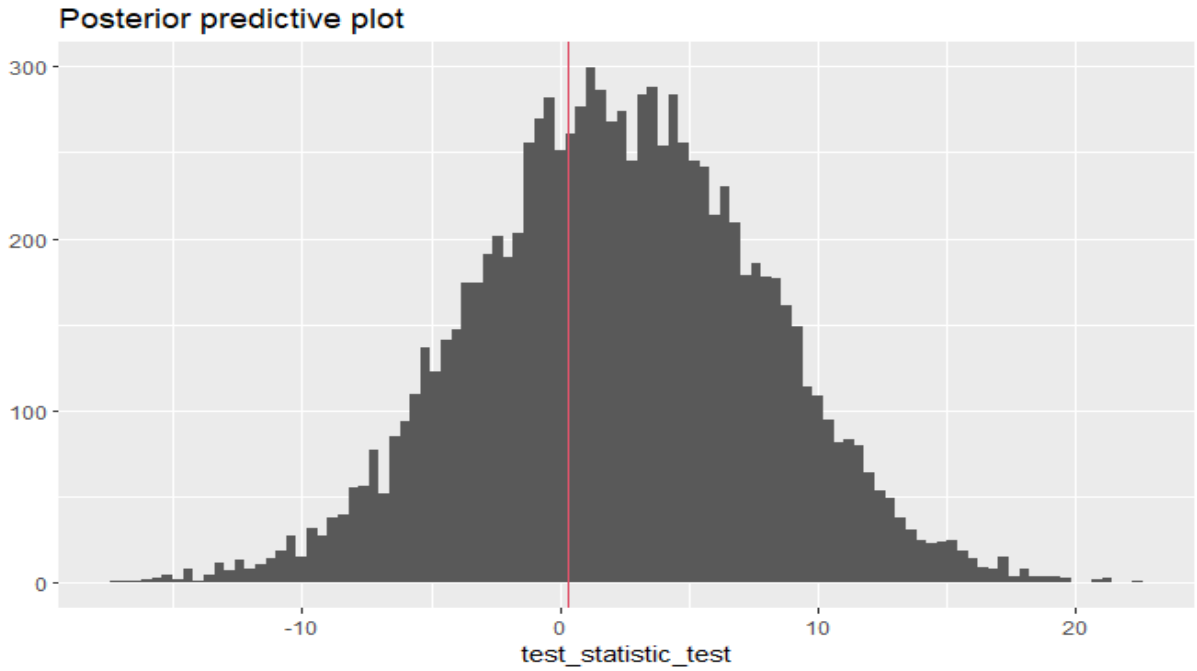

Figure 4: Posterior predictive plot

Considering the credible intervals, we can say the player has been consistent player throughout his test career. Playing the game for 17 years at the international level is certainly commendable but to be consistently performing well at that level speaks volumes about the skill and ability of Rahul Dravid. In cricketing circles, Rahul Dravid is highly regarded and is considered one of the best player to step on the cricketing field.

7 Future Work

In this age of data analytics and big data, there are specific tools developed which are specifically used in Sports as there are so many factors which affect the performance of a player in any sports to be specific. In terms of cricket some of the factors which can affect the performance is the opposition team, the location where the match is played as the pitch on which the game is played impacts the performance, the weather conditions to name a few. If we incorporate the opposition and the location where the match is played in the model, the variation would surely be better explained and in turn, the intervals can be expected to be on the shorter side. A model with linear regression for the mean parameter with the specified variables we just discussed can be considered for future work.

8 References

Internet
Dataset : https://www.kaggle.com/anirbna/sachin-tendulkar-batting-stats
Bayesian Data Analysis (Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin)

Appendix A

The posterior density plots for Test format can be seen in figure 5 below

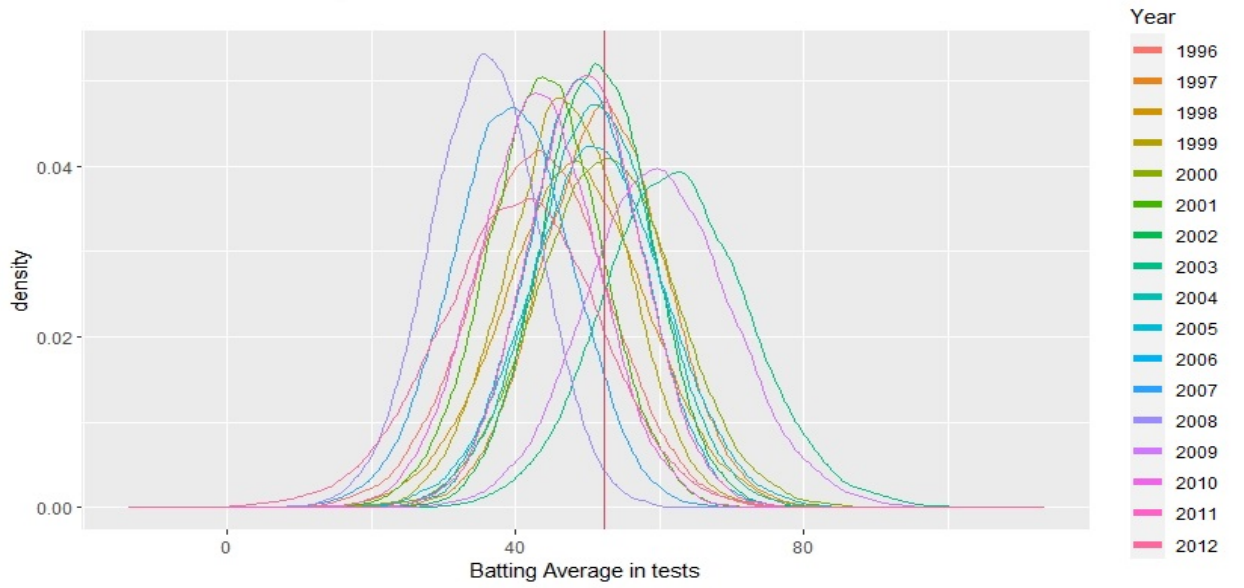

Figure 5: Posterior density plots of averages by year in Test Career

The trace plots for the mean of years 1998, 2003 and 2010 (μ_{1998}, μ_{2003} and μ_{2010}) from the Test career are in figure 6 below. It indicates proper mixing and convergence.

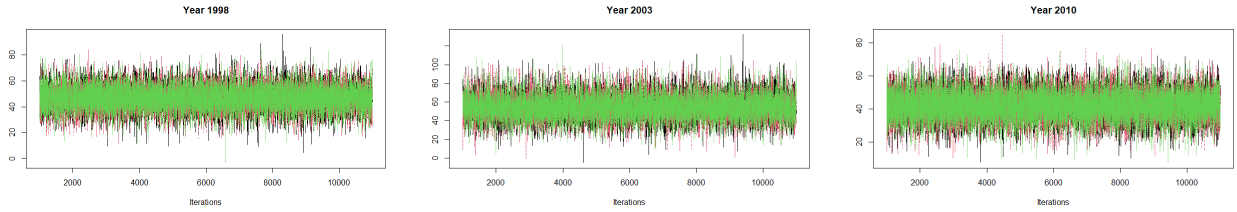

Figure 6: trace plot for μ_{1998}, μ_{2003} and μ_{2010} of test career

Table 1 below contains the 95% credible intervals for the mean parameter μ_{year} in the model for Test career

Year	μ_{year}	σ_{year}	95% Credible interval for μ_{year}
1996	39.71	29.21	(25.71 , 54.30)
1997	52.24	41.38	(36.33 , 68.04)
1998	46.75	37.94	(28.61 , 64.78)
1999	46.63	52.80	(29.18 , 64.36)
2000	50.85	64.19	(28.69 , 73.14)
2001	42.80	42.16	(28.14 , 57.38)
2002	50.38	55.35	(33.74 , 67.29)
2003	54.98	85.94	(30.41 , 81.55)
2004	49.77	67.41	(29.68 , 70.38)
2005	50.84	42.69	(32.81 , 68.86)
2006	49.11	42.69	(34.30 , 63.93)
2007	35.50	30.57	(23.28 , 48.24)
2008	32.85	35.91	(20.54 , 45.30)
2009	58.79	53.15	(36.82 , 81.36)
2010	42.23	48.67	(25.35 , 59.02)
2011	49.17	43.28	(34.48 , 64.11)
2012	27.10	19.94	(13.27 , 45.39)

Table 1: Parameter posterior 95% credible intervals