Inferring the contribution of ground conditions variability to the variability of first innings scores in One Day International cricket[*+]

Scott Brooker and Seamus Hogan

Economics Department, University of Canterbury, Christchurch, New Zealand

Abstract

This paper is part of a wider research programme using a dynamic-programming approach to modelling the choices about the amount of risk to take by batting and bowling teams in One Day International cricket. An important confounding variable in this analysis is the ground conditions (size of ground, nature of pitch and weather conditions) that affect how many runs can be scored for a given amount of risk. This variable does not exist in our historical data set and would regardless be very difficult to accurately observe on the day of a match.

In this paper, we consider a way of estimating a distribution for the ground conditions using only the information contained in the first-innings score and the result of the match. The approach uses this information to estimate the importance of ground conditions in the determination of first innings total scores. We assume a functional form for a model of first innings scores and we estimate the parameters of our model using Monte Carlo methods. We test the impact of a significant rule change and we apply our findings to selected matches before and after the new rules came into play.

---

[*]This paper was prepared especially for the New Zealand Association of Economists (NZAE) Conference.

[+] We are grateful to New Zealand Cricket, Sport and Recreation New Zealand (SPARC) and the SAS Institute (NZ) for providing us with the means, opportunity and required data to conduct this research.

1 The influence of ground conditions

The ground conditions present on the day of a match play an important role in the sport of cricket, but they are certainly not easy to measure. In this paper, we develop a method of inferring the contribution of the variability of ground conditions to the total variability of the score. We use this information to construct conditional distributions for ground conditions based on the outcomes of matches. We are able to extract a measure of ground conditions where no direct measure is available in our data set. We assume that the reader has a basic understanding of the structure of a game of One Day International (ODI) cricket; should this not be the case we recommend the reading of Appendix 1 before continuing.

The outcomes that take place on a sports field are related closely to the performance and ability of the players or athletes taking part in the sport; however, these are not the sole determinants. The sporting world contains many examples where factors unrelated to player and team ability have an impact on the type of game played and the result. The main factors of this type relate usually to weather conditions either prior to or during the time of competition, as well as the characteristics of the venue. The impact varies significantly from sport to sport. In sports mostly played indoors, such as basketball, the impact of weather conditions should be close to zero but “stadium” factors such as the quality of the lighting may have an impact. In rugby union, wet and muddy conditions often lead to a more conservative game, involving less lateral movement of the ball and lower scores. In sprinting, athletes are able to run faster with the wind, but the administrators of the sport decide that world record times will not be counted if there is deemed to be too much wind assistance. In the extreme, a sporting event may not even take place

because of weather conditions; for example, the postponement of sailing events due to insufficient wind.

1.1 Factors contributing to the outcome of a game of ODI cricket

We outline five main factors that influence the first innings score as well as the likelihood of each score being a winning one. These factors are:

- The skill levels of the players on both teams;
- Luck;
- Ground size;
- Pitch conditions;
- Weather conditions.

The skill measure includes the overall strength of the teams as well as the relative strengths of the players in bowling, fielding and batting. While a first innings score of 280 might produce a close contest between two strong batting teams, a score of 220 could produce an equally close contest between two strong bowling and fielding teams, given the same conditions.

Luck plays a role in the outcome of a match; for example, poor umpiring decisions can have a marked influence, as can uncontrolled aerial shots that fall safely rather than going directly to the fielder.

On a small ground, it is relatively easier for the batsmen to hit the ball out of the playing field for boundaries and for this reason scores tend to be higher on small grounds than on large grounds.

Pitches are extremely variable in their nature. The moisture content, the type of soil used, the hardness, the amount of grass and any cracking present on the pitch all have an impact on how the ball behaves when it bounces on the pitch. Any

movement or change of direction of the ball after hitting the pitch makes batting more difficult, as does inconsistent bounce, extreme pace off the pitch and extreme lack of pace off the pitch. Pitches are very individual; therefore, it is not appropriate to assume that all pitches at a particular ground will behave in the same way.

A fascinating aspect of the game of cricket is the tendency of the ball to "swing", or change direction, in the air after it has been bowled. This swing, if present, makes batting significantly more difficult and is likely to lead to lower scores. On a cloudy or humid day the ball generally swings significantly more than on sunny dry days. For this reason the weather is our final factor influencing the outcome of the game.

Our analysis of the game of cricket is limited if we ignore the variability of ground conditions. We explain this by way of example. One of the models in our research programme predicts the average additional runs scored from any possible point in the first innings. If we do not include a variable for ground conditions in our model, we are effectively assuming that all ground conditions are the same. It seems intuitive that this would be a model for what would happen in average ground conditions, but on closer inspection this is not the case. Consider a team that makes a very poor start to a match, perhaps two batsmen are out on the first two balls of a match. Given this start, it is more likely than not that this match is being played in worse than average ground conditions, from the point of view of the batting team. This likelihood, implicitly built into the model, means that the predicted average additional runs for this situation will incorporate the fact that we have a higher probability of being in poor batting conditions than good batting conditions. If we are, in fact, on an average pitch and the poor start was due to bad batting, good bowling or simply luck then our model is going to underestimate the expected number of future

runs. The opposite holds for situations where the batting team makes a very good start. Including a ground conditions variable in our model has two effects; improving the accuracy of the model for average ground conditions and identifying the sensitivity of the model in various situations to different ground conditions.

Directly observing a ground conditions variable is extremely difficult. While the size of a ground is generally constant, weather conditions and the nature of the pitch are certainly not. Some grounds are more likely to have certain weather and pitch conditions than others, but there is significant variation due to the time of year and beyond this a large random component. Measuring the pitch conditions would be difficult and measuring the effect of the weather conditions would be almost impossible, as there are such a variety of factors that can make a ball swing. There may be, to the naked eye, two identical days and the cricket ball may swing one day and not the next. A further issue is that our data set is historical and therefore determining the nature of the pitch, in particular, in games played several years ago is problematic. We decide that an indirect approach to estimating the ground conditions is required. In the remainder of this paper we present a possible approach.

2 The theoretical model

We create two variables by separating the factors influencing the outcome of a game into two groups. Those factors that are specific to the ground conditions on the day are ground size, pitch conditions and weather conditions. We combine these factors into a variable “Conditions”. The remaining factors, skill and luck, ought to be independent of the ground conditions on the day and we combine these factors into a new variable “Performance”. Note that we will consider performance to be a positive function of the batting team’s skill and luck and a negative function of the bowling

team's skill and luck; therefore, an above average value for performance will imply a better performance by the batting team than the bowling team, but not a particular level of either batting or bowling performance.

2.1 The assumed relationship between the two factors

Let $S = \rho + \chi$, where S is the first innings score, ρ is a measure of "Performance" and χ is a measure of "Conditions". We assume an additive relationship between our two right-hand-side variables since we do not expect the deviations from the value of conditions of the total scores achieved to vary between different sets of conditions.

We further assume that ρ and χ are independent and normally distributed variables having distributions $\rho \sim N(0, \sigma_\rho^2)$ and $\chi \sim N(\mu_\chi, \sigma_\chi^2)$, where $\mu_\chi = \mu_S$. This normality assumption is based upon an understanding of the game of cricket. The performance measure is a combined measure of batting team performance and fielding team performance. This means that the most extreme performances would require an extremely good performance from one team and an extremely poor performance from the other team. This would be less likely than an average total performance, which could be caused by almost unlimited combinations of good batting / bad bowling, or vice versa, completely cancelling each other out. This is true even if the separate batting and bowling performance distributions were uniform.

Conditions have an element of repeated sampling from the same distribution, as there are some constant factors associated with playing multiple matches at repeated venues. These include soil type, ground size and predominant climate. This again means that extreme values of conditions are going to be less likely than the values in the middle, assuming that there is greater variability of conditions within a ground than between grounds.

Since the sum of two normally distributed independent random variables is normal, we are also implicitly assuming that the first innings scores have a normal distribution[1]. We note that assuming normality for performance and conditions raises the prospect of a negative total score; however, this is extremely unlikely over the range of the data. The log-normal distribution, while having the desirable property of being bounded at zero, does not fit the data well.

We centre the conditions variable around the mean first innings score and the performance variable around zero in order to create the interpretation that a performance is a certain number of runs more or less than the conditions are worth. This approach, however, is simply a normalising assumption that we make without loss of generality.

2.2 Calculating the conditional distributions for conditions

Let ω be a binary variable taking a value of one if the team batting first wins the match and zero otherwise. Using Bayes' theorem, we can create conditional distributions for conditions, given the first innings score and the result of the game.

Let $f(\chi)$, $g(S)$, $k(\rho)$ and $h(S, \omega)$ denote the density functions of χ, S, ρ and the joint density function of S and ω, respectively. Additionally, let $\Pr(\omega)$ be the probability of observing outcome ω. We define the conditional density function for conditions in match i as

$$f(\chi \mid S = S_i, \omega = \omega_i) = \frac{f(\chi) g(S = S_i \mid \chi = \chi_i) \Pr((\omega = \omega_i \mid S = S_i) \mid \chi = \chi_i)}{h(S = S_i, \omega = \omega_i)}. \quad (1)$$

[1] We test this assumption in a later section.

We need to determine Pr($\omega|S$), $f(\chi)$ and $k(\rho)$ before we can estimate equation (1). Estimating Pr($\omega|S$) is a simple probit model which we will define later in the paper. To determine $f(\chi)$ and $k(\rho)$ we will firstly need to determine the values of σ_ρ^2 and σ_χ^2.

2.3 Inferring the values of σ_ρ^2 and σ_χ^2

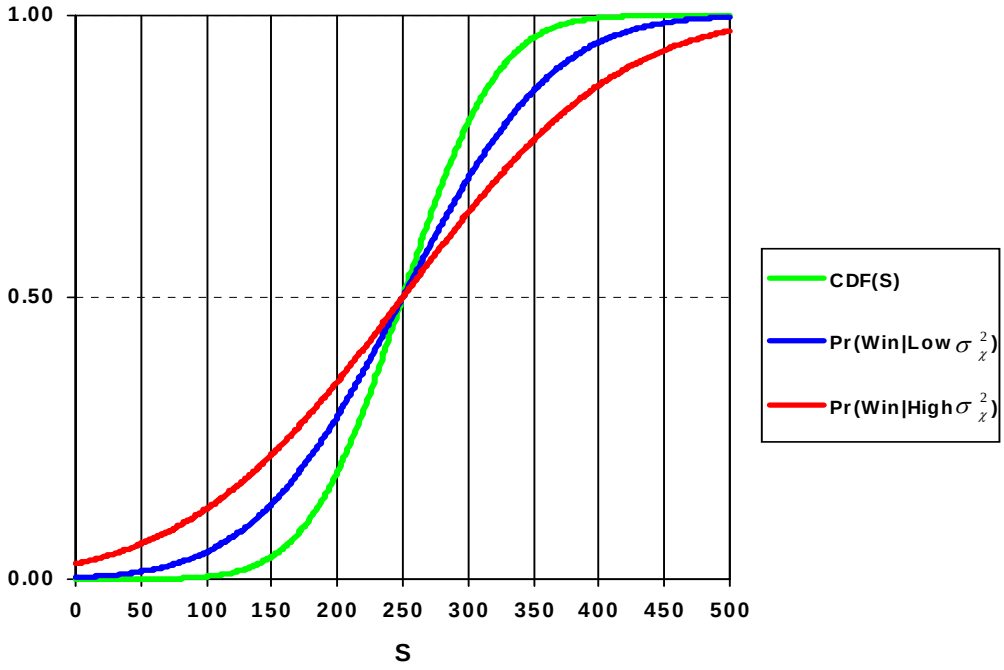
In order to show how we might go about estimating σ_ρ^2 and σ_χ^2, consider two hypothetical games. In hypothetical game one, we assume constant conditions ($\sigma_\chi^2 = 0$). Team one draws a number from the performance distribution ρ and team two then draws a number from the same distribution. The team drawing the higher value of ρ wins and the first innings score *S* is equal to $\chi + \rho$. In this game, the probability of achieving a score in the first innings is exactly the same as the probability of successfully beating that score in the second innings. That is, the graph of the cumulative distribution of first-innings scores will be identical to the graph showing the probability that a team with a given score in the first innings will win the game.

In hypothetical game two, we allow conditions to have a positive variance ($\sigma_\chi^2 \succ 0$). This time nature draws a value for χ before the game begins and teams one and two subsequently draw values for ρ. As with hypothetical game one, the team drawing the higher value of ρ wins and the first innings score *S* is equal to $\chi + \rho$. In this game, however, the presence of variability in conditions will affect both the observed distribution of *S* and the probability of each score being a winning one. If we only observe the distribution of first innings scores; that is, we do not observe any information about performance or conditions, the scores achieved contain information about the conditions. If we observe a relatively low first innings score, it is more

likely that this game was played under a relatively low draw from the conditions distribution. The opposite holds for a high first innings score.

This conditions variance affects the second innings probability of winning function. Given that a low score is likely to have come from a low conditions draw, the probability of team one successfully defending the score is higher than the a priori probability of scoring that score in the first innings when nothing is known about the conditions. The probability of winning function is therefore flatter than the cumulative density of scores function where we have a non-zero variance of conditions. The higher is the variance of the conditions, the flatter is the probability of winning function, assuming a constant total variance of scores. We illustrate this in Figure 2.1.

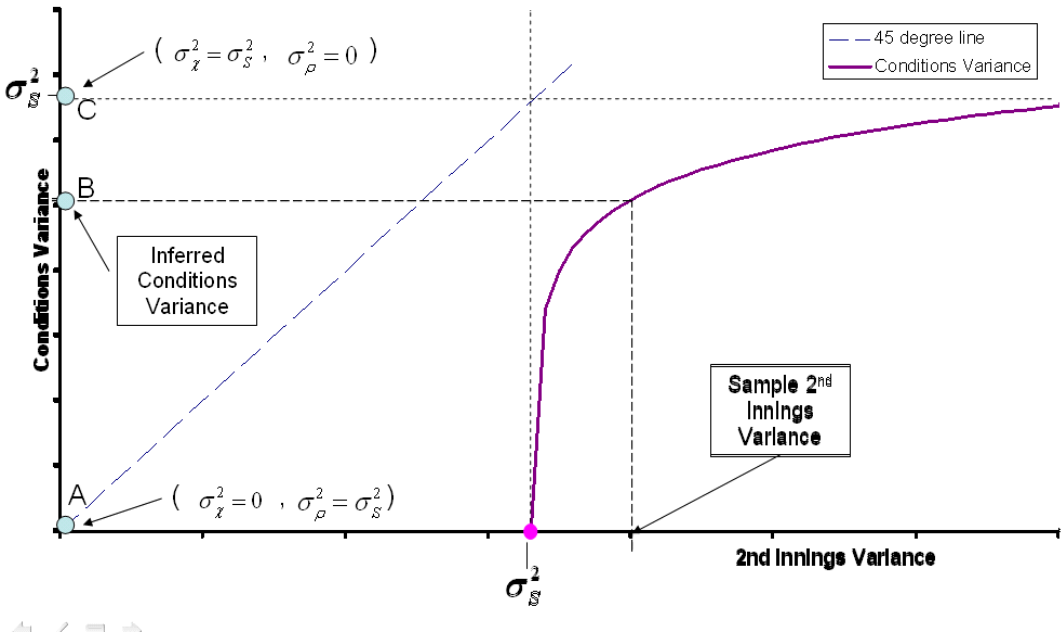
Figure 2.1: First Innings Score and Probability of Winning



At this point, we need to define the second innings distribution as the function whose cumulative density function is identical to the probability of winning function.

Note that this is a very different concept to the distribution of the actual scores observed in the second innings, which we do not use in this paper other than in determining the value of ω_i for each game. It follows that we can infer the contributions of σ_χ^2 and σ_ρ^2 to σ_S^2 by comparing the variances of the first and second innings distributions. We show this in Figure 2.2 where we plot combinations of conditions variance and second innings variance for a given value of the first innings score variance.

Figure 2.2: Inferring the contribution of conditions variance for a given σ_S^2



A conditions variance of zero (see Point "A") will lead to the second innings variance being equal to the first innings variance; this is the case in hypothetical game one, where the variation of performance explains the total variation of score. At the other extreme, a conditions variance tending to the first innings score variance (see Point "C") will lead to the second innings variance tending to infinity. In this case, the

entire variation in score is due to the conditions and the level of performance is constant. If we know the value of the second innings variance then we can read the value of the implied conditions variance from the vertical axis of the graph (see Point "B").

2.4 Accounting for the second innings advantage

We need to make an additional adjustment before we can begin to estimate the values for σ_{χ}^2 and σ_{ρ}^2. Hypothetical game two assumes that the two teams draw from a distribution of ρ whose mean and variance are exogenous; that is, both teams are drawing from the same distribution so there is no advantage in the order of drawing. However, there is a second-mover advantage in ODI cricket. This is due to the team batting second having a known target score, resulting in them being able to adjust their risk strategy depending on the target. While the team batting first wishes to, in most situations, maximise their expected total score and therefore chooses the distribution of ρ with the highest mean, the team batting second wishes to maximise the probability of scoring a higher total than the team batting first achieved. The binary nature of the outcome of the second innings ensures that a team chasing a high total optimises by drawing from a high variance distribution, while a team chasing a low total optimises by drawing from a low variance distribution. In both cases the optimal distribution from which the second mover draws ρ would likely have a lower mean than the optimal distribution from which the first mover draws ρ. A team chasing 200 runs in conditions which are worth 250, for example, might optimise by drawing from a distribution with a mean of 240 and standard deviation of 10 runs, rather than the optimal first innings distribution which might have a mean of 250 and standard deviation of 30 runs.

We assume that the second innings advantage is the difference between the means of the first and second innings distributions and that it is a constant number of runs regardless of the first innings score. We will incorporate this second innings advantage into our Pr(ω) functions in our conditional probability formula for conditions given score and result.

3 The Data

3.1 Sources and timeframe

The data used in the majority of analyses in our wider research programme is a set of 311 matches over the period 20 July 2001 to 25 January 2008**.** It consists of ball-by-ball information collected by New Zealand Cricket. The research described in this paper only requires three pieces of information; the date that the match was played, the first innings score and the result of the match. This information is publicly available on www.cricinfo.com. We therefore use the complete set of matches played by the top eight ranked cricket countries[2] over the period 9 January 2001 to 4 July 2008, totalling 591 matches. Our ball-by-ball data set is a subset of this data set. The wider data set gives us greater estimation power without venturing too far from the date range of our ball-by-ball data. Using this data set also eliminates a selection bias towards matches played by New Zealand.

[2] Rankings are calculated using a system that allocates points for each win and loss. The number of points is adjusted for the strength of the opposition and the average points per match played determines the team ranking.. The top eight ranked ODI teams as at 17 February, 2009 are, in order: South Africa, Australia, India, New Zealand, Pakistan, England, Sri Lanka and West Indies. While the ordering within this top eight often changes, the countries making up the top eight do not. The range in points between the top team (South Africa) and the eighth team (West Indies) is 34 points, while the ninth-placed team (Bangladesh) is 45 points short eighth place. Any team in the top eight would feel that they are a reasonable chance to beat anyone else; however, it is a major surprise when a top eight country loses to a country outside the top eight. For this reason we exclude matches involving non top eight countries for fear that they may distort the data.

3.2 A potential structural break

A major rule change occurred in ODI cricket during the period of our data set. All matches played prior to 1 July, 2005 required that the fielding side could place no more than two fielders outside an approximate oval (know as the "circle") drawn 30 yards from either end of the pitch for the first 90 balls of each innings. This is a fielding restriction as compared to the five fielders allowed outside the circle for the remainder of the innings. In contrast, matches played between 1 July, 2005 and 30 September, 2007 required the above restriction to be in place for the first 60 balls of an innings and for two additional periods of 30 balls, the timing of which were decided by the fielding captain. These 30-ball periods are known as "power plays". A smaller rule change occurred on 1 October 2007, from when fielding sides were allowed three fielders, rather than two, outside the restricted area during the second power play[3].

The increased presence of fielders close to the batsman and the lack of fielders patrolling the boundary serve to increase both scoring rates and the risk of a batsman getting out. There are generally more runs available but it is more difficult to score these runs without hitting the ball over the top of the fielders, rather than along the ground, resulting in the batsman risking hitting a catch. Before we move forward with our analysis, we assume that the minor rule change allowing three fielders in the restricted area during the second power play has no significant effect. By far the more significant rule change is the extension of the fielding restrictions from 90 balls to 120 balls in total. This enables us to divide our data set into two subsets, matches played without the power play rule (Era 1) and with the power play rule (Era 2). Era 1 contains 344 matches while Era 2 contains 247 matches.

[3] As from 1 October 2008, this rule has significantly changed again as now the batting side is responsible for electing the timing of one of the power plays and both power plays allow three fielders outside the circle.

3.3 Testing for the significance of a rule change

We have, from our 591 matches, a distribution of first innings scores as well as the match result for each one of those 591 scores. We initially want to estimate the probability of winning for a given first innings score. For each era, we construct a probit model where we regress the outcome of the game on the score achieved by the team batting first. We exclude tied matches, as they are so rare that they cannot be accurately included in the estimation. We define the probability of winning given first innings score S function $\Pr(\omega|S)$ as a simple probit model in equation (2). Let

$$\omega = \begin{cases} 1 & \text{if the team batting first wins the game} \\ 0 & \text{if the team batting second wins the game} \end{cases}$$

$$\Pr(\omega = 1 | S = S_i) = \Phi(\alpha + \beta S) \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution. It follows that

$$\Pr(\omega = 0 | S = S_i) = 1 - \Pr(\omega = 1 | S = S_i). \tag{3}$$

In Era 1, $\alpha = 3.5168$ and $\beta = 0.0145$, while in Era 2, $\alpha = 2.7168$ and $\beta = 0.0104$. These probit models reveal the probability of winning for the team batting first, given that they scored a particular total. We establish second innings distributions in terms of score by a Monte Carlo method. We sample 10000 random numbers ~ U(0, 1) as probabilities of winning. Each probability corresponds to a score in the probit model and we create a distribution for second innings score in each era. The probability of a particular score being a winning score is the percentile of this created "second innings distribution" at which the score occurs.

We plot the cumulative distribution of first innings scores under both sets of rules in Figure 3.1. To the naked eye there does not seem to be a large difference between the two distributions. We also plot the cumulative distribution of the

probability of defending each score in Figure 3.2. Clearly, there is a difference between the probit models created for the games played under each set of rules.

Figure 3.1: The first innings distributions

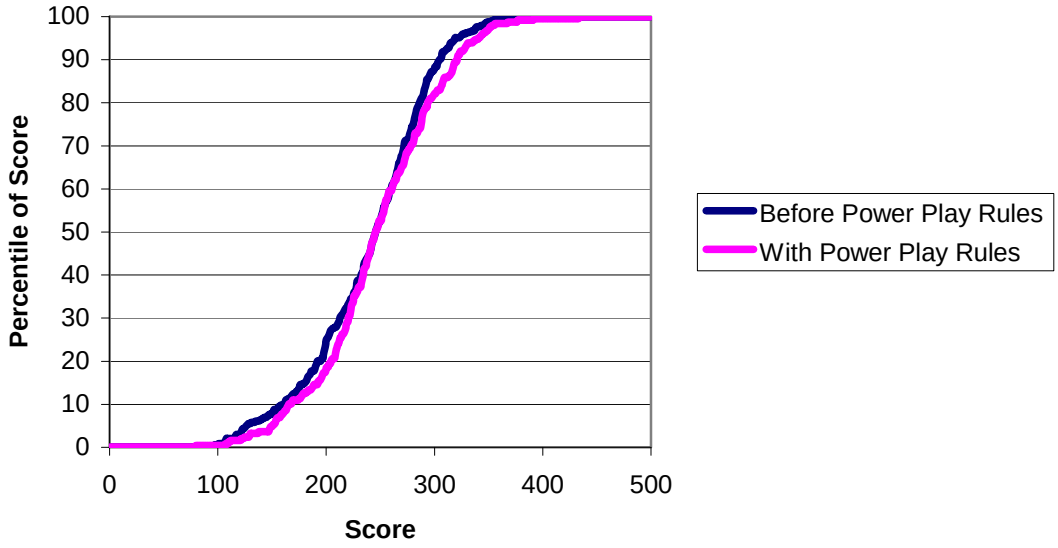
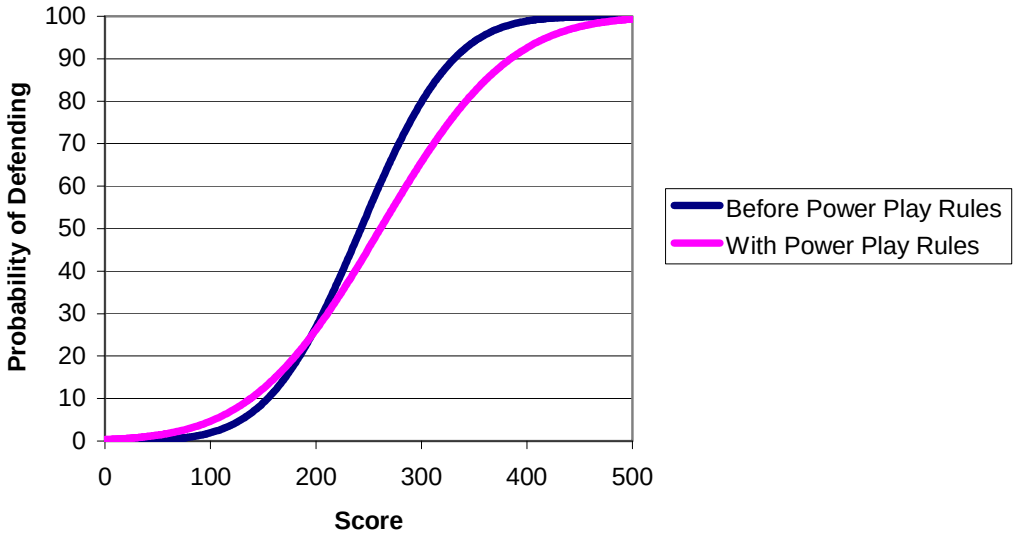


Figure 3.2: The second innings distributions



3.4 The Monte Carlo test procedure

Table 3.1 shows the key summary statistics of the first and second innings distributions, both before and after the implementation of the power play rules.

Table 3.1: Means and Variances of the distributions

	N	Mean 1st Innings	Variance 1st Innings	Mean 2nd Innings	Variance 2nd Innings
Before Power Play Rule	344	239.68	3144.03	242.52	4757.15
With Power Play Rule	247	247.89	3300.87	261.08	9250.81
Total	591	243.11	3220.50	249.54	6499.13

Our first and second innings distributions have a different variance before and after the implementation of the power play rules. This makes standard t-tests of the difference between two samples inappropriate. We test for significant differences in the mean and variance of the pre power play and post power play distributions by conducting a Monte Carlo simulation. Our null hypothesis is that the 591 games all come from the same distribution; therefore, we randomly allocate the games to the two eras and compare the resulting means and variances with the actual means and variances in each era. Our Monte Carlo simulation design is included as Appendix 2.

Since we assume normality for our first and second innings distributions, the mean of each distribution is identical to the median and therefore the mean of the second innings distribution has the interpretation of being the score resulting in a 50% chance of winning. The Monte Carlo simulation generates 10,000 observations from the distributions of the following eight sample statistics:

- Mean score, first innings distribution, with 344 games.
- Variance of score, first innings distribution, with 344 games.
- Mean score, second innings distribution, with 344 games.

- Variance of score, second innings distribution, with 344 games.
- Mean score, first innings distribution, with 247 games.
- Variance of score, first innings distribution, with 247 games.
- Mean score, second innings distribution, with 247 games.
- Variance of score, second innings distribution, with 247 games.

We now determine the percentiles of our simulated distributions at which the means and variances of our actual distributions occur. In table 3.2, we repeat the means and variances information from table 3.1, this time including the percentiles.

Table 3.2: Percentiles of the simulated distributions for calculated parameters

	N	Mean 1st Innings	Variance 1st Innings	Mean 2nd Innings	Variance 2nd Innings
Before Power Play Rule	344	239.68	3144.03	242.52	4757.15
Percentile		4.28%	31.47%	3.38%	3.79%
With Power Play Rule	247	247.89	3300.87	261.08	9250.81
Percentile		95.71%	64.05%	98.25%	93.74%

The interpretation of the percentiles is straightforward. Using the first innings mean score as an example, the simulated distribution is telling us that we expect a mean first innings score of 239.68 or lower in only 4.28% of randomly selected samples of 344 games. Likewise, we would expect a mean first innings score of 247.89 or lower in 95.71% of randomly selected samples of 247 games.

At the 5% significance level, the results of our simulation study suggest that there is a difference between the mean first innings scores as well as a difference in both the mean and variance of the second innings distributions of the matches played under the different sets of rules. The only parameter that we cannot conclude is

significantly different under the two sets of rules is the variance of the first innings scores.

The variance of the second innings distribution is significantly different in Era 2 to Era 1 while the variance of the first innings does not seem to be significantly different. This may be because conditions have genuinely become more variable in Era 2, or it may be because the rule change has led to conditions becoming more important in the second innings. A possible explanation is that, in good batting conditions, a team batting second is in a better position to take advantage of the extra fielding restrictions and is therefore able to chase higher totals successfully. Note that the opposite should not hold for teams chasing low scores. In this case, the fielding captain will want the batting team to engage in a high risk strategy, which he encourages by positioning more fielders closer to the batsmen who attempt to stop the easy, safe scoring opportunities for the batsmen. It is certainly possible that a poor batting pitch will allow the bowling team to create relatively more pressure with this strategy than a good batting pitch, therefore increasing the bowling team's chances of winning. The rule, however, is a fielding restriction on how defensive the bowling side can be. There has never been a restriction on how attacking the bowling team can be so a possible conclusion is that if the Era 2 rules have made it easier for teams to defend low totals, then the rule change might simply be forcing sub-optimal bowling captains into the optimal strategy.

The differences in the means can be attributed to the extra restriction on the bowling team making it easier for batting teams to achieve higher scores. There has been a large increase in the second innings advantage, which we suggest is due to batting teams having an additional ability to engage in a high risk strategy when the situation requires it.

Although we cannot conclude with any certainty that the first innings distributions have different variances between Era 1 and Era 2, we decide that the difference in means as well as the variances of the second innings distributions is sufficient for us to proceed by treating the two eras as separate data sets.

3.5 Testing the distributions for normality

For our method of determining the share of first innings score variance attributable to conditions variance, it is convenient to assume that the first innings scores follow a normal distribution. We test the data for this normality in this section.

Our Era 2 data set of first innings scores fails to reject normality for both a Jarque-Bera (P-value=0.771) and a Kolmogorov-Smirnov test of normality (P-value=0.150). However, both tests reject the null hypothesis of normality at the 0.05% significance level when we assess the Era 1 data.

Figure 3.3 shows the cumulative distribution function of first innings scores for the Era 1 data as well as the cumulative normal distribution function with mean 239.68 and variance 3144.03. Figure 3.4 is the same graph but this time comparing the Era 2 data with the cumulative normal distribution function with mean 247.89 and variance 3300.87. It is apparent that a normality assumption would be the best option available here and we proceed despite our imperfect normality.

Figure 3.3: Comparison of Era 1 scores with a normal distribution

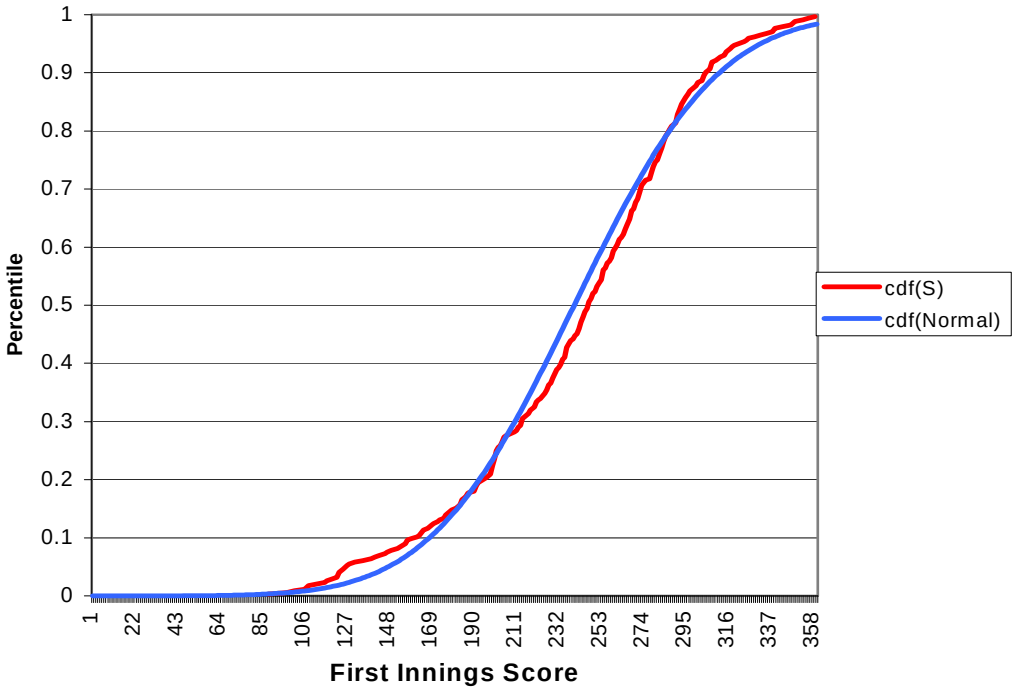
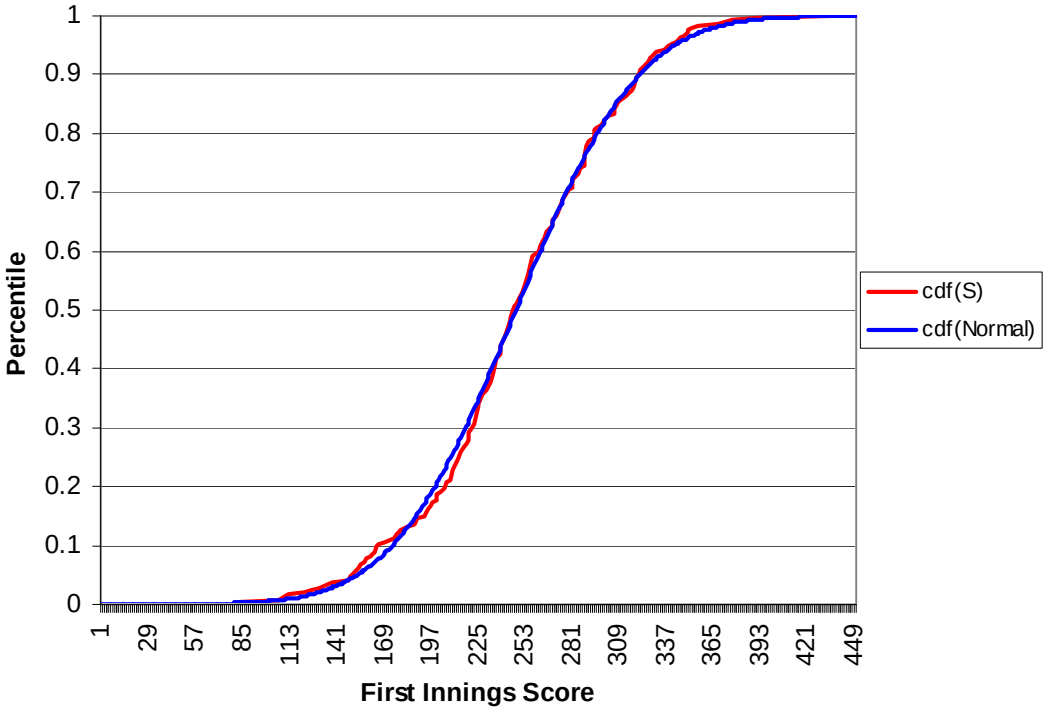


Figure 3.4: Comparison of Era 2 scores with a normal distribution



4 Analysis of the data

We split our data into games played before (Era 1) and after (Era 2) the implementation of the power play rules. We provide a reminder of table 3.1 as table 4.1. Note that there is a difference between the first innings mean and the second innings mean of approximately three runs in Era 1 and approximately 13 runs in Era 2. Also, note that the second innings variance is significantly higher than the first innings variance, more markedly in Era 2. These differences will play a major rule in determining our measures of conditions and performance.

Table 4.1: Means and Variances of the distributions

	N	Mean 1st Innings	Variance 1st Innings	Mean 2nd Innings	Variance 2nd Innings
Before Power Play Rule	344	239.68	3144.03	242.52	4757.15
With Power Play Rule	247	247.89	3300.87	261.08	9250.81
Total	591	243.11	3220.50	249.54	6499.13

Figures 4.1 and 4.2 show the differences between the first and second innings distributions. We have linearly adjusted the second innings distributions to remove the second innings advantage (three runs in Era 1 and 13 runs in Era 2). It is clear, particularly in Era 2, that the cumulative distribution functions cross over at approximately the 50% mark[4]. The implication of this crossover is that, after the removal of the second innings advantage, scores in the upper ranges of the distributions are easier to chase successfully than they are to score in the first innings. Conversely, scores in the lower ranges of the distributions are more difficult to chase successfully than they are to score. This is due to the second innings distribution having a greater variance.

[4] The crossover point is higher in the distributions of games without the power play rules due to the minor amounts of non-normality present in the first innings distribution under these rules.

Figure 4.1: Adjusted Era 1 cumulative distributions

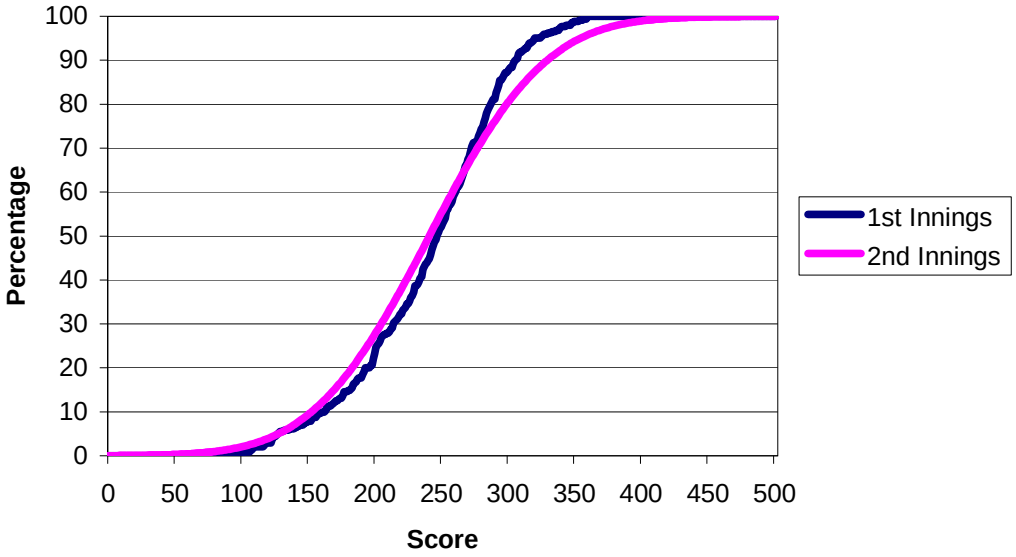
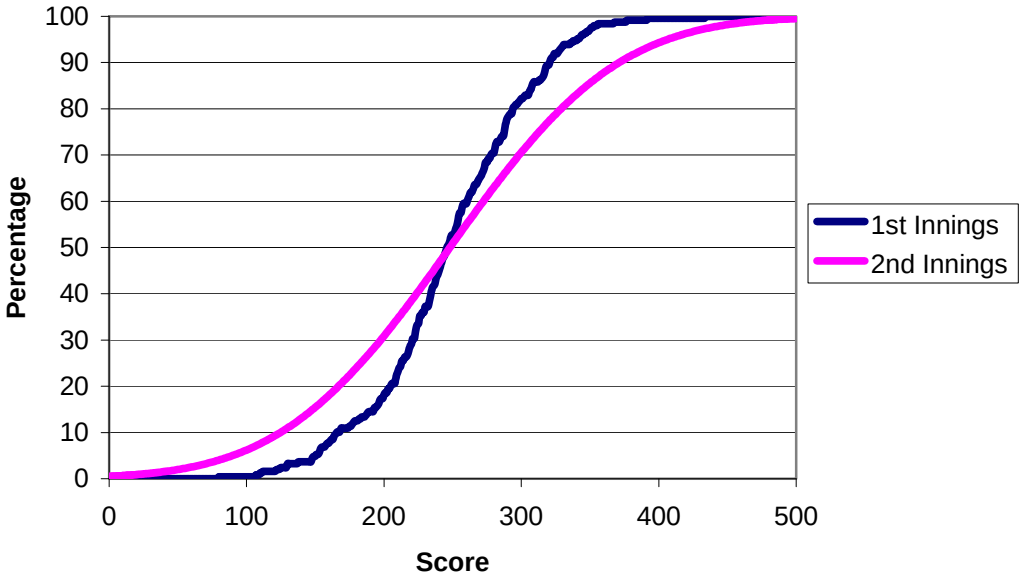


Figure 4.2: Adjusted Era 2 cumulative distributions



4.1 The procedure for splitting the total first innings variance

At this point, we change our notation slightly in order to incorporate the separate data sets from each era.

$$Let\ j \in \begin{Bmatrix} 1 & where\ the\ power\ play\ rules\ are\ not\ in\ place \\ 2 & where\ the\ power\ play\ rules\ are\ in\ place \end{Bmatrix}$$

$Let\ \rho_{ij}\ be\ the\ performance\ variable\ for\ game\ i\ in\ data\ set\ j$

$Let\ \chi_{ij}\ be\ the\ conditions\ variable\ for\ game\ i\ in\ data\ set\ j$

$Let\ S_{ij}\ be\ the\ first\ innings\ score\ for\ game\ i\ in\ data\ set\ j$

$$\rho_{ij} \sim N(0, \sigma_{\rho j})$$

$$\chi_{ij} \sim N(\mu_{\chi j}, \sigma_{\chi j})$$

$$S_{ij} = \rho_{ij} + \chi_{ij}$$

Next, we investigate combinations of $\sigma_{\rho j}$ and $\sigma_{\chi j}$ in order to find a combination that would result in a second innings distribution with similar variance to that obtained from the model of the actual data observed. We assume that χ_j, the conditions factor, is constant throughout a game. This enables us to construct a distribution of first innings score S_j as a function only of performance ρ_j, where this new distribution represents the level of performance that is required on average to achieve each score. As this distribution assumes that conditions are unknown, this is exactly the situation that we are faced with when we estimate the second innings distribution. It follows that this distribution should approximate the second innings distribution; since we are assuming that the conditions are the same for both teams then if ρ_j is higher in the second innings then the team batting second should win.

We set up a macro in SAS to split the first innings variance for each era into performance variance and conditions variance sixty[5] different ways and run a Monte Carlo simulation to determine the most appropriate split for rule set j. Each iteration of the Monte Carlo simulates values for χ_j and ρ_j and calculates the variance of the second innings distribution implied by these values. We then select the values of σ_ρ^2 and σ_χ^2 that provide the closest second innings variance to that observed in our data set. The full Monte Carlo steps are described in Appendix 3**.**

The 12th iteration of the Monte Carlo study using the first innings variance for games played in Era 1 provides the nearest second innings variance to the true Era 1 second innings variance. This implies that 80% of the first innings variance can be attributed to performance and the remaining 20% attributed to conditions. In Era 2, the 29th iteration of the Monte Carlo provides the nearest second innings variance to the true value. Here, performance accounts for 51.67% of the first innings variance with conditions accounting for the remaining 48.33%. It is clear that conditions have become relatively more important in Era 2 than they were in Era 1.

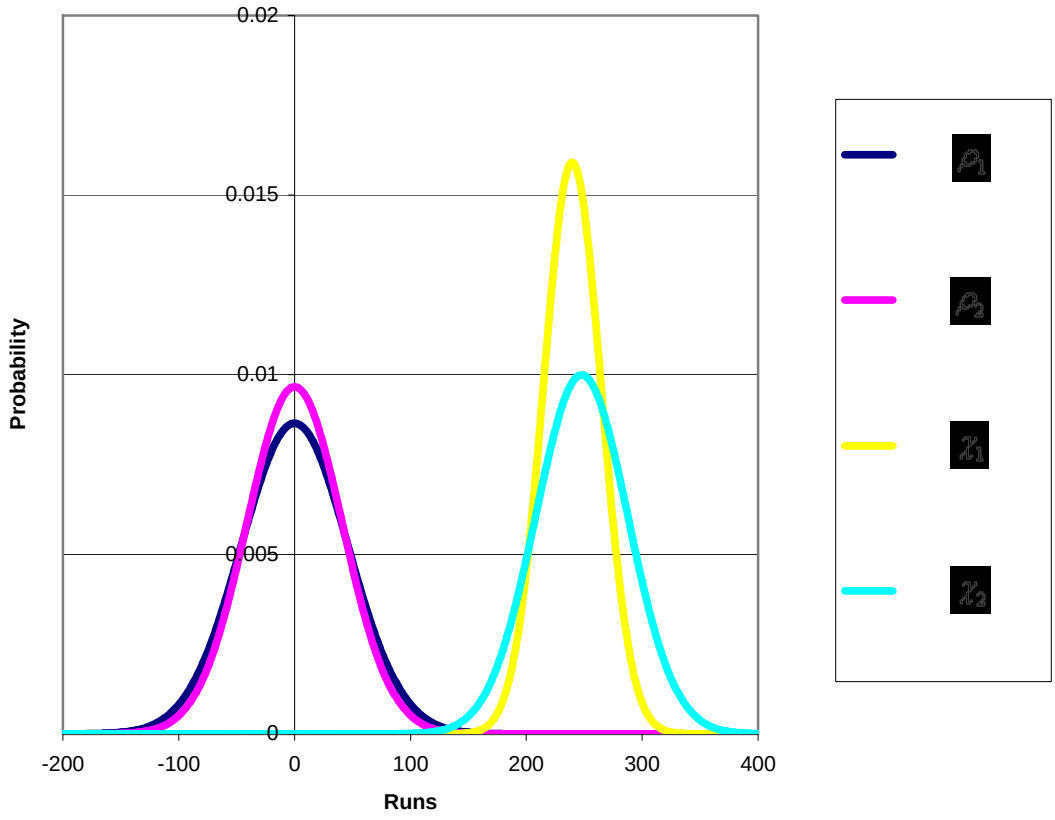
Table 4.2 outlines the variances of performance and conditions and their shares of the total first innings variance in each era, while Figure 4.3 displays a graphical representation of the performance and conditions distributions from each era.

[5] This enables us to increase the variance of performance and decrease the variance of conditions by about 50 each iteration. The complex nature of the simulation combined with the processing power and time available parameters of the experiment led us to select this value. Note that without vastly increasing estimation time, using a higher number of splits/iterations led to us being unable to select a single best iteration with confidence as resulting second innings variances were no longer ordinal in conditions variance.

Table 4.2: Splitting the first innings variance

Era (j)	var(S_j)	var(ρ_j)	var(χ_j)	var(ρ_j) share	var(χ_j) share
1	3144.03	2515.22	628.81	80.00%	20.00%
2	3300.87	1705.45	1595.42	51.67%	48.33%

Figure 4.3: Graphing the performance and conditions variables



4.3 Establishing the conditional distributions using Bayes' Theorem

We now have the distribution functions from which conditions and performance are drawn. We can use these functions and conditional probability rules to determine the distribution from which conditions are drawn for an individual game. We need to calculate the probability of obtaining each value for conditions, given the first innings

score and the result of the game. First we provide a reminder of our conditional density function for conditions, updated to take into account the two eras.

$$f(\chi \mid S = S_{ij}, \omega = \omega_{ij}) = \frac{f(\chi) g(S = S_{ij} \mid \chi = \chi_{ij}) \Pr((\omega = \omega_{ij} \mid S = S_{ij}) \mid \chi = \chi_{ij})}{h(S = S_{ij}, \omega = \omega_{ij})} \quad (4)$$

$$f(\chi) \quad (5)$$

and

$$g(S = S_{ij} \mid \chi = \chi_{ij}) \; = \; k(S_{ij} - \chi_{ij}) \quad (6)$$

can be directly estimated from the conditions distribution and performance distribution respectively. $h(S = S_{ij}, \omega = \omega_{ij})$ can be expanded as follows:

$$h(S = S_{ij}, \omega = \omega_{ij}) = g(S = S_{ij}) \Pr(\omega = \omega_{ij} \mid S = S_{ij}) \quad (7)$$

and $\Pr(\omega = \omega_{ij} \mid S = S_{ij})$ can be directly estimated from the second innings probability of winning distribution. This leaves us with $\Pr((\omega = \omega_{ij} \mid S = S_{ij}) \mid \chi = \chi_{ij})$ still to estimate. Since $S = \rho + \chi$, we can write

$$\Pr((\omega = \omega_{ij} \mid S = S_{ij}) \mid \chi = \chi_{ij}) = \Pr(\omega = \omega_{ij} \mid \rho = \rho_{ij}).$$

The second innings performance distribution is the same as the first innings performance distribution and there exists a second innings advantage of three runs in Era 1 and 13 runs in Era 2. We can then write

$$\Pr(\omega = 1 \mid \rho = \rho_{ij}) = \begin{Bmatrix} \int_{-\infty}^{\rho_{ij}-3} k(\eta)d\eta, & Era1 \\ \int_{-\infty}^{\rho_{ij}-13} k(\eta)d\eta, & Era2 \end{Bmatrix} \tag{8}$$

$$\Pr(\omega = 0 \mid \rho = \rho_{ij}) = 1 - \Pr(\omega = 1 \mid \rho = \rho_{ij})$$

Now we have estimated all the components of equation (4). We simply substitute in equations (2), (3), (4) and (5) to determine the probability of a certain value of conditions given the first innings score and the outcome of the game, two variables that are observable in the data set.

4.4 Selected Results

We plot selected examples of the conditional distributions of conditions given score and result in Figures 4.4 to 4.6. In our three examples, we show the effect of a different first innings score, a different game result and a different set of game rules.

Figure 4.4: Conditional pitch distribution for *j=1*, ω_{ij} =1

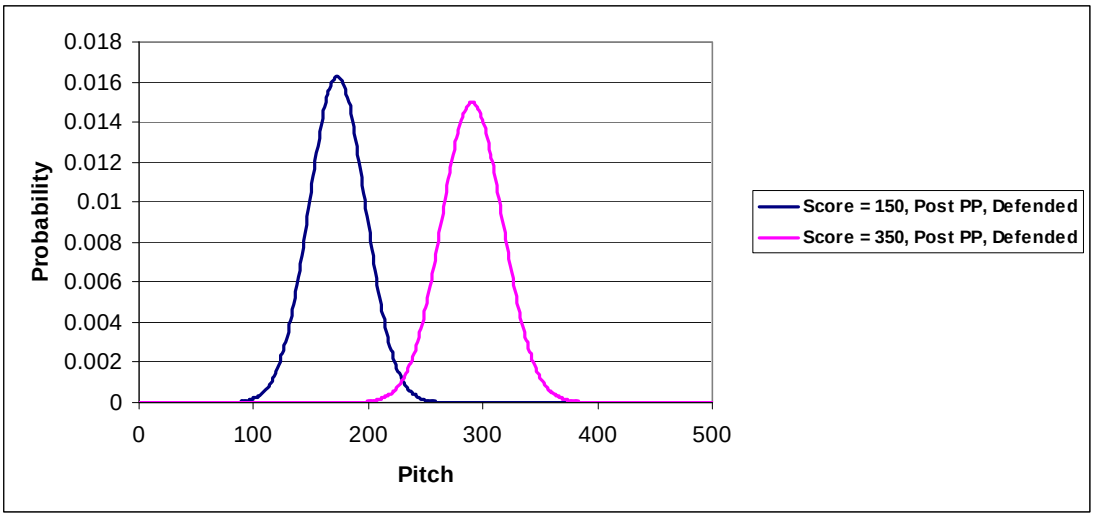


Figure 4.5: Conditional pitch distribution for *j=1*, S =250

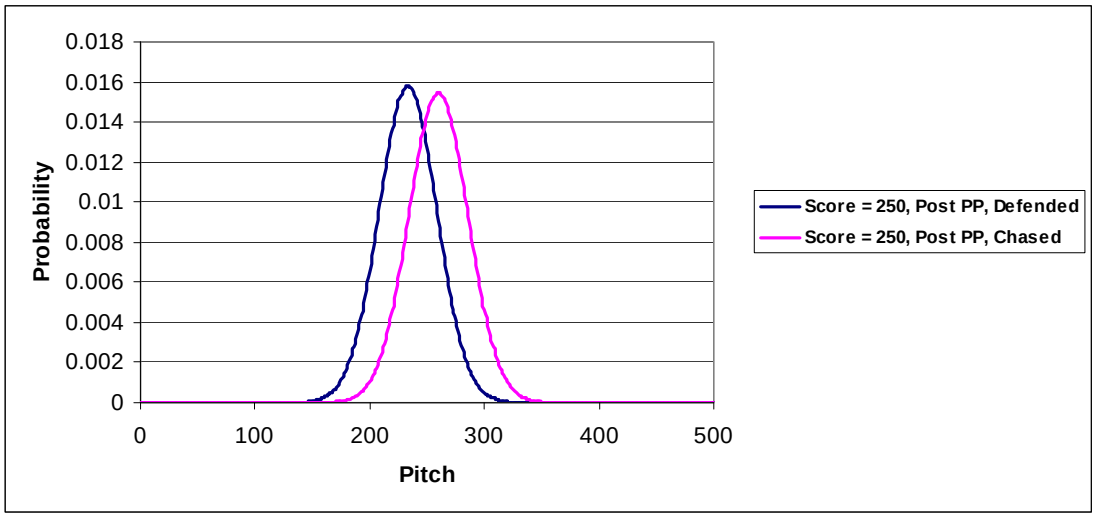
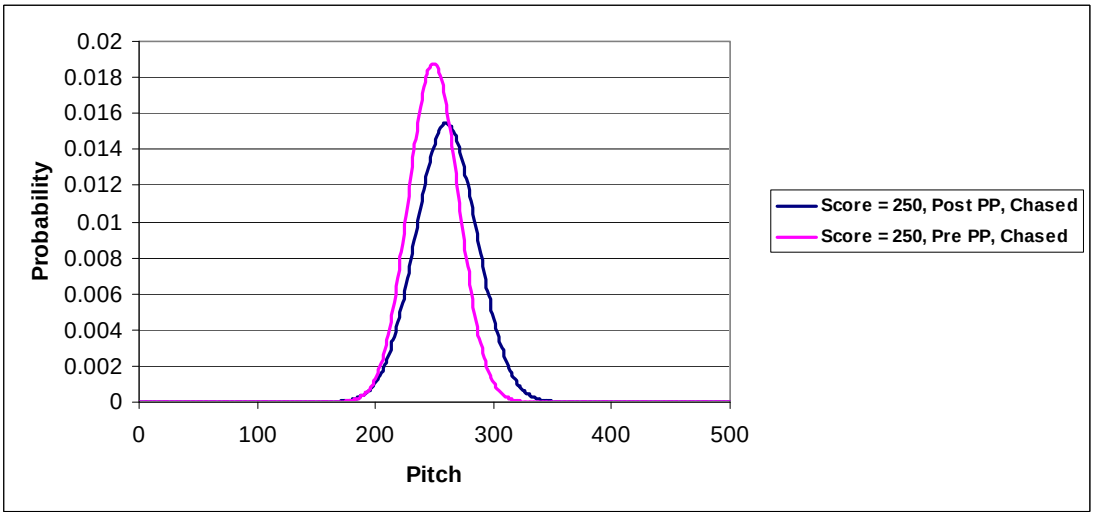


Figure 4.6: Conditional pitch distribution for ω =0, S =250



5 Assessing the fit of the conditional distributions to the data

Theoretically, matches played in conditions with a particular value should result in an average first innings score of that value. We test the accuracy of our distributions by employing several Monte Carlo simulations. Initially we test our method by randomly

drawing one value from the distribution of χ_j and two values from the distribution of ρ_j. We add the first draw of ρ_j to χ_j in order to determine a first innings score, S_j, which we round to the nearest integer. If the first draw of ρ_j is greater than the second, this is a win to the team batting first, otherwise it is a loss. We obtain 10000 scores and results by repeating these steps. We then can apply the appropriate conditional distribution for conditions to each game and we draw 5000 conditions values from this distribution, again rounding to the nearest integer. This gives us a generated data set with 50000000 observations of score and conditions and we can subsequently determine the average score achieved for each (rounded) value of conditions. We plot the results in Figure 5.1 and figure 5.2 below, showing the 2.5th and 97.5th percentiles of the overall conditions distributions to show the conditions that are most likely to be experienced.

Figure 5.1: Average Score in generated data set, Era 1

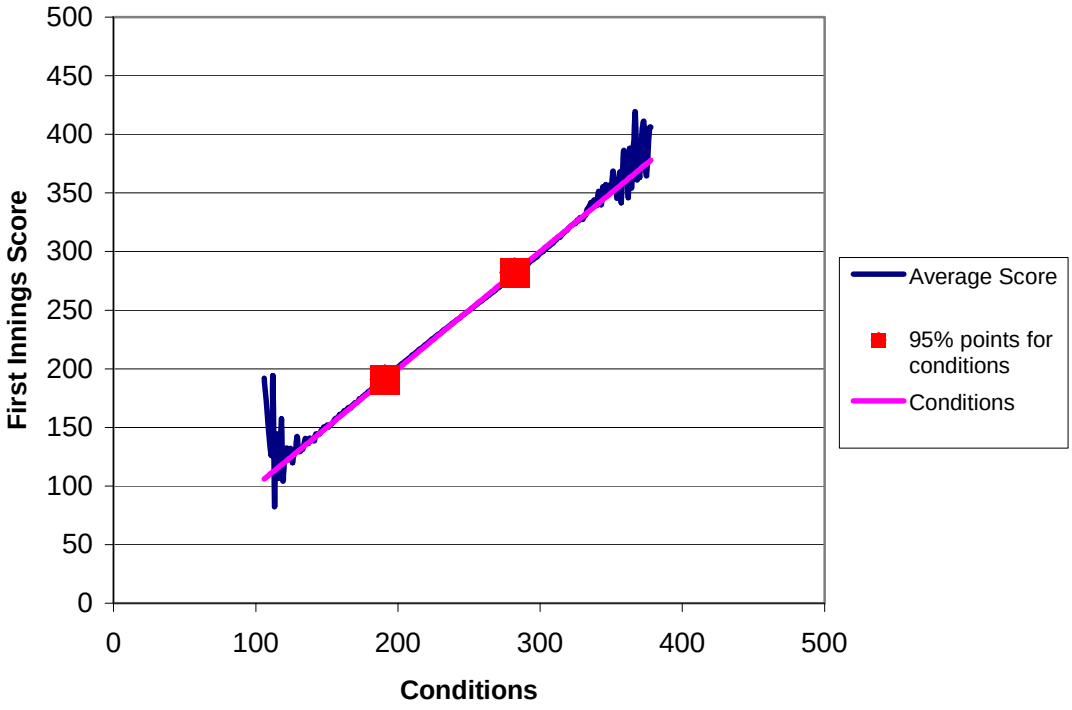
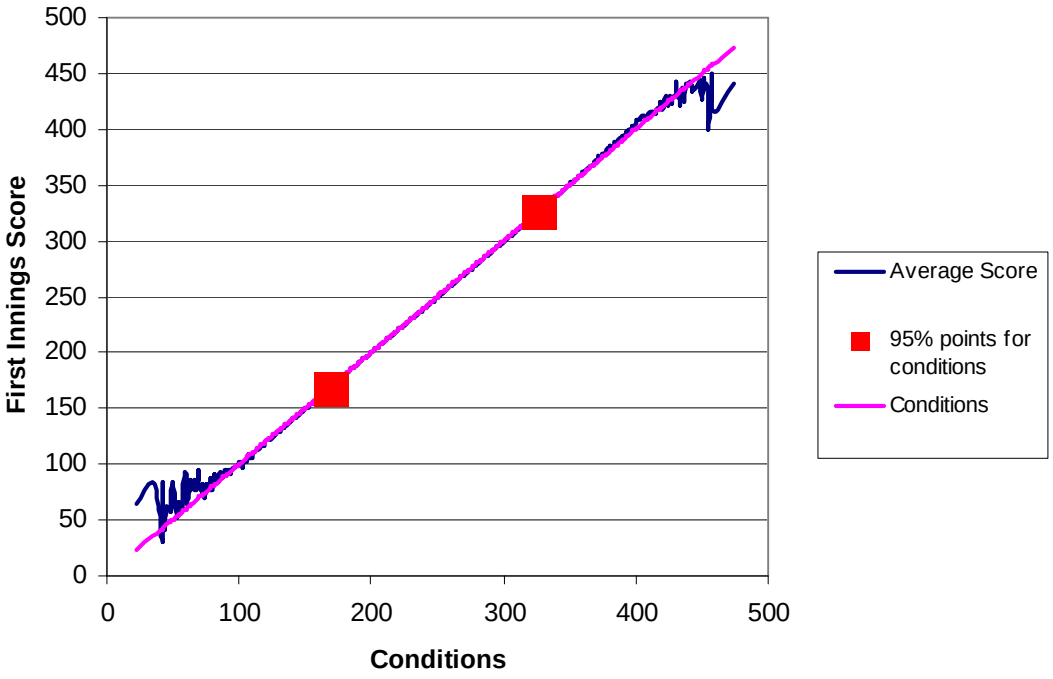


Figure 5.2: Average Score in generated data set, Era 2



It is clearly apparent that the average first innings score in a given set of conditions closely approximates the value of those conditions. We have, to this point, simply tested whether our method works in theoretical games and we need to check whether we can apply it accurately to our actual data sets of matches. As before, we generate 5000 values for conditions from the conditional distribution for each match. There are four graphs showing the results; Figures 5.3 and 5.4 use the matches in our wider data set, while Figures 5.5 and 5.6 use the matches for which we have ball-by-ball data. The latter set provides the most stringent test as the ball-by-ball data differs from the data that was used to create the conditional distributions, as it is a subset of this data.

Figure 5.3: Average Score in wider data set, Era 1

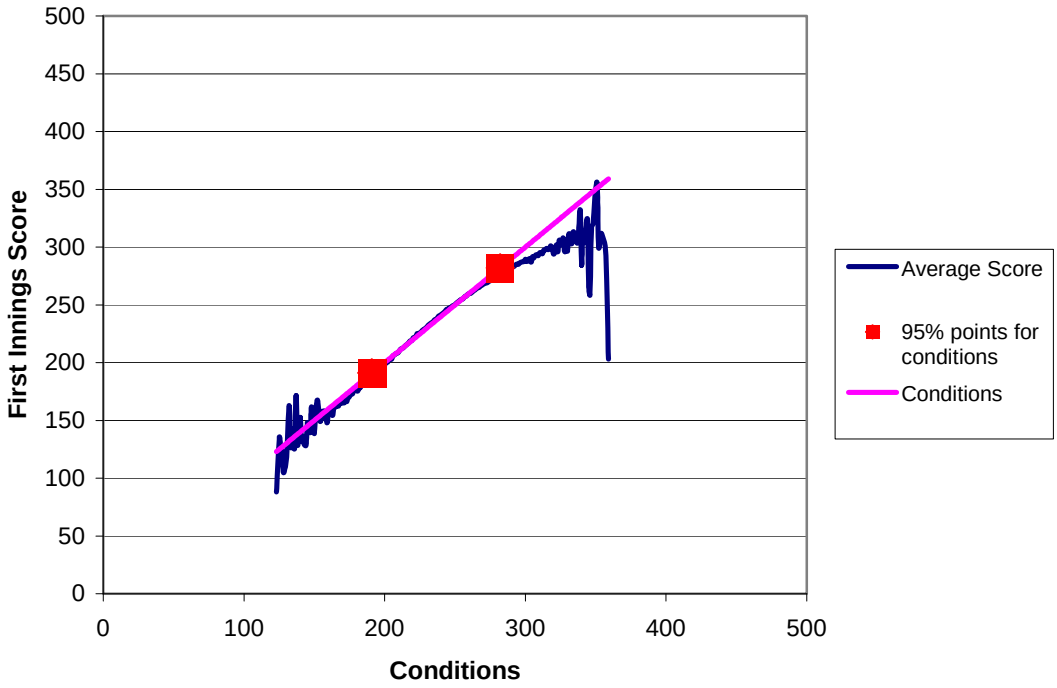


Figure 5.4: Average Score in wider data set, Era 2

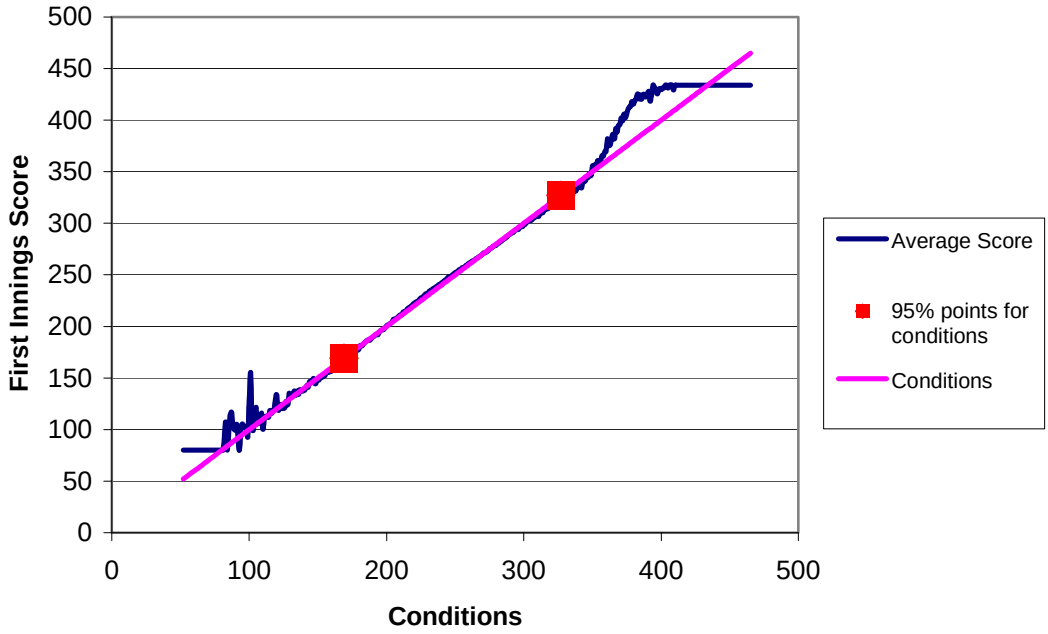


Figure 5.5: Average Score in ball-by-ball data set, Era 1

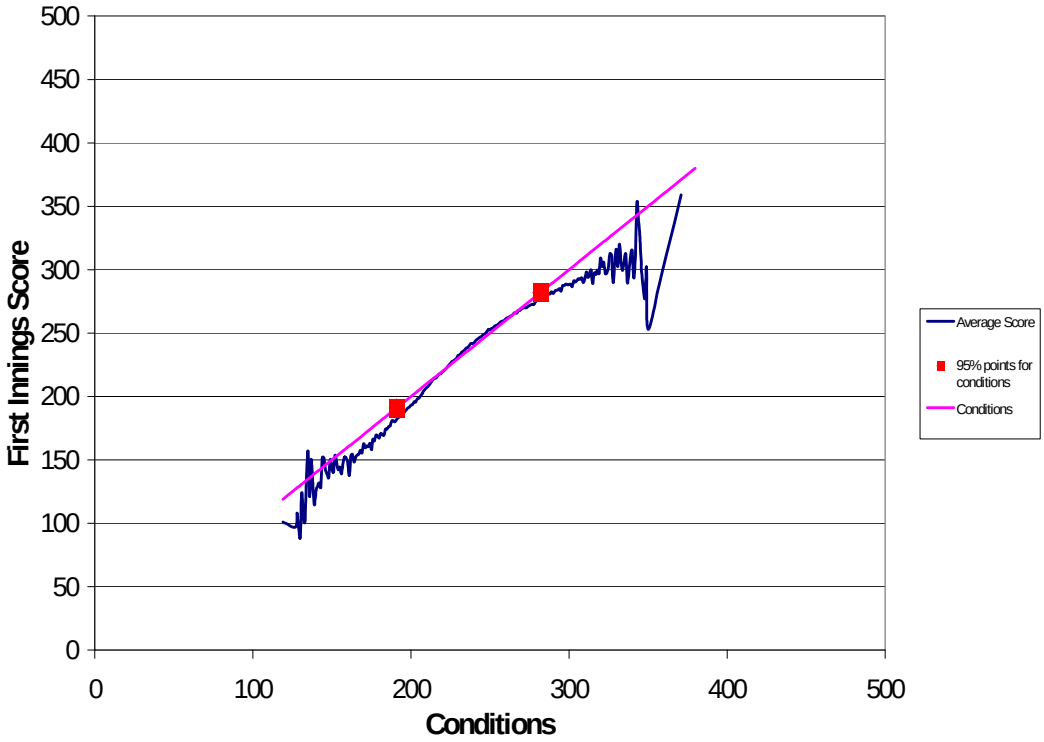
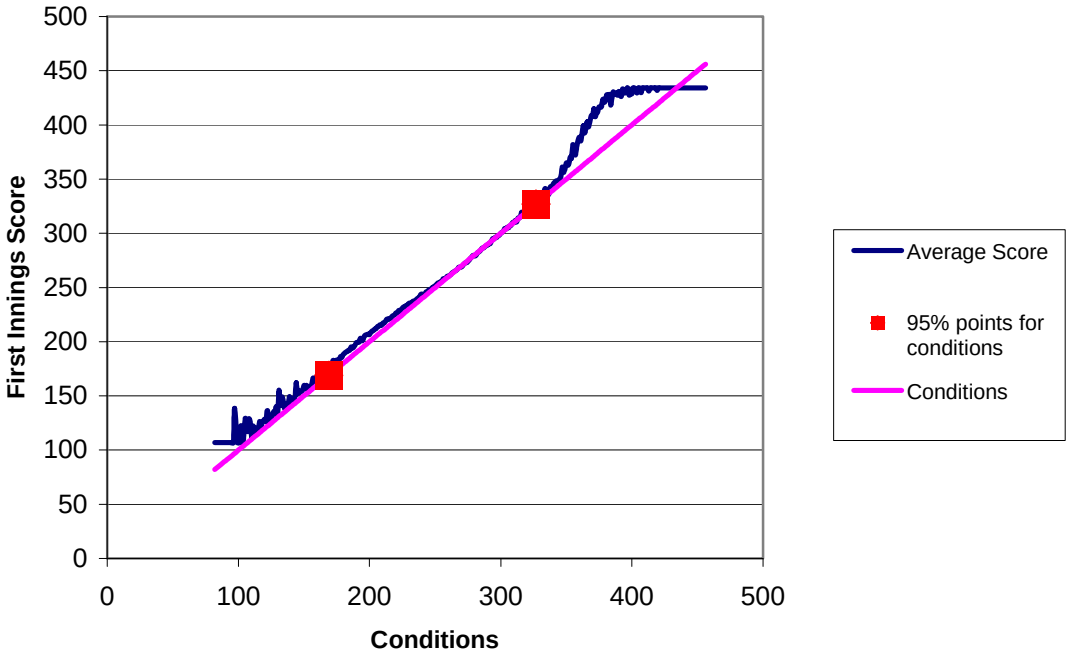


Figure 5.6: Average Score in ball-by-ball data set, Era 2



It is clear that the conditional distributions for conditions are doing a good job of predicting the first innings total. The average score deviates significantly from the conditions value only in conditions that are unlikely to be experienced. The part of the Era 2 graphs where the average score is much higher than the conditions occurs due to the influence of one match between Australia and South Africa on 12 March 2006. In this match, Australia scored 434, the highest first innings score achieved against top eight opposition; remarkably, South Africa won this game to create a record for the highest score successfully achieved to win batting second.

6 Conclusions

By assuming a functional form for a model of first innings score, determining the contribution to the total score variance of each component in the model and applying Bayes' Theorem, we have obtained information pertaining to a critical but unobservable variable. This information is in the form of a distribution that is conditional on the first innings score and the result of the game. In our wider research programme, we are able to randomly draw values from these conditional distributions in order to include conditions as a right-hand-side variable, greatly enhancing the predictive ability of our other models.

Appendix 1: The necessary basics of the game of cricket

Cricket is a sport played between two teams of 11 players on a large, approximately circular field with a 22-yard-long strip of pressed clay, soil and grass known as a "pitch" in the centre. One team will initially be the bowlers and the other team will be the batsmen. All 11 members of the bowling team are on the field while only two members of the batting team are on the field at any one time. The basic idea of the game is relatively simple. A bowler bowls a ball from one end of the pitch by releasing it with a straight arm action in the direction of the batsman. The ball will usually bounce once before reaching the batsman. The two main goals of a batsman are to score "runs" and avoid getting "out". A run is scored each time a batsman, having hit the ball with his bat, running to swap ends of the pitch with the other batsman. Alternatively, a batsman may score an automatic four or six runs by hitting the ball so far that it leaves the playing field. These automatic runs are known as "boundaries", with four being scored if the ball bounces before leaving the playing field and six otherwise. If a batsman is "out" then his turn at batting is over and he must leave the field to be replaced by a team mate.

The batting side may continue batting until ten of the 11 members of their side are out, then the two teams switch roles. A team's turn at batting is called an innings and each team will have either one or two innings depending on the type of game. In general, the team that scores the highest number of runs wins the game.

There are three main versions of the game. In test cricket, the traditional form of the game, each team bats for two innings and a match lasts a maximum of five days, with the match being declared a draw if it is not finished in this time. One Day International (ODI) cricket allows each team to bat for one innings but with a limit of 300 balls per innings. The innings finishes when ten batsmen are out or the 300 balls

are up. As the name suggests, this type of game is all over in a day, running for approximately eight hours. Twenty20 cricket is the newest form of the game and is similar to ODI cricket except that the limit is 120 balls per innings and the game takes approximately three hours. In this paper, we consider only ODI cricket.

Appendix 2: Monte Carlo study design for randomly allocating the matches in each era.

1. Generate a random number for each of the 591 games.
2. Rank the 591 games according to their random number.
3. Assume that the 344 highest-ranked games were played before the power play rules came into effect, regardless of the actual date of the game.
4. Assume that the remaining 247 games were played under the power play rules.
5. Calculate the means and variances of the first innings totals in our two randomly split distributions and store these values.
6. For each randomly split distribution, run a probit regression to determine the probability of winning function.
7. Simulate from each probit regression by generating 10000 random numbers and determining the scores associated with each random number.
8. Calculate the means and variances of the scores generated from each probit and store these values as the means and variances of the second innings distributions.
9. Repeat 10 000 times steps 1 to 8.

Appendix 3: Monte Carlo study design for determining the shares of conditions and performance variance.

1. For the first iteration, choose and $\sigma_\rho^2 = \sigma_S^2$, $\sigma_\chi^2 = 0$
2. Generate 100 000 pairs of random numbers x and y to represent the percentiles of the performance distribution and the conditions distribution respectively. $x, y \sim U(0,1)$
3. Use the inverse normal distribution function to determine the values of ρ_j and χ_j for each x and y. Note that χ_j will equal $\mu_{\chi j}$ in the first iteration as $\sigma_\chi^2 = 0$.
4. Add each pair of ρ_j and χ_j to get the first innings score S_j.
5. Round these values of S_j to the nearest whole number. These 100 000 rounded sums will give us the discrete distribution for S_j, the first innings scores.
6. Group the observations by value of S_j and calculate the mean value of x for each value of S_j. Denote this number $\bar{x}_j(S_j)$ and store it for each S_j. This tells us the mean percentile of the performance distribution that the teams have to be at to achieve a particular score.
7. Sort the stored values of S_j and $\bar{x}_j(S_j)$ in ascending order of $\bar{x}_j(S_j)$. For a large enough number of simulated pairs of x and y, S_j should also appear in ascending order.
8. Generate 100 000 more random numbers $z \sim U(0,1)$.
9. Compare each z to the stored values of $\bar{x}_j(S_j)$. For each z, store the value of S_j with the highest value of $\bar{x}_j(S_j)$ less than z.

10. Calculate and store the variance of the distribution of S_j created in step 9.

11. Let $\sigma_\rho^2 = \sigma_\rho^2 - \frac{\sigma_S^2}{60}$, σ_χ^2 and $\sigma_\chi^2 = \sigma_\chi^2 + \frac{\sigma_S^2}{60}$.

12. Repeat steps 2 to 11, for 60 iterations.

13. Find the closest value of the stored variances obtained from each iteration of step 10 to the observed variance of the simulated second innings distribution. Store the associated values of σ_ρ^2 and σ_χ^2 as the performance and conditions shares, respectively, of the total variance of the first innings distribution σ_S^2.