

# ISSUES IN CRICKET AND GOLF

Derek R. Bingham<sup>1</sup>, Basil M. de Silva<sup>2</sup> and Tim B. Swartz<sup>3</sup>

## Abstract

This paper considers statistical issues related to cricket and golf data. It is a review of the recent work in cricket by de Silva and Swartz [1] and in golf by Bingham and Swartz [2]. With respect to cricket, it is shown that winning the coin toss at the outset of a match provides no competitive advantage in one-day international cricket matches. It is also estimated that playing on one's home field increases the log-odds of the probability of winning a cricket match by approximately .5. In golf, it is shown that the weaker golfer has an advantage in net medal play according to the United States Golf Association handicap system. An alternative procedure which leads to "fairer" golf competitions is also presented.

## 1. ISSUES IN CRICKET

Over the past 10 years, there have been several papers that have considered various statistical aspects concerning the game of cricket. These include Crowe and Middledorp [3], Kumar [4], Ganesalingam, Kumar and Ganeshanandam [5], Kimber [6], Clarke [7] and Danaher [8].

As presented in de Silva and Swartz [1], we consider one-day international (ODI) matches involving games between the 9 nations belonging to the International Cricket Council (ICC). These games represent the game of cricket played at the highest level with relative stability amongst the teams. We have collected data on the 427 matches played during the 1990's up until the Asia Cup concluding in July 1997. This time period captures the modern game of cricket where the rules have been relatively uniform. It is also the case that recent data is more extensive and reliable. To keep strategies constant, we have limited the data to full 50-over matches and have ignored matches decided by run rates. The data was collected from the comprehensive CricInfo<sup>4</sup> web page.

At the beginning of a match, a coin is tossed and the team that wins the toss is granted the choice of batting first or second. Some people believe that a team should bat first, establish a number of runs and produce a psychological hurdle for the second team to overcome. Others believe that there is an advantage in batting second as this team knows what score its opponent has produced. This additional information allows the team batting second to adjust their strategy accordingly. Still others feel that the

---

<sup>1</sup> Department of Mathematics, and Statistics, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6

<sup>2</sup> Department of Statistics and Operations Research, RMIT University, GPO Box 2476V, Melbourne, Victoria, Australia 3001

<sup>3</sup> Department of Mathematics, and Statistics, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6

<sup>4</sup> see [www.cricket.org](http://www.cricket.org)

choice between batting first or second should depend on auxiliary and subjective variables such as the weather, the pitch (ie. field) conditions, the team's health, the team's morale, the opponent, whether the team will bat in daylight or under floodlights, etc. Clearly, this is a topic of considerable interest.

As a preliminary study, Table 1 provides summary data on the 427 ODI matches involving the 9 ICC nations. We see from column  $B_1$  that there is great disparity amongst the various teams with respect to their decision to bat either first or second. For example, upon winning the coin toss, Australia chooses to bat first 87% of the time whereas Sri Lanka chooses to bat first only 36% of the time.

Consider then the data  $(x_i, y_i)$ ,  $i = 1, \dots, n$  where  $n$  is the number of games played by the team of interest,  $x_i = 1(0)$  if the team wins(loses) the coin toss in the  $i^{\text{th}}$  game and  $y_i = 1(0)$  if the team wins(loses) the game. We have the statistical model  $y_i | x_i = 1 \sim \text{Bernoulli}(p_i)$  and  $y_i | x_i = 0 \sim \text{Bernoulli}(q_i)$  where  $P(x_i = 1) = P(x_i = 0) = 1/2$  for  $i = 1, \dots, n$ . We are therefore interested in comparing the strategy  $\sum_{i=1}^n p_i$  versus the strategy  $\sum_{i=1}^n q_i$ . Of the 427 matches in the data set, 8 games resulted in ties. We exclude these matches from the analysis.

**Table 1: Summary data for the 9 ICC nations. Here  $B_1$  is the proportion of time that a team chooses to bat first upon winning the coin toss,  $W_0$  is the overall winning proportion and  $W_h$  is the winning proportion in games played on a home field. The quantities in parentheses are the number of cases.**

Nation	$B_1$	$W_0$	$W_h$
Australia	.87 (68)	.63 (127)	.67 (69)
England	.65 (20)	.36 (45)	.75 (4)
India	.56 (59)	.49 (106)	.72 (39)
New Zealand	.55 (51)	.37 (99)	.50 (44)
Pakistan	.47 (66)	.57 (131)	.60 (20)
South Africa	.69 (42)	.61 (94)	.71 (42)
Sri Lanka	.36 (56)	.46 (100)	.71 (21)
West Indies	.36 (39)	.53 (90)	.57 (21)
Zimbabwe	.58 (26)	.17 (46)	.36 (11)

Our first analysis requires the assumption that  $p_i = p$  and  $q_i = q$  for all  $i = 1, \dots, n$ . This simplistic assumption is clearly unrealistic as it assumes that the opponents are all of equal strength and that the playing conditions are constant over time. However, it is a good starting point and the test of  $H_0: p \leq q$  versus  $H_1: p > q$  is easily carried out using a two-sample Binomial test. The first binomial variable is the number of wins having won the coin toss and the second binomial variable is the number of wins having lost the coin toss. The p-values are .53, .53, .21, .97, .59, .85, .46, .23 and .50 for each of the ICC nations as listed in alphabetical order. These p-values are very high; to be significant, p-values must be small (typically  $< .05$ ). Therefore, using this method, we observe no evidence of successful strategies for any of the 9 ICC teams. In de Silva and Swartz [1], 3 additional analyses based on weaker underlying

assumptions each provide results in the same general direction. We therefore conclude that winning the coin toss has no impact on the outcome of ODI cricket matches.

We now turn to the existence of the home team advantage in ODI cricket matches. From columns ,  $W_h$  and  $W_0$  of Table 1, we see that every ICC nation has a higher winning percentage during home games. Using the sign test, this is convincing evidence of the existence of a home team advantage (ie. p-value =  $1/2^9 \approx .002$ ).

To investigate the effect of the home team advantage, we modify our notation and let  $p_{ijk}$  be the probability that team  $i$  defeats team  $j$  at site  $k$  where  $i, j, k = 1, \dots, 9$ , and, in addition,  $k = 0$  denotes a non-ICC site. We introduce the model  $\text{logit}(p_{ijk}) = \tau_i - \tau_j + \gamma_{ijk}$  where  $\sum_{i=1}^9 \tau_i = 0$  and

$$\gamma_{ijk} = \begin{cases} \gamma & \text{if team } i \text{ is the home team} \\ 0 & \text{if the game is played on a neutral site} \\ -\gamma & \text{if team } j \text{ is the home team} \end{cases}$$

This is a 9-parameter model where  $\tau_i$  is a measure of the differential strength of team  $i$ . Therefore, the offset  $\tau_i - \tau_j$  represents the advantage in log-odds that team  $i$  has over team  $j$ . The model also assumes that the home team advantage  $\gamma$  is constant over all ICC teams. Note that the logit transformation of  $p_{ijk}$  is natural in two respects. Firstly, for teams of equal strength that play on a neutral site, we have  $\text{logit}(p_{ijk}) = 0$  which implies  $p_{ijk} = .5$ . Therefore, there is no need for an intercept term in the model. Secondly, it is sensible to quantify the home team advantage on the log-odds scale since we should expect small relative improvements for strong teams that win most of their games. Conversely, we should expect large relative improvements for weak teams that lose most of their games.

We again exclude the 8 tied games from the 427 matches and fit the model using logistic regression. We obtain  $\hat{\gamma} = .53$  with standard error .14. To put this quantity in perspective, a team with a winning percentage of 50% would increase its winning percentage to 63% when playing at home. Therefore, playing on one's home field provides a considerable edge to the home team.

## 2. ISSUES IN GOLF

Previous studies on handicapping in golf (eg. Scheid [9] and Pollock [10]) have established that it is the better golfer who has an advantage in net matches between 2 golfers. This sentiment is also echoed by the United States Golf Association (USGA) where they state in section 10-2 of the USGA Handicap Formual manual<sup>5</sup>, “As your Handicap Index improves (gets lower), you have a slightly better chance of placing high or winning a handicap event”. As presented in Bingham and Swartz [2], we consider net medal play between 2 golfers when they are both playing well. From a tournament perspective, this is a most practical question. For a golfer does not expect to win a prize when he plays poorly, but when he plays well, at the very least, he expects a fair chance of winning a prize. Therefore, statistically, our problem reduces to looking at the tails of distributions used to model golf scores.

Data were collected from the computer handicap system at the Pemberton Valley Golf and Country Club, Pemberton, British Columbia, Canada during the 1997 golf season. To keep conditions as constant as possible, only rounds played by male members at Pemberton Valley were considered. We limit our analysis to the 49 male members who had completed 40 or more rounds during the year. We use the first 20 rounds as a tuneup period to allow the golfer to reach “mid-season” form. We also restrict our study to the immediate 20 rounds following the tuneup period. We hope that by using a shortened period, golfers will not experience dramatic changes in their skill levels. Each golfer will also have completed the same number of rounds of golf. Therefore our data analysis is based on  $49(20) = 980$  scores.

For each golfer, we choose their best  $m$  net scores amongst the 20 rounds immediately following the initial tuneup period. Here, the net scores are obtained by subtracting a golfer's handicap as determined by the USGA's slope system from the golfer's gross (i.e. actual) score. With  $m$  scores for each golfer, there are  $\binom{49}{2} m^2$  possible matches between 2 golfers that can be simulated. The matches are simulated in the sense that the 2 golfers have not directly competed against one another. We consider  $m = 2, 3, 4$  as this represents the best 10%, 15% and 20% of net scores (i.e. occasions when the golfers play well). We exclude from the analysis the 5 pairs of golfers that have the same handicap index. In Table 2, we give the results of the

**Table 2: Simulated matches between 2 golfers based on their best  $m$  out of 20 net scores. The percentages refer to matches won, lost and tied by the lower handicap (i.e. better) golfer.**

$m$	Matches	Wins	Losses	Ties
2	4,684	33.4%	58.2%	8.4%
3	10,539	34.2%	57.4%	8.4%
4	18,736	35.4%	56.0%	8.6%

simulated matches and observe that the weaker golfer enjoys a dramatic advantage when both golfers are playing well. For example, with  $m = 2$  the weaker golfer wins

---

<sup>5</sup> see [www.usga.org/handicap/manual](http://www.usga.org/handicap/manual)

58.2% of the matches. These results may be surprising as they are in the opposite direction of the existing literature.

We now corroborate these empirical findings with theoretical support. Consider then 2 golfers with independent gross scores  $X_1$  and  $X_2$  where  $H_{X_1}$  and  $H_{X_2}$  are the respective handicap strokes determined by the USGA's slope system. Without loss of generality, we assume that  $H_{X_1} < H_{X_2}$  so that  $X_1$  refers to the gross score of the better golfer. Pollock [10] argues that the normal distribution can be used in modelling gross scores for golfers of varying skill levels. Despite obvious deficiencies in the model such as the underlying assumption of independence between golfers and the approximation of a discrete distribution by a continuous distribution, Pollock's normal model provides insight on a number of handicap issues. In this analysis, we further assume that  $X_i \sim \text{Normal}[\mu(H_{X_i}), \sigma^2(H_{X_i})]$  where  $\mu()$  and  $\sigma()$  are increasing functions. Whereas it is obvious that  $\mu$  is an increasing function, plots in Bingham and Swartz [2] provide evidence that  $\sigma$  is also an increasing function.

Our interest lies in the investigation of

$$\begin{aligned} P_k &= \text{Prob}(\text{the better golfer wins} \mid \text{both golfers play well}) \\ &= \text{Prob}(X_1 - H_{X_1} < X_2 - H_{X_2} \mid X_i - \mu(H_{X_i}) - k\sigma(H_{X_i}), i = 1,2) \end{aligned}$$

where  $k > 0$ . Here,  $X_i - H_{X_i}$  represents the net score of golfer  $i$  and we condition on both golfers playing better than  $k$  standard deviations below their average gross score. Using the formula for conditional probability, we have that

$$P_k = \frac{\int_{x_1=-\infty}^{\mu(H_{X_2})-k\sigma(H_{X_2})+H_{X_1}-H_{X_2}} \int_{x_2=x_1-H_{X_1}+H_{X_2}}^{\mu(H_{X_2})-k\sigma(H_{X_2})} f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1}{\int_{x_1=-\infty}^{\mu(H_{X_1})-k\sigma(H_{X_1})} \int_{x_2=-\infty}^{\mu(H_{X_2})-k\sigma(H_{X_2})} f_{X_1}(x_1)f_{X_2}(x_2)dx_2dx_1}$$

for sufficiently large  $k$  where  $f_{X_i}(x_i) = \frac{1}{\sigma(H_{X_i})} \phi\left(\frac{x_i - \mu(H_{X_i})}{\sigma(H_{X_i})}\right)$  and  $\phi$  is the density of the standard normal distribution.

Bingham and Swartz [2] show that  $\lim_{k \rightarrow \infty} P_k = 0$ . Therefore, as both golfers play better (i.e.  $k \rightarrow \infty$ ), it becomes impossible for the better golfer to win a match based on net scores. This conclusion is in the same direction as the empirical results.

Mosteller and Youtz [11] considered the scores of professional golfers during the final 2 rounds of PGA tournaments under ideal weather conditions. Under these homogeneous conditions, they found that the scores could be well approximated by a base score plus a Poisson variate. On the other hand, we are faced with heterogeneous conditions (i.e. data involving golfers of varying skill levels playing under various conditions). Furthermore, little is at stake for our golfers and we therefore do not expect their effort to be constant over all rounds. Consequently, we do not expect the Poisson model to provide outstanding fit. Rather, we use it as a rough approximation to reality.

Consider then 2 golfers with independent gross scores  $X_1$  and  $X_2$  where  $H_{X_1}$  and  $H_{X_2}$  are the respective handicap strokes determined by the slope system. Without loss of generality, let  $H_{X_1} < H_{X_2}$ . We then assume that the net score  $X_i - H_{X_i}$  is such that

$$X_i - H_{X_i} = B_i + W_i$$

where  $B_i$  is the constant base score and  $W_i \sim \text{Poisson}(\theta_i)$ ,  $i = 1, 2$ . The base score  $B_i$  is meant to represent the  $i^{\text{th}}$  golfer's idealized or perfect net score.

Under net medal play, let  $P$  be the probability that the better golfer wins when both golfers are playing their very best rounds of the year. Assuming that golfer  $i$  plays  $n_i$  rounds,  $i = 1, 2$ , we have that

$$P = \text{Prob}(X_{1,\min} - H_{X_1} < X_{2,\min} - H_{X_2})$$

where the quantity  $X_{i,\min}$  is the lowest of the  $n_i$  scores corresponding to the random variable  $X_i$ .

Under reasonable conditions, it is shown in Bingham and Swartz [2] that  $P \rightarrow 0$  as  $n_1 \rightarrow \infty$  and  $n_2 \rightarrow \infty$ . This establishes again that the better golfer has no chance of winning when the 2 golfers play their best rounds of golf. Bingham and Swartz [2] go on to estimate the probabilities  $P$  for golfers of various handicaps. This study also confirms that it is the weaker golfer who has the advantage when both golfers are playing well.

In Bingham and Swartz [2], a new net score

$$T^* = \frac{113(X - R) / S - 2.10 - 1.082I}{2.74 + 0.053I}$$

is proposed based on the normal model where  $X$  is the golfers gross score,  $R$  is the course rating,  $S$  is the slope rating and  $I$  is the golfer's handicap index. In Table 3, we repeat the analysis of Table 2 using the new performance measure  $T^*$ . We see that the outcomes of the simulated matches are far more balanced than when using traditional net scores. For example, with  $m = 4$ , the better golfer wins 50.3% of the matches. This is much closer to the idealized value 50% than the value 35.4% which is obtained using traditional net scores.

**Table 3: Simulated matches between 2 golfers based on their best  $m$  out of 20 scores using the statistic  $T^*$ . The percentages refer to matches won, lost and tied by the lower handicap (i.e. better) golfer.**

$m$	Matches	Wins	Losses	Ties
2	4,684	47.9%	52.1%	0.0%
3	10,539	49.0%	50.9%	0.0%
4	18,736	50.3%	49.7%	0.0%

## REFERENCES

- [1] de Silva, B.M. and Swartz, T.B., Winning the coin toss and the home team advantage in one-day international cricket matches, *New Zealand Statistician*, 32, 16-22 (1997).
- [2] Bingham, D.R. and Swartz, T.B., Equitable handicapping in golf, Manuscript, (1998).
- [3] Crowe, S.M. and Middeldorp, J., A comparison of leg before wicket rates between Australians and their visiting teams for test cricket series played in Australia, 1977-94, *The Statistician*, 45, 255-262 (1996).
- [4] Kumar, K., Is cricket really by chance?, in *Mathematics and Computers in Sport*, Bond University, Australia, 69-76 (1996).
- [5] Ganesalingam, S., Kumar, K. and Ganeshanandam, S., A statistical look at cricket data, in *Mathematics and Computers in Sport*, Bond University, Australia, 89-104 (1994).
- [6] Kimber, A., A graphical display for comparing bowlers in cricket, *Teaching Statistics*, 15, 84-86 (1993).
- [7] Clarke, S.R., Consistency in sport - with particular reference to cricket, *NZOR Proceedings*, 27th Annual Conference, Victoria University of Wellington, 30-35 (1991).
- [8] Danaher, P.J., Estimating a cricketer's batting average using the product limit estimator, *New Zealand Statistician*, 24(1), 2-5 (1989).
- [9] Scheid, F., An evaluation of the handicap system of the United States Golf Association, in *Optimal Strategies in Sports*, S.P. Ladany and R.E. Machol, editors, North Holland: Amsterdam, 151-155 (1977).

[10] Pollock, S.M., A model of the USGA handicap system and "fairness" of medal and match play, in *Optimal Strategies in Sports*, S.P. Ladany and R.E. Machol, editors, North Holland: Amsterdam, 141-150 (1977).

[11] Mosteller, F. and Youtz, C., Professional golf scores are Poisson on the final tournament days, in *Proceedings of the Section on Statistics in Sports*, American Statistical Association, 39-51 (1992).