On Determining Probability Forecasts from Betting Odds

Erik Štrumbelj

University of Ljubljana, Faculty of Computer and Information Science
Tržaška 25, 1000 Ljubljana, Slovenia
erik.strumbelj@fri.uni-lj.si
tel: +386-1-4768459

Abstract

We show that probabilities determined from betting odds by using Shin's model are more accurate forecasts than probabilities determined with basic normalization or regression models. This applies to both fixed-odds bookmakers and the world's largest betting exchange. We also provide empirical evidence that some bookmakers are significantly different sources of probabilities in terms of forecasting accuracy and that betting exchange odds are not always the best source. In fact, the use of Shin's model decreases, on average, the advantage of betting exchange odds to the point of reversing the order and significance of some of the differences in accuracy. Therefore, previous findings in favor of betting exchanges that did not rely on Shin's model to determine probabilities from bookmakers should be revisited.

Keywords: Sports forecasting, Probability forecasting, Fixed-odds, Betting exchange, Shin's model, Betfair, Calibration

1. Introduction

The scientific literature has been interested in the accuracy of betting odds-based probability forecasts both directly, comparing them to other sources of probability forecasts, and indirectly, through their use in betting strategies and as explanatory variables in statistical models. Probabilities from betting odds are also used as a basis for other economics research, such as betting market efficiency and competitive balance of sports competitions. We do not give a detailed review of all the uses of betting odds. For a historical survey, we refer the reader to Stekler et al. (2010), Vaughan Williams (2005), and Humphreys and Watanabe (2012).

The widespread use of betting odds is not surprising as there is substantial empirical evidence that betting odds are the most accurate publicly available source of probability forecasts for sports. With the growth of online betting, betting odds are also readily available for an increasing number and variety of sports competitions. However, we believe that there are issues with using betting odds as probability forecasts that have not been sufficiently addressed:

(a) does it make a difference, which bookmaker or betting exchange we choose, when two or more are available, and, more importantly

(b) which method should be used to determine probability forecasts from raw betting odds?

In this paper we address these two issues in the context of fixed-odds betting, with emphasis on evaluating different methods for determining probability forecasts from odds. Empirical evaluation is performed using data from several different online bookmakers across 37 competitions in 5 different team sports (basketball, handball, ice hockey, soccer, and volleyball).

1.1. Relevant related work

As a matter of brevity and convenience, we focus on the most relevant results for fixed-odds betting, which is prevalent in team sports[1].

Empirical evidence suggests that betting odds are the most accurate source of sports forecasts. Odds-based probability forecasts have been shown to be better or at least as good as statistical models using sports-related input variables (Forrest et al., 2005; Song et al., 2007; Štrumbelj and Vračar, 2012), expert tipsters (Song et al., 2007; Spann and Skiera, 2009), and (aggregated) lay predictions (Pachur and Biele, 2007; Scheibehenne and Broder, 2007).

A special subset of betting odds are odds from betting exchanges. Unlike fixed-odds, which are formed by bookmakers, betting exchange odds are formed by bettors. That is, betting exchanges facilitate both backing and laying bets and can be considered a form of prediction market.

In many different domains, forecasts from prediction markets are more accurate than those produced by traditional forecasting approaches and single forecasters (Arrow et al., 2008; Graefe and Armstrong, 2011; Tziralis and Tatsiopoulos, 2007). In sports forecasting, the term betting exchange in most cases means Betfair, the world's largest betting exchange. There is substantial empirical evidence that probabilities determined from Betfair

[1] We omit a substantial subset of literature on racetrack betting that focuses primarily on parimutuel markets and their efficiency (see Hausch et al. (2008) for a review).

odds are more accurate forecasts than those from fixed-odds bookmakers (Franck et al., 2010; Smith et al., 2009; Spann and Skiera, 2009; Štrumbelj and Vračar, 2012). This suggests that Betfair odds should be the preferred source of probabilities, when available.

Little work has been done in comparing different fixed-odds bookmakers. An exception is the work by Štrumbelj and Robnik-Šikonja (2010) who showed that there are significant differences between online fixed-odds bookmakers in terms of the accuracy of the forecasts determined from their odds. They did not include Betfair or any other betting exchange in their analysis.

In the next section, we provide the necessary background for the focus of this paper - methods for determining probabilities from betting odds. In Section 3 we describe our experiments and provide empirical evidence in favor of using Shin's model, after which we explore if any bookmaker (or Betfair) should be preferred when selecting a source of probability forecasts. Section 4 concludes the paper.

2. Determining outcome probabilities from betting odds

To make a profit, bookmakers set unfair odds. That is, the sum of inverse odds (booksum) is greater than 1. Therefore, inverse odds can not be directly interpreted as probabilities. In order to use betting odds as probability forecasts, we have to normalize the inverse odds or apply some other method of transforming them into probabilities. The same applies to betting odds from betting exchanges, although the betting exchange booksum is, on average, lower than fixed-odds bookmakers' booksums[2].

Most of the studies mentioned in the introduction use basic normalization (dividing the inverse odds by their sum). In fact, this approach has become almost synonymous with the use of betting odds, although it is not clear if bookmakers indeed add their take proportionately across all possible outcomes. The widespread use of basic normalization can only be attributed to its simplicity.

Alternatively, we can view the outcome as a categorical variable and model the probabilities using a historical data set of betting odds and corresponding match outcomes (see, for example, Forrest and Simmons (2002),Forrest et al. (2005),Goddard et al. (2005)). Due to the nature of the dependant variable, logistic (probit) regression or multinomial regression is used, depending on the number of outcomes. An ordered model is preferred if there is a natural order to the outcomes.

One of a few notable exceptions is the work by Smith et al. (2009) who used a theoretical model of how bookmakers set their odds proposed by Shin (1993). Shin's model can be used to reverse-engineer the bookmaker's underlying probabilistic beliefs from the quoted betting odds. An earlier use of Shin's model are the works by Cain et al. (see Cain et al. (2002, 2003); Smith et al. (2009) and references therein). They show that Shin's model-based approach improves on basic normalization. We adopt their term *Shin probabilities* for probabilities determined from betting odds by using Shin's model.

[2] Betting exchanges profit from taking a commission out of winning bets. This commission is small relative to average fixed-odds bookmaker take.

Surprisingly, Shin probabilities have not been widely adopted and what little use they have seen has focused almost exclusively on racetrack betting. A logical question that follows is, can normalization based on Shin's model improve on basic normalization in individual and team sports.

First, we review basic normalization, Shin's model, and the regression model-based approach.

2.1. Basic normalization

Let $\mathbf{o} = (o_1, o_2, ..., o_n)$ be the quoted decimal odds for a match with $n \geq 2$ possible outcomes and let $o_i > 1$ for all $i = 1..n$. The inverse odds $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_n)$, where

$$\pi_i = \frac{1}{o_i},$$

can be used as latent team strength variables, but do not represent probabilities, because they sum up to more than 1. The sum

$$\Pi = \sum_{i=1}^{n} \pi_i$$

is the booksum and the amount of 'excess' probability $\Pi - 1$ is called the bookmaker take or bookmaker margin.

Dividing by the booksum

$$p_i = \frac{\pi_i}{\Pi}$$

we obtain a set of values that sum up to 1 and can be interpreted as outcome probabilities. We refer to this as *basic normalization.*

2.2. Shin's model

Shin (1991, 1992, 1993) proposed a model based on the assumption that bookmakers quote odds which maximize their expected profit in the presence of uninformed bettors and a known proportion of insider traders.

The bookmaker and the uninformed bettors share the probabilistic beliefs $p = (p_1, p_2, ..., p_n)$, while the insiders are assumed to know the actual outcome before the actual experiment (race, match, etc...).

Without loss of generality, we can assume the total volume of bets is 1 of which $1 - z$ comes from uninformed bettors and z from insiders. Conditional on outcome i occurring, the expected volume bet on the i-th outcome is

$$p_i(1 - z) + z.$$

If the bookmaker quotes $o_i = \frac{1}{\pi_i}$ for outcome i, then the expected liability for that outcome is

$$\frac{1}{\pi_i} \left(p_i(1 - z) + z\right).$$

By assuming that the bookmaker has probabilistic beliefs $\mathbf{p}$, we get the bookmakers unconditional expected liabilities

$$\sum_{i=1}^{n} \frac{p_i}{\pi_i} \left(p_i(1 - z) + z\right)$$

and the total expected profit

$$T(\pi) = 1 - \sum_{i=1}^{n} \frac{p_i}{\pi_i} \left(p_i(1 - z) + z\right).$$

The bookmaker sets $\boldsymbol{\pi}$ to maximize expected profit, subject to $0 \leq \pi_i \leq 1$ and $\Pi < \Pi_{max}$, where Π_{max} is a market-dependent restriction on the booksum in order for the bookmaker to remain competitive. This leads to the solution (see Shin (1993))

$$\pi_i = \sqrt{zp_i + (1-z)p_i^2} \sum_{j=1}^{n} \sqrt{zp_j + (1-z)p_j^2} \tag{1}$$

We have the reverse task of determining the probabilistic beliefs $\mathbf{p}$ given the quoted $\boldsymbol{\pi}$. Jullien and Salanié (1994) showed that Eq. (1) can be inverted to give the *Shin probabilities*

$$p_i = \frac{\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\Pi}} - z}{2(1-z)}. \tag{2}$$

What remains is to compute the proportion of insider trading z. We can use the condition $\sum_{i=1}^{n} p_i = 1$ to obtain

$$z = \frac{\sum_{i=1}^{n} \sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\Pi}} - 2}{n-2}, \tag{3}$$

which can be solved using fixed-point iteration starting at $z_0 = 0$.

In the special case of two possible outcomes (n = 2) Eq. (3) has a tractable analytical solution

$$z = \frac{(\pi_+ - 1)(\pi_-^2 - \pi_+)}{\pi_+(\pi_-^2 - 1)},$$

where $\pi_+ = \pi_1 + \pi_2$ and $\pi_- = \pi_1 - \pi_2$.

2.3. Regression analysis

We can infer the outcome probabilities from historical betting odds and known outcomes. Unlike basic normalization and Shin probabilities, this approach requires a historical set of betting odds and match outcomes.

In this paper, we use two variations of this regression analysis approach. In the case of two possible outcomes (home [H] and away [A] win in basketball and volleyball), we use a logistic regression model to obtain the actual probabilities

$$p_H = \frac{1}{1 + e^{-U}},$$

where the utility function is modeled as

$$U = \beta_0 + \beta_1 \pi_H + \beta_2 \pi_A + \varepsilon.$$

In the case of three possible outcomes (home [H], draw [D], away [A] in handball, ice hockey, and soccer), we use ordered logistic regression

$$\begin{aligned} p_H &= P(\mu_2 < Y) \\ p_D &= P(\mu_1 < Y < \mu_2) \\ p_A &= P(Y < \mu_1) \end{aligned}$$

where the underlying continuous process is modeled as

$$Y = \beta_0 + \beta_1 \pi_H + \beta_2 \pi_D + \beta_3 \pi_A + \varepsilon.$$

The coefficients β and cutoff points μ are parameters to be estimated from historical data.

3. Empirical evidence

We compiled a data set of sports matches with outcomes and betting odds. These include the Betfair betting exchange and the following major online fixed-odds bookmakers: bet.at.home, Bet365, Betclic, bwin, Betway, DOXXbet, Expekt, Interwetten, Jetbull, Ladbrokes, NordicBet, and Unibet. A summary of the sports and competitions is found in Table 1. Note that not every league and/or game is covered by every bookmaker.

3.1. Scoring sports forecasts

Applying different methods for determining probabilities to the same set of betting odds might result in different probabilities. One of the main goals is to verify if they result in significantly different probabilities. More precisely, are there any significant differences in terms of forecasting accuracy. And, if so, which method produces the most accurate forecasts.

First, we evaluate the forecasting accuracy of probabilities using the standard score for categorical forecasts, the Ranked Probability Score (RPS) (Epstein, 1969). We prefer the RPS to the Brier score (Brier, 1950) due to ordinal outcomes in some sports (Home, Draw, Away outcomes in soccer or ice hockey). With ordered outcomes it makes sense to look at the difference between the cumulative probabilities and cumulative actual outcome (see Constantinou and Fenton (2012) for a more detailed argument[3]). Note that for binary forecasts (basketball, volleyball) the RPS simplifies to the Brier score.

[3] Prior to Constantinou and Fenton the RPS had been used in the context of evaluating sports forecasts by Štrumbelj and Robnik-Šikonja (2010).

Sport	Competition	N_{games}
basketball	Greek A1	694
	Italian Lega A	944
	Russian Superleague A	105
	Spanish ACB	884
	Turkish TBL	928
	NBA (USA)	4657
handball	Danish Jack Jones Ligaen	704
	French Division 1	621
	German Bundesliga	1191
	Polish Ekstraklasa	505
	Portugese LPA*	366
	Spanish Liga Asobal	950
hockey	Czech Extraliga	1454
	Finish SM Liga	1648
	German DEL	1561
	Norvegian Eliteserien*	665
	Russian KHL	2556
	Swedish Eliteserien	1320
	Swiss NLA	1200
	NHL (USA)	4899
soccer	Brazilian Campeonato*	1140
	English Championship (2nd tier)	2206
	English League One (3rd tier)	2206
	English League Two (4th tier)	2208
	English premier	1517
	French Ligue 1	1510
	German Bundesliga	1220
	Italian Serie A	1519
	Dutch Erdedivisie	1220
	Scottish Premier League	792
	Spanish Primera Division	1518
	MLS (USA)	1094
volleyball	Belgian League*	251
	French Pro A	631
	German Bundesliga*	396
	Italian Seria A1*	572
	Polish Plusliga*	274
	Total	48126

Table 1: Sports leagues by type. Data from leagues with four seasons worth of data span from 2008/09 to 2011/12. Leagues marked with $*$ have only three seasons and are missing the 2008/09 data. Unless otherwise noted, all leagues are the top tier competitions in their respective countries.

We compare four different methods for determining probabilities from betting odds: basic normalization (*Basic*), Shin probabilities (*Shin*), and two variants of the regression-based approach. Unlike basic normalization and Shin probabilities, regression-based approaches require a historical set of betting odds and actual outcomes to estimate the model, before that model can be used to transform new betting odds into probabilities.

In order to obtain ex-ante forecasts for the regression model-based approaches, probabilities for a match must be determined with a model estimated only on matches that precede it in time. However, this would result in very small samples for the oldest matches and would put the regression-based approach at a disadvantage. Instead, we set aside oldest third of the matches for every bookmaker/league pair and do not use them in the evaluation.

The first variant of the regression-based approach (*Logit*) uses this third of data to estimate a model. The second variant (*LogitR*) updates the model in a rolling manner. That is, for every match, the probabilities are determined using a model estimated on the first third of the data and all matches from the evaluation part of the data that precede the current match. Note that we omit the π_A predictor from all regression models to avoid having linearly dependent predictors.

Results of the comparison of the four different methods are found in Table 2. Shin probabilities have the lowest (best) score in all cases except the median score for hockey, where basic normalization is better. Both regression-based approaches are worse than basic normalization and Shin probabilities for all sports.

We tested the significance of these differences in two steps. First, using

Sport	N		Shin	Basic	LogitR	Logit	Friedman test
basketball	57	mean	0.1791	0.1800	0.1815	0.1829	Q = 61.7, df = 3
		median	0.1784	0.1790	0.1812	0.1800	p = 2.60×10^{-13}
handball	51	mean	0.1631	0.1638	0.1665	0.1653	Q =67.7, df = 3
		median	0.1641	0.1649	0.1653	0.1651	p = 1.32×10^{-14}
hockey	104	mean	0.2175	0.2177	0.2198	0.2184	Q = 54.3, df = 3
		median	0.2249	0.2245	0.2250	0.2248	p = 9.57×10^{-12}
soccer	162	mean	0.2014	0.2016	0.2021	0.2019	Q = 88.3, df = 3
		median	0.2013	0.2013	0.2029	0.2023	p = 5.07×10^{-19}
volleyball	38	mean	0.1743	0.1750	0.1796	0.1791	Q = 53.77, df = 3
		median	0.1836	0.1846	0.1884	0.1866	p = 1.25×10^{-11}

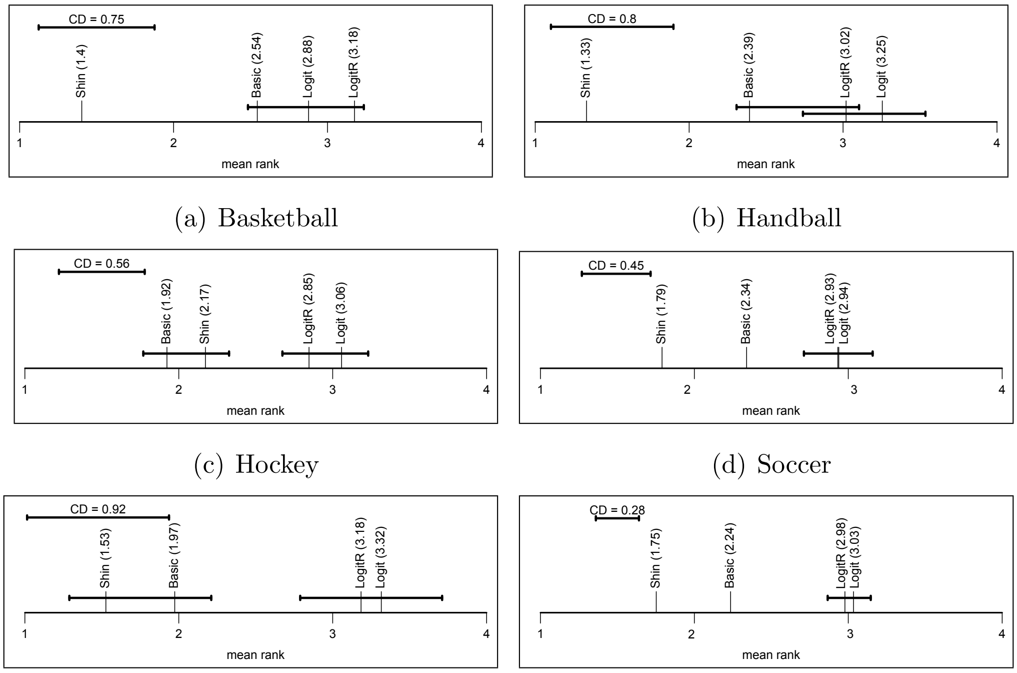
Table 2: Number of bookmaker/league pairs for each sport (N) and mean/median scores by sport and method. Results of the Friedman test for significance of differences between methods' scores are provided for each sport.

the standard non-parametric equivalent to one-way ANOVA - the Friedman test. A non-parametric test is preferred because forecasting scores violate the assumption of normality.

Results of the Frieadman test are shown in Table 2 and reveal that the differences are significant for all sports (all p-values $< 10^{-10}$). That is, for every sport, at least one method produces probabilities that rank, in terms of forecasting accuracy, significantly higher than another method's.

Second, we used the Nemenyi test (see Demšar (2006) for details) to calculate the critical distance (CD) for pairwise post-hoc comparison of all methods for each sport separately and all sports combined. The chosen critical distances control for family-wise error-rate at the 0.01 significance level. Average ranks and results of post-hoc comparison are summarized in Figure 1.

Shin probabilities are significantly better than basic normalization overall and on three sports (basketball, handball, and soccer). Shin probabilities are also better than basic normalization on volleyball and, as noted before, basic normalization produces more accurate forecasts for hockey, but these two differences are not found to be significant. Furthermore, Shin probabilities are in all cases significantly better than both regression-based methods, which are in turn not significantly different from each other on any of the sports.



(a) Basketball

(b) Handball

(c) Hockey

(d) Soccer

(e) Volleyball

(f) All sports

Figure 1: Mean ranks of methods' scores broken down by sport and across all sports. Differences beyond the Nemenyi post-hoc test critical distance (CD) are statistically significant at the 0.01 level. For example, for basketball (a) the average rank of Shin probabilities' RPS scores is more than the critical distance (0.75) better than other methods. Therefore, this difference is statistically significant. The remaining three methods' ranks differ less than the critical distance and are not significant at the 0.01 level.

The advantage of using Shin probabilities instead of basic probabilities does not depend on choice of bookmaker (see Table 3 for all sports combined). In fact, Shin probabilities improved on basic normalization not just for every bookmaker across all sports but for every bookmaker/league pair. The mean (min) improvements are $6.54 \times 10^{-4}(7.1 \times 10^{-5})$ and $5.44 \times 10^{-3}(5.18 \times 10^{-4})$ for the mean and median RPS scores, respectively.

Bookmaker	Shin		Basic	
	mean	median	mean	median
bet.at.home	0.2109	0.1818	0.2114	0.1857
bet365	0.2106	0.1820	0.2110	0.1837
Betclic	0.2110	0.1841	0.2116	0.1882
Betfair	0.2100	0.1810	0.2101	0.1833
Betsafe	0.2107	0.1828	0.2112	0.1861
bwin	0.2109	0.1803	0.2113	0.1833
DOXXbet	0.2107	0.1858	0.2113	0.1895
Expekt	0.2110	0.1837	0.2115	0.1870
Interwetten	0.2113	0.1895	0.2123	0.1963
NordicBet	0.2107	0.1842	0.2112	0.1871

Table 3: Mean and median RPS by bookmaker and method. Results are for matches covered by all listed bookmakers (N = 33981). Ladbrokes, Jetbull, and Betway are excluded from the comparison, due to limited coverage.

As hypothesized and in accordance with related work (Cain et al., 2002, 2003) Shin probabilities are better than basic normalization and regression model-based approaches. It is more surprising that Shin probabilities also improve on basic normalization for Betfair odds. Although the improvement is small compared to fixed-odds bookmaker odds, this suggests that there are at least some commonalities between how bookmakers set odds and how betting exchange odds are formed.

3.2. Tests for calibration and decomposition of forecasts

Calibration plots in Figure 2 provide more insight into sports forecasts. For each sport and outcome we binned the forecasts across all matches and bookmakers into 8 bins of equal size. The bins are based on the quantiles of the inverse betting odds for match and bookmaker. Therefore, for a given sport, the bins are same for all four methods.

Within-bin biases are relatively small, but none of the methods appear to be perfectly calibrated-in-the-small. We can test if these departures from calibration are significant by observing that under the null-hypothesis of perfect calibration the standardized within-bin difference between predicted and observed proportion approximates a standard normal distribution (see, for example, Lahiri and Wang (2013) for details).

Sport	Shin	Basic	LogitR	Logit
soccer	197.1	448.2	194.9	323.7
basketball	108.5	214.0	503.4	546.8
handball	513.8	726.3	345.9	386.9
hockey	352.5	198.2	223.3	484.3
volleyball	62.7	121.6	287.6	376.5

Table 4: Chi-squared values for tests for calibration-in-the-small. All sport, outcome, and method tuples significantly depart from calibration (the 0.99 upper-tail critical values for the chi-squared distributions with 8 and 24 degrees of freedom are approximately 20.1 and 43.0, respectively).

Therefore, under the null-hypothesis of calibration-in-the-small the sum of these statistics across all 8 bins approximates a chi-squared distribution with 8 degrees of freedom and, for sports with 3 outcomes, a chi-squared distribution with 24 degrees of freedom. The resulting chi-squared statistics

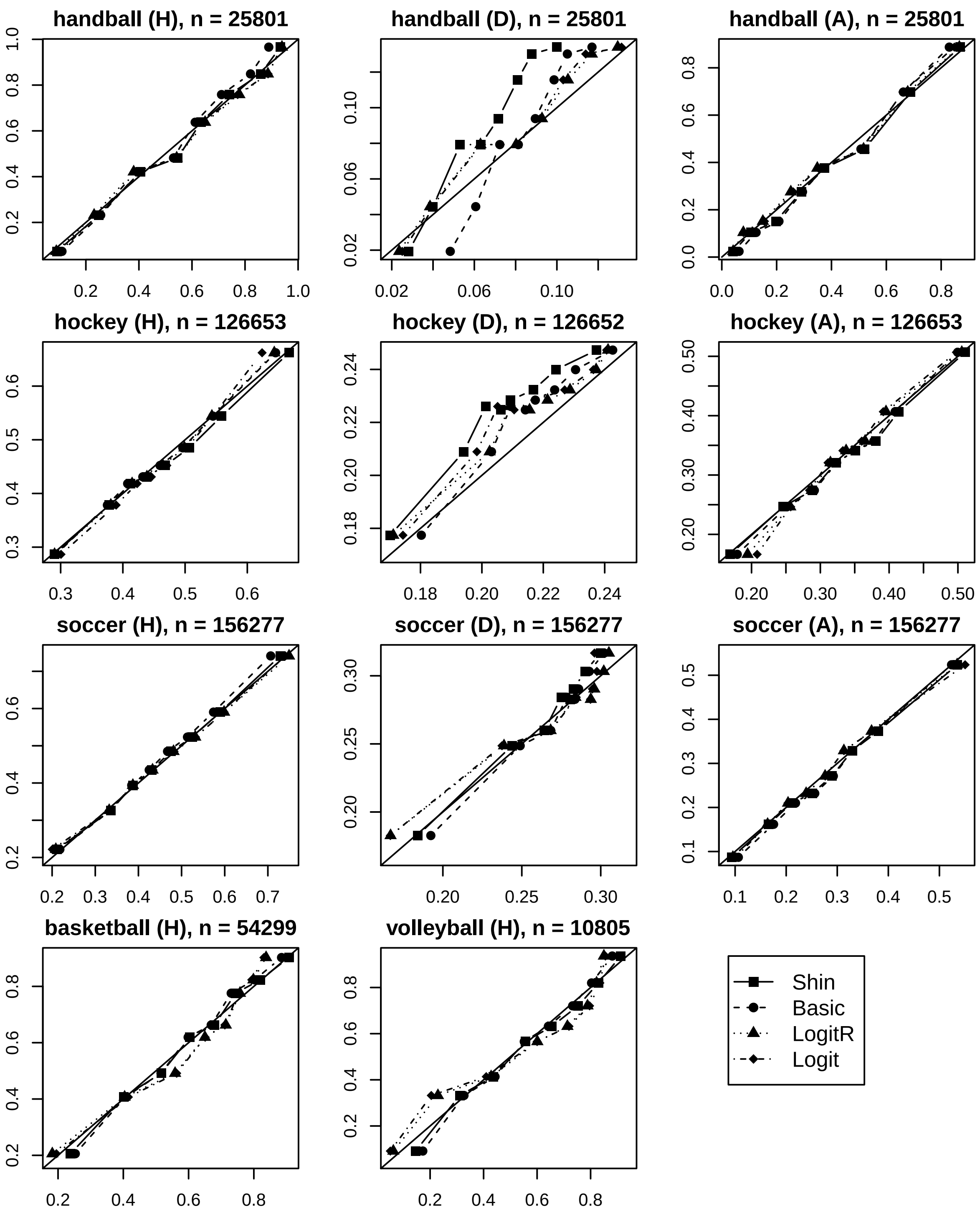


Figure 2: Plots of observed relative frequencies against average forecast.

for each sport and method are shown in Table 4 and confirm that none of the methods are calibrated-in-the-small.

While these departures from calibration are significant due to a large set of data, they are, in practical terms, small. The high uncertainty components given below (Table 5) show that the outright outcome of a sports match is relatively difficulty to forecast. Furthermore, when setting odds, resolution (distinguishing between more/less likely outcomes) is more important than calibration and the observed departures from calibration are well within the margin-of-error allowed for by a $5-10\%$ bookmaker take. As such, it would be reasonable to assume that bookmakers' underlying probabilistic beliefs are not well-calibrated. An alternative explanation is that the departure from calibration is a consequence of the methods for determining probabilities not accurately capturing how bookmakers set the odds. Not knowing how bookmakers determine the odds, we are unable to distinguish between the two.

Further insight into the accuracy of the forecasts provided by decomposing them into calibration and resolution components. For this purpose, we use a generalized decomposition of the Brier score for situations where observations are stratified into bins of probabilities rather than directly on the issued probabilities (Stephenson et al., 2008):

$$\begin{aligned} BS = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj}-o_{kj})^2 &= \frac{1}{n}\sum_{k=1}^{m} n_k(\overline{f}_k-\overline{o}_k)^2 - \frac{1}{n}\sum_{k=1}^{m} n_k(\overline{o}_k-\overline{o})^2 + \overline{o}(1-\overline{o}) \\ &+\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj}-\overline{f}_k)^2 - \frac{2}{n}\sum_{k=1}^{m}\sum_{j=1}^{n_k}(f_{kj}-\overline{f}_k)(o_{kj}-\overline{o}_k), \end{aligned} \tag{4}$$

Sport	Method	REL	GRES	UNC	BRIER
basketball	Shin	3.985E-04	4.831E-02	2.387E-01	0.1908
	Basic	7.429E-04	4.827E-02		0.1912
	LogitR	1.703E-03	4.785E-02		0.1926
	Logit	1.871E-03	4.812E-02		0.1925
handball	Shin	7.865E-04	5.451E-02	1.858E-01	0.1320
	Basic	1.179E-03	5.448E-02		0.1325
	LogitR	6.576E-04	5.324E-02		0.1332
	Logit	6.885E-04	5.270E-02		0.1338
hockey	Shin	1.790E-04	7.675E-03	2.132E-01	0.2057
	Basic	1.041E-04	7.662E-03		0.2057
	LogitR	1.095E-04	7.275E-03		0.2061
	Logit	2.494E-04	6.912E-03		0.2066
soccer	Shin	8.279E-05	1.434E-02	2.140E-01	0.1998
	Basic	1.776E-04	1.432E-02		0.1999
	LogitR	7.497E-05	1.377E-02		0.2003
	Logit	1.259E-04	1.364E-02		0.2005
volleyball	Shin	8.128E-04	6.934E-02	2.462E-01	0.1776
	Basic	1.507E-03	6.931E-02		0.1784
	LogitR	4.181E-03	6.894E-02		0.1814
	Logit	5.106E-03	6.922E-02		0.1820

Table 5: Decomposition of Brier scores of binned forecasts into reliability, generalized resolution, and uncertainty components. The values are averaged across all three outcomes for handball, hockey, and soccer. Best values for each sport are underlined. Note that the uncertainty component $\bar{o}(1-\bar{o})$ in Eq. (4) depends only on the sport and is therefore the same for all methods for a given sport.

where m is the number of bins, n_k the number of forecasts in the $k-$th bin, n the total number of forecasts, f_{kj} the $j-$th forecast in the $k-$th bin and o_{kj} the observed outcome, $\overline{f}_k$ and $\overline{o}_k$ the within-bin averages, and $\overline{o}$ the overall relative frequency.

The five components in the decomposition of Eq. (4) are Reliability (REL), Resolution (RES), Uncertainty (UNC), within-bin variance (WBV), and within-bin correlation (WBC), respectively. The generalized decompositon can be written as BS = REL - GRES + UNC, where GRES = REL - WBV + WBC is the generalized resolution.

The two within-bin terms help compensate the decreasing resolution component when the bin size is increased, making the generalized decomposition less sensitive to the choice of bin size (Stephenson et al., 2008).

In the special case when all forecasts within a bin are the same (that is, when the observations are stratified directly on the issued probabilities) the two within-bin terms vanish and we are left with the standard decomposition of the Brier score into Reliability, Resolution, and Uncertainty components (Murphy, 1973).

The results of the decomposition are shown in Table 5. Shin probabilities have the best Brier score for each sport, which is due to having the best (highest) generalized resolution (GRES). The reliability (calibration) components of the Shin probabilities' Brier scores are best for basketball and volleyball only. Again, this implies that either Shin's model is not accurate model for how bookmakers set odds for matches with draws or that bookmakers find it more difficult to forecast sports with draw outcomes.

3.3. Comparing bookmakers

The data used for the results reported in Table 3 also facilitate comparison of the forecasting accuracy of probabilities determined from different fixed-odds bookmakers and Betfair, while controlling for variance between sports and leagues.

Similar to testing the significance of differences between methods, we first apply the Friedman test and then post-hoc analysis using the Nemenyi test. We omit the details of initial Friedman tests(p values $< 10^{-8}$ for all sports). The post-hoc pairwise comparison results are shown in Figure 3. The procedure was performed for Shin probabilities and basic normalization separately, in order to investigate the potentially different conclusions reached.

First, we discuss the results when Shin probabilities are used to determine probabilities from odds. Overall, Betfair is the best source (see Figure 3(f)), but not found to be significantly better than bet365. With the exception of soccer, the Betfair betting exchange is not the obvious best choice for any of the sports. In fact, bwin and bet365 give significantly more accurate forecasts for handball and volleyball, respectively. The fixed-odds bookmaker Interwetten stands out as the worst source of probability forecasts.

Using basic normalization instead of Shin probabilities changes how bookmakers compare to each other (Figure 3, below the axis in gray color). If basic normalization is used, Betfair stands out as the significantly best source overall and on all five sports (tied with bwin for handball). When interpreting this result, we have to take into account the previous result that Shin probabilities are uniformly better across all bookmakers. That is, this discrepancy is not a result of Shin probabilities being worse than basic normalization for determining probabilities from Betfair odds, but of basic normalization being less appropriate for determining odds from fixed-odds bookmakers.

Therefore, any comparison of probabilities from bookmakers and betting exchanges that relies on basic normalization or regression model-based methods to determine probabilities is unfair to fixed-odds bookmakers might produce misleading results.

Clearly, some bookmakers are, on average, a source of significantly more accurate probability forecasts. In practice, these bookmakers can be identified by measuring their accuracy on historical data. However, it would be more convenient to have for any given sports match a bookmaker selection criterion that does not depend on having such a historical set of data. A natural property to consider is the bookmaker's take or booksum. The reasoning being that a smaller take indicates that the bookmaker competes in a more efficient market, which in turn forces the bookmaker to be a better forecaster.

Figure 4 explores the relationship between bookmakers' takes (booksums) and their forecasting quality. While there is some correlation between the two, booksums do not completely explain why some bookmakers (or betting exchanges) are sources of better forecasts.

4. Conclusion

Empirical evidence suggest that Shin probabilities are, on average, more accurate than probabilities based on basic normalization, which are in turn better than commonly used regression-based approaches. For all methods, the determined probabilities significantly depart from calibration-in-the-small, but the biases are small. Basic normalization might still be preferred in some situations, due to its simplicity. However, most uses of probabilities from betting odds aim at maximizing forecasting accuracy, so Shin probabilities should be used.

Furthermore, the results are in accordance with what has already been suggested, that Shin's model is a more accurate model of how fixed-odds bookmakers set odds. Using basic normalization instead, would lead us to different conclusions when comparing fixed-odds bookmakers with other sources. Based on this, it would be worth revisiting the results from some of the previous studies that relied on basic normalization and concluded that betting exchanges odds are a better source of probability forecasts compared to fixed-odds bookmakers' betting odds.

We found that probability forecasts from Betfair odds are the most accurate for soccer. On two of the remaining four sports (handball, volleyball), we get significantly better results by using one of the two largest online fixed-odds bookmakers (bwin, bet365). We hypothesize that this is due to a much lower volume of bets traded on Betfair on less popular non-soccer sports, while fixed-odds bookmakers still have to offer competitive odds.

We have shown that two of the largest fixed-odds bookmakers (bwin, Bet365) and the world's largest betting exchange Betfair are, on average, sources of better probability forecast, while the fixed-odds bookmaker Interwetten is, on average, the worst. Average booksum (or bookmaker take) partially explains why some bookmakers are sources of better forecasts. For a better understanding of this phenomenon, we require market data beyond just betting odds. We delegate this to further work.

References

Arrow, K.J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J.O., Levmore, S., Litan, R., Milgrom, P., Nelson, F.D., Neumann, G.R., Ottaviani, M., Schelling, T.C., Shiller, R.J., Smith, V.L., Snowberg, E., Sunstein, C.R., Tetlock, P.C., Tetlock, P.E., Varian, H.R., Wolfers, J., Zitzewitz, E., 2008. The promise of prediction markets. Science 320, 877–878.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 75, 1–3.

Cain, M., Law, D., Peel, D., 2002. Is one price enough to value a state-contingent asset correctly? evidence from a gambling market. Applied Financial Economics 12, 33–38.

Cain, M., Law, D., Peel, D., 2003. The favourite-longshot bias, bookmaker margins and insider trading in a variety of betting markets. Bulletin of Economic Research 55, 263–273.

Constantinou, A.C., Fenton, N.E., 2012. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. Journal of Quantitative Analysis in Sports .

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30.

Epstein, E.S., 1969. A scoring system for probability forecast of ranked categories. Journal of Applied Meteorology 8, 985–987.

Forrest, D., Goddard, J., Simmons, R., 2005. Odds-setters as forecasters: The case of English football. International Journal of Forecasting 21, 551 – 564.

Forrest, D., Simmons, R., 2002. Outcome uncertainty and attendance demand in sports: The case of english soccer. Journal of the Royal Statistical Society, Series D (The Statistican) 51, 229–241.

Franck, E.P., Verbeek, E., Nuesch, S., 2010. Prediction accuracy of different market structures – bookmakers versus a betting exchange. International Journal of Forecasting 26, 448–459.

Goddard, J., Beaumont, J., Simmons, R., Forrest, D., 2005. Home advantage and the debate on competitive balance in professional sports leagues. Journal of Sports Sciences 23, 439–445.

Graefe, A., Armstrong, J.S., 2011. Comparing face-to-face meetings, nominal groups, delphi and prediction markets on an estimation task. International Journal of Forecasting 27, 183–195.

Hausch, D.B., Lo, V.S., Ziemba, W.T. (Eds.), 2008. Efficiency of Racetrack Betting Markets. World Scientific.

Humphreys, B.R., Watanabe, N.M., 2012. The Oxford Handbook of Sports Economics : The Economics of Sports Volume 1. Oxford University Press. chapter Competitive Balance.

Jullien, B., Salanié, B., 1994. Measuring the incidence of insider trading: A comment on shin. Economic Journal 104, 1418–19.

Lahiri, K., Wang, J.G., 2013. Evaluating probability forecasts for gdp declines using alternative methodologies. International Journal of Forecasting 29, 175–190.

Murphy, A.H., 1973. A new vector partition of the probability score. Journal of Applied Meteorology 12, 595–600.

Pachur, T., Biele, G., 2007. Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. Acta Psychologica 125, 99–116.

Scheibehenne, B., Broder, A., 2007. Predicting wimbledon 2005 tennis results by mere player name recognition. International Journal of Forecasting 23, 415–426.

Shin, H.S., 1991. Optimal betting odds against insider traders. Economic Journal 101, 1179–85.

Shin, H.S., 1992. Prices of state contingent claims with insider traders, and the favourite-longshot bias. Economic Journal 102, 426–35.

Shin, H.S., 1993. Measuring the incidence of insider trading in a market for state-contingent claims. Economic Journal 103, 1141–53.

Smith, M.A., Paton, D., Williams, L.V., 2009. Do bookmakers possess superior skills to bettors in predicting outcomes? Journal of Economic Behavior & Organization 71, 539 – 549.

Song, C., Boulier, B.L., Stekler, H.O., 2007. The comparative accuracy of judgmental and model forecasts of american football games. International Journal of Forecasting 23, 405–413.

Spann, M., Skiera, B., 2009. Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. Journal of Forecasting 28, 55–72.

Stekler, H., Sendor, D., Verlander, R., 2010. Issues in sports forecasting. International Journal of Forecasting 26, 606 – 621.

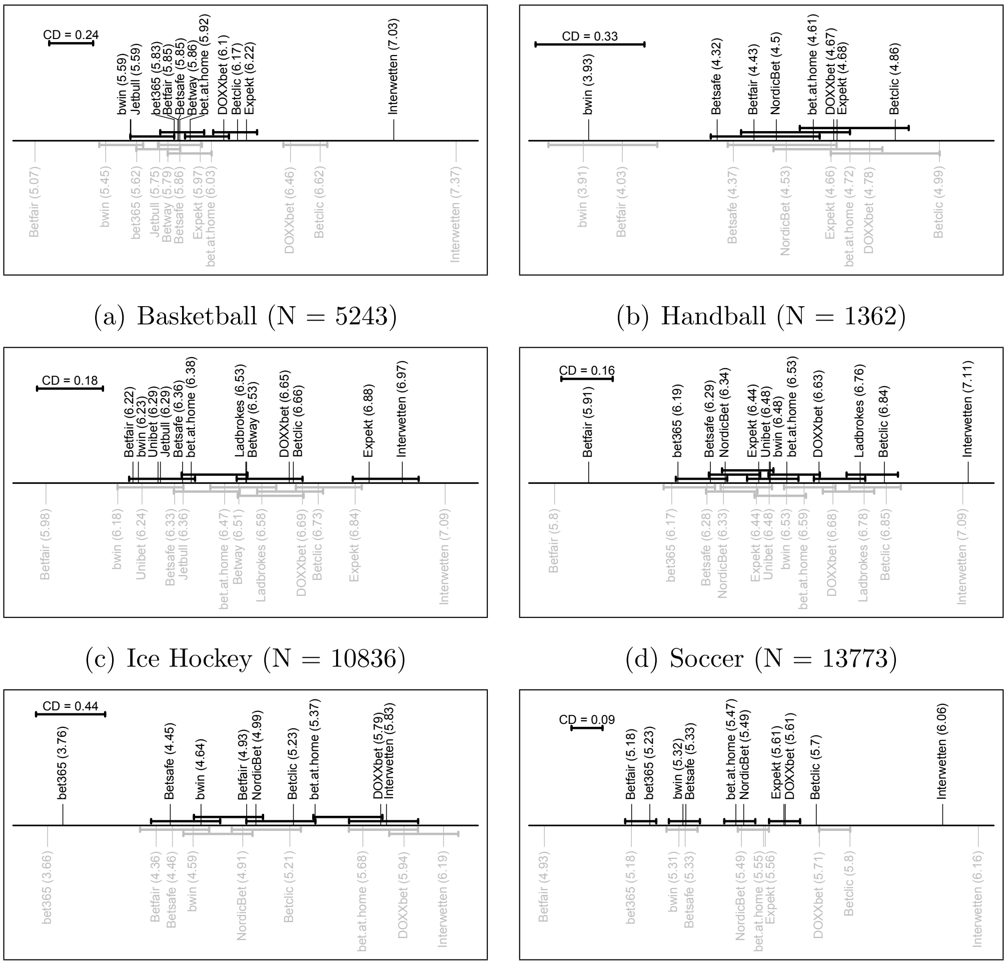
Stephenson, D.B., Coelho, C.A., Jolliffe, I.T., 2008. Two extra components in the brier score decomposition. Weather and Forecasting 23, 752–757.

Tziralis, G., Tatsiopoulos, I., 2007. Prediction markets: An extended literature review. The journal of prediction markets 1, 75–91.

Vaughan Williams, L. (Ed.), 2005. Information Efficiency in Financial and Betting Markets. Cambridge University Press.

Štrumbelj, E., Robnik-Šikonja, M., 2010. Online bookmakers' odds as forecasts: The case of european soccer leagues. International Journal of Forecasting 26, 482–488.

Štrumbelj, E., Vračar, P., 2012. Simulating a basketball match with a homogeneous markov model and forecasting the outcome. International Journal of Forecasting 28, 532–542.



(a) Basketball (N = 5243)

(b) Handball (N = 1362)

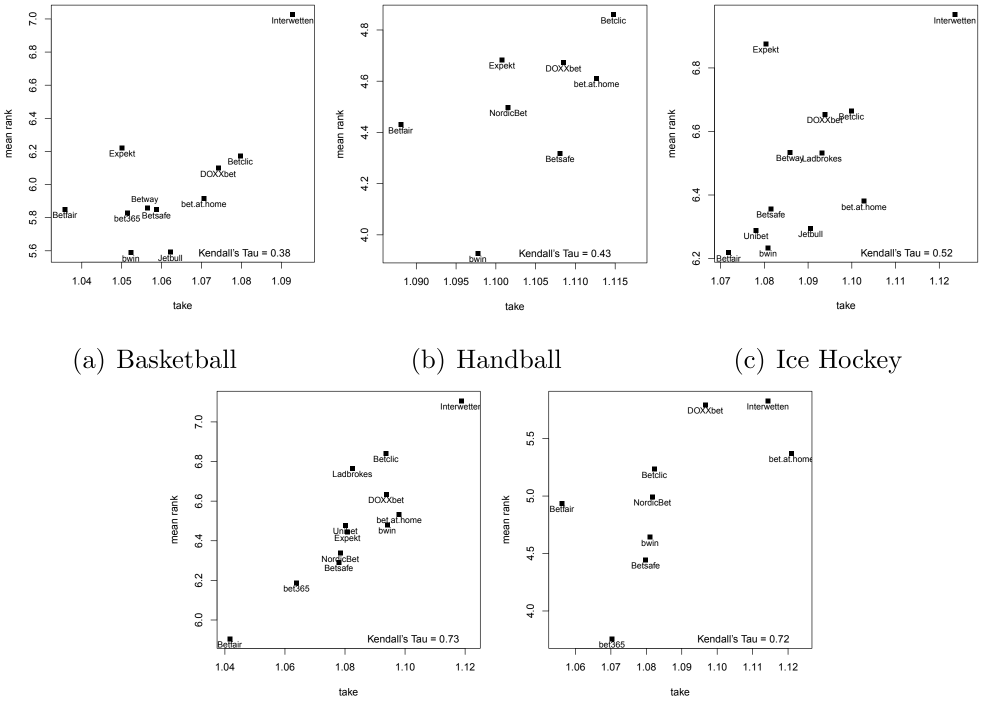
(c) Ice Hockey (N = 10836)

(d) Soccer (N = 13773)

(e) Volleyball (N = 989)

(f) All sports (N = 33981)

Figure 3: Mean ranks of bookmakers' scores broken down by sport. Differences beyond the critical distance (CD) are significant at the 0.01 level. Results in black are for Shin probabilities, results in gray for basic normalization.



(a) Basketball

(b) Handball

(c) Ice Hockey

(d) Soccer

(e) Volleyball

Figure 4: Relationships between a bookmaker's take (booksum) and its forecasting score rank. Correlation coefficients are provided.