Statistical Tests of Real-Money versus Play-Money Prediction Markets

E. S. ROSENBLOOM AND WILLIAM NOTZ

Abstract

Prediction markets are mechanisms that aggregate information such that an estimate of the probability of some future event is produced. It has been established that both real-money and play-money prediction markets are reasonably accurate. An SPRT-like test is used to determine whether there are statistically significant differences in accuracy between the two markets. The results establish that real-money markets are significantly more accurate for non-sports events. We also examine the effect of volume and whether differences between forecasts are market specific.

Keywords: prediction markets, sequential probability ratio test

Authors

E. S. Rosenbloom
(rosenbl@ms.umanitoba.ca) is Professor of Management Science at the I. H. Asper School of Business at the University of Manitoba. His primary research interests are in decision theory, analytic hierarchy process and mathematics of gaming.

William Notz
(notzww@ms.umanitoba.ca) is Professor of Organizational Behaviour at the I. H. Asper School of Business at the University of Manitoba. His primary research interests are in the extension and application of theories of cognition, conflict and motivation to understanding the behaviour of organizational participants.

Copyright © 2006 Electronic Markets
Volume 16 (1): 63–69. www.electronicmarkets.org DOI: 10.1080/10196780500491303

INTRODUCTION

Prediction markets are mechanisms that aggregate information such that the estimate of the probability of some future event is produced. The probability of some future event is evoked by contract payoffs; for example, a contract might pay $100 if the New England Patriots win the Super Bowl, or zero if they do not. The price at which this contract trades therefore represents the collective consensus of its expected value, or the subjective probability that the New England Patriot will win the Super Bowl.

There are two types of prediction markets: real-money markets and play–money markets. Examples of real-money exchanges are the Iowa Electronic Markets (http://www.biz.uiowa.edu/iem) and TradeSports.com (http://www.TradeSports.com). Examples of play-money exchanges are the Hollywood Stock Exchange (http://www.hsx.com) and NewsFutures World News Exchange (http://www.us.NewsFutures.com). In real-money prediction markets the participants risk their own money. In play-money prediction markets participants bear no financial risk. As an incentive to participate in play-money markets, participants are initially given a stake of play money that they can eventually use to bid on real prizes. If a participant loses his or her stake of play money, the participant can replenish or refill their account back to their original stake. There is considerable evidence that such prediction markets, using real money or play money, generate reasonably accurate probability forecasts (Berg *et al.* 2000, 2003; Forsythe *et al.* 1999; Surowiecki 2004).

To date, however, there has only been one study that examined directly the question of whether real-money or play-money made any difference in market predictive performance. Servan-Schreiber *et al.* (2004) compared the predictions of NewsFutures with those of TradeSports regarding the outcomes of NFL football games during the fall-winter 2003 season. They found by using a randomization test that there was no statistically significant difference in accuracy between the TradeSports' predictions and the NewsFutures' predictions. In our research we directly tested the predictive accuracy of the NewsFutures play-money market with the TradeSports real-money market by using an SPRT-like statistical test. In addition, we examined the effects of differences in trading volume, since the two markets differ substantially in that respect, as well as in their currencies. The results of that experiment suggested that the results were market related. We then tested the forecasts of TradeSports and NewsFutures in specific markets.

WHICH MARKET SHOULD PERFORM BETTER?

The dominant theoretical position that bears on the operation of markets is the proposition that market prices summarize trader information so as to produce efficient outcomes. Prices, in other words, represent the aggregation of all information in the system. Moreover, since the market reacts almost instantaneously and correctly to new information, this theoretical position would predict no difference between the two markets (assuming the information would be equally available to participants across markets).

There are arguments to suggest that the real-money market would be more accurate. It is not unreasonable to assume that the prospect of making money as well as the risk of losing money will make participants more motivated in seeking accurate information. Many of the play-money participants may not be very serious in their choices. In addition, there is a Darwinian aspect to real-money markets. Eventually weak players lose their money and no longer participate in the market. Real-money markets become a competition between stronger players.

However, there are counter arguments to suggest that the play-money market could perform better. Kahneman and Tversky's (1979) prospect theory posits that the way in which a decision maker frames a problem as either a loss or a gain will produce systematic deviation from what would be predicted from both expected value and expected-utility theory. More specifically, problems framed in terms of potential losses will generally evoke risk-taking behavior, while the same problem framed in terms of possible gains will usually produce risk aversion behavior. In addition, prospect theory posits that we tend to overweigh the probability of low probability events, under-weigh the probability of moderate and high probability events and that our response to loss is more extreme than our response to gain. However, this theory is based on gains and losses of real money. These behavioral biases may diminish when dealing with play-money rather than real-money.

A second factor that impacts predictive accuracy is differential volume. In general, higher volumes are associated with greater predictive accuracy, and this relationship extends to prediction markets (Berg *et al.* 1997). However, Forsythe *et al.* (1999) reported impressive findings from both election stock markets and laboratory markets that call into question just how much of a necessary condition volume is to accuracy. Their data indicates that only a small core of traders was necessary to drive markets to efficient outcomes. These so-called marginal traders tended to be more experienced, more educated and more knowledgeable market makers, as well as being relatively few in number. Thus, the performance of prediction markets would seem to depend less on large numbers of decision makers than it does on a core of active, informed and motivated traders. While it is not obvious that either of our prediction markets would have a monopoly on such marginal traders, it is nevertheless clear that their mere presence in the TradeSports market would remove any comparative disadvantage in predictive accuracy caused by the greater volume in NewsFutures.

SPRT-LIKE TEST

In order to determine whether the real-money market or the play-money market is more accurate, an SPRT-like hypothesis test was performed. An advantage of an SPRT-like test is that it requires no assumption about which model is the null hypothesis. It can treat each model equally and allow the data to determine which model is more appropriate or alternatively conclude that there is no statistically significant difference in accuracy between the two models. The corresponding hypotheses are:

H_1: Probabilities generated by NewsFutures (play-money) are accurate
versus
H_2: Probabilities generated by TradeSports (real-money) are accurate.

SPRT or Sequential Probability Ratio Test, due to Wald (1947), is a sequential test for a simple hypothesis H_1 against an alternative simple hypothesis H_2. At the end of each stage in the sampling the likelihood ratio L_1/L_2 is computed where the suffixes 1 and 2 refer to the H_1 and H_2 hypotheses respectively and L is the likelihood function of all sample members so far drawn. If $B<L_1/L_2<A$ the sampling is continued to another stage. If $L_1/L_2 \leqslant B$, the H_2 hypothesis is accepted. If $L_1/L_2 \geqslant A$, the H_1 hypothesis is accepted. The two positive constants, A and B, are determined by reference to prescribed requirements concerning the two types of errors made in testing hypotheses, the rejection of H_1 when it is true and the acceptance of H_1 when it is false.

Defining α and β by; α=Probability of accepting H_2 given that H_1 is true (the probability of a Type I error) and β=Probability of accepting H_1 given that H_2 is true (the probability of a Type II error), Wald established the relationship between (α,β) and (A,B). He showed that an upper bound for A is $(1-\beta)/\alpha$ and that a lower bound for B is $\beta/(1-\alpha)$.

An actual determination of A and B is a difficult computational problem. Therefore, the constants A and B are almost always approximated by $(1-\beta)/\alpha$ and $\beta/(1-\alpha)$ respectively. This will mean that when the SPRT terminates, the probability of a Type I error is at most α and the probability of a Type II error is at most β.

Wald proved that under certain regularity conditions SPRT will terminate in a finite number of steps with probability one provided that the data are independent and identically distributed. The data in this experiment consisted of the results of events simultaneously forecasted by NewsFutures and TradeSports. While it is reasonable to assume that the events are independent, they are not identically distributed. Therefore, an SPRT may not terminate. In order to guarantee termination in the comparison of the two probability forecasting methods, a maximum sample size M is required. If the maximum sample size M is reached the test is inconclusive. Finally, since the goal is to determine whether the NewsFutures or the TradeSports probabilities are more appropriate, it is reasonable to treat the two hypotheses equally. That is to select $\alpha=\beta$.

The resulting SPRT-like experiment is as follows:

1. A constant α $(0<\alpha<1)$ and a maximum sample size M are chosen.
2. At any stage m of the experiment the ratio L_1/L_2 is calculated where L_1 and L_2 are the likelihoods of the data under each of the hypotheses. (Note: There is no need to directly compute L_1 and L_2. These numbers will become extremely small and subject to round-off error. However, it is easy to store L_1/L_2 and update this number with each new data point.)
3. If $L_1/L_2 \geqslant (1\text{-}\alpha)/\alpha$, the experiment is terminated with the acceptance of H_1 and the rejection of H_2. If $L_1/L_2 \leqslant \alpha/(1\text{-}\alpha)$ the experiment is terminated with the acceptance of H_2 and the rejection of H_1. If $\alpha/(1\text{-}\alpha) < L_1/L_2 < (1\text{-}\alpha)/\alpha$ and $m<M$ the experiment is continued by taking an additional observation. Finally, if $\alpha/(1\text{-}\alpha) < L_1/L_2 < (1\text{-}\alpha)/\alpha$ and $m=M$, the experiment is terminated with an inconclusive result.

This SPRT-like test was developed by Rosenbloom (2000) and (2003) and can be generalized to choose the best of k probability forecasting models. The test can also be employed in a non-sequential situation, that is, a situation where the data have already been obtained. All the data should be used in the statistical test. The computations are still done sequentially, although the order is irrelevant. If the final likelihood ratio is at least $(1\text{-}\alpha)/\alpha$, the H_1 hypothesis is accepted. If the final likelihood ratio is below $\alpha/(1\text{-}\alpha)$, the H_2 hypothesis is accepted. If the final likelihood ratio is between $\alpha/(1\text{-}\alpha)$ and $(1\text{-}\alpha)/\alpha$, the test is inconclusive.

For the test of NewsFutures versus TradeSports, a significance $\alpha=.01$ and the maximum sample size M=1,000 was chosen. The experiment began 2 June 2004. Whenever NewsFutures and TradeSports offered markets on the same event, the probability forecasts from NewsFutures and TradeSports were simultaneously recorded. Events sampled included baseball games, basketball games, hockey games, tennis matches, golf tournaments (Would Tiger Wood win the US open?), direction of financial markets (Would the Dow be up on a specific day?) and political events (Would John Edwards be chosen as the Vice Presidential candidate?).

As an example of the calculations needed to update the likelihood ratio, consider the first three data points (see Table 1)

For the first event, the NASDQ was down on 2/6/04 so the likelihood ratio L_1/L_2 was equal to (1–.37)/(1–.12) or 0.7159. For the second event, Houston defeated the Chicago Cubs so the likelihood ratio became (.55/.54) * (0.7159) or 0.7292. For the third event, Baltimore lost to the New York Yankees and the likelihood ratio became (.64/.61)* (0.7292) or 0.7650. Sampling continues until either the likelihood ratio exceeds 99, or falls below 1/99 or until 1000 data points are sampled.

On 20 July 2004, after 522 observations, the likelihood ratio reached 0.0097. Since this is below 1/99, the H_2 hypothesis is accepted. That is, at a 1% significance level we concluded that the real money TradeSports model was a more appropriate model than the play money NewsFutures model.

The progress of the likelihood ratio L_1/L_2 can be seen in Figure 1.

Although the TradeSports probabilities are more accurate than the NewsFutures probabilities at the 1% significance level, the differences between them were slight. The correlation between TradeSports probabilities and NewsFutures probabilities was 0.955. A strategy of buying exactly one contract of the NewsFutures price if the TradeSports price is greater (or selling one contract at the NewsFutures price if the TradeSports price is smaller) yielded a return of 5.71%. However, in practice, the return would be lower than this since there are transaction costs such as the bid–ask spread and commissions. The opposite strategy of buying exactly one contract of the TradeSports price if the NewsFutures price is greater (or selling one contract

Table 1. Example of calculations involved in updating the likelihood ratio. First 3 data points

Event	$_p$*NewsFutures*	$_p$*TradeSports*	*Result*	L_1/L_2
NASDQ up	0.370	0.120	0	0.7159
Houston over Chicago Cubs MLB	0.550	0.540	1	0.7292
Baltimore over New York Y MLB	0.360	0.390	0	0.7650

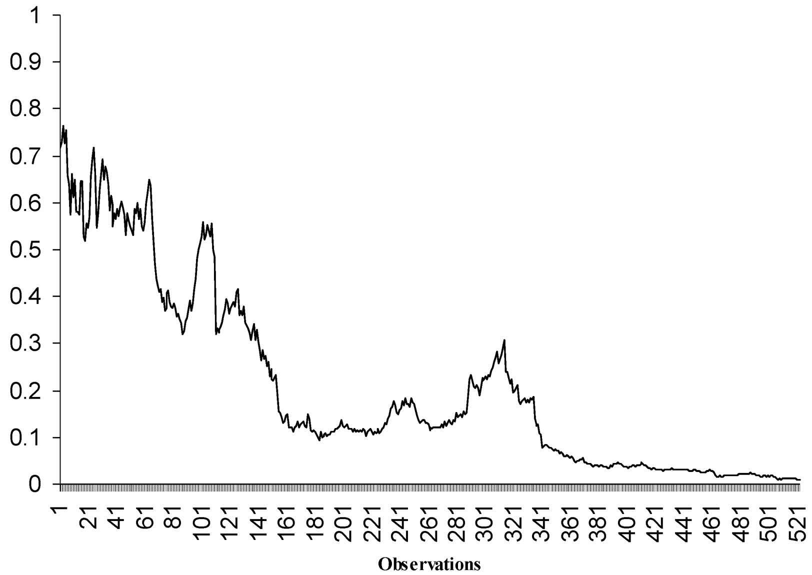

Figure 1. Likelihood ratio L_1/L_2 during experiment

at the TradeSports price if the NewsFutures price is smaller) lost 0.09%.

Earlier we discussed how the number of participants might influence the performance of prediction markets. Both NewsFutures and TradeSports provided data on how many contracts were traded but not on how many people participated in the market. However, if say m contracts were traded in a particular market, it could be that m individuals had taken one side of the contract and m individuals had taken the other side, or it could mean that one person had bought m contracts and someone else had sold m contracts. Further complicating a comparison between NewsFutures and TradeSports is the fact that the contracts are different in the two markets. In NewsFutures a contract expires at either a price of $100 of play money or $0. A contract in TradeSports expires at a price of $10 of US currency or $0. Despite these differences, it appeared that there were far more participants in NewsFutures than in TradeSports. During our experiment, volume in NewsFutures ranged between 95 contracts and 157,891 contracts with a mean of 7,600 contracts and a median of 4,746 contracts. Volume in TradeSports ranged between 1 contract and 21,771 contracts with a mean of 201 contracts and a median of 41 contracts. The real-money TradeSports performed better despite the lower volume. However, if we partition the data set the effect of volume can be seen. We partitioned the data set by separately analysing the 261 observations with the lowest volume in TradeSports and the 261 observations with the highest volume in TradeSports. For the 261 observations with lower volume in TradeSports, the likelihood ratio L_1/L_2 was 0.2112. Therefore, although TradeSports performed better than NewsFutures in this subset, the results were not statistically significant. However, for the 261 observations with higher volume in TradeSports, the likelihood ratio L_1/L_2 was 0.0459. Thus for this higher volume subset, TradeSports performed much better and the results were statistically significant at the 5% level.

Servan-Schreiber *et al.* (2004) did not perform an SPRT-like test in their study. Instead they calculated a number of summary statistics such as mean absolute error, root mean squared error, average quadratic score and average logarithmic score. We also calculated these summary statistics in Table 2.

In the Servan-Schreiber *et al.* paper NewsFutures performed slightly better on these statistics. In our study TradeSports performed slightly better.

The Servan-Schreiber *et al.* study was strictly on NFL football games. Since our study was in June and July there were no NFL games. However, we found surprising results when we partitioned our data set between North American team sports (baseball, basketball and hockey) and the remaining events (tennis, golf, political events, financial events, etc.). In the experiment, 458 out of the 522 data points were from games involving North American team sports. The likelihood ratio L_1/L_2 for these 458 events was 0.252. Although the real-money TradeSports model performed better than the play-money NewsFutures model for these events, the results were not statistically significant. However, for the remaining 64 data points, the likelihood ratio L_1/L_2 was 0.039 and therefore statistically significant at a 5% level.

These results suggested to us that differences between TradeSports and NewsFutures might be market specific, and that events such as North American team sports might be different from other events. We explored this issue with an SPRT-like test on the 2003 NFL season, on the daily direction of the Dow Jones Industrial Average and on North American team sports.

Table 2. Summary statistics for prediction accuracy of real-money and play-money markets

	TradeSports (real-money)	*NewsFutures (play-money)*	*Difference (TS-NF)*
Mean Absolute Error =lose_price	0.474	0.477	−0.003
(lower is better)	(.005)	(.005)	(.007)
Root Mean Square Error =$\sqrt{}$Average(lose_price}^{2}	0.487	0.491	−0.004
(lower is better)	(.021)	(.021)	(0.030)
Average Quadratic Score =100–400*(lose_price^2)	5.165	3.555	1.610
(higher is better)	(1.807)	(1.939)	(2.650)
Average Logarithmic Score =Log(win-price)	−0.666	−0.675	0.009
(higher (less negative) is better)	(.010)	(.010)	(.014)

Notes: win_price=winning proposition price/100
lose_price=losing proposition price/100
Best score for each metric shown in bold.
Standard errors are shown in parentheses.

SPRT-LIKE TEST ON NFL DATA SET

Our SPRT-like test on a representative sample of TradeSports and NewsFutures forecasts found TradeSports to be significantly better than NewsFutures. Servan-Schreiber *et al.*'s study of TradeSports and NewsFutures forecasts for the 2003 NFL season found no statistical differences between the models, but they did not use the SPRT-like test. Since the authors of Servan-Schreiber *et al.* were gracious enough to provide their 2003 NFL dataset to us, we performed a SPRT-like test on their data. The underlying hypotheses were:

H_1: Probabilities for NFL games generated by NewsFutures (play-money) are accurate
versus
H_2: Probabilities for NFL games generated by TradeSports (real-money) are accurate.

The likelihood ratio L_1/L_2 on the 208 games from the Servan-Schreiber *et al.* dataset was 1.599. So although NewsFutures performed slightly better than TradeSports, the results were not close to being statistically significant. Thus the SPRT-like test produced conclusions consistent with those of Servan-Schreiber *et al.* with respect to NFL games.

In order to explore the possibility that the choice of markets affects the performance of the prediction market, we performed two additional studies. The first was on the daily direction of Dow Jones Industrial Average (DJIA) and the second was on North American team sports.

SPRT-LIKE TEST ON DJIA

In addition to the comparison of NewsFutures with TradeSports, the existence of base rate information permitted a third comparison. The SPRT-like method can be applied to multiple hypotheses by calculating for each hypothesis j, $R_j=[\Sigma_i\ (L_i/L_j)]^{-1}$ where L_i/L_j is the likelihood ratio associated with the i and j hypotheses respectively. Data are collected until either the maximum sample size M is reached or one of the R_j's is above 1-α, where α is the significance level of the test. If one of the R_j's is above 1-α, hypothesis H_j is accepted. If all the R_j's remain below 1-α, and the maximum sample size M is reached, the test is inconclusive. We chose a significance level α=.01 and a maximum sample size M=500. If we assume before data collection that each hypothesis has an equal chance of being correct, then R_j can be viewed as the posterior probability of Hypothesis j being correct. With this interpretation the expression for R_j is simply Bayes' formula.

The SPRT-like method was applied to forecasts of the daily direction of the DJIA beginning on 23 February 2004. The probability that the DJIA would be up on a given day was recorded at 9:00 a.m. Eastern Standard Time (one-half hour before the US markets open) for both the TradeSports and NewsFutures models. The base rate model was estimated by calculating the relative frequency of the DJIA up days (9,639) vs. down days (8,872) between 3 February 1930 and 22 February 2004. This base rate [9639/(9639+8872)]=0.5207 was then updated daily during the experiment.

The hypotheses for this experiment were:

H_1: Probabilities for direction of Dow generated by NewsFutures (play money) are accurate
versus
H_2: Probabilities for direction of Dow generated by TradeSports (real money) are accurate
versus
H_3: Probabilities for direction of Dow generated by the base rate are accurate.

This experiment ended on 13 October 2004 with 152 data points. The final R_j's were R_1=.00899, R_2=.99095 and R_3=.00006. Since R_2 was above 1-α or 0.99, the H_2

hypothesis that the real-money TradeSports model was the most appropriate for forecasting the direction of the DJIA, should be accepted. In addition, both prediction markets were significantly better than the base rate model.

SPRT-LIKE TEST ON TEAM SPORTS

Since the results reported earlier suggested the possibility of market-specific effects, we tested the performance of the two markets on yet another dataset that we collected on North American team sports. The hypotheses were:

> *H_1: Probabilities generated by NewsFutures (play money) are accurate*
> *versus*
> *H_2: Probabilities generated by TradeSports (real money) are accurate.*

The significance level α was again chosen to be 0.01 and the maximum sample size was chosen to be 500. The experiment, involving US college basketball, NBA basketball and Major League Baseball, began on 11 March 2005 and ended on 17 April 2005 when the maximum sample size was reached. The final likelihood ratio L_1/L_2 was 0.166. Therefore, although the real-money TradeSports performed better than the play-money NewsFutures, the ratio was not statistically significant.

Altogether, we had three independent datasets involving North American team sports. The original 2004 data set had 458 games, the Servan-Schreiber *et al.* 2003 data set had 208 games and the 2005 data set had 500 games. The respective likelihood ratios L_1/L_2 were .252, 1.599, and .166. Therefore, we had a total of 1,166 points with a likelihood ratio of L_1/L_2 of (.252)(1.599)(.166) or .067. Despite this large data, the differences between NewsFutures and TradeSports for North American team sports were insignificant.

COMBINING TRADESPORTS AND NEWSFUTURES FORECASTS

The participants in the TradeSports and NewsFutures prediction markets generate probability estimates for one-time events. The SPRT-like test established that at a 1% significance level the real-money TradeSports model had generated more accurate probabilities then the play-money NewsFutures model. However, this does not mean that the TradeSports probabilities were the 'true' probabilities. For most one-time random events it is unlikely that there will ever be a technique for generating true probabilities. Nevertheless, Servan-Schreiber *et al.* suggested that probability forecasts might be improved by synthesizing the TradeSports and NewsFutures using linear regression. A drawback of using linear regression with a dichotomous response variable is that there is no guarantee that the regression probability forecast would be between 0 and 1. We did apply linear regression on the original data set of 522 representative events and obtained the regression equation $p=.002-1.134p_1+2.049p_2$ where p_1 is the NewsFutures probability forecast and p_2 is the TradeSports probability forecast.

However, we suggest two alternate approaches for synthesizing the two probability forecasts. One is to find the maximum likelihood convex combination $\alpha p_1+(1-\alpha)p_2$ of the NewsFutures forecast p_1 and the TradeSports forecast p_2, with α between 0 and 1. We applied this approach to our original data set of 522 events. The result was that the maximum likelihood convex combination occurred at $\alpha=0$; in other words, use the TradeSports probability forecast by itself.

A second approach is to use logistic regression. With logistic regression, a probability forecast of the form $\exp(\beta_0+\beta_1p_1+\beta_2p_2)/(1+\exp(\beta_0+\beta_1p_1+\beta_2p_2))$ is generated where p_1 is the NewsFutures forecast, p_2 is the TradeSports forecast, and β_0, β_1, and β_2 are the maximum likelihood estimators. This probability forecast will always be between 0 and 1. We applied this approach to our original data set of 522 events. The maximum likelihood estimators were $\beta_0=-2.209$, $\beta_1=-5.888$, and $\beta_2=10.307$.

To determine whether this logistic regression model is reasonable, we tested it on the 2004 DJIA data set and the 2005 North American team sports data set. We used the SPRT-like approach to test the logistic regression model versus the TradeSports probability forecasts.

For the 2004 DJIA data set, the resulting hypothesis test was:

> *H_1: Probabilities for the daily direction of the DJIA generated by TradeSports are accurate*
> *versus*
> *H_2: Probabilities for the daily direction of the DJIA generated by the logistic regression model synthesizing TradeSports and NewsFutures are accurate.*

The final likelihood ratio L_1/L_2 was 0.355. The results were not statistically significant. Although the test on this data set was inconclusive, the logistic regression model performed slightly better than the TradeSports model.

For the 2005 team sports data set, the resulting hypothesis test was:

> *H_1: Probabilities for North American sports games generated by TradeSports are accurate*
> *versus*
> *H_2: Probabilities for North American sports games generated by the logistic regression model synthesizing TradeSports and NewsFutures are accurate.*

On this data, TradeSports performed better. The final likelihood ratio L_1/L_2 was 36.32, significant at a 5% level.

The fact that the logistic regression model was competitive with TradeSports on the DJIA data set suggests that there is some merit in synthesizing the two probability forecasts. Likely, it is better to calibrate the logistic regression model for a specific market rather than use the 2004 representative sample.

CONCLUSIONS

The SPRT-like approach is a natural and simple method of testing different probability forecasting models for one-time events. Rather than having to interpret a host of summary statistics, it provides a decision on whether one of the forecasting techniques is superior.

When the SPRT-like method was originally employed on a representative sample of forecasts, the real-money TradeSports market was significantly more accurate than the play-money NewsFutures market, despite the lower volume in the TradeSports market. However, a closer look at the data strongly suggests that the results are market related. Results from North American team sports indicated no statistical differences between TradeSports and NewsFutures. Results from forecasting the DJIA showed the real-money TradeSports market as clearly superior. We can only speculate on why there might be differences between popular sports events and other events. For popular sports events there are many sources for the approximate odds such as casinos and sports books. Participants in TradeSports and NewsFutures are getting cues from these other sources that will influence their assessment of the probabilities. For events such as whether Edwards would be chosen as the Vice Presidential candidate, or whether a particular financial market will be up, participants are much more on their own in assessing the probabilities, and it is in these events that the real-money TradeSports market performed significantly better. Of course, the objective of prediction markets is the forecasting of real world events, and these usually do not have cues from casinos and sports books.

Both real-money and play-money prediction markets provide reasonably accurate probability forecasts. In forecasting the daily direction of the DJIA, both the real-money and play-money probability forecasts were far more accurate than the base probability. However, the SPRT-like experiment determined that the real-money market was the more accurate, particularly for non-sports events. This effect may be attributable to either a decrease in predictive accuracy associated with the retention of losers in the NewsFutures (play-money) population, or an increase in accuracy as a function of the disproportionate numbers of (and more rational?) marginal traders who were lured to the real-money market. Indeed, both effects may have occurred and only additional research, probably experimental, can address this issue successfully.

Notwithstanding theoretical explanations, the superior accuracy of the real-money market for non-sports events was clear, and if this general result is replicated, the implications for the utility of prediction markets may be rather negative. It is easier to set up a prediction market using play-money than one using real-money. There are many legal and technical hurdles that must be overcome to establish a real-money prediction market. In addition, an important use of prediction markets is to obtain better sales forecasts. It would be problematic for companies to force their employees to risk their own money in a real-money prediction market. However, the results in this paper establish that increased accuracy is obtained in real-money prediction markets.

References

Berg, J., Forsythe, R. and Rietz, T. (1997) 'What Makes Markets Predict Well? Evidence From the Iowa Electronic Markets', in W. Arbers, W. Guth, P. Hammerstein, B. Moldouanu and E. Van Damme (eds) *Understanding Strategic Interaction: Essays in Honor of Reinhard Selten*, New York: Springer, pp 444–63.

Berg, J., Forsythe, R., Nelson, F. and Rietz, T. (2000) 'Results from a Dozen Years of Election Futures Market Research', Technical Report, University of Iowa.

Berg, J., Nelson, F. and Rietz, T. (2003) 'Accuracy and Forecast Standard Error of Prediction Markets', Working Paper, Tippie College of Business.

Forsythe, R., Rietz, T. and Ross, T. (1999) 'Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets', *Journal of Economic Behavior and Organization* 39: 83–110.

Kahneman, D. and Tversky, A. (1979) 'Prospect Theory: An Analysis of Decision Under Risk', *Econometrics* 47: 263–93.

Rosenbloom, E. S. (2000) 'Selecting the Best of k Multinomial Parameter Estimation Procedures Using SPRT', *Sequential Analysis* 19(4): 177–92.

Rosenbloom, E. S. (2003) 'A Better Probability Model for the Racetrack Using Beyer Speed Numbers', *OMEGA* 31(5): 339–48.

Servan-Schreiber, E., Pennock, D., Wolfers, J. and Galebach, B. (2004) 'Prediction markets: Does money matter?', *Electronic Markets* 14(3): 1–10.

Surowiecki, J. (2004) *The Wisdom of Crowds*, New York: Doubleday.

Wald, A. (1947) *Sequential Analysis*, London: John Wiley & Sons.