Prediction Market Prices as Martingales: Theory and Analysis

David Klein
Statistics 157

Introduction

With prediction markets growing in number and in prominence in various domains, the construction of a modeling framework for the behavior of prices on traded contracts has become an increasingly important endeavor. In this paper, we present such a theoretical framework, as we attempt to use martingale theory in the analysis of prediction market price fluctuations. The application of this theory to prediction market prices generates certain predictions regarding, in particular, win probabilities, the distribution of maximum and minimum prices, and the distribution of interval crossings, which we test using empirical data on contract prices for baseball matches from the online prediction marketplace Tradesports.

Background

For the purposes of this paper, we define a prediction market as a venue at which contracts whose ultimate value depends on the occurrence or failure to occur of some specified event (presumably with a limited time horizon) are publicly traded. Classic examples of such contracts are those whose value is tied to the event that a specific candidate (e.g., Barack Obama) becomes president of the United States, or (as will be particularly relevant for this paper) the event that a sports team wins a given match. From the moment contracts are initially put up for bid by the hosting party until the time at which the contracts pay out, they may be bought and sold by individual traders. In this sense, prediction markets function as an admixture of traditional betting markets and stock markets: Like stock markets (and unlike betting markets), prediction market contracts may be sold by individual participants; unlike most stock markets, however, there is a clear termination point for the contract.

In general, this paper will assume the Tradesports model: Contracts vary between the (arbitrary) values of 0 and 100; a contract is initially offered at some value between 0 and 100, and may be traded until the termination point for the contract, at which point its value is either 100 (in which case it pays out $10) or zero (in which case it pays out nothing). During the trading period for the contract, its value may fluctuate as investor beliefs about the outcome change. In this paper, we concern ourselves principally with these

price fluctuations; our central tool in the analysis of these movements is an artifact from probability theory known as a martingale.

A sequence $Y = Y_0, \ldots, Y_n$ is a martingale with respect to a random sequence $X = X_0, \ldots, X_n$ if for all $n \geq 0$ the equality $\mathbb{E}(Y_n | X_0, \ldots, X_{n-1}) = Y_n$ holds. For prediction markets, if we let X be a random sequence of price perturbations, then we assert that if we define Y such that $Y_n = \sum_{i=0}^{n} X_i$, then the price sequence Y is a martingale. This follows from the principle that the price at any given point represents the consensus probability that the event in question will occur, and is thus the fair price for the gamble. Thus, the expectation of the future price based on currently available information will be equal to the current price.

One important property of a martingale that follows directly from the definition is the fact that $\mathbb{E}(Y_n) = \mathbb{E}(Y_0)$ for all $n \geq 0$. This is easily shown using the tower property of expectation:

$$\begin{aligned} \mathbb{E}(Y_n) &= \mathbb{E}(\mathbb{E}(Y_n \mid X_0, \ldots, X_{n-1})) \\ &= \mathbb{E}(Y_{n-1}) \end{aligned}$$

Repeated iteration of this process gives the desired equality. Though this result applies only to a fixed time n, the Optional Stopping Theorem asserts that it can be extended to a random time T given that T is a stopping time. (T is defined to be a stopping time if it is decidable whether or not $n = T$ for a given value of n based on the information contained in $X_0, \ldots, X_n$. For example, if we define T to be the time when a gambler first achieves positive profits, then T is a stopping time; if we define T to be the time immediately prior to the gambler's first loss, T is not a stopping time.) Formally, the Optional Stopping Theorem states that, for a stopping time T, $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$ given that $\mathbb{P}(T < \infty) = 1$, $|\mathbb{E}(Y_T)| < \infty$, and $\mathbb{E}(Y_n I_{\{T>n\}}) \to 0$ as $n \to 0$. The Optional Stopping Theorem provides the basis for the following equality, from which the key theoretical results of this paper derive.

Consider a price x such that $Y_0 = x$ (x is the starting price), and prices a and b such that $0 \leq a < x < b \leq 100$. Let T be the first time the price reaches either a or b, given that it starts at x. T is clearly a stopping time, and it is intuitively plausible, though we omit the formal proof, that Y_T satisfies the conditions of the Optional Stopping Theorem. Thus, $\mathbb{E}(Y_T) = \mathbb{E}(Y_0) = x$.

Additionally, if we define π_b to be the probability that the price reaches b before it reaches a, then we have that $\mathbb{E}(Y_T) = (1 - \pi_b)a + \pi_b\, b$ (since Y_T can take only the values a or b, and it takes the former with probability $1 - \pi_b$ and the latter with probability π_b). Setting the two expressions for $\mathbb{E}(Y_T)$ equal to each other gives

$$\begin{aligned} x &= (1 - \pi_b)a + \pi_b\, b \\ x &= a + \pi_b(b - a) \\ \pi_b &= \frac{x - a}{b - a} \end{aligned} \tag{1}$$

It follows that

$$\pi_a = \frac{b - x}{b - a} \tag{2}$$

where π_a is the probability that the price reaches a before b.

A fundamental entailment of this formula is that if we suppose that the contract pays out if Team 1 wins and fails to pay out if Team 1 loses, then we may evaluate the probabilities that Team 1 wins or, alternately, that Team 2 wins for a given starting price x by setting $a = 0$ and $b = 100$. (We assume here and in all cases to follow that Team 1 wins if and only if the terminal price of the contract is 100.) These probabilities, respectively, are

$$\mathbb{P}(\text{Team 1 wins}) = \frac{x}{100} \tag{3}$$

$$\mathbb{P}(\text{Team 2 wins}) = \frac{100 - x}{100} \tag{4}$$

Additionally, we can derive certain formulae regarding m_Y and M_Y, random variables representing the minimum and maximum prices recorded for a given traded contract. Clearly, if $m_Y \leq a$ for some given price a, it must be the case that the price of the contract reaches a before it reaches 100. Thus, $\mathbb{P}(m_Y \leq a) = (100 - x)/(100 - a)$, which follows from (**??**), with $a = a$, $b = 100$. Similarly, using (**??**) with $a = 0$, $b = b$, we have that

$\mathbb{P}(M_Y < b) = (b - x)/b$. Note that if we partition the price sequence Y for a given traded contract into non-overlapping subsequences, these subsequences are martingales as well. We use this observation in conjunction with formulae (**??**) and (**??**) and Bayes' Theorem to compute the cumulative distribution function of the minimum conditional on the outcome that Team 1 wins.

$$\begin{aligned}\mathbb{P}(m_Y \leq a \mid \text{Team 1 wins}) &= \frac{\left(\frac{a}{100}\right)\left(\frac{100-x}{100-a}\right)}{\left(\frac{a}{100}\right)\left(\frac{100-x}{100-a}\right) + (1)\left(\frac{x-a}{100-a}\right)} \\ &= \frac{a(100-x)}{x(100-a)} \end{aligned} \tag{5}$$

The first equation makes use of the following facts: $\mathbb{P}(\text{Team 1 wins} \mid m_Y \leq a) = a/100$; $\mathbb{P}(m_Y \leq a) = (100-x)/(100-a)$; $\mathbb{P}(\text{Team 1 wins} \mid m_Y > a) = 1$; and $\mathbb{P}(m_Y > a) = (x-a)/(100-a)$. The first, second, and fourth equalities are simple applications of (**??**) and (**??**) with appropriate choices for a, b, and x, while the third follows from the condition that $a \geq 0$; if the event has terminated and the price has not reached a, then it has not reached zero, and therefore, it must be the case that Team 1 has won. Given these equalities, we arrive at (**??**) via basic algebra. Applying the same approach, we may derive a similar formula with regard to the conditional cumulative distribution function of the maximum price in the case that Team 2 wins:

$$\mathbb{P}(M_Y < b \mid \text{Team 2 wins}) = \frac{100(b-x)}{b(100-x)} \tag{6}$$

Another random variable of interest is Z, the number of crossings the price makes of a given interval $[a, b]$. (The price sequence Y crosses $[a, b]$ when it reaches b, having started at a, or vice-versa.) For a general interval $[a, b]$, we compute the probability of a single crossing ($Z = 1$) as follows: First, we note that in order for a crossing to occur, it must be the case that the price sequence reaches either a or b. For $x \in [a, b]$, the probability that a single crossing from a to b occurs is equal to $\left(\frac{b-x}{b-a}\right)\left(\frac{a}{b}\right)\left(\frac{b-a}{100-a}\right)$ (i.e., the probability that the price sequence reaches a before b, reaches b before zero (starting from a), and then reaches 100 before reaching a again (starting from b)), while the probability of a single crossing from b to a is $\left(\frac{x-a}{b-a}\right)\left(\frac{100-b}{100-a}\right)\left(\frac{b-a}{b}\right)$, derived similarly. The probability of a single crossing for $x \in [a, b]$ is the sum of these two probabilities, since they represent disjoint events. (Note that if it is not the case that $x \in [a, b]$, it must either be true

that $x \leq a$ or $x \geq b$; in these cases, a single crossing from b to a or from a to b, respectively, is impossible, and thus $\mathbb{P}(Z = 1 \mid x \leq a) = \left(\frac{a}{b}\right)\left(\frac{b-a}{100-a}\right)$, and $\mathbb{P}(Z = 1 \mid x \geq b) = \left(\frac{100-b}{100-a}\right)\left(\frac{b-a}{b}\right)$.) For $Z \geq 2$, the approach is similar; we simply add terms prior to the end term to account for each subsequent crossing.

In the case where the interval $[a, b]$ is symmetric about 50, the formula is considerably simpler. Information about the first endpoint of the interval the price sequence reaches is irrelevant, since the price is just as likely to cross up from a to b as it is to cross down from b to a. (If we write $b = 100 - a$, it is easily seen that $\frac{a}{b} = \frac{100-b}{100-a} = \frac{a}{100-a}$.) Thus, the general formula for the probability of k crossings given that the price sequence ever enters $[a, 100-a]$ is

$$\mathbb{P}(Z = k) = \left(\frac{a}{100-a}\right)^k \left(\frac{100-2a}{100-a}\right) \tag{7}$$

Note that this is the formula for a (shifted) geometric distribution with $p = (100 - 2a)/(100 - a)$.

Methodology and Results

As the foregoing analysis makes clear, the presumption that prediction market prices may be described as martingales generates a number of predictions that we may test empirically. To this end, we collected price data on Tradesports contracts for 91 baseball games played between August 7 and October 27, 2008. For each such game, data consisted of the price sequence from the opening bid price (the starting price) until the price at termination (either 100 or 0). We used these data to assess the accuracy of the three main theoretical predictions described above, namely: 1) The starting price reflects the probability that a given team will ultimately prevail; 2) The conditional distributions of the minimum and maximum are those given in (**??**) and (**??**), respectively; and 3) The distribution of the number of crossings of an interval that is symmetric about 50 is given by (**??**).

For the purposes of testing these predictions, it is clearly desirable that we may treat the games in the data set as independent, identically distributed realizations of a particular random variable. While the assumption of inde-

pendence is not difficult to justify, the identical distribution condition poses a slight problem. In particular, the formulae which generate predictions (1) and (2) depend on x, the starting price, which may vary from game to game. Thus, if we consider the achievement of a given minimum or the failure to achieve a given minimum, for example, as a random indicator variable, our data set is like a series of coin flips where the coins may have different values for p. Thus, it was necessary to adopt strategies to standardize p.

For the purposes of testing prediction (1), games were grouped according to starting price; all games whose starting price was within a given range (e.g., $50 \leq x < 60$) were placed in the same group, and all groups of equal size with sufficiently many (i.e., more than 10) games were tested. We created two separate partitions by starting price – one had price groups $[50, 60)$ and $[60, 70)$ (each of which contained 39 games), while the other had groups $[50, 55)$, $[55, 60)$, and $[60, 65)$. (The groups in the second partition contained 19, 20, and 31 games, respectively.) For each group in a given partition, the mean starting price was computed. This mean price divided by 100 was taken to be the success probability p for a series of Bernoulli trials (success in this case is the event that Team 1 wins). Thus, the number of games won by Team 1 in the group as a whole was considered to be a binomially distributed random variable. Using the binomial distribution, we were able to compute the endpoints of the critical interval (that is, the interval in which 95 percent of values would be expected to fall) for each price group.

The critical intervals for the two groups in the first partition were $[15, 27]$ and $[18, 30]$, respectively, while the critical intervals for the three groups in the second partition were $[6, 14]$, $[7, 16]$ and $[14, 24]$. The observed values of the number of victories by Team 1 in each group were 23 and 24 for the first partition, and 10, 13, and 19 for the second. (See Figs. 1 and 2 in the Appendix for a visual representation of this data.) Thus, all critical intervals contained the sample estimates for the parameter, and thus there is no strong evidence to reject the null hypothesis that prediction market prices may be modeled as martingales based on this criterion.

With regard to the conditional minima and maxima, we chose to consider all contracts that passed through the value 50. The subsequence beginning at 50 is itself a martingale, and so we take $x = 50$ to be the starting price for each contract. (The choice of 50 was arbitrary, based primarily on simplicity

and symmetry: Each team won half of the 64 games whose price reached 50 at some point.) The martingale theory described above is presumed to apply equally well to these truncated trading periods. Thus, substituting 50 for x in (**??**) and (**??**), we have that $\mathbb{P}(m_Y \leq a \mid \text{Team 1 wins}) = \frac{a}{100-a}$ and $\mathbb{P}(M_Y < b \mid \text{Team 2 wins}) = \frac{2(b-50)}{b}$. Using these probabilities and the binomial formula as above, we were able to construct the critical intervals for the minimum value 40 conditional on victory by Team 1, and, respectively, for the maximum value 60 conditional on victory by Team 2. These intervals were then compared with the actual number of games won by Team 1 (respectively, Team 2) whose minimum (maximum) price was below 40 (60). These intervals were $[16, 26]$ and $[6, 16]$. The number of games won by Team 1 in which the minimum price reached after 50 was below 40, 21, was equal to the number of games won by Team 2 in which the maximum price achieved after 50 was 21; while this number is contained in the critical interval for the minimum, it is beyond the range of the critical interval for the maximum. In fact, under the null hypothesis that the maximum probability is as given in (**??**), the likelihood of getting a sample of 32 games in which 21 or more had a post-50 maximum price less than 60 was virtually zero. This result thus casts serious doubt on whether prediction market prices may in fact be modeled as martingales in the manner described above.

Additionally, for each price less than or equal to 50, we tallied the number of games whose post-50 minimum price was less than or equal to the given price. In this way, we were able to generate the empirical cdf for the minimum. A graph of the empirical and theoretical distribution functions (see Fig. 3) shows a high degree of consonance, and suggests that the martingale model describes such minimum prices quite well. The results are not so agreeable, however, for the empirical cdf of the maximum: The observed number of games where the post-50 maximum price is less than a given price is consistently higher than the predicted number of such games. This is driven in particular by the fact that 14 of the 32 games won by Team 2 after the price reached 50 never reached a price above 50 after hitting 50 for the first time.

Finally, we examine the difference between the observed and expected numbers of crossings of a given (symmetric) interval. This requires no fix to omit an x from the relevant formula, since a is the only parameter in the

expression. We arbitrarily selected this a to be 40, which gives the interval $[40, 60]$. Eighty-two of the 91 games contained a point in this interval and were thus suitable for analysis. Using (**??**), for which $p = 1/3$ for $a = 40$, we computed the vector of expected crossings to be approximately (27, 18, 12, 8, 16) for 0, 1, 2, 3, and 4 or more crossings, respectively. (Note that the sum of the elements in the vector is only 81 due to rounding error.) The vector of observed crossings was tabulated to be (38, 22, 13, 6, 1, 2). With this data, we administered a chi-square goodness of fit test comparing the observed and expected counts to test the hypothesis that the shifted geometric distribution with $p = 1/3$ in fact describes the data. The p-value for this test was 0.0026, which implies that the proposed distribution is a bad fit for the data. In particular, it predicts many fewer games with zero crossings and many more with four or more than were in fact observed.

Discussion

The results of this analysis are mixed. At a basic level, it appears that the starting contract price is a fairly accurate predictor of the likelihood that the event will in fact occur. However, predictions regarding the conditional maximum price for a given contract are not supported by these data, nor are those concerning the number of crossings of an interval. In particular, for the markets analyzed in these data, it appears that there are fewer large fluctuations than one would expect using martingale-based theory. We note additionally that the data set contained a disproportionately large number of games (84 of 91) whose starting price was greater than 50. Thus, if it is the case that contracts tend to follow a given trend line more closely than the theory implies, the failure of the theoretical predictions regarding maximum prices may possibly be due to the large number of games that started above 50 and drifted down to zero in a fairly consistent manner. Obviously, it is not clear from this analysis whether this is specific to baseball matches or Tradesports or whether it applies to prediction markets in general, and thus it remains undecided whether martingales may actually be used to generate useful predictions for prediction market price movements.

Appendix: Graphs

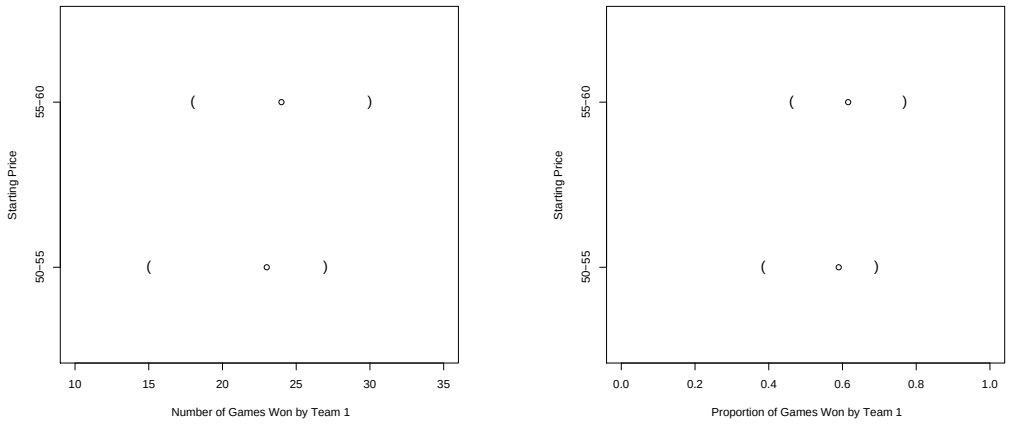

Figure 1: Observed counts (left) and proportions (right) of games won by Team 1 for starting price groups $[50, 60)$, $[60, 70)$. (Note that the parentheses mark the critical interval.)

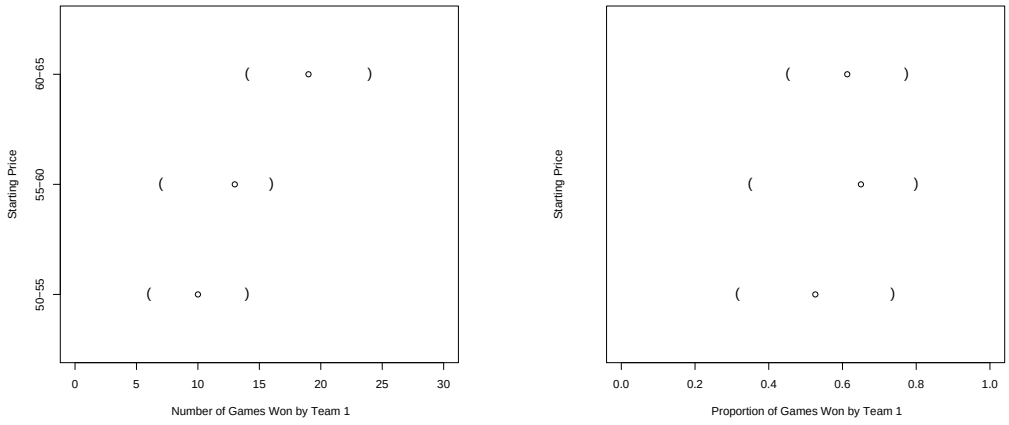

Figure 2: Observed counts (left) and proportions (right) of games won by Team 1 for starting price groups $[50, 55)$, $[55, 60)$, and $[60, 65)$ and their critical intervals.

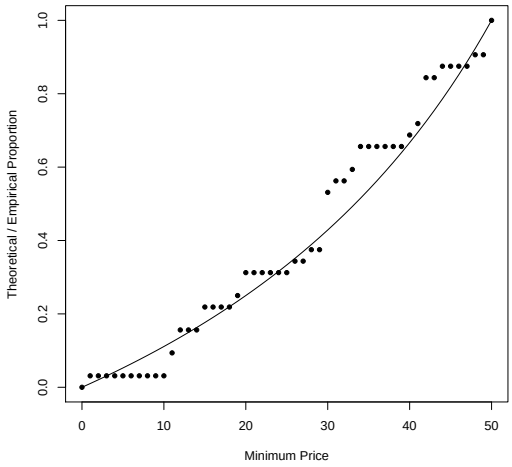

Figure 3: Empirical/theoretical cdf for the minimum (post-50) price conditional on a victory by Team 1. The points are the observed values, while the curve represents the theoretical values.

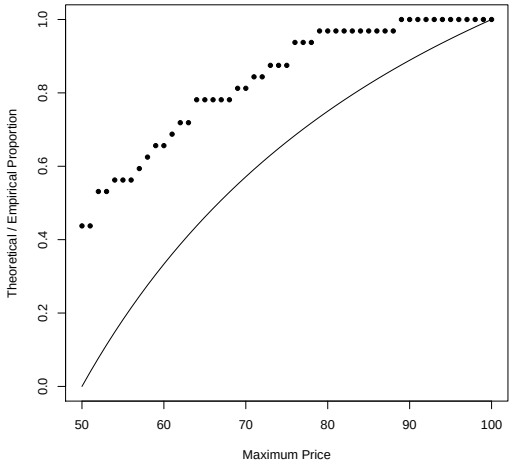

Figure 4: Empirical/theoretical cdf for the maximum (post-50) price conditional on a victory by Team 2.

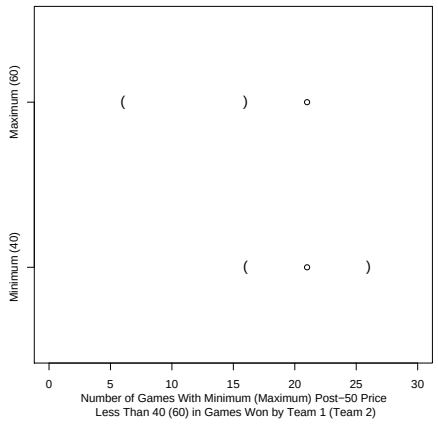

Figure 5: Observed counts of numbers of games won by Team 1 (Team 2, respectively) in which the minimum (maximum) price reached after 50 was below 40 (60) and critical intervals.