Estimating the Frequency Distribution of the Numbers Bet on the California Lottery

Mark Finkelstein*

November 15, 1993

*Department of Mathematics, University of California, Irvine, CA 92717. Running head: Estimating the Numbers Bet on the Lottery. AMS classification, Primary: 62P99, Secondary: 62E99,62G99, 65C99, 65U05. Key words: Lottery, Estimation, Frequency Distribution, Betting Strategy.

Abstract

Several schemes are proposed to estimate the relative frequencies of the individual numbers bet on the (California) lottery. The hypothesis that the small numbers are bet more frequently than the large ones is tested, and evidence supporting the hypothesis is presented. A computation is proposed which would take 6.6 months; instead an approximation to the function involved is used, reducing the computation to 0.2 seconds. A favorable (but impractical) strategy for betting on the lottery is proposed.

On February 15, 1992, as a result of previous games with no winner, the Virginia State Lottery jackpot grew to an amount of $27,007,364, resulting in an unprecedented volume of ticket sales. A consortium of Australian investors came to Virginia in an attempt to purchase all 7,059,052 possible ticket combinations, and were successful in winning the lottery, although they had been able to purchase only roughly 85% of the possible tickets because of logistical difficulties [3]. It was rumored their strategy was based in part on the idea that although they might have to share their winnings with other winning ticket-holders, individual purchasers who select their own numbers tend to purchase the numbers from 1-12, and from 1-31 more heavily than the remaining numbers, as these numbers represent dates: birthdays and other special days; consequently, the distribution of numbers bet would not be not uniform, whereas the distribution of winning numbers is uniform (at least in principle; see below), and hence, one could obtain an advantage.

Stern and Cover [4] studied the Canadian Lotto "6/49" (choose 6 numbers from the integers from 1 to 49), which publishes the distribution of numbers selected by ticket-purchasers weekly, and they demonstrate that a bet on the 6 least popular numbers has a positive expectation because the distribution of numbers bet is not uniform.

We study herein the California State Lottery [1], for which the distribution of the numbers bet is not published. We will estimate this distribution instead, to address the question of whether the smaller numbers are bet more heavily, or whether the numbers are bet uniformly.

A "ticket" on the (California) lottery (Super Lotto) is the selection and purchase of a combination of 6 integers (sextet) chosen without replacement from the integers $\{1, 2, ..., 51\}$. The $\left(\begin{array}{c} 51 \\ 6 \end{array} \right)$ possible sextets will be indexed by the variable $i = \{n_1 < n_2 < \ldots < n_6\}$, the sextet of numbers, and occasionally to emphasize this we will write $\vec{\imath}$ for i. The "play" on

the i-th sextet in the n-th game of the lottery is the sum of all the \$1.00 tickets purchased on the sextet i. Noting that some individuals decide on some random basis to play or not to play the lottery on the n-th game, and some individuals choose their tickets by picking the individual numbers in them, while others allow the vending machine to select a ticket "at random" (meaning uniformly over all sextets i), we denote the play on the sextet i in the n-th game by the non-negative integer-valued random variable X_i^n. We assume that for each i, $\{X_i^1, X_i^2, \ldots\}$ are independent, identically distributed (F_{X_i}). (We make some remarks about removing this assumption at the end of the paper.)

The winning sextet for the n-th game is a selection drawn at random by the Lottery Commission, uniformly over all sextets i; we denote the winning sextet by the random sextet S^n. It is also clear that $\{S^1, S^2, \ldots\}$ are i.i.d. (F_S). We assume that since the process by which the winning sextet is drawn is literally "balls drawn without replacement from an urn," for each i and each n, S^n is i.i.d. uniform and independent of $\{X_i^1, X_i^2, \ldots\}$. (The assumption of "equally likely outcomes" may not be met in practice. Johnson and Klotz [2] conclude otherwise in their study of the Multi-State Lottery.) The number of winners (i.e. purchasers of a winning ticket) on the n-th game is $X_{S^n}^n$.

We wish to know the frequency distribution of the numbers bet, i.e. the numbers $\{p_1, \ldots, p_{51}\}$, $\sum_{j=1}^{51} p_j = 6$, as each bet consists of selecting 6 numbers. Note that

$$p_j = \frac{\sum_{\{\vec{i}: j \in \vec{i}\}} EX_{\vec{i}}}{\sum_{\vec{i}} EX_{\vec{i}}} = \frac{\sum_{\vec{i}} EX_{\vec{i}} I_{[j \in \vec{i}]}}{\sum_{\vec{i}} EX_{\vec{i}}}$$

This information is not available directly, as the Lottery Commission will not divulge this (as opposed to the Canadian Lottery [4].) We propose to infer this distribution based on publicly available information, namely: for each game of the Lottery, the number of tickets

sold, the winning sextet, the number of tickets purchased with the same 6 numbers as the winning sextet, and the numbers of tickets purchased with 5, 4 or 3 numbers in common with the winning sextet, respectively. We prove below that the estimators of these 51 numbers based upon the number of jackpot winners (6 out of 6), or the estimators based upon 5, 4 or 3 out of 6, respectively, converge almost surely to the respective parameters, and then make some comments about our ability to make such inference in practice. These four sets of estimators will prove not sufficiently sensitive to produce the results we seek, and we will then introduce another method of estimating the parameters which will yield improved results.

For each of the N games, observe $< S^n, X^n_{S^n} >, \;\; n = 1, \ldots, N$. Let

$$\hat{p}^N_j = \frac{\sum_{n=1}^N X^n_{S^n} I_{[j \in S^n]}}{\sum_{n=1}^N X^n_{S^n}} / \frac{\sum_{n=1}^N I_{[j \in S^n]}}{\frac{6}{51}N}, \quad j = 1, \ldots, 51.$$

Note that the denominator(s) are eventually positive. The rationale for these estimators is as follows: $\sum_{n=1}^N X^n_{S^n} I_{[j \in S^n]} / \sum_{n=1}^N X^n_{S^n}$ represents the observed proportion of winners who selected the number j in their sextet; in the limit the winning sextets will be uniformly distributed over all sextets. The quantity $\sum_{n=1}^N I_{[j \in S^n]} / \frac{6}{51}N$ is an adjustment to reflect the fact that, for small values of N, the observed proportion of winners who selected the number j in their sextet is weighted toward those j's which were actually selected in winning sextets.

From the Strong Law of Large Numbers,

$$\begin{aligned}
(1/N) \sum_{n=1}^N X^n_{S^n} I_{[j \in S^n]} &\stackrel{a.s}{\to} E(X_S I_{[j \in S]}) \\
&= E(E(X_S I_{[j \in S]} | S)) \\
&= \sum_{\vec{i}} E(X_S I_{[j \in S]} | S = \vec{i}) P[S = \vec{i}] \\
&= \sum_{\vec{i}} E(X_{\vec{i}} I_{[j \in \vec{i}]} | S = \vec{i}) P[S = \vec{i}]
\end{aligned}$$

which, by the independence of S and X_i

$$= \sum_{\vec{\imath}} E(X_{\vec{\imath}} I_{[j\in\vec{\imath}]}) / \begin{pmatrix} 51 \\ 6 \end{pmatrix}$$

Similarly, we have that

$$(1/N)\sum_{n=1}^{N} X_{S^n}^n \stackrel{a.s.}{\to} E(X_S)$$

$$= \sum_{\vec{\imath}} E(X_{\vec{\imath}}) / \begin{pmatrix} 51 \\ 6 \end{pmatrix}$$

and

$$(1/N)\sum_{n=1}^{N} I_{[j\in S^n]} \stackrel{a.s}{\to} EI_{[j\in S]} = \tfrac{6}{51}.$$

Thus, we have that

$$\hat{p}_j^N \stackrel{a.s.}{\to} p_j, \quad j = 1, \ldots, 51.$$

Although this procedure will produce estimators which converge a.s. to the desired parameters, they do not appear to be very efficient. Note that while the quantity $\sum_{n=1}^{N} X_{S^n}^n$, the total number of lottery winners in the N games, increases a.s. to infinity, it is clear that the rate is slow (currently averaging about 1/2 winner per game.)

We can make use of additional information to obtain some alternative estimators. The lottery also pays lesser amounts to the ticket-holders whose tickets have 5, 4, or 3 numbers in common with the winning sextet, and publishes the number of such ticket-holders.

Let $T_i^{(5)}$ denote the set of all sextets from $\{1, 2, \ldots, 51\}$ which have exactly 5 numbers in common with the sextet i (and similarly, define $T_i^{(4)}$, $T_i^{(3)}$). There are $\begin{pmatrix} 6 \\ 1 \end{pmatrix} \cdot 45 = 270$ such sextets. The number of ticket-holders whose tickets have exactly 5 numbers in common with the winning sextet S^n is therefore

$$\sum_{k\in T_{S^n}^{(5)}} X_k^n.$$

Then, as we did above, for each of the N games, observe $< S^n, \sum_{k \in T^{(5)}_{S^n}} X^n_k >$. Again, as above, we denote

$$\hat{p}^{(5)N}_j = \frac{\sum_{n=1}^N \sum_{k \in T^{(5)}_{S^n}} X^n_k I_{[j \in S^n]}}{\sum_{n=1}^N \sum_{k \in T^{(5)}_{S^n}} X^n_k} / \frac{\sum_{n=1}^N I_{[j \in S^n]}}{\frac{6}{51} N}$$

($p^{(4)N}_j$ and $p^{(3)N}_j$ are similarly defined, using $T^{(4)}_i$ and $T^{(3)}_i$.)

As before, from the Strong Law of Large Numbers,

$$\begin{aligned}
(1/N) \sum_{n=1}^N \sum_{k \in T^{(5)}_{S^n}} X^n_k I_{[j \in S^n]} &\stackrel{a.s.}{\to} E\left(\sum_{k \in T^{(5)}_S} X_k I_{[j \in S]} \right) \\
&= \sum_{\vec{i}} \sum_{k \in T^{(5)}_i} E(X_k I_{[j \in \vec{i}]}) / \binom{51}{6} \\
&= \sum_{\{\vec{i}: j \in \vec{i}\}} \sum_{k \in T^{(5)}_i} E X_k / \binom{51}{6},
\end{aligned}$$

while

$$\begin{aligned}
(1/N) \sum_{n=1}^N \sum_{k \in T^{(5)}_{S^n}} X^n_k &\stackrel{a.s.}{\to} E\left(\sum_{k \in T^{(5)}_S} X_k \right) \\
&= \sum_{\vec{i}} \sum_{k \in T^{(5)}_i} E X_k / \binom{51}{6}
\end{aligned}$$

which from the symmetry of the summation

$$= 270 \sum_{\vec{i}} E X_i / \binom{51}{6}, \quad (1)$$

as will be formally verified below, together with (2) and (3). Hence we obtain

$$\hat{p}^{(5)N}_j \stackrel{a.s.}{\to} \frac{\sum_{\{\vec{i}: j \in \vec{i}\}} \sum_{k \in T^{(5)}_i} E X_k}{270 \sum_{\vec{i}} E X_i}.$$

To analyze the numerator of this expression, when $j \in \vec{i}$, for fixed i, j, write

$$T^{(5)}_i = A^j_i \cup B^j_i,$$

where A^j_i consists of the 225 sextets which have exactly 5 numbers in common with i, one of which is j, and where B^j_i consists of the 45 sextets which have exactly 5 numbers in common

with i, and do not contain j. From the symmetry of the summation,

$$\sum_{\{\vec{\imath}:j\in\vec{\imath}\}}\sum_{k\in A_i^j} EX_k = 225\sum_{\{\vec{\imath}:j\in\vec{\imath}\}} EX_i. \tag{2}$$

Noting that $\#\{i : j \in i\}\times\#\{B_i^j\} = 6\#\{i : j \notin i\}$ and again the symmetry of the summation,

$$\sum_{\{\vec{\imath}:j\in\vec{\imath}\}}\sum_{k\in B_i^j} EX_k = 6\sum_{\{\vec{\imath}:j\notin\vec{\imath}\}} EX_i. \tag{3}$$

Thus

$$\sum_{\{\vec{\imath}:j\in\vec{\imath}\}}\sum_{k\in T_i^{(5)}} EX_k = 219\sum_{\{\vec{\imath}:j\in\vec{\imath}\}} EX_i + 6\sum_{\vec{\imath}} EX_i$$

from which we conclude that

$$p_j^{(5)N} \overset{a.s.}{\to} \tfrac{73}{90}p_j + \tfrac{1}{45}.$$

Analogous definitions of $\hat{p}_j^{(4)N}$ and $\hat{p}_j^{(3)N}$ and similar reasoning lead to

$$p_j^{(4)N} \overset{a.s.}{\to} \tfrac{28}{45}p_j + \tfrac{2}{45}$$

and

$$p_j^{(3)N} \overset{a.s.}{\to} \tfrac{13}{30}p_j + \tfrac{1}{15}.$$

Verification of (1), (2) and (3):

Observe that

$$k \in T_i^{(5)} \iff i \in T_k^{(5)}.$$

Then

$$\begin{aligned}
\sum_{\vec{\imath}}\sum_{\vec{k}\in T_i^{(5)}} EX_{\vec{k}} &= \sum_{\vec{\imath}}\sum_{\vec{k}} EX_{\vec{k}} I_{[\vec{k}\in T_i^{(5)}]} \\
&= \sum_{\vec{k}}\sum_{\vec{\imath}} EX_{\vec{k}} I_{[\vec{k}\in T_i^{(5)}]} \\
&= \sum_{\vec{k}}\sum_{\vec{\imath}} EX_{\vec{k}} I_{[\vec{\imath}\in T_k^{(5)}]}
\end{aligned}$$

$$\begin{aligned} &= \sum_{\vec{k}} EX_{\vec{k}} \left(\sum_{\vec{\imath}} I_{[\vec{\imath} \in T_k^{(5)}]} \right) \\ &= 270 \sum_{\vec{k}} EX_{\vec{k}} \\ &= 270 \sum_{\vec{\imath}} EX_{\vec{\imath}}. \end{aligned}$$

This establishes (1). To verify (2), observe that

$$[j \in \vec{\imath}, \vec{k} \in A_i^j] \iff [j \in \vec{k}, i \in A_k^j].$$

Then

$$\begin{aligned} \sum_{\{i:j \in i\}} \sum_{k \in A_i^j} EX_k &= \sum_{\vec{\imath}} \sum_{\vec{k}} EX_{\vec{k}} I_{[j \in i]} I_{[k \in A_i^j]} \\ &= \sum_{\vec{k}} \sum_{\vec{\imath}} EX_{\vec{k}} I_{[j \in \vec{k}]} I_{[i \in A_k^j]} \\ &= 225 \sum_{\vec{k}} EX_{\vec{k}} I_{[j \in \vec{k}]} \\ &= 225 \sum_{\{i:j \in i\}} EX_i. \end{aligned}$$

This establishes (2). To verify (3), observe that

$$[j \in \vec{\imath}, \vec{k} \in B_i^j] \iff [j \in \vec{\imath}, j \notin \vec{k}, \vec{k} \in T_i^{(5)}] \iff [j \in \vec{\imath}, j \notin \vec{k}, i \in T_k^{(5)}],$$

from which it follows that

$$\begin{aligned} \sum_{\vec{\imath}} \sum_{\vec{k}} EX_{\vec{k}} I_{[j \in \vec{\imath}]} I_{[k \in B_i^j]} &= \sum_{\vec{k}} \sum_{\vec{\imath}} EX_{\vec{k}} I_{[j \in \vec{\imath}]} I_{[j \notin k]} I_{[i \in T(5)_k]} \\ &= \sum_{\vec{k}} \left(EX_{\vec{k}} \right) I_{[j \notin \vec{k}]} \sum_{\vec{\imath}} I_{[j \in \vec{\imath}]} I_{[i \in T_k^{(5)}]} \end{aligned}$$

and when $[j \notin \vec{k}]$,

$$\sum_{\vec{\imath}} I_{[j \in \vec{\imath}]} I_{[\vec{\imath} \in T_k^{(5)}]} = 6.$$

The Analysis

At its inception, the California Lottery was in the format "pick 6 numbers from the integers from 1 to 49" ("6/49"), which was subsequently changed, to "6/53", and finally on December 18, 1991, to "6/51". The data for this study were taken from the $N = 176$ games in the current format, from December 18, 1991 through August 21, 1993 [1, pp.122-144].

The estimators $\hat{p}_j^N, \hat{p}_j^{(5)N}, \hat{p}_j^{(4)N}$, and $\hat{p}_j^{(3)N}$ were computed, with results shown in Table 1. To test the hypothesis that the numbers 1-12, and the numbers 1- 31 are bet more heavily than the numbers above 31, we pooled the estimators $\hat{p}_j^N$ for $1 \leq j \leq 12, 13 \leq j \leq 31$, and $32 \leq j \leq 51$, obtaining means 0.1146, 0.1169, and 0.1156, respectively, which clearly do not support the hypothesis. For each of the estimators $\hat{p}_j^{(5)N}, \hat{p}_j^{(4)N}$, and $\hat{p}_j^{(3)N}$ we also pooled the estimators for groups 1-12, 13-31, and 32-51. The means for the four pooled estimators, $\hat{p}^{(3)N}, \hat{p}^{(4)N}, \hat{p}^{(5)N}, \hat{p}^N$, are displayed in the first four columns of Table 2. Applying the Kruskal-Wallis rank sum test (in these cases the statistic is distributed approximately χ^2, $2d.f.$) to each of these pooled estimators, we see that we cannot reject the null hypothesis, that the three sets of numbers come from the same distribution.

As a further test of the hypothesis that the numbers bet are uniformly distributed, we calculated the statistic

$$\sum_{j=1}^{51} (\hat{p}_j^N - \tfrac{6}{51})^2 = 0.09799. \tag{4}$$

Since the distribution of this statistic is not elementary, we performed a simulation of $N = 176$ games of the lottery. We took the ticket sales in each game as the published ticket sales for that game, and generated a random number of winners for each game according to a Poisson distribution with parameter $1/\begin{pmatrix} 51 \\ 6 \end{pmatrix} = 1/18009460$. We then re-computed the

statistic (4). We repeated the simulation 1000 times, and found that the value 0.09799 was exceeded 208 times, and so conclude that using this test we cannot reject the hypothesis that the numbers are bet uniformly.

Since we failed in the above tests to reject the hypothesis of uniform p_j's, we shall now consider an alternate method of estimating them. Given values of $\{p_1, \ldots, p_{51}\}$, assume that bettors purchase tickets by selecting 6 numbers without replacement from the integers from 1 to 51 randomly, independently and weighted according to $\{p_1, \ldots, p_{51}\}$. Then, for the n-th game, given the 6 winning numbers observed, an individual bettor's probability of selecting a ticket with exactly 3 out of 6 numbers in common with the winning sextet is p, which depends only upon the winning sextet and $\{p_1, \ldots, p_{51}\}$. If the total wager on the n-th game is W_n, then it follows that the number of \$5.00 winners (holders of tickets with exactly 3 out of 6 numbers in common with the winning sextet) will be binomially distributed, and since W_n is large (approx. $5 \cdot 10^6$) and $p \approx 0.01$, the observed number of \$5.00 winners, suitably normed and centered, will be distributed $\mathcal{N}(0, 1)$, standard normal, a z-score. These 176 z-scores (denoted $z_1, \ldots, z_{176}$) will have a sum of squares S which will be distributed chi-square. This defines a function $S = S(p_1, \ldots, p_{51})$. We define our estimators $\hat{p}_j$ as that set of p_j's which minimize S. These can be found numerically by a search in the 50-dimensional simplex: $\sum p_j = 1$ intersected with the positive orthant of 51-dimensional p-space, and then multiplying each p_j by 6 so that $\sum p_j = 6$.

This computation and search procedure is easier said than done, however. Putting aside the issue of the search, let us consider merely the issue of computing the value of $S(p_1, \ldots, p_{51})$, given $\{p_1, \ldots, p_{51}\}$. To compute each z-score, we need to compute, given the winning six numbers, the probability of selecting a ticket with exactly 3 out of 6 in common

with the winning ticket. There are $\binom{6}{3}\binom{45}{3}$ such tickets. Each can occur in $6! = 720$ possible orders. We must repeat this calculation for 176 z-scores. The probability of selecting $i_1, \ldots, i_6$ in exactly that order is

$$p_{i_1} \frac{p_{i_2}}{1 - p_{i_1}} \cdots \frac{p_{i_6}}{1 - p_{i_1} - \ldots - p_{i_5}}$$

which requires 25 arithmetic operations. Thus, one evaluation of $S(p_1, \ldots, p_{51})$ requires $\binom{6}{3}\binom{45}{3}(6!) \cdot 176 \cdot 25 = 9 \cdot 10^{11}$ arithmetic operations.

All hope is not lost, however, for if we were selecting 6 numbers from M numbers, then clearly as M tends to infinity, the probabilities of ticket selection for a sampling-with-replacement scheme will converge to the probabilities of ticket selection for sampling-without-replacement. Since $M = 51$ in our case, we will approximate the function S by a "with replacement" scheme. To calculate the probability of selecting a ticket of 6 numbers with exactly 3 numbers in common with the winning ticket, given $\{p_1, \ldots, p_{51}\}$, we calculate $\sum p_i p_j p_k$, summed over the $\binom{6}{3}$ triplets $< i, j, k >$ chosen from the winning sextet, multiplied by a constant chosen to make probabilities sum to 1. From this, z_n, and hence S, is easily computed. (The evaluation of S now takes (40 multiplications + 19 additions)$\cdot 176 = 10,384$ arithmetic operations, a factor of $8.5 \cdot 10^7$ improvement.) Computing S in this way it is feasible to search the 50-dimensional simplex $\{(p_1, \ldots, p_{51}) : \sum p_j = 1, p_1 \geq 0, \ldots, p_{51} \geq 0\}$, minimizing S one variable at a time, and cycling through all 51 variables repeatedly until a (local) minimum of S is found. (Using a step size of $p_j/100$, this search requires 1910 evaluations of S. If we had calculated the exact probabilities by sampling without replacement, our search would have used $1910 \cdot 9 \cdot 10^{11} = 1.7 \cdot 10^{15}$ arithmetic operations, which on a 100 MIPS machine would require about 6.6 months of machine time, as opposed to

$1910 \cdot 10384 \approx 2 \cdot 10^7$ operations, or 0.2 seconds of machine time.) Observation of various 1-dimensional sections of S suggest that this minimum is actually a global minimum. The values of p_j at this local minimum, each multiplied by 6 so that $\sum p_j = 6$, are denoted $\hat{p}_j$ in Table 1. The indices of the order statistics of the $\hat{p}_j$ are shown in column 2 of Table 1. The most popular numbers are 9, 7, 3, 8, 11, and 6, which bear considerable similarity to the most popular numbers purchased in the Canadian lottery (reported as 3, 7, 9, 11, 25, and 27 in [4]). The means of the sets $\{\hat{p}_1, \ldots, \hat{p}_{12}\}$, $\{\hat{p}_{13}, \ldots, \hat{p}_{31}\}$, and $\{\hat{p}_{32}, \ldots, \hat{p}_{51}\}$ are 0.1314, 0.1194, and 0.1074 respectively, and are displayed in Table 2. The Kruskal-Wallis rank sum test applied to these 3 sets produce a statistic whose value is is 37.18, and whose distribution is approximately χ^2, 2 d.f., and we therefore reject the hypothesis that the three sets of p's are all observations from the same distribution at all significance levels. Thus we have a viable method for estimating the numbers bet on the Lottery which is sensitive enough to reveal the non-uniformity in the numbers $\{p_1, \ldots, p_{51}\}$.

As a check of the method, we performed the following simulation. We generated random winning sextets equiprobably over all sextets, for each of the 176 games. We assumed the true values of the p's were as follows: $p_1 = p_2 = \ldots = p_{12} = 0.1314$, $p_{13} = \ldots = p_{31} = 0.1194$, $p_{32} = \ldots = p_{51} = 0.1074$. For each of the 176 games, using these p-values, the "winning" sextet, and the observed total number of tickets purchased, we generated a random number of \$5.00 winning tickets, according to the binomial distribution discussed above, but with greater variance. Then we performed the same minimization search procedure discussed above, starting at equiprobable p's. The set of minimizing p's is denoted $\tilde{p}_j$ in Table 1 and in Table 2. For the minimizing set, the means of the $\tilde{p}$'s for 1-12, 13-31, and 32-51 were 0.1304, 0.1195, and 0.1081 respectively, in good agreement with the assumed p_j's.

Now that we have the p_j's in hand, let us explore whether we can exploit the non-uniformity in betting patterns to overcome the "house advantage": If we bet the 6 least likely $\hat{p}$'s, we observe from Table 1 that these numbers are 2.52 times as unlikely to be selected by other bettors as the "average" numbers are, and these in turn are 2.33 times as unlikely to be chosen as the 6 most likely p's. With this information, we can compute our expected returns:

Given that we have won the jackpot (an event which occurs in any case with probability 1/18009460) our expected payoff depends upon the number of other winners with whom we have to share. If the jackpot is J, the expected payoff is $\frac{J}{1}[\text{Probability of } 0 \text{ other winners}] + \frac{J}{2}[\text{Probability of } 1 \text{ other winner}] + \ldots = \frac{J}{1}[e^{-\lambda}\frac{\lambda^0}{0!}] + \frac{J}{2}[e^{-\lambda}\frac{\lambda^1}{1!}] + \ldots = \frac{J}{\lambda}(1 - e^{\lambda})$ and so our expected return is $\frac{J}{\lambda}(1 - e^{\lambda})/18009460$, where λ is the rate parameter obtained by multiplying the probability of selecting for purchase the winning sextet by the number of tickets purchased. It is clear, then, that the advantage, if any, comes from holding a winning ticket when λ is as small as possible. For the data we have, in the case of purchasing the 6 least popular numbers, 50, 46, 49, 43, 48, and 51, we find the expected return for a \$1.00 bet to be \$0.42, while in the case of purchasing the 6 most popular numbers, 9, 7, 3, 8, 11, and 6, the expected return is \$0.27. Choosing "average" numbers has an expected return of \$0.36. Thus, unfortunately, we cannot gain an advantage from this strategy, if we bet on *every* game of the Lottery. This makes sense, intuitively: even if we were to find 6 numbers that no one else bet, the probability of our winning would be only 1/18009460. It is clear that this is a losing proposition, unless the payout from the lottery is at least \$18,009,460.

Suppose we modify our strategy, then, by playing the 6 least popular numbers but only entering the lottery when the payoff is at least \$18 million (which occurs once every 7 weeks,

on average). In this case, based on the data we have, our expected returns for a \$1.00 bet are as follows: for the 6 least popular numbers, \$1.14; for 6 "average" numbers, \$0.89; and for the 6 most popular numbers, \$0.59. Thus, a favorable strategy appears to be at hand; the only drawback to it (other than its widespread adoption) is the expected waiting time for a win: 2.3 million years.

Notes: Recall that we assumed that $\{X_i^1, X_i^2, \ldots\}$ are independent, identically distributed (F_{X_i}). This assumption is not met in practice, as we observe that more tickets are purchased in games in which the payoff is large. These games occur after a sequence of games in which there is no winner. We could, however, normalize these random variables by dividing each by the total wager in that game, obtaining the random variables $\{X_i^1/\sum_k X_k^1, X_i^2/\sum_k X_k^2, \ldots\}$ which (under the assumption that bettors do not change their betting strategy according to the size of the jackpot, but rather only whether to bet or not) would be i.i.d., and we could then repeat the analysis we did with these random variables instead. The results would be the same.

Thanks to Professors George McCarty, Howard Tucker, Jerry Veeh, and Robert Whitley for extensive discussions and many helpful suggestions all along the way.

REFERENCES

[1] Lotto Scorecard, California Lottery Commission, September 1, 1993.

[2] R. Johnson and J. Klotz (1993), Estimating Hot Numbers and Testing Uniformity for the Lottery, *J.A.S.A.* **88**, 662-668.

[3] *New York Times*, February 25, 1992, pp. A1, A18.

[4] H. Stern and T. M. Cover (1989), Maximum Entropy and the Lottery, *J.A.S.A.* **84**, 980-985.

TABLE 1

j	(j)	$\sum_{n=1}^{N} I_{[j\in S^n]}$	$\sum_{n=1}^{N} X^n_{S^n} I_{[j\in S^n]}$	$\hat{p}_j^{(3)N}$	$\hat{p}_j^{(4)N}$	$\hat{p}_j^{(5)N}$	$\hat{p}_j^{N}$	$\tilde{p}_j$	$\hat{p}_j$
1	9	17	5	0.119	0.123	0.123	0.068	0.133	0.129
2	7	21	9	0.110	0.110	0.109	0.099	0.130	0.130
3	3	24	11	0.129	0.132	0.135	0.105	0.130	0.135
4	8	21	11	0.117	0.116	0.115	0.121	0.130	0.127
5	11	27	22	0.154	0.154	0.151	0.187	0.130	0.130
6	6	21	11	0.154	0.156	0.158	0.121	0.128	0.131
7	2	20	7	0.134	0.138	0.136	0.081	0.133	0.135
8	5	29	20	0.104	0.102	0.104	0.159	0.128	0.134
9	1	16	4	0.106	0.108	0.108	0.058	0.133	0.143
10	12	20	7	0.117	0.116	0.112	0.081	0.126	0.122
11	19	27	17	0.107	0.110	0.115	0.145	0.131	0.133
12	16	18	12	0.111	0.113	0.115	0.153	0.133	0.129
13	4	17	12	0.121	0.127	0.129	0.162	0.123	0.126
14	26	22	10	0.102	0.107	0.111	0.105	0.118	0.117
15	13	26	12	0.113	0.115	0.116	0.106	0.121	0.121
16	21	21	11	0.123	0.121	0.124	0.121	0.118	0.129
17	27	20	20	0.153	0.153	0.164	0.230	0.122	0.121
18	22	21	14	0.125	0.128	0.135	0.153	0.116	0.121
19	10	22	15	0.112	0.115	0.119	0.157	0.119	0.129
20	28	29	13	0.104	0.103	0.099	0.103	0.121	0.113
21	15	23	14	0.108	0.110	0.112	0.140	0.121	0.125
22	17	20	5	0.107	0.108	0.109	0.058	0.119	0.122
23	18	17	9	0.139	0.136	0.139	0.122	0.118	0.112
24	25	19	7	0.103	0.100	0.099	0.085	0.119	0.110
25	33	22	10	0.111	0.111	0.105	0.105	0.118	0.119
26	14	18	8	0.137	0.139	0.141	0.102	0.118	0.127
27	42	23	12	0.115	0.117	0.115	0.120	0.119	0.124
28	37	20	13	0.133	0.134	0.134	0.150	0.118	0.121
29	45	19	5	0.107	0.105	0.105	0.061	0.119	0.110
30	30	19	3	0.096	0.096	0.089	0.036	0.122	0.113
31	20	26	12	0.126	0.123	0.122	0.106	0.119	0.112
32	32	20	5	0.106	0.103	0.098	0.058	0.106	0.112
33	23	18	7	0.111	0.108	0.113	0.089	0.107	0.118
34	31	21	8	0.108	0.107	0.106	0.088	0.107	0.107
35	44	18	15	0.125	0.128	0.128	0.192	0.107	0.105
36	29	19	14	0.139	0.136	0.134	0.170	0.106	0.108
37	24	18	11	0.154	0.152	0.153	0.141	0.106	0.115
38	39	14	10	0.106	0.106	0.105	0.164	0.104	0.107
39	36	13	9	0.100	0.103	0.103	0.159	0.108	0.110
40	47	14	1	0.101	0.098	0.101	0.016	0.112	0.105
41	34	16	6	0.112	0.111	0.109	0.086	0.112	0.107
42	38	20	14	0.139	0.139	0.141	0.161	0.108	0.116
43	41	24	7	0.095	0.092	0.092	0.067	0.108	0.103
44	35	22	9	0.110	0.103	0.104	0.094	0.112	0.111
45	40	18	6	0.128	0.126	0.128	0.077	0.114	0.115
46	51	21	12	0.105	0.104	0.102	0.131	0.109	0.098
47	48	24	19	0.111	0.110	0.105	0.182	0.106	0.108
48	43	28	12	0.111	0.111	0.107	0.099	0.108	0.103
49	49	23	15	0.127	0.125	0.121	0.150	0.107	0.103
50	46	28	16	0.111	0.108	0.107	0.131	0.109	0.096
51	50	12	3	0.100	0.100	0.100	0.058	0.109	0.104

TABLE 2						
Group	$\hat{p}^{(3)N}$	$\hat{p}^{(4)N}$	$\hat{p}^{(5)N}$	$\hat{p}^{N}$	$\tilde{p}$	$\hat{p}$
$1 \leq j \leq 12$	0.1218	0.1237	0.1235	0.1146	0.1304	0.1314
$13 \leq j \leq 31$	0.1179	0.1185	0.1193	0.1169	0.1195	0.1194
$32 \leq j \leq 51$	0.1149	0.1136	0.1128	0.1156	0.1081	0.1074
Kruskal-Wallis test statistic	1.62	3.52	4.50	0.07	43.76	37.18
Signficance probability (P-value)	0.45	0.33	0.11	0.97	10^{-9}	10^{-8}