Efficiency of betting markets and rationality of players: evidence from the french 6/49 Lotto

Patrick ROGER[*], Marie-Hélène BROIHANNE[†]

September 2003

Abstract

We analyze the existence of preferred numbers on the french lotto market and prove that this market is not strongly efficient in the sense of Thaler-Ziemba (1988). The preference for low numbers is first put to light by means of stochastic dominance tests. The specific features of the french lotto game allow us to build a simple estimate of the probability distribution of numbers actually played. The results are compared to the (highly time-consuming) maximum likelihood estimator used by Farrell *et al.(2000)*. It is shown that the two methods give very close results.

1 Introduction

Pari mutuel betting markets such as the Lotto game are well-suited to investigate efficiency issues and the rationality of players. The essential reason is that pari mutuel gamblers bet against each other; the expected return in such a game not only depends on the realized state of nature, that is to say, the result of the draw, but also on two other variables. First, the total amount invested by bettors in a given draw influences the expected return since a fixed part of the stakes returns to players. Second, the numbers chosen by the other bettors determine the return of a given grid because prizes are shared among all winners at a given rank.

If we assume that numbers are randomly drawn by the sponsor of the game, the distribution of combinations selected by the players should be also randomly drawn

[*]Corresponding author: University Louis Pasteur, PEGE, 61 Avenue de la Forêt Noire, 67085 Strasbourg CEDEX, France. roger@cournot.u-strasbg.fr. We are grateful to Gauthier Lanot who provided some of the gauss estimations programs. We also thank Roberto Garvia for interesting discussions on the history of lotteries and the participants of seminars in Strasbourg for useful suggestions.

[†]University Louis Pasteur, PEGE, 61 Avenue de la Forêt Noire, 67085 Strasbourg CEDEX, France. mhb@cournot.u-strasbg.fr.

at the aggregate level of the population of players. If it was not the case, a simple strategy would be to pick unpopular numbers to get an abnormal expected return. It is necessary to distinguish between positive and abnormal expected returns. In fact, obtaining a positive expected return in such a game is very difficult, if not impossible, due to a usually very high take-out rate (about 50%) in most countries.

Thaler and Ziemba (1988) modified the definition of efficiency to take into account the fact that non monetary incentives can partially justify the players' participation, despite the high take-out rate. They define the weak form of efficiency as the impossibility to obtain a positive expected return. The market is said strongly efficient if all combinations generate the same expected return, that is one minus the take-out rate.

In this paper, we analyze the efficiency of the french lotto market. There is some evidence that the game is not weakly efficient; however, as players must participate to two successive drawings on a given day, it is not clear if inefficencies can really be profitable. We also prove that strong efficiency is clearly rejected. In other words, the french lotto market exhibits the same inefficiencies as the ones pointed out in many other countries[1]; some numbers are clearly preferred by players and others are largely unpopular. It also confirms the experimental results presented by Boland and Pawitan (1999) showing the difficulty for people to choose numbers randomly.

The french lotto market has some specific features; in particular, the rules governing rollovers are designed to prevent "lottomanias", as depicted in Beenstock and Haitovsky (2001). Moreover, since 1997, there are seven ranks of gains with three ranks including the bonus number. This feature allows us to define a simple way to estimate the probability distribution of numbers chosen by players. We compare these estimates with the ones provided by sophisticated econometric methods, using importance sampling, presented in Farrell *et al,* and show that the two methods give close results.

The paper is organized as follows. Section 2 describes the evolution of the rules of the french lotto game, especially the ones concerning the sharing of prizes and the rollover rules. Section 3 contains the theoretical part of the work in which we present the estimation methods. Section 4 contains our empirical results. It starts by the description of the data our empirical analysis is based on ; then, preliminary tests concerning the efficiency issue. We then develop the results obtained with the maximum likehood estimator and the "bonus number" estimator. Finally, the weak form of efficiency is tested by using a simple econometric model allowing us to define popular and unpopular numbers; this approach leads to the valuation of the expected return of the "most profitable" combinations.

We show that choosing unpopular numbers is not clearly sufficient to compensate the large take-out rate, even if some numbers exhibit positive expected returns.

[1]See for example Ziemba *et al.*(1986) for the Canadian lotto, Scott and Gulley (1995) for various U.S lotto games and Papachristou and Karamanis (1998) for the Greek lotto.

2 French lotto rules and publicly available information

2.1 The rules of the game

The french lotto is a usual 6/49 game with a complementary (or bonus) number; it started in may 1976 with one weekly draw on wednesdays. In March 1984 a second draw was introduced on saturdays. Tickets played for the wednesday draw could also participate to the saturday game (obviously with a doubled ticket price). Since september 1990, two draws are performed each wednesday and each saturday.

On 9/13/1992, the "Système Flash" was introduced. It is the french equivalent of the "Lucky Dip" or the "Quick Pick", that is a way to randomly generate lotto tickets for players.

Today, buying a ticket gives the right (and the obligation) to participate to the two draws of one specified day, either wednesday or saturday. For regular draws, no jackpot is announced; it is a pure pari mutuel game. About one half of the stakes returns to the players; the take-out rate being now 49.5%. It was fairly stable through time, varying from 46% at the starting date, to the current level, with a not too surprising positive slope.

The french lotto is now managed by "La Française des Jeux" (referred to as FdJ in the following), a state-owned firm created in 1978[2].

Unfortunately, the rules of the game were periodically modified, especially concerning the sharing of prizes among the winning ranks. At the beginning, there were 5 winning ranks, numbered from 1 to 5, corresponding to 6 to 3 matched numbers plus the possibility 5+B consisting in finding 5 matched numbers plus the bonus number (rank 2). The winning combination 3+B was introduced in 1984 when the second weekly draw started; this winning possibility was only valid on saturday. The corresponding monetary prize was ten times the one for 3 matched numbers. This winning combination disappeared when the number of weekly draws rose to four. Finally, it was reintroduced in October 1997 with another one, the 4+B combination. Now, the seven winning combinations seem to be stabilized. Table 1 provides the winning probabilities according to the period considered. Finally, it must be precised that combinations 3+B and 4+B pay twice as much as the 3 and 4 combinations.

TABLE 1 around here

The price of a ticket was 1 french franc up to 6/19/1996 and doubled by this date, a wednesday. It is interesting to notice, on figure 1 (week 962), that the amount of sales did not jump on this occasion; people instantaneously adjusted the number of tickets played without devoting more money to the game. It means that a lower number of tickets was sold for each draw, then increasing the probability of a rollover.

[2]All the laws governing the game are available on the official website www.sojah.com.

In fact, focusing on the year of change, we observe that the mean number of jackpot winners (rank 1) was 2.49 before the price change (96 draws) and 1.92 after (112 draws). Obviously, this modification is important to evaluate the expected return. An other minor change came at the beginning of 2002 when tickets started to be sold in euros, the price becoming 0.3 euro instead of the french franc equivalent of 0.304 (1 euro=6.55957 french francs).

FIGURE 1 AROUND HERE

Many papers study the influence of rollovers on demand. For the french case, such a study is very difficult for several reasons. First, the rollover rules evolved through time, not always in a clear way. Prizes not won at the first rank were first distributed among the winners of the following ranks. Then, rules changed and 50% of the amount devoted to the first rank of the second draw were carried over to the first rank of the next draw. In 2002, a new rule appeared ; the way amounts not won are shared is decided by the chief executive of the FdJ! In fact, all the prizes not distributed at the second draw of a given day are put in what we call a Rollover Fund, whose official name is "Fonds de Réserve et de Super Cagnotte". Contrary to what happens in many countries, it is not mandatory that the next draws benefit from the rollover. The chief executive of the FdJ decides what to do with the money of the Rollover Fund; more precisely, he decides when future jackpots will be fed by the Fund and how much money will be used in these situations. All these informations become public by means of announcements in the "Journal Officiel de la République Française"; consequently, when players buy tickets, they are aware of extra jackpots. Obviously, there is also advertising on radio and TV when such a case occurs. The most common announced Jackpot in this case is 4 millions euros. Finally, since april 2003, the jackpot prize of the first draw, if not won, is transferred to the two lowest ranks (3 and 3+C).

On the informational point of view, there is then no significant difference between the french lotto and the corresponding game in other countries, if we assume that players immediately integrate all the new information concerning rollovers, maybe a rather strong assumption. In fact, the power given to the chief executive to allocate amounts of the Rollover Fund prevents lottomania, such as the one observed, for example, on the U.S Powerball when 17 rollovers appeared in a row, leading to a $200 million jackpot prize (Hartley *et al., 2000).*

3 Theory and estimation methods

In this section, we present three methods to test the efficiency of the lotto market. The two first are aimed to test strong efficiency by estimating the probability distribution of numbers actually played. The maximum likelihood estimator used by Farrell *et al.* (2000) is briefly described. The implementation of this method needs to simulate combinations by means of importance sampling. It is then highly time-consuming. A specific feature of the french lotto market (three winning ranks including the bonus number) allows us to build a simple estimator of this probability distribution. The third method presented here is aimed to test the possibility of getting positive returns by playing unpopular numbers. It was already used in Ziemba *et al.* (1986) and Papachristou and Karamanis (1998).

3.1 Maximum likelihood

Sponsors of lotto games do not publish the probability distribution of numbers actually played. It must be estimated through the use of public information, consisting in results of official drawings, especially the number of winners at each rank and the individual gains.

In the following, $\mathbf{p}$ refers to the vector of probabilities that numbers are picked first; thereby, the probability that a particular combination is chosen differs from the product of the estimated probabilities of numbers in the combination. This implicit assumption is made in the model of Farrell *et al.*(2000) because they consider that a combination play is equivalent to a draw without replacement of 6 numbers out of 49 with non uniform weights. As this assumption is sometimes observed in practice (the number 7 is chosen for its own luck attributes independently of the other numbers in a combination), there is evidence that players prefer combinations forming sequences or, at least, that they do not choose numbers independently when, for example, they play birthday dates. Using a sample of the actual distribution of combinations for a single draw in the UK National Lottery, Simon (1999) directly showed the extent of conscious selection. This expression was first employed by Clotfelter and Cook (1993) to define non random choices of numbers.

Hereafter, we describe briefly the methodology, details can be found in Farrell *et al.*.

Let $p(i)$ the probability that number i is picked first among the 49 numbers; the $p(i)$ are positive numbers summing to 1. A random choice would mean $p(i) = \frac{1}{49}$ for each number i. The probability of drawing six given ordered numbers $\{a_1, ..., a_6\}$ is :

$$P\left(\mathbf{a};\mathbf{p}\right) = \frac{p(a_1)p(a_2)...p(a_6)}{\Pi_{i=1}^{5}\left(1 - \sum_{j=1}^{i} p(a_j)\right)}$$

The denominator takes into account the “without replacement” feature of the drawing. Let $\boldsymbol{\sigma}$ the selection of these six numbers without order, that is to say, the subset of all ordered sets containing $\{a_1, ..., a_6\}$. The probability of choosing $\boldsymbol{\sigma}$ is then given by :

$$P\left(\boldsymbol{\sigma};\mathbf{p}\right) = \sum_{\rho\in\mathcal{P}(\boldsymbol{\sigma})} \frac{p(a_1)p(a_2)...p(a_6)}{\Pi_{i=1}^{5}\left(1 - \sum_{j=1}^{i} p(\rho(a_j))\right)}$$

where ρ is a permutation of $\mathbf{a}$ and $\mathcal{P}\left(\boldsymbol{\sigma}\right)$ is the set of the 720 permutations of $\boldsymbol{\sigma}$.

A similar expression for subsets of three numbers can be obtained by the following formula :

$$P_3(\boldsymbol{\sigma};\mathbf{p}) = \sum_{\{j_1,\ldots,j_6\}\in\mathcal{S}(3;\boldsymbol{\sigma})} \frac{p(j_1)p(j_2)\ldots p(j_6)}{\Pi_{i=1}^{5}\left(1-\sum_{k=1}^{i} p(j_k)\right)}$$

where $\mathcal{S}(3;\boldsymbol{\sigma})$ is the set of all arrangements of six numbers having exactly three common numbers with $\boldsymbol{\sigma}$.

The parameters $\mathbf{p}$ can then be estimated through the maximization of likelihood, using the numbers of winners at the first (6 correct numbers) and last rank (3 correct numbers) and the number of tickets sold, denoted respectively as N_6^k, N_3^k and Q^k for the k-th drawing. Knowing that our sample contains $m = 3638$ drawings, the likelihood is written as :

$$\sum_{k=1}^{m}\left[N_6^k \ln\left(P(\boldsymbol{\sigma}_k;\mathbf{p})\right) + N_3^k \ln\left(P_3(\boldsymbol{\sigma}_k;\mathbf{p})\right) + \left(Q^k - N_6^k - N_3^k\right)\ln\left(1 - P(\boldsymbol{\sigma}_k;\mathbf{p}) - P_3(\boldsymbol{\sigma}_k;\mathbf{p})\right)\right]$$

As pointed out by Farrell *et al.*, the set $\mathcal{S}(3;\boldsymbol{\sigma})$ contains millions of elements. To obtain unbiased estimates of $P_3(\boldsymbol{\sigma};\mathbf{p})$, Farrell *et al.* apply the technique of importance sampling, that is they use a Monte Carlo simulation method to generate combinations according to the estimated probability distribution. This procedure is briefly described here[3]. For each draw $\boldsymbol{\sigma}_k$, $H(=500)$ combinations without replacement of 3 numbers out of $\boldsymbol{\sigma}_k$ and $H(=500)$ combinations without replacement of 3 numbers out of the complementary of $\boldsymbol{\sigma}_k$ are drawn. Finally, for each draw $\boldsymbol{\sigma}_k$, $H(=500)$ ordered combinations of numbers out of $\{1,2,3,4,5,6\}$ are drawn to determine the positions of the three numbers drawn out of $\boldsymbol{\sigma}_k$ and to define simulated vectors k_h of the set $\mathcal{S}(3;\boldsymbol{\sigma}_k)$. These vectors are used to compute $\hat{P}_3(\boldsymbol{\sigma}_k;\mathbf{p}) = \frac{1}{500}\sum_{h=1}^{500}\frac{P(k_h;\mathbf{p})}{q(k_h)}$ where $q(k_h)$ are probabilities of the simulated vectors k_h. An antithetic variance-reduction argument is also used to give asymptotically consistent estimates.

Though this estimator has interesting convergence properties, the simulation method takes much computer time, in fact several complete days for our complete sample of 3638 drawings (Farrell *et al.* computed estimators with 116 draws). Consequently, we seized the opportunity given by the specific features of the french lotto to build a much more simple estimator. We show in section 4 that the two estimators provide very close results.

3.2 The bonus number estimator

The specific feature of the french lotto market, that is to say, the seven ranks of gains (since october 1997) including three ranks with the bonus number, allows to largely simplify the estimation of the probability distribution of numbers actually played. The essential problem mentioned earlier was that the set $\mathcal{S}(3;\boldsymbol{\sigma})$ is too large to allow complete coverage in the maximum likelihood procedure. In fact, the weakness of

[3] A complete technical explanation can be found in Farrell *et al.'s* Appendix B.

public data relies on the impossibility to infer the exact numbers chosen by a player when the only information you know is that the player notched three correct numbers. However, when 3+B is a winning combination, you know for sure that a winner at this rank selected the bonus number. The intuitive idea in the following is to consider the subset of 3 and 3+B winners as a sample of the population of tickets. The ratio of the number of 3+B-winners and 3-winners, compared to the theoretical ratio arising under the uniform distribution, is a good indicator of the popularity of the bonus number under consideration.

Let Q^k still denote the number of tickets played in the k-th drawing. If we knew the numbers notched by the players, the estimation of $\mathbf{p}$ would be obtained simply by calculating, for each number i the ratio $\frac{M_i^k}{6Q^k}$, M_i^k being the number of tickets containing number i. As $\sum_{i=1}^{49} M_i^k = 6Q^k$, we could get the probability distribution of numbers for a given drawing. The method would be the same in aggregating several successive drawings as long as it is assumed that the probability distribution is stable through time. Unfortunately, the M_i^k are unknown and were indirectly estimated in the preceding section.

Considering the set $\mathcal{N}_{3B}^k$ of winning combinations at ranks 6 and 7, that is the grids matching either 3 numbers or 3 numbers plus the bonus number in drawing k, and denoting $N_{3B}^k = Card\left(\mathcal{N}_{3B}^k\right)$, we can write :

$$N_{3B}^k = N_3^k + N_B^k$$

where N_3^k and N_B^k are the numbers of winning tickets at ranks 7 (subset $\mathcal{N}_3^k$) and 6 (subset $\mathcal{N}_B^k$), respectively.

Let i be the bonus number in drawing k: N_{3B}^k can be rewritten as :

$$N_{3B}^k = q^k(i)N_{3B}^k + (1 - q^k(i))N_{3B}^k$$

where $q^k(i)$ is the proportion of tickets containing i in the subset of winning tickets at ranks 3 and 3+B.

Knowing the N_3^k and N_B^k, we easily obtain $q^k(i)$ when i is the bonus number in drawing k. Under the uniform distribution, the proportion of grids, denoted as α, containing 3+B correct numbers in the subset $\mathcal{N}_3^k \cup \mathcal{N}_B^k$ is $\alpha = \frac{C_{42}^2}{C_{42}^2 + {}_{42}^3} = 6.976\,7 \times 10^{-2}$. We then estimate, $p^k(i)$ by $\widehat{p}^k(i) = \frac{q^k(i)}{49 \times \alpha}$. The ratio $\frac{q^k(i)}{\alpha}$ may be interpreted as a measure of popularity of number i in drawing k and $\frac{1}{49}$ is, obviously, the probability of choosing i first under the uniform distribution.

$p(i)$ is then estimated by :

$$\frac{1}{m_i} \sum_{k=1}^{m} \widehat{p}^k(a^k) \mathbf{1}_{\{a^k = i\}} \tag{1}$$

where m_i is the number of drawings with i as the bonus number and a^k is the bonus number in drawing k.

It must be noticed that this approach could also be used with 4+B and 5+B winning combinations (ranks 4 and 2). However, the number of winners at these ranks is considerably lower than the one observed at the 3+C rank. Consequently, aggregating the three ranks doesn't improve the estimation[4]. This remark also explains why this estimator cannot be used for most lotto games since they only offer the 5+B winning rank, with a low number of winners.

3.3 Estimation of the expected returns of numbers

Numbers selection is seemingly not random and this kind of behavior could generate violations to the weak form of efficiency. As a matter of fact, drawings of unpopular numbers are associated with larger unit payoffs because there exist only a few players who choose these particular combinations. The popular numbers then give rise, when they are drawn, to lower unit payoffs since the pool is shared among numerous winning players.

The evidence of conscious selection in this subsection is based on the simple econometric model of Ziemba and al. (1986). This methodology allows to rank numbers by their unpopularity. Moreover, it helps to reveal those numbers which are sufficiently unpopular to be systematically associated with positive payoffs.

The expected payoff per euro is defined as the sum of prizes weighted by the probability that they are won:

$$R_k = \frac{\sum_{i=1}^{n} P_i \pi_i}{c}$$

where π_i is the prize of rank i (from 1 to 7 in the french lotto system since 1997), P_i is the probability that a random ticket wins the rank i's prize, and c is the euro price[5] of a ticket.

The regression equation is:

$$R_k = \prod_{i=1}^{49} \alpha_i^{x_{ki}} e^{u_k}$$

where x_{kj} is a dummy variable which is set to 1 if the i-th number appeared in the k-th drawing and zero otherwise; u_k is the corresponding disturbance term.

The regression of the expected payoff against the lotto numbers allows us to evaluate each number's contribution to the total pot of the drawing. In this equation, parameters α_i are the profitability factors. The regressing equation is first linearized by taking logarithms on both sides:

[4] The results for this case are available upon request.

[5] Since 19th of june 1996, the ticket price has doubled (0.304898 euro or 2 french francs) and has been rounded to 0.3 euro at the beginning of 2002.

$$\ln R_k = \sum_{i=1}^{49} a_i x_{ki} + u_k$$

Numbers are defined as unprofitable if $H_o(\alpha_i > 1)$ or equivalently $H_o(a_i > 0)$ is rejected at conventional probability levels by means of a one-tail t-test. Conversely, numbers are defined as profitable if $H_o(\alpha_i < 1)$ or equivalently $H_o(a_i < 0)$ is rejected at conventional probability levels by means of a one-tail t-test. Finally, numbers are defined as average if they are neither profitable nor unprofitable, i.e. $H_o(\alpha_i = 1)$ or equivalently $H_o(a_i = 0)$ is not rejected at conventional probability levels by means of a two-tail t-test.

4 Empirical results

4.1 The data

All the lotto results since the origin can be freely downloaded on the official website of the FdJ[6] . For each drawing, the six drawn numbers, the bonus number, the number of winners at each rank and, finally, the unit prizes are available. The game started in may 1976 but, in the first weeks, tickets were sold in the Paris region only. The game was subsequently extended to the whole country. However, as the probability of a winner at the first rank is strongly linked to the number of tickets sold, there were many rollovers in the starting period. We then decided to start the study at the beginning of 1978, deleting one year and a half, more precisely 84 weeks (and draws). The database then consists in 3638 draws on a period of 25 years, ending on 4/26/2003. It is a large database, compared to the ones used in related studies. For example Farrell *et al.* analyzed 116 drawings, Finkelstein (1995) 176, Forrest *et al.*(2002), 254 drawings and Papachristou and Karamanis (1998), 414 drawings.

Our data, associated with the sharing rules and the take-out rate, allow to recover the number of tickets sold for each draw. It is worth mentioning that the level of sales is not publicly announced by FdJ, contrary to what is observed for some foreign lotteries[7].

Statistics about the combinations chosen by french players are not publicly available; it is a common practice among lotto organizers, this information being considered as a strategic one. Simon (1999) obtained, for a single draw, the distribution of combinations, that is to say, the number of combinations played one time, two times and so on. He observed that this distribution was highly non standard and far from the uniform distribution. For the specific draw he considered, about 70 millions

[6] www.fdjeux.com

[7] For example, Camelot, the firm organizing the british lotto, publishes the level of sales on its website and the belgian lottery as well.

tickets were sold, implying that the mean number of occurrences of any combination should be around 5 under the uniform distribution hypothesis. In fact, the distribution was highly skewed. On one side, the number of combinations not played at all in this draw was really high, compared to the expected number that should occur under random selection. On the other side, the 1% most popular set of combinations accounted for 10% of the grids, to be compared to the 2.5% expected in a random selection.

4.2 The estimation of sales

To estimate the number of tickets sold, we used the data concerning the sharing of prizes. As the way gains are shared across ranks has evolved through time, the data concerning the successive regimes were collected on the official website devoted to the legislation on games.

To avoid the problem of rollovers in this estimation, we only considered the percentage of stakes devoted to the ranks 6 and 7 (combinations with three correct numbers or three numbers plus the bonus number).

Let G_j^k the unit gains at rank j in the k-th draw ($k = 1, 2$) on a given day and N_j^k the corresponding number of winners. The total gains for ranks 6 and 7 are given by :

$$\sum_{k=1}^{2} \sum_{j=6}^{7} G_j^k N_j^k$$

If τ is the take-out rate, π_{67} the percentage of stakes devoted to ranks 6 and 7 and P the price of a ticket, we obtain the number of tickets sold, denoted as T by :

$$T = \frac{\sum_{k=1}^{2} \sum_{j=6}^{7} G_j^k N_j^k}{2 \times P \, (1 - \tau) \, \pi_{67}}$$

Prizes were rounded to the lower franc before 2002 and to the lower tenth of euro since then. Without any other information, we added 0.5 franc to the prizes up to the end of 2001 and 0.05 euro since the beginning of 2002 to estimate the level of sales. For each draw, this correction corresponds to 200000 to 400000 tickets.

The coefficient 2 in the denominator is related to the fact that tickets are valid for the two draws of the day under consideration. Consequently, the ticket price is the one for one draw, that is 0.3 euro today. Figure 2 illustrates the evolution of the number of tickets sold for each draw since the beginning of 1978. Several regimes can be observed. The changes of regime occur essentially at three points in time. First, in 1984, when a saturday draw was introduced, second in 1990 when two draws per day were introduced and, finally, on 6/19/1996 when the price of the ticket was doubled.

FIGURE 2 AROUND HERE

4.3 Summary statistics and preliminary tests

The number of first-rank winners stays in the range $[0; 103]$. However, as the number of draws leading to more than 16 winners is low, they are grouped in the same category in table 3.

TABLE 3 AROUND HERE

More than half of the draws end up with less than two winners. Moreover, in 16% of the draws, there is no winner at the first rank.

The first question to be answered concerns the randomness of draws. Each drawing is taken in a population of $\mathbf{C}_{49}^{6}$ possible combinations. The usual χ^2 test cannot be used since successive numbers within a draw are not independent. Joe (1993) provided a test statistic to take this case into account[8], an adjustment already used in Stern and Cover (1989) in their study of the Canadian 6/49 Lotto. The variable χ_J^2 defined as

$$\chi_J^2 = \frac{N-1}{N-n}\sum_{i=1}^{N}\frac{(O_i - E)^2}{E}$$

is asymptotically distributed as a χ^2 random variable with $N-1$ degrees of freedom. $N = 49$ is the number of balls, $n = 6$ is the number of balls to be drawn, O_i is the observed frequency of number i and $E = m\frac{n}{N}$ is the expected frequency under random selection, $m = 3638$ being the sample size. The computed value of χ_J^2 is 49.02 on our sample, to be compared to the critical value at the 5% level, $\chi^2_{48;0.05} = 65.2$.

We also tested randomness by means of the usual Kolmogorov-Smirnov test. The critical value d_N at the α-level is defined by :

$$P\left(\sup_i |F_N(i) - F(i)| < d_N\right) = 1 - \alpha$$

For $\alpha = 5\%$, the critical value is $d_N = 0.19028$ when our observed value is 0.006.

These results suggest that the drawing mechanism used by FdJ is a fair one. In the following, we will then focus on the conscious selection of numbers by players, assuming that the drawing mechanism is random.

4.4 Evidence of preferred subsets of numbers?

The usual arguments in favor of conscious selection are based on the observation that people prefer to choose numbers according to their own rules. A non-uniform distribution is likely to appear as soon as these rules may be shared by many players.

[8]Another way to implement a test of randomness of lotto combinations is to write the χ^2 statistic at the combinations' level. Genest *et al.* (2002) show that this variable may be approximated by a weighted sum of six χ^2-distributed random variables. It allows to test combinations of subsets of numbers in the six numbers draws.

The most well known example concerns birthday dates. In other words, people picking numbers with such heuristics will generally notch low numbers in a 6/49 game. This hypothesis has been confirmed by Finkelstein (1995) for the California lottery and by Papachristou and Karamanis (1998) for the Greek lotto. In this section, we first analyze the relationship between the number of winners at the first rank and the drawn numbers.

4.4.1 Mean of draws and number of winners

The "birthday date" effect can be roughly summarized by saying that players prefer low numbers. Then, a simple way to test such an assumption is to link the number of first-rank winners to the mean of the drawn numbers. Figure 4 illustrates the relationship between the two variables. In case of a random choice by players, the number of winners at the first rank would only be influenced by the volume of tickets sold. On the contrary, if players share some rules to choose numbers, the distribution of drawn numbers, conditional on the number of winners, is likely to be different from the uniform distribution.

FIGURE 4 AROUND HERE

It appears clearly that the number of first-rank winners is a decreasing function of the mean of the draw, suggesting that people have a tendency to select low numbers. This preliminary observation doesn't take the level of sales into account, this variable being a natural explanatory variable. To deal with this problem, let X be the mean of the draw and N_6^k the number of winners at the first rank in drawing k (the index refers to the number of matched numbers). The null hypothesis of random selection implies that N_6^k may be explained by the level of sales. If we denote by N_D the number of possible draws (13983816 in a 6/49 game), and Q the number of tickets sold, the distribution of the number of winners is binomial $\left(Q; \frac{1}{N_D}\right)$. Taking into account the size of these parameters, it can be approximated by a Poisson distribution with parameter $\lambda = \frac{Q}{N_D}$. Let Z be the vector defined as the standardized deviation of the number of winners :

$$Z = \left(\frac{N_6^k - \lambda^k}{\sqrt{\lambda_k}}, k = 1, ..., m \right)$$

We then perform the following regression :

$$Z = aX + b\mathbf{1} + \varepsilon$$

where $\mathbf{1}$ is a vector of ones. Table 2 summarizes the results. It appears that the mean of draws is highly significant to explain the deviation of the number of winners at the first rank. The t-value of a is -18.49 and the R^2 is 0.086. This result gives a strong indication of the tendency of players to choose low numbers. This effect is reinforced by the fact that some players (unfortunately an unknown proportion) buy tickets by means of the “Système Flash”, then getting random numbers on their grids.

4.4.2 Alternative tests of conscious choice

The preceding paragraph showed that preferred combinations contain smaller numbers than unpopular ones. However, the relation concerned the mean of the draws, that is a very rough statistic in this case. To get more insights in the relationship between the number of winners and the numbers actually selected by the players, we refer now to the usual concept of first-order stochastic dominance, denoted as $\succ$. Let F the cumulative distribution of the variable "number selected", denoted as X, taking values in the set of integers between 1 and 49. Keeping the preceding notations, there are $M = 6 \times m = 21828$ numbers drawn in the sample. As illustrated before, the uniform distribution assumption cannot be rejected on the whole sample of official draws. Let now N_6^k denote the random variable "number of winners at the first rank in drawing k". If people choose randomly their numbers, the variables X and N_1^k are independent. Let F_j the distribution function of numbers, conditional on the event $\left\{N_6^k = j\right\}$ and F_{j+} the distribution function conditional on the event $\left\{N_6^k \geq j\right\}$.

Figure 5 shows the differences $F_1 - F_0$ and $F_{1+} - F_0$ which are always positive. It implies that the conditional distribution of numbers when there is no winner dominates the corresponding distribution when there is exactly one winner or when there are more than one winner at the first rank.

FIGURE 5 around here

It clearly illustrates the conscious choice of small numbers by players. This result is remarkable because we didn't take into account the level of sales which may be important in the explanation of rollovers. However, there is no reason to think that a relationship exists between the level of sales and the numbers appearing in the official drawing.

To confirm these relationships, we divide the set of 49 numbers in seven subsets of equal range. Table 4 contains the frequencies in each group of numbers when there are 0,1,2 and more than five winners. The last line in the table gives the χ^2 when these distributions are compared to the uniform one. The critical values of the χ^2 test are respectively 12.6 and 18.5 at the 95% and 99.5% levels. It appears that the null hypothesis is rejected in three of the four cases. However, in the two first, the large numbers are over-represented in the sample. The opposite is observed on the last category, corresponding to draws with more than five winners. It then confirms the conscious selection of numbers and the dependence between the number of first-rank winners and the numbers drawn.

4.5 Maximum likelihood results

The maximum likelihood estimation described in section 3.1 is first applied to our whole sample. As in the french case, ranks 6 and 7 refer to 3 winning numbers with or without the complementary number, we use $N_{3B}^k = N_3^k + N_B^k$ instead of N_3^k in the likelihood expression. Moreover, the numbers of tickets sold in each draw is computed according to the methodology of subsection 4.1.

Figure 6 shows the estimated probability distribution of numbers chosen by players. As most of the studies in other countries, we observe a non uniform distribution[9]. Numbers above 29, roughly speaking, numbers that are not birthday dates, appear to be clearly less popular than lower numbers. This is also the case of numbers 1, 2, 20, 21 and 23 but for these numbers there is no explanation of such a behavior. The more popular numbers are those to which people usually impart chance attributes (7, 12, 13), though it is unclear why 12 is more popular than 13.

FIGURE 6 around here

On the opposite, 32, 39,40 and 41 are the most unpopular numbers; these numbers are likely to generate positive returns or, at least less negative returns. It will be confirmed later on in the test of weak efficiency.

4.5.1 Comparison of maximum likelihood and bonus number estimations

As said before, 3+B is a specific winning combination in the french lotto since October 1997. Our database contains 1156 drawings since this date. We can then compare the two estimation methods on this subset. The number of times a given bonus number appears on these 1156 drawings varies from 14 to 35. For each number, we calculated the estimated probability of occurrence of each number by means of equation 1. Figure 7 summarizes the results. The average difference between the two estimators is 0.15% (calculated with respect to MLE values). The absolute individual differences are in the range $[-13.6 \times 10^{-4}; 12.58 \times 10^{-4}]$. Consequently, it can be seen on figure 7 that the two estimations are very close, identifying the same popular and unpopular numbers. The greatest advantage of the bonus number estimator is that it can be instantaneously obtained when the maximum likehood estimator needs several hours of CPU time. However, as mentionned before, this technique cannot be used with profit when only the 5+B combination is a winning one. Figure 8 compares this estimation with the one obtained by maximum likelihood. Due to the low number of winners at rank 5+B, large differences appear between the two, even if the most popular and unpopular numbers are correctly identified. For example, the largest difference is observed for number 24; this number appears 25 times as a bonus number in the sample. But the mean number of 5+B winners is 15.8 which is very low to get a correct estimate of the probability.

[9] Our results cannot be directly compared to those obtained by Farrell *et al.* because their estimation takes into account the number of tickets that use the lucky dip facility. However, their estimates with and without the facility are similar and the authors claim that this device only affects the likelihood of the sample so that allowing for it is important.

The other interesting point lies in the comparison of figures 6 and 7. The first one concerns the whole period, starting in 1978 while the other concerns the last six years. The low differences between the two is impressive. It means that preferences for "lucky" numbers are very stable through time. An other interpretation is that players do not learn anything from the results of the successive drawings. They continue to play "low return" combinations. In the following, we address this question to evaluate if some combinations can generate positive returns.

4.5.2 Regression results

Table 5 shows the regression coefficients of the 49 numbers (values in parentheses are t-values), computed over the full sample of 3638 drawings ($R^2 = 0.122$), using the model presented in section 3.3. A vast majority of coefficients are negative ; it indicates that, obviously, most numbers are not profitable. However, some numbers exhibit a significant positive return, also confirming that choices are not random[10].

Overselected numbers in the french lotto market are the following: 7, 12, 9, 13, 11, 5, 10, 8, 24, 25. In this list, numbers are ranked according to their negative contribution to expected payoffs. This is a non exhaustive list but we can see that these numbers may be date numbers that were previously identified as potential popular numbers. In fact, they are all popular enough to generate a significantly negative expected payoff at conventional probability levels. On the other side, underselected numbers present less significant coefficients but are clearly profitable and, for most of them, non-date numbers. 32 and 41 are the two most unpopular numbers and generate significant positive returns. Associated with 39, 40, 38, 43, they are likely to generate positive returns, casting doubt on the weak efficiency of the french lotto market.

However, we have to be careful for a simple reason. When a player buys two grids (the minimum stake), he pays 1.2 euro because the grids are valid for the two drawings of the day under consideration. Knowing that more than 98% of tickets are loosing ones, the player choosing the unpopular numbers can wait quite a long time before realizing the expected profit.

5 Concluding remarks

As it has been shown that all three methodologies show up the same patterns with regards to the numbers popularity, an explanation of the persistence of some numbers profitability is needed. It is indeed unclear why people do not identify profitable numbers and always choose unprofitable ones. A simple structural explanation deals with the information release of the lotto market. Authorities do not publish statistics on the numbers actually played so that people can't use this information to improve

[10]The LR statistic, equal to 198,639 allows us to significantly reject the hypothesis $H_o : a_1 = a_2 = ... = a_{49}$.

the expected value of their tickets. However, as pointed out by Farrell *et al.*, nothing allows us to assert that people would exploit the opportunity because they suspect that there exists a kind of utility gain associated with playing one's own lucky numbers. Moreover, to explain why profit opportunities remain unexploited, they suggest that this gain ranks above the utility loss associated with the higher probability of sharing the jackpot prize. Psychological phenomena like the illusion of control or the gambler's fallacy appear convincing to give the intuition of the utility gain sources though there exists no direct test of these hypotheses. Many heuristics that lead to a false perception of randomness may also induce people to put weights on probabilities. This kind of behavior may be captured by an alternative description of preferences, for example by cumulative prospect theory, first presented by Kahneman and Tversky (1979) and developed in Tversky and Kahneman (1992). Moreover, such a description allows to transform objective probabilities in decision weights, especially in putting more weight on low-probability events. Jullien and Salanié (2000) already showed that racetrack bettors are well described by cumulative prospect theory and Bradley (2003) also recently illustrated that this approach is well-suited to characterize betting behavior.

References

[1] Beenstock, M. and Y. Haitovsky (2001), Lottomania and other Anomalies in the Market for Lotto, *Journal of Economic Psychology,* 22, 721-44.

[2] Boland, P.,J. and Pawitan, Y. (1999), Trying to be random in selecting numbers for lotto, *Journal of Statistics Education*, 7, n°3.

[3] Bradley, I. (2003), The Representative Bettor, Bet Size and Prospect Theory, *Economics Letters,* 78, 409-13.

[4] Clotfelter, C., T. and P., J. Cook (1993), The 'gambler's fallacy' in lottery play, *Management Science*, 39, 1521-5.

[5] Farrell, L., Hartley, R., Lanot, G. and I. Walker (2000), The Demand for Lotto: The Role of Conscious Selection, *Journal of Business & Economic Statistics*, 18, n°2, 228-41.

[6] Finkelstein, M. (1995), Estimating the Frequency Distribution of the Numbers Bet on the California Lottery, *Applied Mathematics and Computation,* 69, n°2-3, 195-207.

[7] Forrest, D., R. Simmons and N.Chesters (2002), Buying a Dream : Alternative Models of Demand for Lotto, *Economic Inquiry*, 40, n°3, 485-96.

[8] Genest, C., Lockhart, R.A. and M.A. Stephens (2002), χ^2 and the lottery, *Journal of the Royal Statistical Society:The Statistician*, 51:2, 243-257.

[9] Hausch, B. and W.,T.,Ziemba (1995), Efficiency of sport and lottery betting markets, in Handbooks in Operational research and Management Science, Vol 9, R.Jarrow Eds, Elsevier Science.

[10] Henze, N. (1998), "Drawings Since Hit Tables in Lotteries" and a New Multivariate Geometric Distribution, *Statistics and Probability Letters*, 40, 321-27.

[11] Joe, H. (1993), Tests of Uniformity for Sets of Lotto Numbers, *Statistics&Probability Letters,* 16, n°3, 169-251.

[12] Johnson, R. and J. Klotz (1993), Estimating hot numbers and testing uniformity for the Lotto, *Journal of the American Statistical Association*, 88, n°42, 662-8.

[13] Jullien, B. and B. Salanié (2000), Estimating Preferences under Risk : the Case of racetrack Bettors, *Journal of Political Economy,* 108, 503-530.

[14] Kahneman, D., A.Tversky (1979), Prospect Theory : an Analysis of Decision under Risk, *Econometrica,* 47, n°2, 263-91.

[15] Papachristou G. and D. Karamanis (1998), Investigating Efficiency of Betting Markets : Evidence from the Greek 6/49 Lotto, *Journal of Banking and Finance,* 22, 1597-1615.

[16] Scott, F., A. and O., D. Gulley (1995), Testing for Efficiency in Lotto Markets, *Economic Inquiry*, 33, 175-188.

[17] Simon, J. (1999), An Analysis of the Distribution of Combinations Chosen by UK National Lottery Players, *Journal of Risk and Uncertainty*, 17:3, 243-76.

[18] Stern, H. and Cover, T. (1989), Maximum Entropy and the Lottery, *Journal of the American Statistical Association,* 84, 980-85.

[19] Thaler, R. and W. Ziemba (1988), Pari Mutuel Betting Markets : Racetracks and Lotteries, *Journal of Economic Perspectives,* 2, 161-74.

[20] Tversky, A. and Kahneman, D. (1992), Advances in Prospect Theory : Cumulative Representation of Uncertainty, *Journal of Risk and Uncertainty,* 12,297-323.

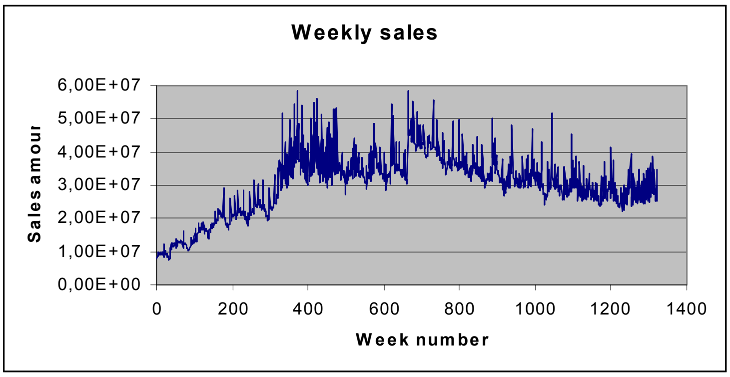
[21] Walker, I. and J. Young (2001), An Economist's Guide to Lottery Design, *The Economic Journal,* 111 (November), 700-22.

[22] Ziemba, W., S. Brumelle, A. Gauthier and S. Schwartz (1986), *Dr Z's Lotto Guidebook*, Dr Z Investments, Los Angeles.

Periods	Number of drawings	Ranks and probabilities (w)=wednesday, (s)=saturday						
		1	2	3	4	5	6	7
		$6N$	$5N{+}B$	$5N$	$4N{+}B$	$4N$	$3N{+}B$	$3N$
1978/01/04 - 1984/02/29	322	$\frac{1}{C_{49}^{6}}$	$\frac{6}{C_{49}^{6}}$	$\frac{252}{C_{49}^{6}}$	$n.s$	$\frac{C_6^4 C_{43}^2}{C_{49}^6}$	$n.s$	$\frac{C_6^3 C_{43}^3}{C_{49}^6}$
1984/03/07 - 1990/09/08	680	$\frac{1}{C_{49}^{6}}$	$\frac{6}{C_{49}^{6}}$	$\frac{252}{C_{49}^{6}}$	$n.s$	$\frac{C_6^4 C_{43}^2}{C_{49}^6}$	$n.s$(w) $\frac{C_6^3 C_{42}^2}{C_{49}^6}$ (s)	$\frac{C_6^3 C_{43}^3}{C_{49}^6}$(w) $\frac{C_6^3 C_{42}^3}{C_{49}^6}$(s)
1990/09/12 - 1997/10/11	1480	$\frac{1}{C_{49}^{6}}$	$\frac{6}{C_{49}^{6}}$	$\frac{252}{C_{49}^{6}}$	$n.s$	$\frac{C_6^4 C_{43}^2}{C_{49}^6}$	$n.s$	$\frac{C_6^3 C_{43}^3}{C_{49}^6}$
1997/10/15 - 2003/04/26	1156	$\frac{1}{C_{49}^{6}}$	$\frac{6}{C_{49}^{6}}$	$\frac{252}{C_{49}^{6}}$	$\frac{C_6^4 * 42}{C_{49}^6}$	$\frac{C_6^4 C_{43}^2}{C_{49}^6}$	$\frac{C_6^3 C_{42}^2}{C_{49}^6}$	$\frac{C_6^3 C_{42}^3}{C_{49}^6}$

Table 1: Winning ranks and corresponding probabilities

Figure 1: French Lotto Weekly Sales



Coef	Value	St dev	t-value
a	-2.75	0.149	-18.49
b	8.5	0.477	17.83
R^2	0.086		

Table 2: Regression results : $Z = aX + b\mathbf{1} + \varepsilon$

Figure 2: Evolution of ticket sales (in volume)

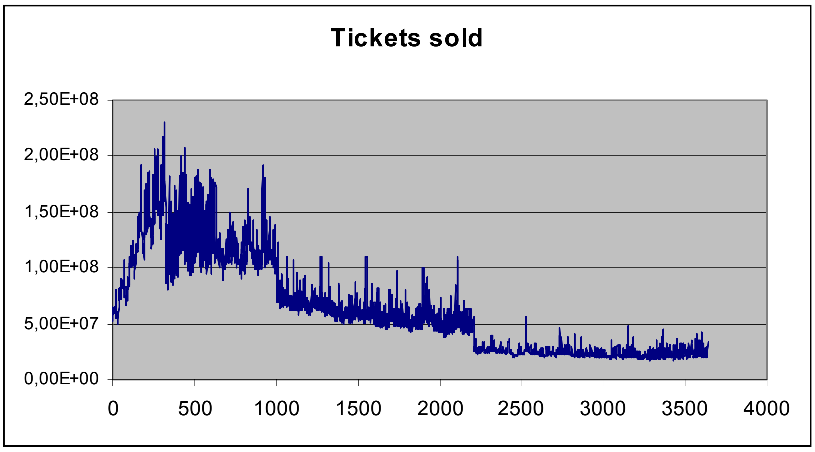


Figure 3: Frequency of drawn numbers

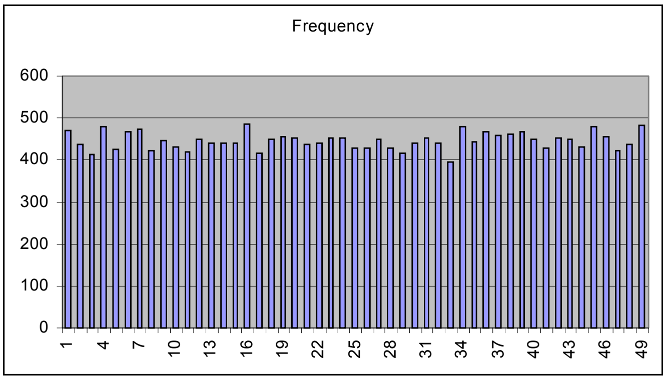


Figure 4: Mean of the draw versus number of winners

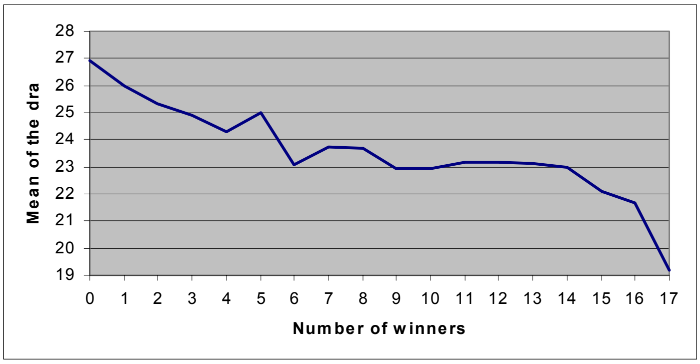


Figure 5: Stochastic dominance tests

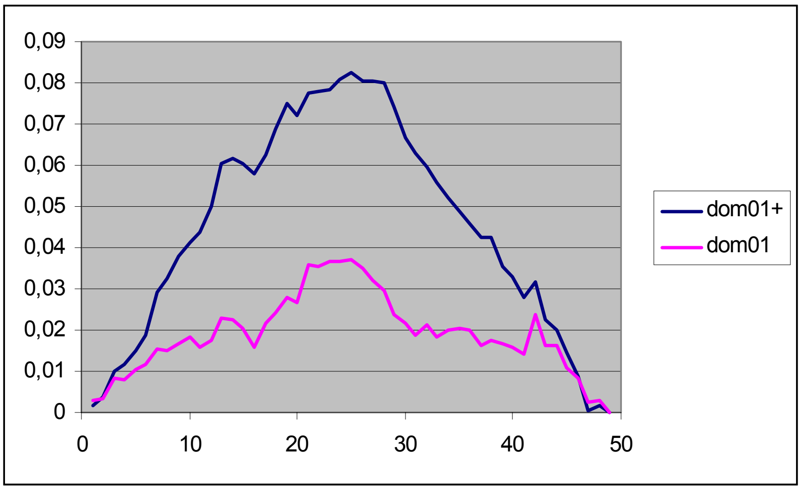


Figure 6: Probability distribution of numbers actually played

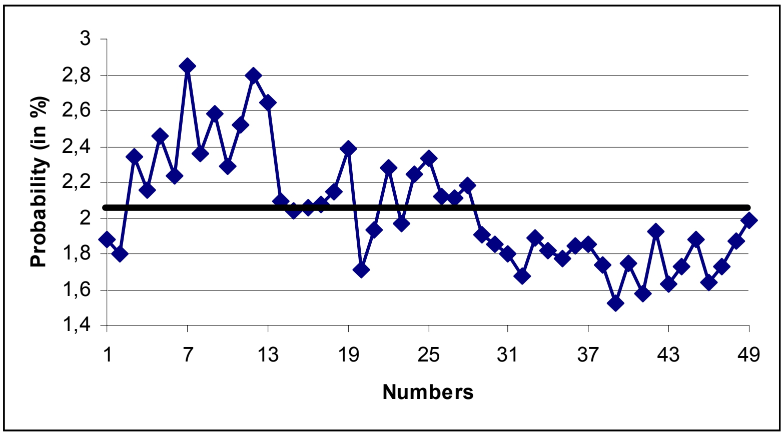
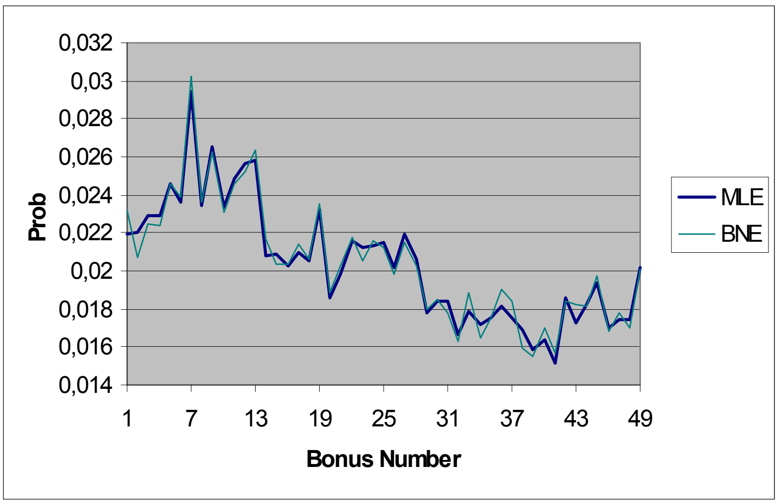


Figure 7: Comparison of MLE and BNE estimators



Number of winners	Number of drawings	in %	Number of winners	Number of drawings	in %
0	578	0.159	9	79	0.022
1	763	0.210	10	50	0.014
2	598	0.164	11	37	0.010
3	423	0.116	12	31	0.009
4	290	0.080	13	35	0.010
5	210	0.058	14	20	0.005
6	167	0.046	15	20	0.005
7	114	0.031	16	22	0.006
8	99	0.027	$\geq$17	102	0.025

Table 3: Number of winners

Numbers	**0 winner**	**1 winner**	**2 winners**	**$\geq$5 winners**
1 to 7	417	621	521	952
8 to 14	390	548	478	1007
15 to 21	451	657	524	879
22 to 28	482	608	486	895
29 to 35	578	723	492	731
36 to 42	556	749	549	722
43 to 49	594	678	538	730
χ^2-value	79.98	43.99	8.78	97.85

Table 4: Frequencies versus number of winners

Figure 8: MLE versus 5+C estimators

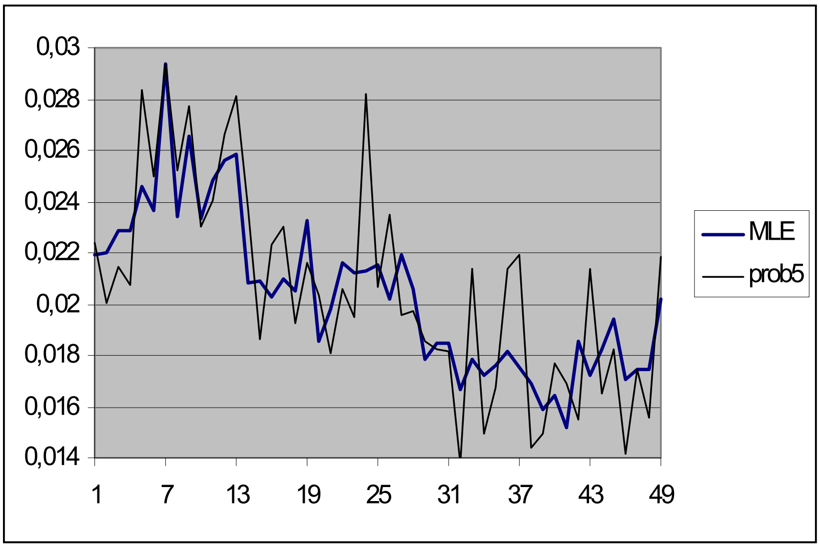


Table 5: Estimation of the expected payoff regression of numbers 1-49

	Regression coefficients		*Regression coefficients*		*Regression coefficients*
1	-0.0548 (-2.3594)	18	-0.0489 (-2.0938)	35	-0.0027 (-0.1173)
2	-0.0810 (-3.4208)	19	-0.1069 (-4.5997)	36	-0.0168 (-0.7303)
3	-0.0982 (-4.0287)	20	-0.0220 (-0.9220)	37	-0.0102 (-0.4349)
4	-0.0898 (-3.8869)	21	0.0107 (0.4539)	38	0.0306 (1.2964)
5	-0.1560 (-6.5593)	22	-0.1216 (-5.1429)	39	0.0701 (3.0187)
6	-0.1214 (-5.2177)	23	-0.0526 (-2.2372)	40	0.0358 (1.5228)
7	-0.2557 (-11.0431)	24	-0.1272 (-5.3599)	41	0.0760 (3.1797)
8	-0.1419 (-5.8404)	25	-0.1238 (-5.1455)	42	-0.0032 (-0.1374)
9	-0.1958 (-8.2144)	26	-0.0613 (-2.5366)	43	0.0155 (0.6617)
10	-0.1434 (-5.9426)	27	-0.0937 (-3.9616)	44	0.0012 (0.0535)
11	-0.1775 (-7.3500)	28	-0.0863 (-3.5987)	45	-0.0557 (-2.4344)
12	-0.2156 (-9.0751)	29	-0.0423 (-1.7311)	46	0.0063 (0.2735)
13	-0.1907 (-8.0320)	30	-0.0259 (-1.0939)	47	-0.0144 (-0.5991)
14	-0.0860 (-3.5939)	31	0.0047 (0.2037)	48	-0.0277 (-1.1545)
15	-0.0895 (-3.7780)	32	0.0934 (3.9407)	49	-0.0483 (-2.0770)
16	-0.0840 (-3.6327)	33	-0.0527 (-2.1508)		
17	-0.0539 (-2.2104)	34	-0.0076 (-0.3278)		