CS-BIGS 4(1): 66-72 http://www.bentley.edu/csbigs/Vol4-1/chen.pdf

A Statistical Analysis of Popular Lottery "Winning" Strategies

Albert C. Chen
Torrey Pines High School, USA

Y. Helio Yang
San Diego State University, USA

F. Fred Chen
Science Applications International Corporation, USA

Many people believe they can win the lottery, either believing in books on how to beat the system or relying on their own methods. They hear stories of people winning the jackpot after buying just one ticket and forget about the countless others who lose money each time. This paper shows the ineffectiveness of popular lottery strategies. The California SuperLotto historical winning draws are used to study various game playing strategies. The research results indicate that the winning lottery numbers and the winning mega number are uniformly distributed in their respective ranges, and all numbers have an equal chance of being picked. We further show that no strategy is significantly better than the others for the California SuperLotto. Although the low frequency strategy, applied to the lottery numbers, has significantly more matches than the other strategies as more historical drawing data are used, it is still extremely unlikely that big winnings are produced by playing the lottery "forever". The presentation is accessible to readers with a basic exposure to statistics.

INTRODUCTION

A lottery is a game that involves the drawing of numbers for a prize. Few people win, while the vast majority does not. As published on the Official California Lottery Financial Report, of the total $59.25 billion the lottery took in from people buying tickets, about $30.54 billion was given back in prizes (CAlottery report to the public 08/09). There is a net loss for the players, so people should not expect to profit from a lottery.

Lotteries are often run by national or local governments as a way of raising funds. The California Lottery was created in 1984 to generate supplemental funding for public schools; about 34% of lottery funds go toward education. According to a 2006/2007 Report of Lottery Expenditures for K-12 Education, prepared by the California Department of Education, on the average 61% of Lottery funds for education are spent on Salaries and Benefits for instructors, 24% on classroom materials such as textbooks, and the rest is spent on other areas (Calottery.com). Every year, millions of people play the California SuperLotto, contributing greatly to the lottery funds. In California SuperLotto, five winning lottery numbers from one to forty seven are chosen using a machine. Then, one mega number is chosen from one to twenty seven. All numbers need to be matched to win the lottery.

We briefly review a few lottery related studies. Hennigan (2009) studied the profile of people playing lottery and found that per capita spending on the lottery goes down as education levels and median income go up. The study also found that lottery participation is roughly equal among demographic groups. Grote and Matheson (2006) investigated the expected value of a lottery ticket as the jackpot rises and more people play. They investigated strategies for the state to maximize its profit. Bradley (2009) described how Euler analyzed the Genoese lottery to determine fair prize amounts.

There is much advice on market on how to win a lottery. Some trade books make suggestions on lottery winning strategies; many people buy these books hoping to "improve their luck" with the author's published, purportedly-effective, strategies. These books claim to be able to tell which numbers are most likely to appear, substantially reduce odds, and avoid numbers that are sure to lose. For example, one book proposes a strategy in which a combination of recently frequent and infrequent (hot and cold) numbers is chosen. A similar strategy is to choose the most frequent lottery numbers. Another strategy is to choose the least frequent lottery numbers. These strategies can both be implemented by looking at the frequency lists on the California State Lotto website. These books advise against choosing random numbers generated by a computer, a procedure called a quick pick. There is little evidence of effectiveness of these published strategies.

There are many possible reasons why millions of people play the California SuperLotto. A substantial portion of those players believe they have a good chance to win the lottery. One main reason for this misconception is that players trust the authors of books on how to win the lottery and are too willing to believe them in hopes of easy money. Perhaps some players have superstitions or lucky numbers, or believe in some bias in the lottery system that would favor some numbers over others. They may also not understand probability, which shows that there is a one in $\binom{47}{5}\binom{27}{1}$ chance, that is a probability of 1/(41,416,353) of winning the jackpot in the California SuperLotto. Disregarding these odds, players may point to anecdotal evidence of similar people, small in number of course, who have won as evidence that they themselves will win someday. Ultimately, the trust in these methods lies in faulty logic and incorrect statistical beliefs. Some popularly attempted methods to win the lottery are shown to be ineffective in this paper.

PROBLEM STATEMENT AND HYPOTHESES

In this study, we are interested in finding out the answers to the following questions:

1. Do the California SuperLotto lottery numbers and mega numbers occur with equal probability in the history of the game?
2. Is there a game strategy that outperforms others in the history of the game?
3. Is the performance of a strategy associated with the amount of historical information considered?

Three commonly used strategies are compared in this study. The first strategy is a "random strategy," in which people use quick picks, letting computers generate numbers for them randomly. These people believe lottery is a random event so historical winning numbers do not provide much meaning. The second strategy is called "low frequency strategy," in which people pick the numbers that occur less frequently. The rationale is that people believe the probability of the occurrence of each number is the same. This implies that the frequencies should even out in the long run, so they pick the low frequency numbers, i.e., cold numbers, hoping the tide will change. The third strategy is called "high frequency strategy," in which people pick the numbers that occur often historically. They assume that these popular numbers are likely to continue to appear as a trend, possibly due to the physical lottery machine setup, creating bias. These strategies can be simulated by taking the historical lottery winning numbers posted on the official California Lottery website.

Specifically the following hypotheses are developed for the research questions:

H_1: The winning lottery numbers are uniformly distributed.

H_2: The winning mega numbers are uniformly distributed.

H_3: The three strategies of selecting lottery numbers have no statistically significant difference in their performances.

H_4: The three strategies of selecting mega numbers have no significant statistical difference in their performances.

H_5: The performances of lottery numbers using the three strategies are not associated with the amount of historical information used.

H_{5a}: The performances of lottery numbers using the "random strategy" on lottery numbers are not associated with the amount of historical information.

H_{5b}: The performances of lottery numbers using the "low frequency strategy" on lottery numbers are not associated with the amount of historical information.

H_{5c}: The performances of lottery numbers using the "high frequency strategy" on lottery numbers are not associated with the amount of historical information.

H_6: The performances of the mega number using the three strategies are not associated with the amount of historical information used.

H_{6a}: The performances of the mega number using the "random strategy" on mega numbers are not associated with the amount of historical information.

H_{6b}: The performances of the mega number using the "low frequency strategy" on mega numbers are not associated with the amount of historical information.

H_{6c}: The performances of the mega number using the "high frequency strategy" on mega numbers are not associated with the amount of historical information.

RESEARCH METHODOLOGIES

We analyzed the lottery strategies and their effectiveness on California SuperLotto draws from 1375 to 2274. These nine hundred lotto draws represent the whole history of the lottery under the current, revised game rules. The data are first prepared for the study and the three abovementioned strategies are then simulated using the historical data. The data preparation procedure and the strategy simulation procedure are shown in Appendix A. For this study, the performance of a strategy is gauged by the average number of matches to the winning numbers chosen in thirty draws. Figures 1-3 show the time series plots of simulated results of three strategies for the lottery numbers. Figures 4-6 show the time series plots of simulated results of three strategies for the mega numbers.

Statistical tests were conducted on the simulation results. For Hypotheses 1 and 2, we used a Chi-square Goodness of Fit test to check if the numbers are uniformly distributed. For Hypotheses 3 and 4, we used a one-way Analysis of Variance (ANOVA) test to compare the average performance of the three simulated strategies. For Hypotheses 5 and 6, we used a simple regression to test if the strategy performance is associated with the amount of historical information used to generate the matches.

HYPOTHESIS TESTING RESULTS

Table 1 shows the hypothesis testing results for H_1-H_4. For the Chi-square test of the lottery numbers, the null hypothesis is that the lottery numbers are uniformly distributed from 1 to 47. For the Chi-square test of the

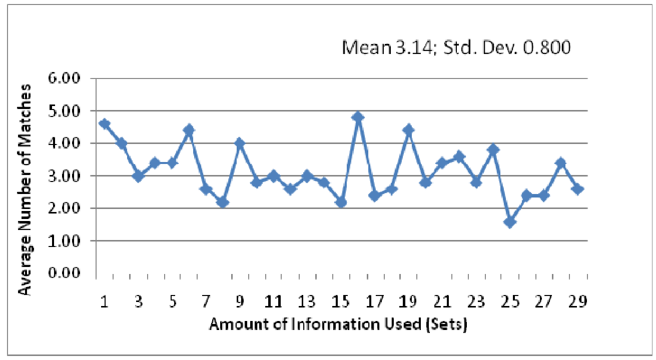

Figure 1. Random strategy performance of lottery numbers.

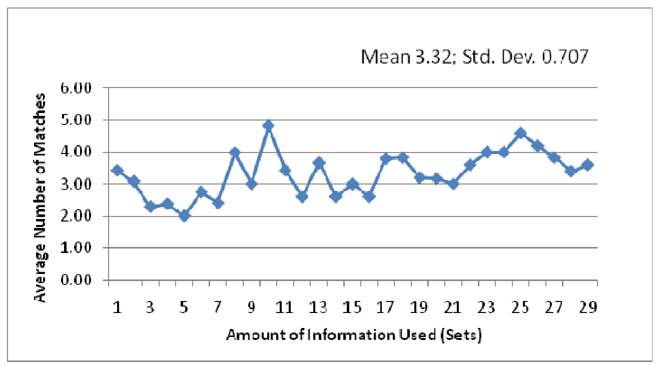

Figure 2. Low frequency strategy performance of lottery numbers.

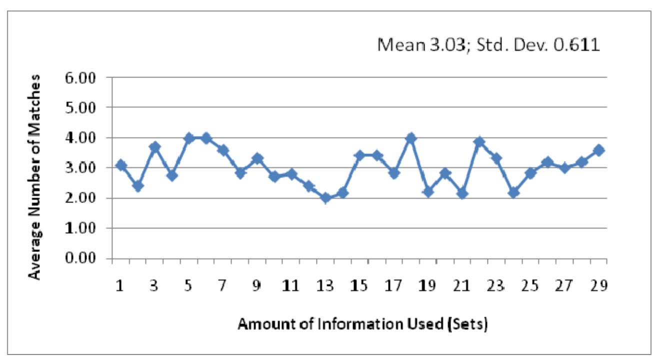

Figure 3. High frequency strategy performance of lottery numbers.

Table 1. Hypotheses H_1-H_4 Testing Results

Hypothesis	Description	Methodology	Result	χ^2 or F Statistic	df	P-value
H_1	Uniformly distributed lottery numbers	χ^2 test	Do not reject	28.231	46	0.982
H_2	Uniformly distributed mega numbers	χ^2 test	Do not reject	30.900	26	0.232
H_3	Lottery number strategy no performance difference	ANOVA	Do not reject	1.237	$df_1=2$ $df_2=84$	0.296
H_4	Mega number strategy no performance difference	ANOVA	Do not reject	.738	$df_1=2$ $df_2=84$	0.481

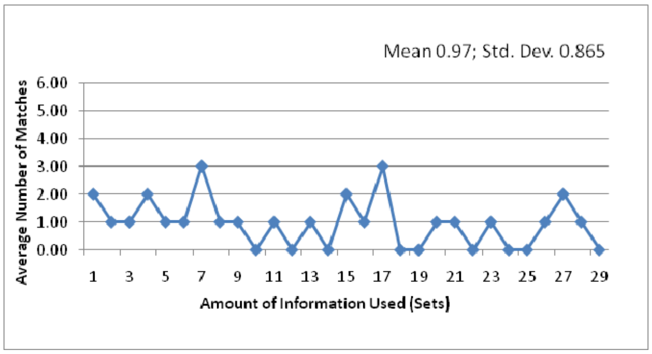

Figure 4. Random mega strategy performance of mega numbers.

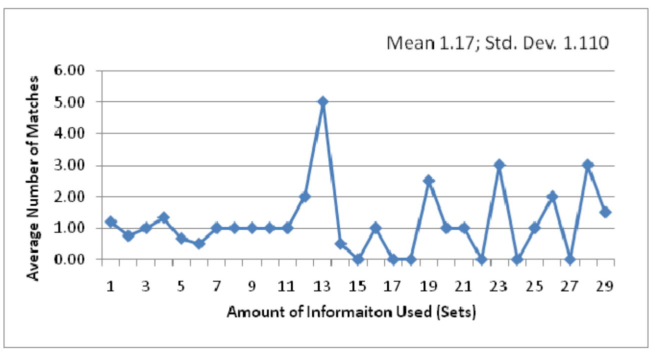

Figure 5. Low frequency strategy performance of mega numbers.

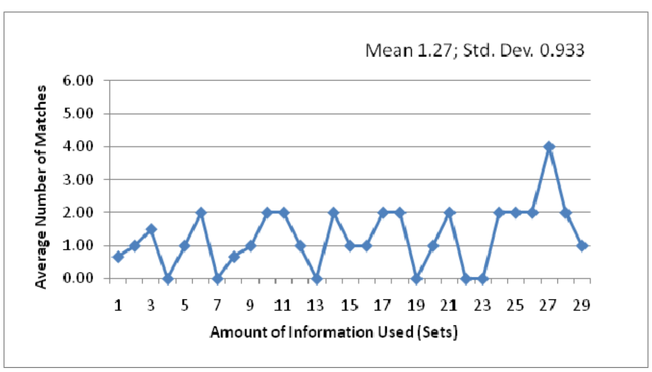

Figure 3. High frequency strategy performance of mega numbers.

mega numbers, the null hypothesis is that the lottery numbers are uniformly distributed from 1 to 27. Both hypotheses cannot be rejected due to high p-values. The simulated results of the three strategies display fairly close performances.

For lottery numbers the random strategy had a total of 91 matches. The low frequency strategy had 96.28 matches, and the high frequent number strategy had 87.90 matches in total. The null hypothesis is that there are no statistically significant differences on strategy performance for the whole lottery history. The one-way ANOVA test has a p-value of 0.296, so the null hypothesis cannot be rejected, indicating that there are no significant statistical differences among "smart" betting and random picking methods. Similar results are demonstrated in mega numbers. The random strategy had 28 matches overall. The low frequency strategy had 33.95 matches and the high frequency strategy had 36.83 matches in total, respectively. The one-way ANOVA test has a p-value of .481 so the null hypothesis cannot be rejected.

Regression is used to test if the strategy performances are associated with the amount of historical information used. Table 2 shows the results. The regression models for H_{5a}, H_{5c}, H_{6a}, H_{6b}, and H_{6c} do not have p-values lower than the 5% threshold. However, the regression model of the low frequency strategy applied to lottery numbers has a p-value of 0.004, which is statistically significant at the 1% level. This rejects H_{5b} and indicates that the amount of historical information is a significant factor associated with the performance of a low frequency strategy. Looking at the positive coefficient, the low frequency strategy's performance appears to improve over the course of the data.

Further Testing Results

The rejection of H_{5b} is an interesting outcome that is worth further investigation. We split the strategy performances into two groups—small amount of historical information used and large amount of historical information used. The "small amount of information"

Table 2. Regression Results for Hypotheses H_5-H_6

Hypothesis	Description	Adjusted R^2	Result	Coefficient	P-value
H_{5a}	Performance of random strategy on lottery numbers not related to information	0.077	Do not reject	-0.031	0.079
H_{5b}	Performance of low frequency strategy on lottery numbers not related to information	0.240	**Rejected**	0.043	**0.004**
H_{5c}	Performance of high frequency strategy on lottery numbers not related to information	-0.032	Do not reject	-0.005	0.715
H_{6a}	Performance of random strategy on mega numbers not related to information	0.060	Do not reject	-0.031	0.107
H_{6b}	Performance of low frequency strategy on mega numbers not related to information	-0.026	Do not reject	0.013	0.597
H_{6c}	Performance of high frequency strategy on mega numbers not related to information	0.078	Do not reject	0.036	0.078

Table 3. ANOVA Tests on All Strategies using Two Levels of Information

Description	Methodology	Result	F statistic	P-value
Lottery number strategy no performance difference with small amount of information	ANOVA	Do not reject	.308	0.737
Lottery number strategy no performance difference with large amount of information	ANOVA	**Rejected**	3.293	**0.048**

df1=2, df2=39

Table 4. Post-hoc tests on performance differences among three strategies with a large amount of information.

	For (I), (J) 1=random 2=low 3=high		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
LSD	1	2	-.55929*	.257	**.036**	-1.0801	-.0385
		3	.02500	.257	.923	-.4958	.5458
	2	1	.55929*	.257	**.036**	.0385	1.0801
		3	.58429*	.257	**.029**	.0635	1.1051
	3	1	-.02500	.257	.923	-.5458	.4958
		2	-.58429*	.257	**.029**	-1.1051	-.0635

Dependent Variable: Average Number of Matches. * The mean difference is significant at the 0.05 level.

group contains result sets #1 to #15, which use information from at most 450 historical draws. The "large amount of information" group contains result sets #16 to #29, which use information from at least 480 historical draws. We wonder if the low frequency strategy performance yields significantly higher matches than other strategies when the large amount of historical information is used, and vice versa. ANOVA analyses are conducted and results are shown in Table 3.

The null hypotheses of both these tests are that there is no statistically significant performance difference among strategies with regards to the amount of historical information used. We found that for the small amount of information level, the p-value is 0.737 so the null hypothesis cannot be rejected. However, for the large amount of information level, the p-value is 0.048, which is statistically significant at 5% level. The null hypothesis is rejected, and the post hoc tests (Table 4) show that the low frequency strategy's average performance is better than the other two strategies. As more information is used, the low frequency strategy appears to be advantageous. The reader must be warned that this does not mean the strategy can yield much bigger winnings, since statistical significance is not the same as practical worth.

DISCUSSION AND IMPLICATIONS

In this study we tested the distribution of lottery numbers and mega numbers from actual California SuperLotto past drawings. We also tested the performances of three lottery strategies.

From the Chi-square results, the winning lottery numbers and the mega numbers appear to be uniformly distributed in their respective ranges, namely 1-47 and 1-27. This means that the chance for each number to be the winning number is the same. Therefore, we can reasonably believe that the process and machines that the State of California uses to generate winning numbers are not biased. The drawing process is executed fairly.

From the ANOVA tests we concluded that no strategy is significantly better that the others in the history of the California SuperLotto game. We found that the random, low frequency, and high frequency strategies have similar average number of matches in the long run. These commonly used strategies yield similar performances. A quick pick strategy is not worse than the other two strategies, contrary to the claims in some trade books on the market. The trust many people have in so called "experts" is unreasonable, as these techniques do not work better than random guessing for the California SuperLotto.

An interesting finding is that the low frequency strategy performance on the lottery numbers shows a slightly upward and statistically significant trend over the amount of the information used. Further ANOVA tests show that the low frequency strategy gives significantly more matches than the other two strategies when a large amount of drawing information is used. The amount of historical information appears to have a positive impact on low frequency strategy performance. However, the slope of the regression line is .043, indicating that in our case the improvement in small. Much time would be needed to make this increase substantial enough to make big winnings, and the linear increase may not continue with time. Practically, there is no real difference, and performance may actually decrease as too much information is used. The lottery is still simply a game of chance. No matter which strategy is used, the probability of winning a large cash prize is extremely low. We suggest that no one plays the lottery unless it is done to have fun or to support education. After all, lottery funding has contributed $21 billion dollars to educational fund since 1985 (Calottery.com). Unfortunately, a Hennigan (2009) study shows that people with lower income and education level are "contributing" to the education fund.

CONCLUSION

We used the California SuperLotto historical winning draws to study common, purportedly effective strategies of playing the game. We concluded that almost all six of the hypotheses are supported by the analyses, with the exception of H_{5b}. The lottery is fair, as shown by the Chi-square tests. The commonly used strategies are no better than random guessing, so "experts" who claim to know how to beat the system should not be trusted. It is likely that those experts are making more money out of selling those books than profiting from lotteries using their own strategies.

Although the low frequency strategy may seem to be better than the others as more time passes, in reality the slight increase in number of matches does not come close to resulting in big lottery winnings. It must be remembered that statistical significance is not the same as practical importance. That is, one would have to play lottery forever, or for a very long period of time, in order to have some minor gains over other strategies which yield very low return anyway. It is unfortunate that a lack of understanding of statistics leads to the hype and misspending of money for lower income people.

REFERENCES

Bradley, R. E. Euler's Analysis of the Genoese Lottery. Math DL. 20 Jan. 2009. <http://mathdl.maa.org/mathDL/46/?pa=content&sa=viewDocument&nodeId=217&bodyId=146>.

California official lottery website http://www.calottery.com/.

Calottery report to the public 08/09. 2010. Retrieved June 9, 2010 from http://www.calottery.com/pv_obj_cache/pv_obj_id_73372A19B7286AD1A97C80621D9951E3F7090800/filename/HowYourLotteryDollarIsSpent2009.pdf

Hennigan, G. 2009. LL Bets Are On: Sales highest in neighborhoods with lower median incomes. *McClatchy - Tribune Business News, March* 2009.

Grote, K. R. and Matheson, V. 2006. In Search of a Fair Bet in the Lottery. *Eastern Economic Journal*, 32, no. 4.

Correspondence: hyang@mail.sdsu.edu

APPENDIX A

Data Preparation Procedures:

The actual winning lottery numbers are organized by dividing them into thirty sets of thirty draws each. Thirty draws is a large enough sample size to find the top five and low five numbers in frequency, and thirty sets give enough data to statistically analyze the performances of the strategies. Then we tally the frequencies of lottery and mega numbers in these sets. Detailed steps are discussed as follows.

1. Obtain the results of 900 past draws from the California SuperLotto from the official lottery web site http://www.calottery.com. (drawing numbers 1375 to 2274).
2. Split the data into 30 non-overlapping sets of 30 draws each from oldest to most recent. (A set consists of thirty lottery draws as a unit).
3. For ease of reference, rename draw number 1375 as draw number 1; 1376 as 2; 1377 as 3, and so on, up to draw number 2274, which corresponds to draw number 900.
4. Let the n^{th} group be the concatenation of the first n sets. Find the frequencies of the lottery numbers 1-47 in California lottery draws 1-30, 1-60, 1-90, …1-900 (G_1,G_2,…,G_{30}) . Do this separately for mega numbers as well.
5. Perform a Chi-square test on the lottery numbers 1-47 in G_{30}. Repeat for mega numbers 1-27.
6. Rank lottery numbers in G_1,G_2,…G_{29} based on frequency. Repeat for mega numbers.
7. Find the five most frequent numbers and five least frequent numbers in each of G_1, G_2,…,G_{29} for lottery and mega numbers.

Strategy Simulation Procedures:

To simulate the random strategy, first random numbers are generated using Microsoft Excel's random number generator to simulate 900 lottery draw "quick picks," in which a computer automatically generates five unique numbers for a lottery ticket. Compare random draws with actual draws. Take random draw #30 and see how many matches it has with California lottery draw #30. Take draw #31 and count the corresponding number of matches. Continue comparing corresponding draws until random draw #900 and actual draw #900 are compared. Then, split these data into twenty-nine sets: 31-60, 61-90, 91-120…871-900. Tally the total number of matches for each set. This simulates the performance of the random numbers, or quick picks.

To simulate the low frequency lottery strategy, find the number of matches of the five least frequent numbers of a group in the next set of thirty draws (i.e. least frequent numbers of G_n in $Set_{n+1,}$ where n is the predicting period). For example, find the number of appearances of the bottom five numbers from group 1-120 in set 121-150. This gives the number of matches if a person took the five least frequent numbers and used them in the next 30 lottery drawings. We have developed a method to break ties, in the event that there are more than five most frequent/least frequent numbers. This rule is discussed in the next section.

To simulate the high frequency strategy, repeat the same procedure in the low frequency strategy, but using the five most frequently occurring numbers instead of the five least frequent numbers to project winning numbers in the next set. Use the same method to break ties. Similar procedures are repeated for the mega numbers.

For this study, the performance of a strategy is gauged by the average amount of matches to the winning numbers chosen in thirty draws. In the lottery, performance is actually measured by the amount of money won. This is determined draw by draw, and depends on number of ticket holders who won the same prize, jackpot size, and the combination of lottery and mega numbers matched. However, these variables change over time and cannot be controlled. Therefore, we use the amount of matches as a general gauge of performance; in general, the more matches, the more money won.

Tie Breaking Procedures:

A problem arose about how to group the numbers in case of a tie. We used the most frequent or least frequent numbers, including tied ones, and found their matches in the next thirty draws. Then we averaged the number of matches, and called this is the performance of the strategy—the average number of hits per frequent/infrequent number. For the random strategy, we counted the total number of matches, then divided by five. This is also the average number of hits per number, so we can compare the three strategies.