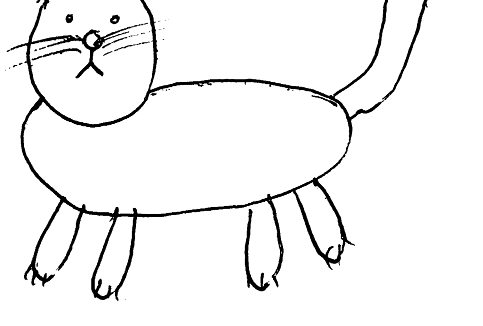
G13PJ1 Third Year Project
Interim Report
Tuesday 5th November 1996

"It Could Be You"
An investigation into the randomness of the National Lottery

Ian Nelson

Hi Mr Supervisor of Nelly's project.



G13PJ1 Interim Report

Introduction

The UK National Lottery has now been in operation for almost two years, with around 65 million tickets being bought each week. With huge fortunes up for grabs, it is thus important to ensure that the winning numbers chosen are selected at random. Further, when looking at lists showing the number of people winning the various prizes each week, there appears to be a huge spread of values. For instance, in week 9 a staggering 133 people shared the jackpot prize, against an expected value of 5. This would seem to suggest that punters' choices of numbers are non-random. It is therefore also interesting to see what the proportion of people winning prizes each week tells us about the manner in which they select their numbers.

In summary then, this project is mostly concerned with answering the following questions:

- Are the numbers chosen by the lottery machines random? If not, why not?

- Are the numbers chosen by gamblers random? If not, in what way are they non-random?

Progress To Date

The initial task was to acquire the data required - lists of winning numbers, sales figures, number of people winning the various prizes, etc. The vast majority of this was directly from the operators of the National Lottery, Camelot[A]. However, some data (most notably the sales figures and details of which ball set and machine have been used each week) were obtained from the World-Wide Web[B].

Having obtained the relevant data, I began performing a number of simple chi-squared tests upon the frequencies with which the 49 numbers have been selected by the lottery machines[1]. Considering all seven balls for every week returns a p-value of 0.78 when compared to a χ^2_{48}, which is far from significant. Using the information regarding which ball set and machine have been used for each draw, I obtained the frequencies for the 49 numbers for each machine (of which there are three[2]) and ball set (of which there are eight). Performing more chi-squared tests failed to reveal anything of great significance - the lowest p-value returned was 0.077, which came about when considering the bonus balls selected by Guinevere.

It has been pointed out (Joe, 1993)[C] that we need to modify the standard goodness-of-fit statistic $\sum_k \{O_k - E_k\}^2 / E_k$ in tests such as those described above. This is to account for the fact that we are sampling several balls at a time, without replacement. I intend to evaluate the modified form of the test statistic and compare the result in the near future.

---

[1] Unless otherwise specified, all calculations and results given refer to the first 100 lottery draws, i.e. from 19 November 1994 through to 12 October 1996.

[2] "Arthur", "Guinevere" and "Lancelot".

I have been fortunate enough to have been provided with a number of papers by John Haigh, one of which includes a list of 31 tests of randomness used by Camelot to test the randomness of the machines. Some, for instance, are particularly dependent on the order in which the balls were drawn, whilst others are more concerned with the length of the gap between a particular number having been drawn. I intend to select a variety of these tests and evaluate the distribution of the test statistic, either manually (for some of the simpler tests, such as "Evens": X = Number of even numbers) or using Minitab to manipulate the ball values (for the more calculation-intensive tests, such as "Variance": $X=\left\{7\sum x_i^2-\left\{\sum x_i^2\right\}\right\}/2$). The null distribution of the statistic can be calculated analytically in some cases, such as the "Evens" test described above. However, for others I feel that it may be simpler to simulate the statistic a large number of times, and judge the extremity of the observed statistic by comparing it to these simulated values. To this end, I have been writing a computer program in C to perform such simulations, and hope to have it completed within the next week.

Future Plans

Given that the winning numbers chosen each week are random (and, uninterestingly, it appears that this is likely to be the case!), I intend to turn my attention to the proportion of winning tickets, and what this tells us about the combinations chosen by gamblers. In order to do this, it is first necessary to calculate the probability of winning each of the five different prizes. This is possible by simple combinatorics:

Prize		Frequency	Probability	Odds
Jackpot		$1 = f_J$	7.15×10^{-8}	1 in 13,983,816
Match 5 + Bonus	(6C5)	$6 = f_B$	4.29×10^{-7}	1 in 2,330,636
Match 5	(6C5 x 42C1)	$252 = f_5$	1.80×10^{-5}	1 in 55,491
Match 4	(6C4 x 43C2)	$13{,}545 = f_4$	9.69×10^{-4}	1 in 1,032
Match 3	(6C3 x 43C3)	$246{,}820 = f_3$	1.77×10^{-2}	1 in 57
TOTALS		$260{,}624 = f$	1.86×10^{-2}	1 in 54

Suppose we assume that punters choices are random, and let the total number of different tickets be $F=\binom{49}{6}=13{,}983{,}816$. Then since we know the probability of winning each prize (f_i /F), and also the number of tickets sold (N), we can model the number of tickets winning each prize (X) by a binomial distribution, X ~ Bin (N , f_i/F). Further, since N is so large and f_i /F so small, this approximates extremely well to a normal distribution, X ~ N (μ , μ), where $\mu = Nf_i/F$. So we would expect $Z=(X-\mu)/\sqrt{\mu}$ to follow close to a standard Normal N (0 , 1) distribution.

The few experimental calculations I have carried out thus far, however, suggest that this model simply does not hold. Considering, for example, the total number of winners in Week 9[3], we obtain a Z value of 704.450! This week may have been an extreme case, but similar calculations for each week also fail to deliver values in the expected (-2 , 2) range. I intend to produce more detailed information regarding this model for inclusion in my final report.

[3] Week 9: 69,846,979 tickets sold; 2,105,521 prizes won.

The spectacular failure of this seemingly plausible model appears to indicate that the choices of combinations made by gamblers are far from random. This may not be as surprising as it initially sounds. There have been a number of occasions now when I have watched the lottery draw in company, and upon the result being announced, it is not unusual for someone to comment that "There probably won't be many winners" or "I expect lots of people will be sharing the jackpot". This in itself suggests that the methods people use to select their entries avoid certain "types" of combination, and favour others. For example, people appear to be wary of selecting consecutive numbers, yet at least two consecutive numbers should be drawn on average every other draw. I intend to devote a large section of this project to suggesting a number of similar factors which could influence the number of winners, and using the data available to see how realistic these hypotheses are. Some of the theories I intend to investigate include:

- **The Birthday Factor.** If gamblers choose numbers based on the birth dates of family and friends, this would result in a large proportion of the numbers selected being in the range [1,12] and [1,31].

- **Minimum Separation.** An extension of the idea regarding consecutive numbers, this gives some idea of the nature of the spread of the selections chosen.

- **Pretty Patterns.** How does the layout of the numbers on the playslip affect the choices made?

- **Lucky Numbers.** Does the appearance in the winning combination of traditionally lucky or unlucky values, such as 7 and 13 affect the number of winners? What about the appearance of those numbers which have occured most (or least) often in previous draws, such as 44 and 39?

- **Lucky Dip.** Has the existence of a "Lucky Dip" facility (which has been operating since draw 71) affected the number of tickets winning prizes? Approximately 8% of tickets are bought in this manner.

I also intend to investigate a number of other incidental matters, such as whether there is any correlation between the number of winners in each category.

[A] The National Lottery
PO Box 1010
Liverpool
L70 1NL

(0645) 100 000

[B] Richard K. Lloyd's Lottery Website: http://www.connect.org.uk/lottery/

[C] Joe H (1993): Tests of uniformity for sets of lotto numbers. Statistics and Probability Letters 16, 181-188