T2 Statistics Project

Patterns and Trends of the National Lottery

Andrew Ferrier
Candidate No: 0844

Royal Grammar School, Guildford
Centre No: 64480

Introduction

Aims

The Aims of this Project are:

- To analyse some of the statistics of the national lottery (especially sales information), and to see whether there are any trends, similarities, patterns, or correlations in the data.
- To check whether some of the statistical variables surrounding the national lottery fit a known statistical distribution.

How the Statistics Were Collected

The statistics were collected from the Internet. I originally found complete information on the first 187 lottery draws, but decided that this was too many to be easily manageable. Therefore I decided to randomly select 50 draws from this set of information. I used my calculator to generate random numbers from 1 to 187 (without generating previous numbers), and used this to select a final table of data which contained 50 randomly selected lottery draws, which I then sorted by draw number. The data will be reproduced in portions throughout the project. No difficulties in collecting data were encountered.

How the Lottery Works

People who play the lottery pick six different numbers on a ticket from the range 1 to 49. Any person can play the lottery as many times as they like, but each ticket sold counts separately in the final statistics.

Six numbers are selected using a a machine which selects six ‘random’ balls from a selection of balls marked 1 to 49. This is done without replacement. The machine also selects a ‘bonus ball’. There are three different machines, and there are ten separate ball sets. The machine and ball set that are used for each lottery draw are selected by two separate people. Camelot, the company that run the lottery, claim that this ensures the randomness of the lottery draw.

After the balls are selected by the machine, people can collect their prizes if they have matched enough balls. Possible valid matches (in descending prize order) are:

- All of the six main numbers matched (jackpot).
- Five of the numbers plus the bonus number matched.
- 5 numbers matched (not including the bonus ball).
- 4 numbers matched (not including the bonus ball).
- 3 numbers matched (not including the bonus ball).

Analysis of the Lottery Statistics

Sales Statistics

The first attempt I made at finding patterns in the national lottery was to look at the statistics for sales. This is because these are fairly easy to predict. A table summarising the sales statistics for all of the randomly chosen 50 draws is below.

Draw No.	Date	Month	Year	Total Ticket Sales	Total Prize Winners	Total Prize Fund
3	3	Dec	1994	48,263,533	888,165	21,718,590
15	25	Feb	1995	61,301,422	1,477,009	27,585,640
17	11	Mar	1995	67,688,637	1,386,419	30,459,887
18	18	Mar	1995	62,479,486	1,350,111	28,115,769
28	27	May	1995	74,771,757	1,430,675	33,647,291
29	3	Jun	1995	64,826,761	972,517	29,172,042
32	24	Jun	1995	64,505,245	1,046,991	29,027,360
34	8	Jul	1995	73,414,008	1,203,747	33,036,304
37	29	Jul	1995	63,812,748	716,986	28,715,737
38	5	Aug	1995	63,405,879	930,988	28,532,646
43	9	Sep	1995	64,604,276	1,106,058	29,071,924
46	30	Sep	1995	65,805,302	1,091,136	29,612,386
48	14	Oct	1995	65,828,800	1,201,199	29,622,960
49	21	Oct	1995	65,759,630	1,302,306	29,591,834
51	4	Nov	1995	65,551,611	1,418,881	29,498,225
52	11	Nov	1995	65,805,646	1,398,789	29,612,541
54	25	Nov	1995	65,820,282	1,086,258	29,619,127
64	3	Feb	1996	78,125,931	1,643,447	35,156,669
66	17	Feb	1996	73,903,189	1,415,035	33,256,435
74	13	Apr	1996	69,652,526	967,078	31,343,637
81	1	Jun	1996	68,733,675	896,342	30,930,154
83	15	Jun	1996	66,461,361	1,009,678	29,907,612
86	6	Jul	1996	68,289,582	978,993	30,730,312
94	31	Aug	1996	67,640,227	1,493,514	30,438,102
96	14	Sep	1996	67,769,246	1,228,158	30,496,161
98	28	Sep	1996	68,644,687	1,094,354	30,890,109
100	12	Oct	1996	70,065,289	1,156,406	31,529,380
103	2	Nov	1996	70,473,976	1,443,900	31,713,289
113	11	Jan	1997	69,141,121	1,290,684	31,113,504
114	18	Jan	1997	69,157,372	1,137,412	31,120,817
121	19	Feb	1997	28,494,520	557,952	12,822,534
124	1	Mar	1997	62,694,583	1,332,598	28,212,562
126	8	Mar	1997	61,970,334	1,271,751	27,886,650
129	19	Mar	1997	26,640,638	449,227	11,988,287
130	22	Mar	1997	64,546,216	1,038,922	29,045,797
136	12	Apr	1997	59,959,500	1,467,495	26,981,775
139	23	Apr	1997	25,715,330	640,194	11,571,899
141	30	Apr	1997	25,339,461	548,769	11,402,757
149	28	May	1997	25,129,823	441,463	11,308,420
152	7	Jun	1997	62,401,085	1,160,799	28,080,488
154	14	Jun	1997	64,550,792	1,298,118	29,047,856
159	2	Jul	1997	38,092,277	702,732	17,141,525

165	23	Jul	1997	27,646,490	568,542	12,440,921
167	30	Jul	1997	27,887,919	652,637	12,549,564
171	13	Aug	1997	27,281,835	582,801	12,276,826
172	16	Aug	1997	57,126,940	1,032,363	25,707,123
173	20	Aug	1997	27,103,614	486,633	12,196,626
176	30	Aug	1997	57,909,697	826,802	26,059,364
177	3	Sep	1997	26,641,840	487,958	11,988,828
179	10	Sep	1997	26,577,126	368,485	11,959,707

There ought to be a correlation between ticket sales and the total number of prizewinners (because the more tickets are bought, the more people should win prizes). To check whether this is the case, I carried out the following correlation calculation[1]:

H_0: There is no correlation between ticket sales and the total number of prizewinners ($\rho = 0$)
H_1: There is a correlation between ticket sales and the total number of prizewinners ($\rho \neq 0$).

$$S_{xx} = 1.443 \times 10^{16}$$

$$S_{yy} = 5.680 \times 10^{12}$$

$$S_{xy} = 2.445 \times 10^{14}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = 0.854$$

(r is the product-moment correlation coefficient)

Using critical value tables, the critical value for a sample size of 50 at a significance level of 5% (a typical significance level used) is 0.2353,

$$0.854 \succ 0.2353$$

The correlation coefficient is in excess of the required critical value, $\therefore$ I reject H_0, and I can conclude that there is a significant positive correlation between ticket sales and the total number of prizewinners. The correlation is more obvious when one looks at a scatter diagram of the data, together with a best-fit line. The calculation of the best-fit line and the scatter diagram are below.

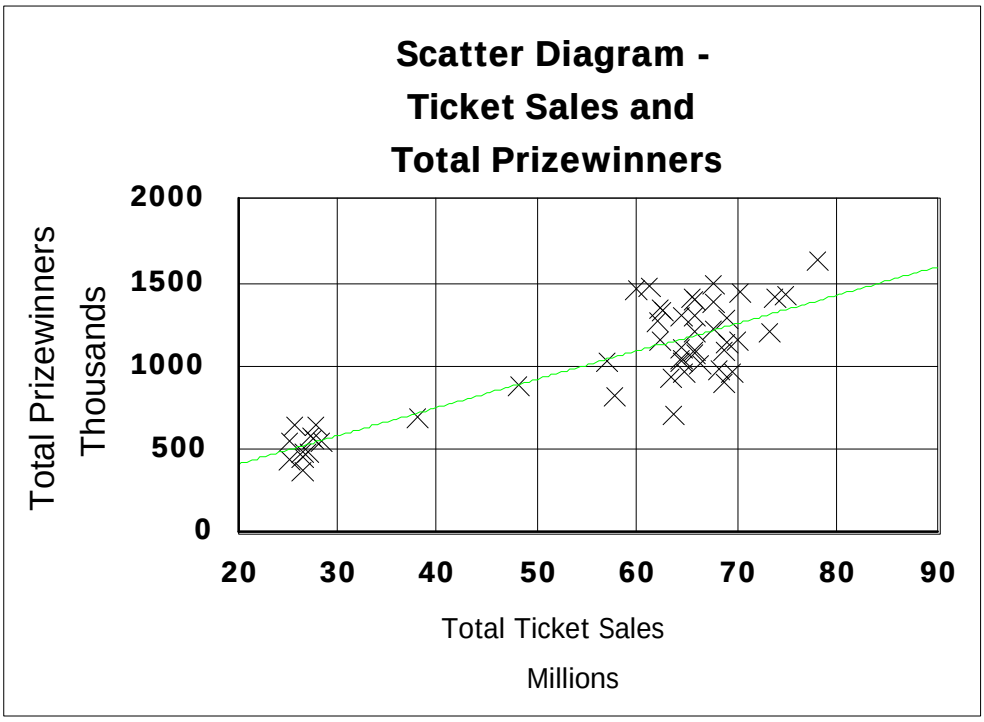
$$y = \alpha + \beta x$$

$$\beta = \frac{S_{xy}}{S_{xx}} = 0.017$$

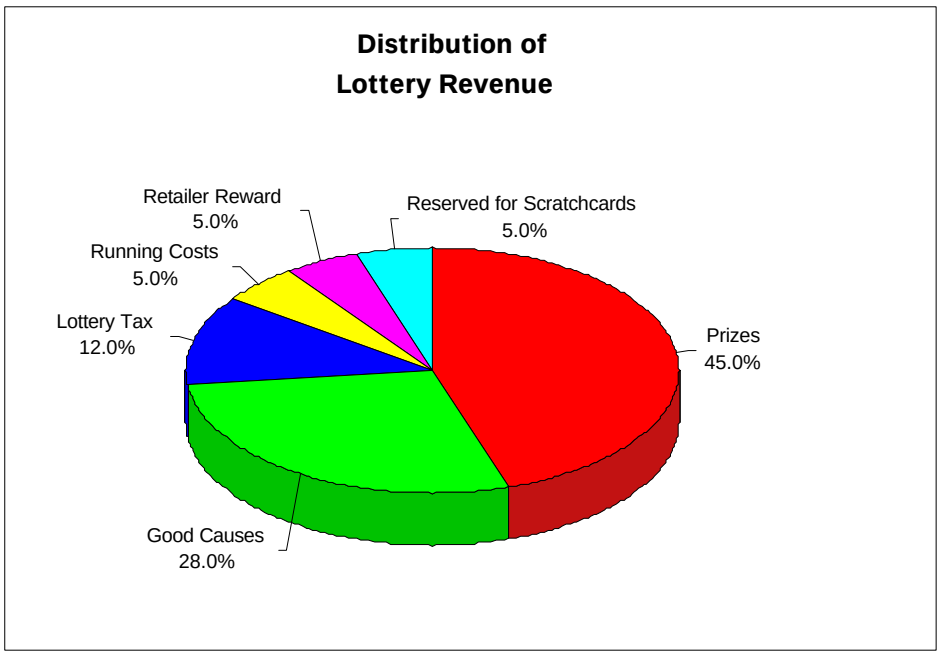
$$\alpha = \bar{y} - \beta\bar{x} = 1.034 \times 10^6 - (0.017 \times 5.671 \times 10^7) = 7.288 \times 10^4$$

$$\therefore y = 7.288 \times 10^4 + 0.017x$$

[1] Note: All calculated figures in this project, unless stated otherwise, are quoted to 3 decimal places. As many significant figures as possible are used in intermediate calculations.



The best-fit line has been drawn on the scatter diagram and it clearly shows the pattern - as ticket sales increase, in general, so do the number of prizewinners.



Because the total prize fund depends entirely on the number of sales (as shown above, 45% of the revenue generated by people playing the national lottery goes towards the prize fund), the correlation coefficient for the number of sales and the prize fund should be 1 (if two variables are in direct proportion, the correlation coefficient between them should be 1).

Thus, I checked the correlation coefficient between the number of sales and the prize fund:

H_0: The number of sales and the prize fund are in direct proportion ($r = 1$).
H_1: The number of sales and the prize fund are not in direct proportion ($r \neq 1$).

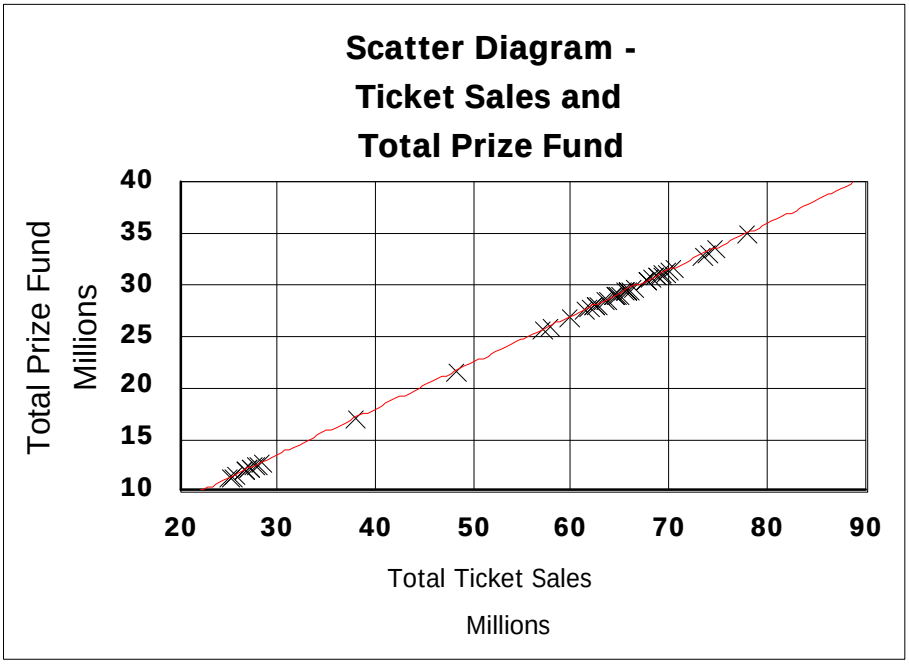
$$S_{xx} = 1.443 \times 10^{16}$$

$$S_{yy} = 2.923 \times 10^{15}$$

$$S_{xy} = 6.495 \times 10^{15}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = 1$$

$r = 1$, $\therefore$ I accept H_0: There is a direct proportion relationship between ticket sales and total prize fund, as I expected. This is shown on as a scatter diagram below, and the diagram is overlaid by a best-fit line, which goes through all the points.



Also, if there is a directly proportional relationship between these two factors, the correlation coefficient of total prizewinners and the total prize fund should be the same as that between total sales and total prizewinners ($r = 0.854$). I tested this:

H_0: There is no correlation between the number of total prizewinners and the total prize fund ($\rho = 0$).
H_1: There is a correlation between the number of total prizewinners and the total prize fund ($\rho \neq 0$).

$$S_{xx} = 2.923 \times 10^{15}$$

$$S_{yy} = 5.680 \times 10^{12}$$

$$S_{xy} = 1.100 \times 10^{14}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = 0.854$$

I reject H_0 - there is a correlation between the number of total prizewinners and the total prize fund, and the correlation coefficient is the same as before.

The Numbers Chosen by the Lottery Machine

Another test I performed was to see whether there was any correlation between the number chosen by the lottery machine for each of balls (1st ball drawn, 2nd ball drawn etc., up to the 6th ball chosen), and the total number of lottery winners that week. One might expect some correlation, if, for example, people in general tend to pick lower numbers. The data for the balls chosen, and the total number of winners each week, is presented in a table below.

Draw	Date	Month	Year	1st Ball	2nd Ball	3rd Ball	4th Ball	5th Ball	6th Ball	Total Prize Winners
3	3	Dec	1994	21	11	17	30	29	40	888,165
15	25	Feb	1995	18	33	8	31	5	10	1,477,009
17	11	Mar	1995	2	22	13	46	29	27	1,386,419
18	18	Mar	1995	41	19	31	18	9	24	1,350,111
28	27	May	1995	45	12	25	37	44	13	1,430,675
29	3	Jun	1995	31	1	29	40	21	32	972,517
32	24	Jun	1995	5	43	45	21	15	42	1,046,991
34	8	Jul	1995	3	14	11	20	1	40	1,203,747
37	29	Jul	1995	41	34	49	28	46	45	716,986
38	5	Aug	1995	35	1	25	30	45	8	930,988
43	9	Sep	1995	12	22	41	2	20	45	1,106,058
46	30	Sep	1995	11	33	40	10	32	29	1,091,136
48	14	Oct	1995	25	30	9	5	4	47	1,201,199
49	21	Oct	1995	17	19	2	21	6	47	1,302,306
51	4	Nov	1995	6	14	18	48	27	44	1,418,881
52	11	Nov	1995	23	28	48	10	7	30	1,398,789
54	25	Nov	1995	46	42	28	16	30	23	1,086,258
64	3	Feb	1996	2	32	44	22	9	26	1,643,447
66	17	Feb	1996	18	14	16	22	4	15	1,415,035
74	13	Apr	1996	38	47	23	44	49	40	967,078
81	1	Jun	1996	35	45	24	37	36	39	896,342
83	15	Jun	1996	47	25	18	44	13	46	1,009,678
86	6	Jul	1996	44	47	45	43	26	13	978,993
94	31	Aug	1996	27	3	5	47	14	44	1,493,514
96	14	Sep	1996	10	9	38	48	11	2	1,228,158
98	28	Sep	1996	19	26	23	39	36	31	1,094,354
100	12	Oct	1996	25	15	45	16	39	30	1,156,406
103	2	Nov	1996	48	35	43	23	32	7	1,443,900
113	11	Jan	1997	16	43	4	11	18	35	1,290,684
114	18	Jan	1997	22	48	3	31	21	26	1,137,412
121	19	Feb	1997	39	36	9	11	19	41	557,952
124	1	Mar	1997	33	23	49	8	2	42	1,332,598
126	8	Mar	1997	12	29	5	39	20	42	1,271,751
129	19	Mar	1997	40	16	48	13	29	17	449,227

130	22	Mar	1997	41	39	27	40	14	43	1,038,922
136	12	Apr	1997	9	37	27	44	42	29	1,467,495
139	23	Apr	1997	23	34	12	8	17	43	640,194
141	30	Apr	1997	16	40	10	17	24	8	548,769
149	28	May	1997	19	36	48	24	25	3	441,463
152	7	Jun	1997	21	30	40	25	14	12	1,160,799
154	14	Jun	1997	10	8	16	43	44	3	1,298,118
159	2	Jul	1997	18	6	41	22	10	45	702,732
165	23	Jul	1997	4	28	11	21	34	27	568,542
167	30	Jul	1997	12	19	3	45	6	44	652,637
171	13	Aug	1997	29	17	10	5	41	43	582,801
172	16	Aug	1997	2	49	20	9	19	38	1,032,363
173	20	Aug	1997	28	8	41	20	27	36	486,633
176	30	Aug	1997	33	46	30	38	6	20	826,802
177	3	Sep	1997	43	22	14	38	10	30	487,958
179	10	Sep	1997	31	20	45	43	35	32	368,485

I found the product-moment correlation coefficient between each ball drawn (1st drawn, 2nd drawn, and so on up to the 6th ball drawn) and the total number of prize winners. The calculation is not given, but it was performed in the same way as before. The resulting correlation coefficients are given in the table below:

Ball No.	$r =$
1	-0.240
2	-0.041
3	-0.084
4	0.127
5	-0.202
6	-0.092

Next I tested the correlation coefficients I had found:

H_0: There is no correlation between each ball drawn and the total number of prizewinners.
H_1: There is a correlation between each ball drawn and the total number of prizewinners.

The critical values for n = 50, at the 5% level, are ±0.2353 (found from tables).

Ball No.	Is the Correlation Coefficient Significant?
1	Significant
2	Not Significant
3	Not Significant
4	Not Significant
5	Not Significant
6	Not Significant

Thus there appears to be a correlation between the first ball drawn every week and the total number of prizewinners, although the correlation is only just valid, because the value of the correlation coefficient is very close to the critical value.

The correlation is negative, which implies that the lower the value of the ball chosen as the 1st ball, the higher the number of prizewinners that week. This may seem strange, but it can be explained.

One simple explanation, which may or may not be the case, is to do with birthdays: if people pick their birthday dates for lottery numbers, as many do, then low numbers (below 31) are more often picked. When these numbers are selected by the lottery machine, more people will win prizes.

Ball Set and Machine Used

After checking whether the number of prizewinners was influenced by the balls drawn, I checked whether there was a correlation between the ball set used and the number of prizewinners, as well as the machine used and the number of prizewinners. The data for the ball sets, the machine used and the total prizewinners for each week are presented below.

Machine Used	Ball Set	Total Winners
1	7	1386419
3	2	1643447
1	6	1032363
2	4	1203747
1	3	568542
2	2	1046991
3	7	1418881
2	3	1467495
1	4	1298118
3	2	1228158
3	8	1091136
1	4	1106058
3	5	1271751
1	10	652637
2	3	1290684
3	8	548769
1	2	1302306
2	1	702732
1	7	1477009
3	4	1415035
1	3	1094354
3	4	441463
2	2	888165
1	4	1160799
2	2	1137412
1	3	1398789
3	4	640194
3	8	1201199
3	1	1156406
1	3	1493514
2	7	486633
2	2	582801
2	4	972517
2	3	368485
3	1	1332598
2	6	826802
1	8	930988
2	2	896342
1	6	967078
3	3	557952
2	2	449227
2	5	1038922

2	7	716986
1	4	1350111
1	1	487958
3	8	978993
2	8	1430675
1	6	1086258
2	7	1009678
3	3	1443900

Again, the working is not given for the calculation of the product-moment correlation coefficient, but the coefficients are given in the table:

Correlation between Ball Set and Total Prizewinners:	$r = -0.011$
Correlation between Machine Used and Total Prizewinners:	$r = -0.045$

As can be seen, these coefficients are almost 0, and certainly do not indicate any relationship between the factors involved.

Is the Lottery Random?

Theoretically, the balls chosen by the national lottery machine should follow a uniform distribution. This is what one would expect if the system used is truly random. One can test how well the balls chosen fit a uniform distribution. First I counted the total number of times each ball occurred in my sample of 50 draws. The data is presented below in a stem and leaf diagram.

0	1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4
0	5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9
1	0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4
1	5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9
2	0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4
2	5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9
3	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4
3	5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9
4	0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4
4	5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9

Testing the numbers picked in the 50 draws at the 5% level,

H_0: The numbers fit a uniform distribution (they are random).
H_1: The numbers do not fit a uniform distribution (they are not random).

$$\text{Expected number of times each ball is picked} = \frac{6}{49} \times 50 = 6.122$$

Ball	Observed (O_i)	Expected (E_i)	$\frac{(O_i-E_i)^2}{E_i}$
1	3	6.122	1.592
2	7	6.122	0.126
3	5	6.122	0.206
4	4	6.122	0.736
5	6	6.122	0.002
6	5	6.122	0.206
7	2	6.122	2.775
8	7	6.122	0.126
9	7	6.122	0.126
10	9	6.122	1.353
11	7	6.122	0.126
12	6	6.122	0.002
13	6	6.122	0.002
14	5	6.122	0.206
15	3	6.122	1.592
16	7	6.122	0.126
17	6	6.122	0.002
18	7	6.122	0.126
19	7	6.122	0.126
20	7	6.122	0.126
21	7	6.122	0.126
22	7	6.122	0.126
23	7	6.122	0.126
24	4	6.122	0.736
25	7	6.122	0.126
26	4	6.122	0.736
27	7	6.122	0.126
28	5	6.122	0.206
29	7	6.122	0.126
30	9	6.122	1.353
31	6	6.122	0.002
32	5	6.122	0.206
33	3	6.122	1.592
34	3	6.122	1.592
35	5	6.122	0.206
36	4	6.122	0.736
37	5	6.122	0.206
38	5	6.122	0.206
39	5	6.122	0.206
40	9	6.122	1.353
41	8	6.122	0.576
42	5	6.122	0.206
43	10	6.122	2.457
44	10	6.122	2.457
45	10	6.122	2.457
46	5	6.122	0.206
47	6	6.122	0.002
48	7	6.122	0.126
49	4	6.122	0.736

$$\Sigma\left[\frac{(O_i-E_i)^2}{E_i}\right] = 29.0$$

$$\text{Degrees of Freedom} = y = 49 - 1 = 48$$

$$\chi^2_{48}(5\%) = 65.17$$

$$29.0 \prec 65.17$$

$\therefore$ I do not reject H_0: there is no evidence to suggest the selection of the first ball is not random.

Conclusions

The Conclusions drawn from this project are summarised below:

- There is a very significant correlation between the number of lottery winners and the number of ticket sales, as would be expected.
- The total prize fund is in direct proportion to ticket sales, as claimed by Camelot, the lottery operators.
- There is a negative correlation between the first ball picked and the total number of prizewinners. The pattern suggests picking high numbers would be a good idea to maximize potential jackpot earnings.
- There is no correlation between the ball set used and the total number of prizewinners.
- There is no correlation between the machine used and the total number of prizewinners.
- There is no evidence that the selection of the lottery balls is not random.