PROBABILITY DISTRIBUTIONS IN LOTTERY GAMES

Jiří Rozkovec

Abstract

The aim of this paper is to analyse a probability distribution of percentiles of a lottery game, Sportka, and a distribution of drawn numbers respectively. This lottery has been run by the Sazka company since 1957. A history of the results of all the draws is published on the Internet, but no other statistical analysis of the distribution of drawn numbers has been performed. There is no information which enables someone making a bet to check whether the results show any statistically significant difference between the empirical and expected frequencies. From the probability theory point of view, the results of a lottery game like Sportka can be described by a discrete random variable which should have a uniform discrete distribution. The first part of this paper compares empirical and theoretical frequencies using the chi-square, goodness-of-fit test with a given significance level (this level will be maintained for all tests performed). Further, the distributions of the minimal, maximal and median values of drawn numbers are analysed, where the chi-square test was used again. The tests were performed for every year of the history of the game, i. e. 1957-2011.

Key words: lottery games, probability distribution, chi-square goodness-of-fit test, distribution of percentiles

JEL Code: C12, C16

Introduction

Nowadays many lotteries exist in the world and a lot of people bet on them. The betting agencies publish the results of all the draws (Simon 251-5), but there is no obligation to perform any tests of the distribution of the drawn numbers. Former studies have mostly tried to analyse the empirical distribution and to set an optimal strategy for bettors (Finkelstein 202-7, Johnson and Klotz 662-8, Stern and Cover 980-5). This paper contains an analysis of this distribution from various points of view using the chi-square goodness-of-fit test (Genest, Lockhart and Stephens 243-8).

1 Historical background

The lottery game Sportka, which has been run by the Czech company Sazka since 1957, was chosen. The data (history of the results) which was used is available on (SPORTKA). First, a brief description of the history of the game.

The lottery started in 1957 with one draw (hereinafter as Draw I) with six drawn numbers from 1 to 49. The draws took place on Sundays. During 1965 Sazka launched a second draw (Draw II) also with six drawn numbers on Sundays. Since 1977 both draws have had seven numbers and from 1995 the lottery has been played also on Wednesdays. Finally, from 2006 there has been one extra draw per year, usually in 37th week. A summary of the history is contained in Tab. 1.

Tab. 1 History of Sportka

Period (year/week)	Draw I	Draw II
1957/16 - 1965/13	6 numbers, one draw per week	Was not performed
1965/14 - 1976/52	6 numbers, one draw per week	6 numbers, one draw per week
1977/1 – 1995/14	7 numbers, one draw per week	7 numbers, one draw per week
1995/15 – ...	7 numbers, two draws per week	7 numbers, two draws per week

Source: (SPORTKA)

2 The Distribution of the Drawn Numbers

It is clear that it should be a discrete uniform distribution where each number from 1 to 49 has the same probability of occurrence The frequencies of particular numbers were evaluated over a long period (for each year) and Draw I and Draw II were separated. There was no point in mixing the draws together as they are in fact two independent games. For such a comparison of the empirical and theoretical frequencies the Chi-Square goodness-of-fit test (hereinafter ChST) is used (for details see Anděl 155-6, Genest, Lockhart and Stephens 243-51) . The used confidence level is still 95%.

Concerning the individual steps of the test, the following calculations were performed:

i) the null hypothesis: for the j^{th} year, the theoretical frequencies *n'(i,j,0)* should be equal for all 49 classes, as the distribution should be uniform. For all drawn values *i=1,…,49* the following formula is obtained:

$$n_{i,j,0} = \frac{f(j) \cdot d(j)}{a(j)} \qquad i=1,...,49,\ j=1957,...,2010 \qquad (1)$$

where *f(j)* is the number of draws per year. It varies from 50 to 53, from 104 to 107 respectively, *d(j)* is the number of drawn numbers within the draw. It is 6 (from 1957), then it is 7 (from 1977), *a(j)* = 49, as they still draw from the set {1, 2, ..., 49}.
ii) the test statistic χ has χ^2 distribution with 48 degrees of freedom, because there are 49 classes, and the critical value is $\chi^2(48) = 65{,}17$. The critical region is the interval $[65{,}17; +\infty)$.

Tab. 2 Chi-Square test Statistics

Year	**1957**	**1958**	**1959**	**1960**	**1961**	**1962**	**1963**	**1964**	**1965**
Draw I	48,06	41,80	39,24	23,78	35,17	43,56	44,82	39,48	31,47
Draw II	---	---	---	---	---	---	---	---	50,33
Year	**1966**	**1967**	**1968**	**1969**	**1970**	**1971**	**1972**	**1973**	**1974**
Draw I	33,80	38,87	39,73	37,60	33,20	37,60	36,97	35,71	45,13
Draw II	36,68	41,03	32,55	48,28	57,70	35,71	59,58	49,22	47,02
Year	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**
Draw I	29,31	31,24	50,62	40,68	52,23	40,12	34,19	42,79	54,92
Draw II	59,82	34,76	32,58	52,30	36,08	38,50	61,65*	42,00	47,92
Year	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**
Draw I	40,35	34,59	40,12	46,67	53,31	56,27	49,54	38,71	34,31
Draw II	54,35	47,22	40,92	59,57	44,42	31,23	53,58	35,96	45,57
Year	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**
Draw I	64,88*	53,85	56,00	50,58	38,63	41,06	49,94	37,02	56,54
Draw II	53,85	45,50	43,69	46,75	34,06	53,71	41,06	35,00	28,40
Year	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Draw I	39,58	35,54	42,13	39,98	38,53	37,25	46,53	39,38	43,47
Draw II	36,75	45,50	45,17	42,67	55,60	55,08	36,27	38,47	36,80

Source: own calculations based on the data from (SPORTKA)

In Tab. 2 are the test statistics of the Chi-Square test for the Draw I and Draw II from the period 1957-2010.

It can be seen that no test statistic falls into the critical region, so it is not possible to reject the hypothesis about the discrete uniform distribution of the drawn numbers at the 95% confidence level. To verify this conclusion the maximal values of the test statistic for the both draws can be found in this table. They are (denoted by *): for Draw I in 1993 (χ = 64,88), for Draw II in 1981 (χ = 61,65).

Nevertheless, neither are in the critical region. Further, it is interesting to note that there are particular years where not all the numbers were drawn. Even though the uniformity was not rejected. These numbers and periods are in the Tab. 3.

Tab. 3 Undrawn Numbers

Year	Number	Draw
1958	41	I
1962	32	I
1965	39	II
1985	2	II
1987	30	II

Source: own calculations based on the data from (SPORTKA)

3 The Distribution of the Drawn Numbers

Now the distribution of the minimum, maximum and median of the drawn numbers is investigated. Such characteristics have their own distribution. Only the distribution of the minimum is described here: the other distributions are analogies. Furthermore, only the results from 1977 onwards are dealt with here since when seven numbers have been drawn.

The range of the minimal value of the drawn numbers is from 1 to 43 (it is impossible to have, for instance, 46 as the minimum, if seven numbers are drawn from the set {1,...,49}). If the minimum value is represented as a random variable $\boldsymbol{X}$, then the probabilities are:

$$P(X=i)=\frac{\binom{49-i}{6}}{\binom{49}{7}} \quad i=1,...,43 \qquad (2)$$

When the empirical and theoretical frequencies are tested (again by ChST), for Draw I the test statistic is χ=180,57. The critical value is $\chi^2(42)$=58,12, so it is possible to reject the null hypothesis that the minimal drawn number has the distribution stated above. The same is true for Draw II as well (χ=133,83).To illustrate this point, Tab. 4 presents both frequencies (for Draw I).

Tab. 4 Minimal Drawn Number – Theoretical and Empirical Frequencies – Draw I

Nr.	Theor. Fr.	Emp. Fr.	Nr.	Theor. Fr.	Emp. Fr.	Nr.	Theor. Freq.	Emp. Fr.
1	0,1429	0,1966	16	0,0129	0,0084	31	0,0002	-
2	0,1250	0,1702	17	0,0105	0,0067	32	0,0001	-
3	0,1090	0,1320	18	0,0086	0,0034	33	0,0001	-
4	0,0948	0,0994	19	0,0069	0,0022	34	0,0001	-
5	0,0822	0,0820	20	0,0055	0,0006	35	0,0000	-
6	0,0710	0,0764	21	0,0044	0,0017	36	0,0000	-
7	0,0611	0,0528	22	0,0034	0,0017	37	0,0000	-
8	0,0523	0,0404	23	0,0027	0,0017	38	0,0000	-
9	0,0447	0,0309	24	0,0021	0,0011	39	0,0000	-
10	0,0380	0,0169	25	0,0016	0,0017	40	0,0000	-
11	0,0321	0,0202	26	0,0012	0,0011	41	0,0000	-
12	0,0271	0,0146	27	0,0009	0,0006	42	0,0000	-
13	0,0227	0,0140	28	0,0006	0,0017	43	0,0000	-
14	0,0189	0,0112	29	0,0005	-			
15	0,0157	0,0090	30	0,0003	0,0006			

Source: own calculations based on the data from (SPORTKA)

Now the results concerning the maximum and median, starting with the probabilities. For maximum the probability function is

$$P(X=i)=\frac{\binom{i-1}{6}}{\binom{49}{7}} \quad i=7,\ldots,49. \qquad (3)$$

And for median

$$P(X=i)=\frac{\binom{i-1}{3}\binom{49-i}{3}}{\binom{49}{7}} \quad i=4,\ldots,46. \qquad (4)$$

The test statistics are in the Tab. 5.

Tab. 5 Maximal and Median of Drawn Numbers – Test Statistics

	Maximum	**Median**
Draw I	74,22	47,36*
Draw II	88,78	30,44*

Source: own calculations based on the data from (SPORTKA)

Thus it can be said that the empirical distribution of the maximum does not fit the theoretical one (the test statistics are greater than the critical value 58,12), but for the median value these statistics are in the region of acceptance (denoted by *). In which case, it is not possible to reject the hypothesis about the distribution of the median represented by formula (4).

Conclusion

An attempt was made to analyze the results of the lottery game Sportka throughout its whole history. The distribution of the drawn numbers was of particular interest because such results are presented very exceptionally. Firstly a hypothesis about their uniform discrete distribution in every year from 1957 to 2010 was tested. The null hypothesis was not rejected at 95% confidence level. Secondly, a probability function of some statistics (maximum, minimum

and median) as constructed to test a second hypothesis. In this case, the null hypothesis about the distribution (see formulas (2), (3) and (4)) only for the median was accepted. For the minimum and maximum of the drawn numbers the null hypotheses were rejected.

References

Anděl, Jiří. *Statistické metody*. Praha: MatfyzPress, 2007.

Finkelstein, Mark. „Estimating the frequency-distribution of the numbers bet on the California lottery“ *Applied Mathematics and Computation*. 69:2-3 (1995): 195-207.

Genest, Christian, Lockhart, Richard A., and Stephens, Michael A. „Chi-square and the Lottery“ *Journal of the Royal Statistical Societ, Series D-The Statistician*. 51 (2002): 243-257.

Johnson, Richard, and Klotz, Jerome. „Estimating hot numbers and testing uniformity for the lottery.“ *Journal of the American Statistical Association*. 88:422 (1993): 662-668.

Simon, Jonathan. “An analysis of the distribution of combinations chosen by UK National Lottery players“ *Journal of Risk and Uncertainity*. 17:3 (1998): 243-276.

SPORTKA. “Statistiky losovaných čísel”. Sazka, a.s. 17.5.2012.
< http://www.sazka.cz/cz/loterie-a-hry/sportka/statistiky/archivy-losovanych-cisel/>.

Stern, Helman, and Cover, Thomas M. “Maximum-entropy and the lottery” *Journal of the American Statistical Association*. 84:408 (1989): 980-985.

Contact

Jiří Rozkovec

Technical University of Liberec, Faculty of Economics, Department of Economic Statistics, Studentska 2, 46117 LIBEREC, Czech Republic

jiri.rozkovec@tul.cz