On the preference for prime numbers: The case of lotto players

Patrick Roger, Tristan Roger

On the preference for prime numbers: The case of lotto players

Patrick Roger* Tristan Roger†

March 9, 2024

Abstract

Lotto players do not usually select their numbers at random. The mental process of selecting numbers is called conscious selection. In this paper, we study the importance of prime numbers in conscious selection. Our study shows that prime numbers are more popular than non primes. It follows that betting on prime numbers is suboptimal in Lotto games, which obey the parimutuel principle. We demonstrate this result on up to 10 years of data from the Belgian National Lottery (Lotto and Euromillions data). Prime numbers are selected significantly more than non prime numbers. We control for potential confounding effects, evidenced in the literature on gambling, such as the small number preference and the lucky number preference (i.e., number 7 in Western countries).

Keywords: Prime numbers, Conscious selection, Games of chance, Lotteries.

Declarations: The authors have no relevant financial or non-financial interests to disclose. No funding was received for conducting this study.

Data: The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contribution statements: The two authors contributed to the study conception, design and execution. All authors participated in writing the manuscript and approved the final version.

*__Corresponding author__, LaRJE, Université de Nouvelle-Calédonie – Mailing address: IAE NC, Campus de Nouville, BP R4, 98851 Nouméa Cedex FRANCE – Email: patrick.roger@unc.nc, and LaRGE Research Center, EM Strasbourg Business School, university of Strasbourg

†ICN Business School, CEREFIGE, Université de Lorraine – Mailing address: 86 rue Sergent Blandan, 54000 Nancy, France – Email: tristan.roger@icn-artem.com

1 Introduction

A prime number is a natural number strictly greater than 1, that is not a product of two smaller natural numbers. In other words, a number is prime when it can be divided only by one and by itself. For centuries, prime numbers have fascinated mathematicians. Understanding their distribution have always been a topic of interest since Euclide have proved (around 300BC) that there exists an infinite number of primes. The attraction surrounding prime numbers is not limited to mathematicians. An article from the New York Times, published on January 26, 2018 explained how a church deacon found the biggest prime number.[1]

Prime numbers do not only captivate mathematicians but also the general public. For instance, for the 50$^{\text{th}}$ anniversary of the Arecibo message sent to potential extraterrestrial intelligence in 1974, a new thirteen page message called "A beacon in the galaxy" has been prepared by NASA's Jet Propulsion Laboratory to update the initial information contained in the Arecibo message. An important part of the basic mathematics included in this message is the list of prime numbers lower than 100, and the largest prime number known when the message was composed ((see page 15 of Jiang et al., 2022)). Other examples of everyday focus on prime numbers can be found in cinema/TV: the TV series *Numb3rs* refers to prime numbers, in particular the episode 5 of the first season which directly quotes the Riemann hypothesis and the distribution of prime numbers. Similarly, the Washington Post (September 26, 2016) titled an article "The magnificent seven and 5 other movies with prime numbers in their titles".[2] Since prime numbers are taught early

[1] *How a church deacon found the biggest prime number yet (it wasn't as hard as you think)*, by Valencia Prashad, New York Times, January 26$^{\text{th}}$ 2018, `https://www.nytimes.com/2018/01/26/science/prime-number-mersenne-church.html`. People interested to participate in this search for the biggest prime number can be guided through the project GIMPS (Great Internet Mersenne Prime Search), `https://www.mersenne.org/`

[2] Unfortunately there is a mistake in this paper because one of the mentioned movies has the number 1 in its title, and 1 is not a prime number.

on in school, (around the 4th grade depending on the country)[3] most people have been exposed to prime numbers to some extent.

In this paper, we test whether the specific role of prime numbers in mathematical education and the image of mystery and magic that prime numbers carry in newspapers articles and movies can influence (possibly unconsciously) the number selection of lottery players. Over the last three decades, the literature on lottery gambling has repeatedly demonstrated that lottery players do not choose their numbers at random, even though official draws of lotto games do are random (Baker and McHale, 2009, 2011; Cook and Clotfelter, 1993; Farrell et al., 2000; Papachristou and Karamanis, 1998; Roger and Broihanne, 2007; Roger et al., 2023; Simon, 1998; Turner, 2010; Wang et al., 2016). Cook and Clotfelter (1993) were the first to name this non-random choice of lottery numbers "conscious selection". Explanations for such conscious selection relate to people being superstitious (e.g., playing lucky numbers) or playing specific combinations of numbers (e.g., birthday dates). In most Western countries, 7 is viewed as a lucky number and therefore overplayed (Roger, 2011; Wang et al., 2016; Polin et al., 2021; Roger et al., 2023). In China, 8 is the typical lucky number (Shum et al., 2014; Brown and Mitchell, 2008; Hirshleifer et al., 2018). The status of the number 13 remains unclear. Although it was historically considered unlucky, the number 13 is overplayed by Lotto players in countries such as France (Roger and Broihanne, 2007), Belgium (Roger et al., 2023) or the Netherlands (Wang et al., 2016). Players can also be victims of conscious selection[4]

[3]See for example `https://curriculum.illustrativemathematics.org/k5/teachers/grade-4/unit-1/lesson-3/lesson.html` for France and `https://www.rtbf.be/embed/m?id=2734620` for Belgium.

[4]The expression "Conscious selection" seems to refer to conscious choices, but preferences for numbers can be unconscious. For example, in the criminal context, Dhami et al. (2020) show that sentencers prefer certain numbers when meting out the sentence length (for custody and community service) and penalty amount (for fines/compensation).

because they bet on birthday dates or prefer some combinations of numbers.[5] While it is difficult to rationalize gambling on state lotteries (Stetzka and Winter, 2023), betting on popular numbers when playing a Lotto game looks even more irrational. Indeed, the parimutuel principle[6] underlying Lotto games implies that the more winners there are, the lower amount of money is earned. It follows that betting on popular numbers decreases the expected value of gains because individual prizes are lower when popular numbers show up in the official draw.

To test whether Lotto players exhibit a preference for prime numbers, we use data from the Belgian National Lottery. We focus on the two most popular lotteries, namely the Lotto game (a 6/45 game with one bonus number) and the Euromillions lottery (a 5/50 game with two bonus numbers). Over our sample period, an average of 595,343 (623,852) players participate in each of the two weekly draws of Lotto (Euromillions). The figures can be compared to 11 million people that compose the national population. We perform univariate and multivariate analyses on a sample of 1044 Lotto draws over a period of ten complete years (2014-2023) and 758 Euromillions draws from September 26, 2016 to the end of the year 2023.[7] Overall, our results indicate that players bet more on prime numbers than on non-prime numbers. We are able to demonstrate this feature both on the Lotto game and the Euromillions game. Our results are also robust to considering potential confounding effects. Specifically, we control, among others, for the small number effect and the lucky number effects. Indeed, since the frequency of prime numbers is higher for small numbers and since lucky numbers are often prime number (at least in western

[5]Baker and McHale (2011) quote the example of a draw of the Canadian lottery (March 19th, 2008), with 23, 40, 41, 42, 44 and 45 as the winning numbers, and 43 as the bonus number. Though the jackpot was not won, an extremely high number of players (239) were second-rank winners (5 correct numbers plus the bonus number). This example reveals that sequences may be considered as optimal choices by some players.

[6]In a parimutuel game, a given amount is devoted to winners and shared among all winners.

[7]We select this specific starting date for the Euromillions game because it is the first draw where two bonus numbers are drawn in the range [1;12]. Before this date, the two bonus numbers were drawn in the [1;11] range; the change also induced a switch between winning ranks 6 and 7 because of the winning probability transformation.

countries), the prime numbers effect we evidence could have been the result of both the small number and the lucky number effects. Our findings are unchanged. We complete our analysis by a specific study of bonus numbers. We confirm our preceding results that the proportion of winners with the bonus number(s) is higher when the bonus number(s) is(are) prime(s).

The remainder of this paper is organized as follows. Section 2 describes the functioning of the games under investigation, presents the data and the descriptive statistics of our two samples (Euromillions lottery and Lotto game). Section 3 presents the methodology and Section 4 the results. Finally, Section 5 concludes the paper.

2 Lotto and Euromillions games: presentation, data and descriptive statistics

2.1 General presentation

In general, playing a lotto game consists in betting on a combination of k numbers out of $N >> k$, without replacement. When the official draw occurs, k numbers are drawn at random, and b bonus numbers are added to the draw (in most games, $b = 1$ or $b = 2$). Therefore, such a game is partly characterized by the triple (N, k, b). Depending on the type of game, the bonus numbers can be drawn from the set of the remaining $N - k$ numbers after the main draw. This is the case for the Belgian Lotto. The bonus numbers can also be drawn from an independent set of numbers, as it is the case for the Euromillions lottery (the two bonus numbers are drawn in the [1;12] range without replacement). In short, the Lotto game is characterized by $(N, k, b) = (45, 6, 1)$ and the Euromillions game by $(N, k, b) = (50, 5, 2)$. In addition to the differences in the triple (N, k, b), the main

difference between the two lotteries relates to the way the bonus numbers are drawn.

These two lotteries obey the parimutuel principle which applies as follows. First, a takeout rate (close to 50%) is applied to the global amount of bets. The remainder is redistributed to the winners, according to prespecified sharing rules over a set of winning ranks. There are currently 9 winning ranks in the Lotto game and 13 in the Euromillions lottery.[8]. The description of the winning ranks and the corresponding probabilities can be found in Table 3. The amount M to be redistributed is shared across ranks, a given percentage of M being devoted to each rank.

2.2 Data

We use unique data provided by the Belgian National Lottery, which provides, for each draw, the global amount of bets and the number of players. As far as we know, the Belgian National Lottery is the only lottery that publicly provides such data. We downloaded the data from the website of the Belgian National Lottery.[9] The data consists, for each game - Lotto and Euromillions - of the date of the draws, the numbers drawn, including bonus numbers, the number of players and the global amount of bets. In addition, the data also contains, for each winning rank, the number of winners and the corresponding individual prize. Our sample period ranges from January 1, 2014 to December 31, 2023 for the Lotto game and from September 26, 2016 to December 31, 2023 for the Euromillions lottery. In total, there are 1044 Lotto draws and 758 Euromillions draws.

2.3 Descriptive statistics

[8]Details can be found in the legal document (Arrêté royal, March 21, 2018) available at `https://www.etaamb.be/fr/arrete-royal-du-01-avril-2016_n2016003257.html`)

[9]`https://www.loterie-nationale.be/`

Table 1: Summary statistics

Panel A: Belgian lotto (1,044 draws)					
	Mean	Median	Standard deviation	Minimum	Maximum
# players	595343.36	632392.00	153819.81	361014.00	1000140.00
Overall bet	3986660.27	4404242.38	1296490.48	2088777.00	8324598.75
# tickets sold	3524365.35	3694550.50	1189963.99	1806075.00	7643020.00
Mean number drawn	23.11	23.17	5.05	7.83	38.50
# prime numbers	1.82	2.00	1.07	0.00	5.00
# prime bonuses	0.30	0.00	0.46	0.00	1.00
# non-prime odds	1.34	1.00	0.97	0.00	4.00
# of Numbers ≤ 30	3.98	4.00	1.10	0.00	6.00
# of Numbers ≤ 12	1.59	2.00	1.03	0.00	5.00
Average bet per player	5.81	5.79	0.56	4.87	7.64
Average popularity	0.02	0.02	0.00	0.01	0.03
Average bonus ratio	0.08	0.07	0.01	0.05	0.12
Panel B: Euromillions lottery (758 draws)					
	Mean	Median	Standard deviation	Minimum	Maximum
# players	623852.43	602010.00	206322.33	346236.00	1653185.00
Overall bet	4901619.87	4553965.00	2130003.87	2401880.00	16278395.00
# tickets sold	1960647.95	1821586.00	852001.55	960752.00	6511358.00
Mean number drawn	25.53	25.80	5.87	9.00	40.40
# prime numbers	1.53	2.00	0.99	0.00	5.00
# prime bonuses	0.88	1.00	0.66	0.00	2.00
# non-prime odds	1.12	1.00	0.92	0.00	4.00
# of Numbers ≤ 30	3.02	3.00	1.02	0.00	5.00
# of Numbers ≤ 12	1.18	1.00	0.89	0.00	4.00
Average bet per player	3.07	3.02	0.26	2.65	3.94
Average popularity	0.05	0.05	0.00	0.04	0.07
Average bonus ratio	0.37	0.37	0.03	0.26	0.45

This Table presents summary statistics.

The descriptive statistics related to the set of variables we will use in our analysis are summarized in Table 1. On average, 595,343 (623,852) players participated in each Lotto (Euromillions) draw. The average amount spent on each draw is 4 million euros for Lotto and 4.9 million euros for Euromillions, corresponding to purchases of 5.8 (3.0) Lotto (Euromillions) tickets per player. The other variables in Table 1 show that our overall sample behaves like a random sample of draws as expected. For example, the average value of numbers drawn should be close to 23 since [1;45] is the range in which numbers are randomly drawn. The average over the 1044 draws is 23.11, not significantly different from 23 ($t - stat = 0.7336$, $p - val = 0.4634$). The same remark applies to Euromillions with a sample average of 25.53 versus an expected value of 25.5 (numbers in the range [1;50]), with $t - stat = 0.1584$ and $p - val = 0.8742$.

Regarding the distribution of numbers in the different draws, our sample has an average of 3.98 numbers less than 30 when the expected value is $6 \times 30/45 = 4$. For Euromillions, the sample average is 3.02 and the corresponding expected value is $5 \times 30/50 = 3$.

For the Lotto game, there are 14 prime numbers (2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43) while there are 15 for the Euromillions lottery (2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47). The expected number of prime numbers in a draw is $6 \times 14/45 = 1.866$ for Lotto and $5 \times 15/50 = 1.5$ for Euromillions. The corresponding sample values are 1.82 and 1.53 which are not significantly different from the expected values ($t = -1.33, p - value = 0.1828$ for Lotto and $t = 0.8848, p - value = 0.3745$ for Euromillions).

Regarding the distribution of prime numbers across draws, Table 2 shows that, for each of the two games, no significant deviation appears with respect to the distribution of primes in random draws. The χ^2 statistics are equal to, respectively, 4.0488 and $6.117 = 68$. The corresponding p-values are 0.6701 and 0.2950.

Table 2: Distribution of prime numbers

	Belgian Lotto			Euromillions lottery		
	Theoretical	Empirical	Statistics	Theoretical	Empirical	Statistics
0 prime number	94.37	100		116.14	113	
1 prime numbers	304.90	323	$\chi^2 = 4.0488$	280.98	263	$\chi^2 = 6.1168$
2 prime numbers	367.01	361		245.86	272	
3 prime numbers	209.72	190		96.85	88	
4 prime numbers	59.66	63	p-val= 0.6701	17.09	20	p-val= 0.2950
5 prime numbers	7.95	7		1.07	2	
6 prime numbers	0.38	0		-	-	

Table 3 provides detailed information on the number of winners per rank. The first key point explaining our methodological choices described in the next section relates to the median number of winners for the first ranks. For Euromillions, more than 50% of draws did not have a winner with 5 correct numbers, corresponding to ranks 1 to 3. For Lotto, the median number of winners with 6 correct numbers is also zero. Therefore we cannot build a popularity index using only the winners who find all correct numbers. A meaningful popularity index must account for winners at lower ranks (as in Turner, 2010). We present a rigorous definition of our popularity index in Section 3.

The second key point related to Table 3 is the large variation between the minimum and the maximum number of winners at a given rank. For intermediate ranks, this ratio is approximately 10. There are two main explanations for these variations. The first is the number of tickets sold; sales are strongly driven by the amount of the jackpot and, even if people were choosing their numbers at random, the number of winners at a given rank would be proportional to sales. However, the second reason for variations in the number of winners is conscious selection, especially the well documented preference for small numbers (Boland and Pawitan, 1999; Otekunrin et al., 2021; Turner, 2010; Wang et al., 2016).

Table 3: Number of winners per rank

Panel A: Belgian lotto							
	Mean	Median	Standard deviation	Minimum	Maximum	# winning combinations	Winning probability
Rank 1 (6+0)	0.44	0.00	0.73	0	5	1	1.2277×10^{-7}
Rank 2 (5+1)	2.48	2.00	2.40	0	21	6	7.3664×10^{-7}
Rank 3 (5+0)	97.80	86.00	51.63	22	507	228	2.7992×10^{-5}
Rank 4 (4+1)	245.18	222.00	120.35	75	1083	570	6.9981×10^{-5}
Rank 5 (4+0)	4544.38	4222.00	1885.43	1603	13187	10545	1.2946×10^{-3}
Rank 6 (3+1)	6072.00	5678.50	2483.81	2216	17314	14060	1.7262×10^{-3}
Rank 7 (3+0)	72904.98	70620.50	26853.98	30188	157363	168720	2.0714×10^{-2}
Rank 8 (2+1)	54655.67	51497.50	20490.62	23264	137380	126540	1.5536×10^{-2}
Rank 9 (1+1)	177222.88	173500.00	56062.02	91991	345694	442890	5.4375×10^{-2}
Panel B: Euromillions lottery							
	Mean	Median	Standard deviation	Minimum	Maximum	# winning combinations	Winning probability
Rank 1 (5+2)	0.02	0.00	0.13	0	1	1	7.1510×10^{-9}
Rank 2 (5+1)	0.26	0.00	0.52	0	3	20	1.4302×10^{-7}
Rank 3 (5+0)	0.61	0.00	0.94	0	7	45	3.2180×10^{-7}
Rank 4 (4+2)	3.22	2.00	2.99	0	20	225	1.6090×10^{-6}
Rank 5 (4+1)	63.15	55.00	34.20	15	227	4500	3.2180×10^{-5}
Rank 6 (3+2)	142.43	120.00	81.48	36	682	9900	7.0796×10^{-5}
Rank 7 (4+0)	142.68	124.00	74.93	44	546	10125	7.2405×10^{-5}
Rank 8 (2+2)	2036.61	1763.50	1087.60	640	9491	141900	1.0147×10^{-3}
Rank 9 (3+1)	2788.25	2451.00	1303.00	1071	9319	198000	1.4159×10^{-3}
Rank 10 (3+0)	6281.70	5509.50	2918.49	2470	22417	445500	3.1858×10^{-3}
Rank 11 (1+2)	10672.03	9237.00	5487.83	3416	45557	744750	5.3258×10^{-3}
Rank 12 (2+1)	39856.91	35982.50	17780.20	17113	124479	2838000	2.0295×10^{-2}
Rank 13 (2+0)	89617.56	80795.00	39699.82	40476	312464	6385500	4.5664×10^{-2}

Columns 1 to 5 report statistics on the number of winners per rank. Panel A (B) refers to the Belgian lotto (Euromillions lottery). Columns 6 and 7 give the number of winning combinations and the winning probability of the corresponding ranks. The total number of combinations is 8,145,060 for the Belgian lotto (Panel A) and 139,838,160 for the Euromillions lottery (Panel B). Over our sample period, there was an important modification for the Belgian lotto: A 9^{th} rank of gain was introduced on May 26, 2018. This 9^{th} rank (1 correct number plus the bonus number) pays a fixed amount equal to the ticket price, which is currently €1.25 (€1 before May 26, 2018). Let N_p and N_s denote the cardinals of the two sets of numbers (principal set and stars/bonus) used to draw the winning combination, n_p (n_s) denote the number of numbers drawn in the principal set (set of stars/bonus), and k_p (k_s) the number of correct numbers (stars/bonus) for a given ticket. The equations below provide the probability of winning with k_p and k_s correct numbers and stars/bonus.
For the Belgian lotto, we have:

$$P(k_p, k_s) = \frac{\mathbf{C}(n_p, k_p)\mathbf{C}(N_p - n_p - n_s, n_p - k_p - k_s)}{\mathbf{C}(N_p, n_p)} \tag{1}$$

For the Euromillions lottery, we have:

$$P(k_p, k_s) = \frac{\mathbf{C}(n_p, k_p)\mathbf{C}(N_p - n_p, n_p - k_p)\mathbf{C}(n_s, k_s)\mathbf{C}(N_s - n_s, n_s - k_s)}{\mathbf{C}(N_p, n_p)\mathbf{C}(N_s, n_s)} \tag{2}$$

where $\mathbf{C}(M, j)$ is the number of combinations of j numbers out of M without replacement.

3 Methodology

The general approach we follow in this paper is based on the fact that a significantly larger proportion of winners in a given draw (compared to what is expected under random choice) means that the numbers that show up in this draw are more popular, on average, than what is expected if players were to choose their numbers randomly. To measure the popularity of a given number in our sample, we build on the popularity index described in Roger et al. (2023). The changes we introduce are justified by the difference of purpose of the two studies. Roger et al. (2023) study variations in popularity across time (of number 19) while our study focuses on a comparison of popularity across subsets of numbers (primes vs. non-primes).

Studying the popularity of a subset of numbers implies to distinguish between the main draw (6 numbers in the Lotto game and 5 numbers in Euromillions lottery) and the draw of bonus numbers. For the main draw, it is enough to determine the aggregate proportion of winners for all the ranks without the bonus number(s). The first column of Table 3 shows that the relevant subset of ranks is {1, 3, 5, 7} for the Lotto game and {3, 7, 10, 13} for the Euromillions game. We will therefore compare the proportions of winners across draws in these subsets of ranks as a function of the number of prime numbers showing up in the draw (without considering here the bonus number).

A second step will reinforce our analysis. If the null hypothesis is that prime numbers are not more popular than non prime numbers (also called composite numbers), there should be no difference between winning ranks where the bonus number is prime, compared to ranks where the bonus number is not prime. In the following subsection, we provide the technical details of the construction of the two popularity indices, without and with bonus.

3.1 The popularity index for the main draw

For the sake of simplicity, we present here the index for the lotto game. It is then adapted easily to the Euromillions game. Figure 2 provides the frequencies of individual numbers for the two games. Each number shows up in more than 100 (50) draws for the Lotto (Euromillions) sample. Hence, averaging the proportion of winners in draws in which a given number $n = 1, 2, ..., 45$ appears provides a good estimate of the popularity of number n. This is the basis of the popularity index described more precisely below.

When a popular number appears in a draw, the proportion of winners deviates from the theoretical probability of winning. We mentioned above that the relevant subset of ranks for the lotto game is {1, 3, 5, 7}. The last column of Table 3 shows that the expected proportion of winners for this subset of ranks is $1.22 \times 10^{-7} + 2.80 \times 10^{-5} + 1.29 \times 10^{-3} + 2.07 \times 10^{-2}$, which is approximately equal to 2.2%. In fact, over the 1,044 draws of our sample, the average proportion of winners for these 4 ranks is 2.198%. However, the range of values goes from a minimum of 1.4505% to a maximum of 3.36907%.

Quantities below are indexed by i corresponding to the draw. The notations and definitions are the following:

- G_i is the number of tickets sold for draw i;
- $\Theta(n)$ is the set of draws in which n appears;
- R is the relevant set of ranks used to calculate the popularity index. As mentioned previously, $R = \{1, 3, 5, 7\}$ for the lotto game ($R = \{3, 7, 10, 13\}$ for the Euromillions game);
- $W_i(R)$ is the cumulated number of winners at draw i for the set of ranks R;
- For $i \in \Theta(n), \Gamma(n, i) = \frac{W_i(R)}{G_i}$ is the cumulated percentage of winners at ranks in R

at draw i. This is an estimate of the popularity of n for a single draw in which n shows up.

- $\Gamma(n) = \frac{1}{\#\Theta(n)} \sum_{i \in \Theta(n)} \Gamma(n, i)$ is the average popularity score for number n over draws in the sample period. It is our popularity index when bonus numbers are not considered.

3.2 The popularity index for bonus numbers

Consider the definition of ranks in Table 3. Winning ranks 2, 4, 6 correspond respectively to 5, 4, 3 correct numbers in the main draw plus the correct bonus number, and ranks 3, 5, 7 correspond respectively to 5, 4, 3 correct numbers in the main draw without the bonus number. Therefore, the last column of Table 3 allows us to calculate the expected proportion of winners with the bonus number in the set of winners with or without the bonus number. We will call this index the Bonus Ratio.

As an illustration, consider ranks 4 and 5; for these two ranks, 4 correct numbers have been found in the main draw but the bonus number is also correct for a rank-4 winner (and not for a rank-5 winner). If players choose randomly their numbers, the proportion of winners at rank 4 (4 correct numbers plus the bonus number) among all winners that found 4 correct numbers in the main draw (rank 4 + rank 5) is the ratio obtained with the corresponding probabilities in the last column of Table 3, that is $7.00 \times 10^{-5}/(7.00 \times 10^{-5} + 1.29 \times 10^{-3})$. In other words, in the set of winners with 4 correct numbers, we expect that 5.15% have also the correct bonus number. If the bonus number is popular (unpopular) and played by a lot of (few) people we expect the calculated percentage to be higher (lower) than 5.15%.

We define below the notations and the process that provide the popularity index of a

bonus number n. The reasoning developed in the preceding paragraph is aggregated over ranks 2, 4, 6. We sum the number of winners with 5, 4, 3 correct numbers and the correct bonus number and we divide by the total number of winners with 5, 4, 3 correct numbers (cumulated number of winners over ranks 2 to 7). Denote $R = \{2, 3, 4, 5, 6, 7\}$ the set of ranks with 3 to 5 correct numbers and $R_b = \{2, 4, 6, \}$ the set of ranks with 3 to 5 correct numbers with the correct bonus number.

- $G_{R,i}$ is the cumulated number of winners over R in draw i;
- $\Theta(n)$ is the set of draws in which n appears as the bonus number;
- $Bonus\Gamma(n) = \frac{1}{\#\Theta(n)} \sum_{i \in \Theta(n)} G_{R_b,i}/G_{R,i}$ is the average popularity score for number n over draws in the sample period (Bonus Ratio).

3.3 Permutation tests

The difficulty to compare the popularity of the 14 (15) prime numbers to the 31 (35) non prime numbers of the Lotto (Euromillions) game comes from the small size of the sets of prime numbers which prevents to use a standard t-test. To address this issue, we use a nonparametric permutation test.[10] We describe the test for the Lotto game. Each of the 45 numbers is characterized by a popularity score. Therefore, the most simple measure of the difference of popularity between prime numbers and non-prime numbers is the difference between the average popularity of the 14 primes and the average popularity of the 31 non primes. To check whether this difference is significant, we simulate 100,000 times a pair of random draws, one draw of 14 numbers in the complete set of 45 numbers and one draw of 31 numbers in the same complete set. We therefore get a simulated distribution of the difference in average popularity between random draws of 14 numbers and random draws

[10] See Ludbrook and Dudley (1998) for a detailed description of the permutation tests.

of 31 numbers. We then position the observed difference (average popularity of primes minus average popularity of non primes) on the distribution of differences to determine the p-value of the test and check whether the difference between the popularity of prime and composite numbers is significant.

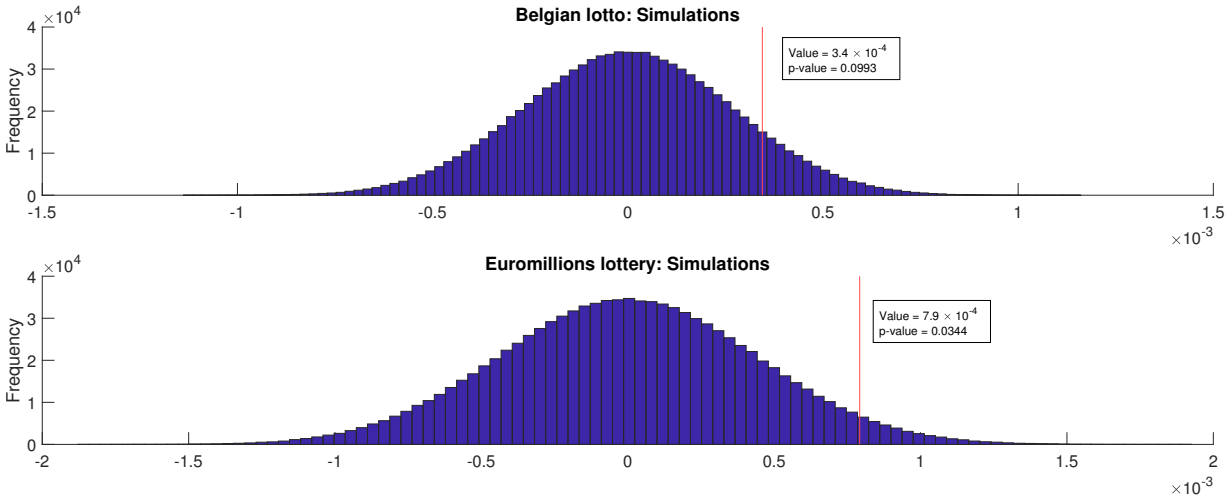
3.4 Econometric analyses

To demonstrate that prime numbers are more popular than non-prime, we need to take into account several confounding factors beyond the univariate analysis of the role of primes as a driver of the proportion of winners. Even if permutation tests provide significant results, they do not control such for confounding factors. The literature on lotteries reveals that many different factors can explain a substantial part of the variance of the proportion of winners (e.g. Papachristou and Karamanis (1998); Roger and Broihanne (2007); Roger et al. (2023)). Moreover, the bonus numbers are not considered in the above permutation tests.

There are at least three control variables that must be taken into account for the Lotto game and four for the Euromillions lottery. First, an important observation from the study of conscious selection is that people prefer small numbers (e.g., Figure 1 in Wang et al. (2016) or Figure 3 in Roger and Broihanne (2007)). This preference is often justified by the tendency of people to bet on birthday dates, leading to overbet on numbers less than 30. This potential confounding factor is important in our case because the average value of prime numbers (20.07) is significantly lower than the average of non-prime numbers (24.32). We will therefore introduce a control variable defined as the the number of numbers lower than 30 in the draw. As a robustness check, we will replace this variable by the average value of numbers in the draw which is an alternative way to take into account the small number effect. Second, there are two draws per week for each game, and the

draws occurring on Fridays and Saturdays (week-end draws) attract more players than those on Tuesdays and Wednesdays (midweek draws). In particular, the number of tickets per player is larger on week-end draws. Such a seasonality may be important for two reasons. We cannot be sure that week-end players are the same as midweek players, and second, buying more tickets means selecting more numbers, possibly allowing players to change their selection process from one ticket to the next (under the Parimutuel principle, it is not optimal to play the same ticket several times). In other words, we expect that two tickets played by two different players are less "different" than two tickets played by the same player. Third, we built two popularity indices, one based on the proportion of winners without bonus, the other being the Bonus Ratio. The number of prime numbers in the set of bonus members can influence the popularity measured on the main draw, at least for the Euromillions lottery for which bonus numbers are drawn from an independent set. In fact we expect the number of prime numbers in the set of bonus numbers to be negatively linked to the popularity index defined on the main draw in subsection 3.1 if our conjecture is correct. In short, if people overbet on prime bonus numbers, the Bonus Ratio increases but the proportion of winners without the bonus numbers decreases. Finally, all prime numbers are odd numbers, except number 2. It could therefore happen that what we identify as a preference for prime numbers is simply a preference for odd numbers. To control this potential effect we use a variable counting in each draw the number of odd numbers that are not prime.

Figure 1: Simulations



4 Empirical study

4.1 Univariate analysis

Figure 1 provides the results of the simulations for the two games. As mentioned above, 100,000 pairs of samples of respectively 14 numbers and 31 numbers are randomly drawn. For each of the 100,000 pair we calculate the difference of average popularity, hence building a dietribution of differences of popularity taking into account size differences. For the Euromillions lottery, the process is the same, except that each pair is composed of respectively 15 and 35 numbers. Not surprisingly, the two distributions are close to normal because of the simulation process. The p-value is just below 10% for the Lotto game and below 5% for the Euromillions game. These positive preliminary results show that taking into account control variables is necessary to assess the existence of the prime number effect.

Table 4: Proportion of winners and prime numbers

	Belgian lotto			Euromillions lottery		
	Proportion of winners (without bonus)	Proportion of winners (without bonus)	Ratio bonus winners over all winners	Proportion of winners (without bonus)	Proportion of winners (without bonus)	Ratio bonus winners over all winners
Mean numbers	-0.0004***			-0.0005***		
	(-29.5100)			(-22.8600)		
# Numbers ≤ 30		0.0018***	0.0004		0.0030***	-0.0003
		(28.3600)	(1.0800)		(24.4600)	(-0.3000)
# prime numbers	0.0003***	0.0004***	-0.0000	0.0008***	0.0011***	0.0001
	(4.0000)	(5.8600)	(-0.0400)	(5.5400)	(8.3500)	(0.1100)
# prime bonus numbers	-0.0002	-0.0001	0.0040***	-0.0011***	-0.0012***	0.0182***
	(-1.0400)	(-0.3600)	(5.1400)	(-5.6400)	(-6.1300)	(11.3900)
# non-prime odds number	-0.0000	-0.0000	-0.0005	-0.0002	-0.0001	0.0005
	(-0.3900)	(-0.1800)	(-1.3300)	(-1.3600)	(-0.6200)	(0.4300)
Bet per player	-0.0003	-0.0003**	0.0016*	0.0006	0.0003	0.0001
	(-1.6200)	(-1.9900)	(1.8400)	(1.1200)	(0.5700)	(0.0200)
Wednesday dummy	-0.0003	-0.0003	0.0019*			
	(-1.4000)	(-1.5800)	(1.9600)			
Tuesday dummy				0.0003	0.0001	0.0001
				(0.9400)	(0.3600)	(0.0300)
Number of observations	1044	1044	1044	758	758	758
R^2	0.4730	0.4540	0.0260	0.4630	0.4930	0.1410

t-statistics are in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10%, respectively.

4.2 Multivariate analysis

Following the explanations given in section 3.4, we ran the following basic regression model:

$$\begin{aligned}\text{Proportion of winners}_t = \alpha &+ \beta_1 \text{Mean numbers}_t + \beta_2 \#\text{Prime numbers}_t \\ &+ \beta_3 \#\text{Prime bonus numbers}_t + \beta_4 \#\text{Non prime odd numbers}_t \\ &+ \beta_5 \text{Bet per player}_t + \beta_6 \text{Day of the week dummy}_t + \epsilon_t \end{aligned} \tag{3}$$

where *Proportion of winners*$_t$ is the proportion of winners for draw t. The proportion of winners is calculated using ranks 1, 3, 5 and 7 for the Lotto game and ranks 3, 7, 10 and 13 for the Euromillions game. All these ranks are ranks without bonus. *Mean numbers*$_t$ is the average value of numbers drawn in draw t, *#Prime numbers*$_t$ is the number of prime numbers in draw t, *#Prime bonus numbers*$_t$ is the number of prime numbers in the bonus numbers at draw t, *#Non prime odd numbers*$_t$ is the number of non prime odd numbers at draw t, *Bet per player*$_t$ is the average bet per player at draw t, and *Day of the week dummy*$_t$ is a dummy variable that is equal to 1 for the midweek draw (Wednesday for the Lotto and Tuesday for the Euromillions) and 0 otherwise. In a second regression model, we consider the variable *# Numbers* $\leq$ *30*$_t$ instead of *Mean numbers*$_t$. The variable *# Numbers* $\leq$ *30*$_t$ corresponds to the number of numbers drawn that are less than or equal to 30.

Table 4 summarizes the results, not only of the above regression model but also of the equivalent model for the bonus ratio (column 4 for the Lotto and 7 for the Euromillions). Columns 2 to 4 are related to Lotto and columns 5 to 7 to Euromillions. For each of the two games, the small number effect is taken into account through two control variables: the average value of numbers drawn (columns 2 and 5) and the number of numbers drawn that are less than or equal to 30 (columns 3 and 6). We will comment only the models corresponding to the above regression (columns 2 and 5), the comments being the same

for the alternate model (columns 3 and 6). We discuss later the regression with the Bonus ratio as dependent variable.

In line with previous literature, we find highly significant coefficients for the average value of drawn numbers. Despite the low average value of prime numbers, compared to non-prime numbers, we observe that the number of prime numbers is highly significant (t-stat of 4.00 for Lotto and 5.54 for Euromillions). An interesting difference appears between the two games. For the Lotto game, the number of prime bonus numbers is not significant. This is not surprising since the bonus number is drawn in the same set as the numbers of the principal draw. On the contrary, the equivalent variable for Euromillions is highly significant with a negative sign. The interpretation is the following. If people prefer prime numbers, they will often choose prime numbers for the two bonus numbers as well. When one or the two bonus numbers are prime, the proportion of winners without bonus decreases even if people play prime numbers on the main draw. Since our popularity index (i.e. proportion of winners) is built using winning ranks without bonus, popular prime bonus numbers will not be captured by our variable. This interpretation is confirmed on columns 4 and 7 where the regression aims at explaining the bonus ratio. The number of prime bonus numbers is highly significant with t-stats equal to 5.14 for Lotto and 11.39 for Euromillions.

This result is a strong signal in favor of our assumption that there is a preference for prime numbers, compared to non prime numbers. Finally, we observe that the number of odd non prime numbers is not significant, moreover with negative coefficients. It reveals that the preference for prime numbers is not a preference for odd numbers.

In the Appendix, we perform robustness checks by introducing two more control variables. First, if the small number effect is caused by players betting on birthday numbers, there could be a difference between numbers less than of equal to 30 (days) and numbers less than or equal to 12 (months). We could expect numbers less than or equal to 12 to

be more represented because of people playing at the same time their birth day and birth month. The second element introduced in the robustness check assessed in Table 5 is a dummy variable equal to 1 if the number 7 shows up in the draw. Of course, 7 is a prime number but we could argue that people bet on number 7 because it is perceived as a lucky number, not as a prime number. Table 5 shows that adding the number 7 dummy variable which is highly significant improves the adjusted R^2 and decreases slightly the significance of the prime number variable, especially for the lotto game when the small number effect is measured by the average value of numbers drawn. However, from a global perspective, the same variables as before are significant, confirming the prime number effect.

5 Conclusion

Our study used Belgian data to test whether Lotto and Euromillions players exhibit a preference for prime numbers. Specifically, we performed statistical and econometric analyses on a sample of 1044 Lotto draws over the period January 2014 to December 2023 and 758 Euromillions draws from September, 26, 2016 to December 2023. These high quality datasets are publicly available and provide information about the number of players and the amount spent by players in each draw. To accurately measure the popularity of a given number in the lottery data, we built a popularity index that is inferred from the actual proportion of winners among a subset of ranks. The main advantage of this methodology is that it enables us to identify variations in conscious selection over time and across games. Our results show that prime numbers are more popular than non prime numbers. We demonstrate this preference in two ways. First, studying the proportions of winners at ranks without bonus numbers, we show that the number of prime numbers in a draw is a significant driver of the proportion of winners, even after controlling for the average value of the numbers drawn, the presence of the number 7 in the draw, the day

of the week and the number of other odd numbers that are not prime. Second, we built a second popularity index only based on bonus numbers. We obtain similar results. The number of prime bonus numbers in the draw is a strong determinant of the proportion of winners in the set of ranks that include at least a bonus number.

References

Baker, R.D., McHale, I.G., 2009. Modelling the probability distribution of prize winnings in the UK national lottery: Consequences of conscious selection. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172, 813–834. doi:10.1111/j.1467-985X.2009.00599.x.

Baker, R.D., McHale, I.G., 2011. Investigating the behavioural characteristics of lottery players by using a combination preference model for conscious selection. Journal of the Royal Statistical Society: Series A (Statistics in Society) 174, 1071–1086. doi:10.1111/j.1467-985X.2011.00693.x.

Boland, P.J., Pawitan, Y., 1999. Trying to be random in selecting numbers for lotto. Journal of Statistics Education 7.

Brown, P., Mitchell, J., 2008. Culture and stock price clustering: Evidence from the peoples' Republic of China. Pacific-Basin Finance Journal 16, 95–120. URL: http://www.sciencedirect.com/science/article/pii/S0927538X0700025X, doi:https://doi.org/10.1016/j.pacfin.2007.04.005.

Cook, P.J., Clotfelter, C.T., 1993. The peculiar scale economies of Lotto. The American Economic Review 83, 634–643. URL: http://www.jstor.org/stable/2117538.

Dhami, M.K., Belton, I.K., Merrall, E., McGrath, A., Bird, S.M., 2020. Criminal sentencing by preferred numbers. Journal of Empirical Legal Studies 17, 139–163. doi:10.1111/jels.12246.

Farrell, L., Hartley, R., Lanot, G., Walker, I., 2000. The demand for Lotto: The role of conscious selection. Journal of Business & Economic Statistics 18, 228–241. doi:10.1080/07350015.2000.10524865.

Hirshleifer, D., Jian, M., Zhang, H., 2018. Superstition and financial decision making. Management Science 64, 235–252. doi:10.1287/mnsc.2016.2584.

Jiang, J.H., Li, H., Chong, M., Jin, Q., Rosen, P.E., Jiang, X., Fahy, K.A., Taylor, S.F., Kong, Z., Hah, J., Zhu, Z.H., 2022. A beacon in the galaxy: Updated arecibo message for potential fast and seti projects. Galaxies 10. doi:10.3390/galaxies10020055.

Ludbrook, J., Dudley, H., 1998. Why permutation tests are superior to t and F tests in biomedical research. The American Statistician 52, 127–132. doi:10.1080/00031305.1998.10480551.

Otekunrin, O.A., Folorunso, A.G., Alawode, K.O., 2021. Number preferences in selected nigerian lottery games. Judgment and Decision Making 16, 10601071. doi:10.1017/S1930297500008081.

Papachristou, G., Karamanis, D., 1998. Investigating efficiency in betting markets: Evidence from the Greek 6/49 Lotto. Journal of Banking & Finance 22, 1597–1615. doi:https://doi.org/10.1016/S0378-4266(98)00071-5.

Polin, B.A., Isaac, E.B., Aharon, I., 2021. Patterns in manually selected numbers in the Israeli lottery. Judgment and Decision Making 16, 1039–1059.

Roger, P., 2011. Testing alternative theories of financial decision making: A survey study with lottery bonds. Journal of Behavioral Finance 12, 219–232. doi:10.1080/15427560.2011.620200.

Roger, P., Broihanne, M.H., 2007. Efficiency of betting markets and rationality of players: Evidence from the French 6/49 Lotto. Journal of Applied Statistics 34, 645–662. doi:10.1080/02664760701236889.

Roger, P., D'Hondt, C., Plotkina, D., Hoffmann, A., 2023. Number 19: Another victim of the covid-19 pandemic? Journal of Gambling Studies 39, 1417–1450. doi:10.1007/s10899-022-10145-3.

Shum, M., Sun, W., Ye, G., 2014. Superstition and "lucky"apartments: Evidence from transaction-level data. Journal of Comparative Economics 42, 109–117.

Simon, J., 1998. An analysis of the distribution of combinations chosen by UK national lottery players. Journal of Risk and Uncertainty 17, 243–277.

Stetzka, R.M., Winter, S., 2023. How rational is gambling? Journal of Economic Surveys 37, 1432–1488. doi:https://doi.org/10.1111/joes.12473.

Turner, N.E., 2010. Lottery ticket preferences as indicated by the variation in the number of winners. Journal of Gambling Studies 26, 421–439.

Wang, T.V., van Loon, R.J.P., van den Assem, M.J., van Dolder, D., 2016. Number preferences in lotteries. Judgment and Decision Making 11, 243–259.

6 Appendix

t-statistics are in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10%, respectively.

Table 5: Robustness check: Number 7 and Numbers less than or equal to 12

	Belgian lotto			Euromillions lottery		
	Proportion of winners (without bonus)	Proportion of winners (without bonus)	Ratio bonus winners over all winners	Proportion of winners (without bonus)	Proportion of winners (without bonus)	Ratio bonus winners over all winners
Mean numbers	-0.0004***			-0.0005***		
	(-29.0200)			(-22.6100)		
# Numbers ≤ 30		0.0016***	0.0003		0.0028***	-0.0004
		(24.3500)	(0.7500)		(21.5600)	(-0.3900)
# Numbers ≤ 12		0.0005***	0.0003		0.0004**	0.0003
		(6.9900)	(0.6400)		(2.5300)	(0.2400)
# prime numbers	0.0001*	0.0002***	0.0000	0.0006***	0.0009***	-0.0000
	(1.7400)	(2.9100)	(0.0300)	(4.5600)	(6.7200)	(-0.0000)
# prime bonus numbers	-0.0001	-0.0001	0.0040***	-0.0011***	-0.0011***	0.0183***
	(-0.9300)	(-0.5300)	(5.1000)	(-5.7600)	(-6.2600)	(11.3800)
Dummy 7 drawn	0.0022***	0.0023***	-0.0006	0.0023***	0.0027***	0.0012
	(10.8900)	(11.5800)	(-0.5600)	(5.1400)	(6.3800)	(0.3300)
# non-prime odds number	-0.0000	-0.0000	-0.0005	-0.0002	-0.0001	0.0005
	(-0.6500)	(-0.4000)	(-1.3100)	(-1.1700)	(-0.4000)	(0.4400)
Bet per player	-0.0002	-0.0002	0.0016*	0.0005	0.0001	-0.0000
	(-1.4000)	(-1.5700)	(1.8400)	(0.8400)	(0.1200)	(-0.0100)
Wednesday dummy	-0.0002	-0.0002	0.0019*			
	(-1.1400)	(-1.1000)	(1.9600)			
Tuesday dummy				0.0002	0.0001	0.0001
				(0.8100)	(0.2100)	(0.0200)
Number of observations	1044.0000	1044.0000	1044.0000	758.0000	758.0000	758.0000
R^2	0.5260	0.5490	0.0240	0.4810	0.5270	0.1390

Figure 2: Frequency of numbers drawn

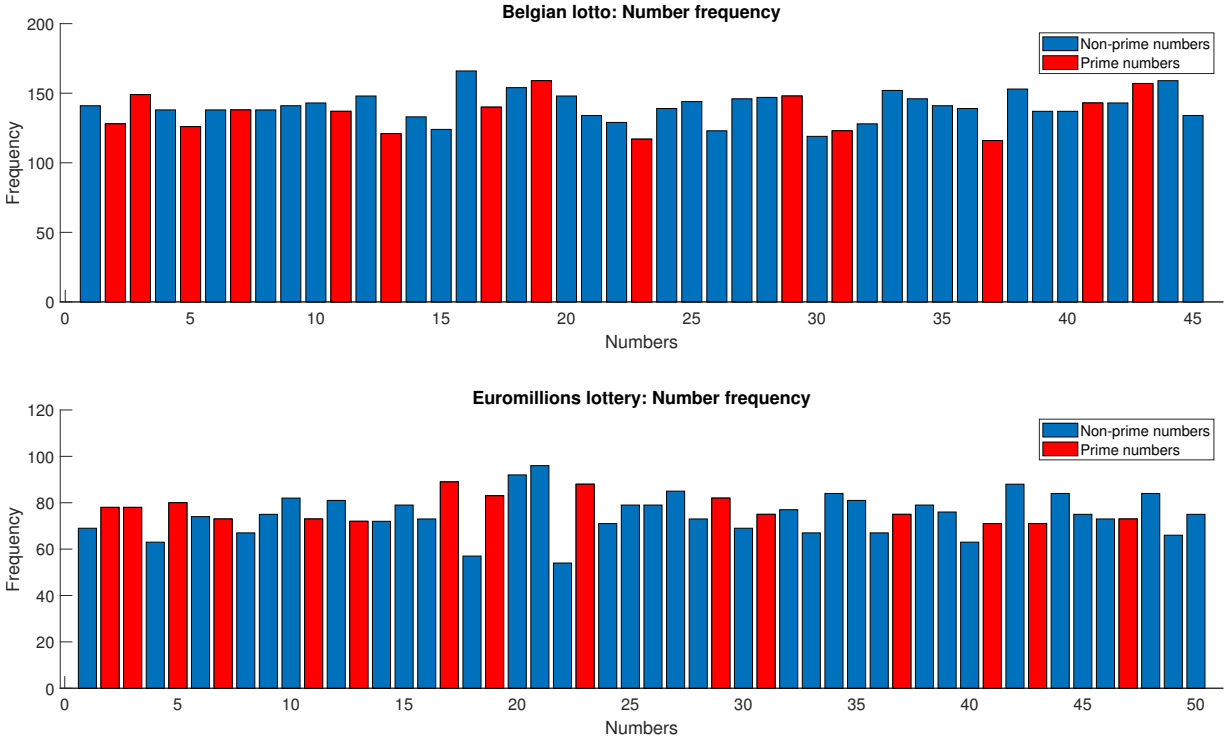


Figure 3: Popularity of numbers drawn