Repartition of Lotto Numbers among ranks

rupp(scd)

May 26, 2012

If $X \sim U(a_1 \dots a_n)$ where $a_1 > \dots > a_n$ are numbers and we pickup N independant realization of X: $X_1, \dots, X_N$ without repetition.

We create $Y_1, \dots, Y_N$ by ordering the X_i's e.g: $Y_1 < Y_2 < \dots < Y_N$.

We want to calculate the laws of the Y_i's.

when $X \sim U(a_1, \dots, a_n)$ then $P(X < a)$ is the sum $1/n \sum_{i=1}^{n} \delta(a_i \geq a)$

where $\delta(x) = 1$ if x=1(or x='TRUE') and 0 else.

In the case of Canadian Lotto, the repartition of each variable is:

$$P(X_i) < a = 1/49 \sum_{i=1}^{49} \delta(i \geq a). \tag{1}$$

We can modelize the lotery by an hypergeometric model:

Let $A(i)$ be the set consisting of the i^{th} number ,$1 \leq i \leq 49$ and E be the set of numbers from 1 to 49.We will also note in what follows $f_j(x)$ the function $x \to P(Y_j = x)$, $1 \leq j \leq 6$.

1 Law of Y_1

We start by the law of Y_1, the first number of the lottery grid which is the minimum number of the 6.

Then the probability that Y_1 is equal to 1 is simply computed by the hypergeometric formula (since no number picked may be inferior to 1).

The law that among a population of cardinality N, a number of specimen k are present from a subpopulation of E which is in proportion p inside E is given by the so called Hypergeometric law

$$P(X = k) = \frac{C_{pN}^{k} \times C_{(1-p)N}^{n-k}}{C_N^n} \tag{2}$$

The current probability is given by the Hypergeometric law with parameters $p = 1/49, N = 49, n = 6$ and $k = 1$.

$$P(Y_1 = 1) = C_1^1 \times C_{48}^{6-1}/C_{49}^{6} = C_{48}^{5}/C_{49}^{6} = \frac{6}{49} \approx 0.122. \tag{3}$$

Let us name u this number.

The probability that Y_1 is equal to 2 is the probability that the number 2 is picked among the winning numbers and that 1 is *not* picked among the other numbers.

We need to be carefully about the set from which the samples are obtained , this is not E but$E\backslash A(i)$, indeed the probability that 2 is picked is already computed (by u) and we need to work with independant factors so the parameters will be such that $card(E\backslash A(i)) = 48$, not 49.Besides the number of samples will be n=5.

This latest probability is easily computed by an hypergeometric law with parameters $p = 1/48, 1-p = 47/48, N = 48, n = 5$ and $k = 0$.

$$P(Y_1 = 2) = u \times \frac{C_1^0 \times C_{47}^5}{C_{48}^5} = u \times \frac{43}{48} \approx 0.109. \tag{4}$$

More generally for $i \geq 2$ and up to $i = 44$ we can compute the probability that $Y_1 = i$ by considering the population of numbers from 1 to i-1 and computing an hypergeometric law of parameters $p = (i-1)/48, N = 48, n = 5$ and $k = 0$.

Indeed the probability that $Y_1 = i$ is the product of two probabilities:

- the first is the probability that among the samples of 6 specimen collected in the population E we find the unique representant of $A(i)$ (this is u)

- the second is the probability that among this same sample we find 0 representant of $\{1, \ldots, i-1\}$.

$$P(Y_1 = i) = u \times \frac{C_{i-1}^0 \times C_{48-(i-1)}^5}{C_{48}^5} = u \times \frac{(49-i)(48-i)(47-i)(46-i)(45-i)}{48 \times 47 \times 46 \times 45 \times 44} \tag{5}$$

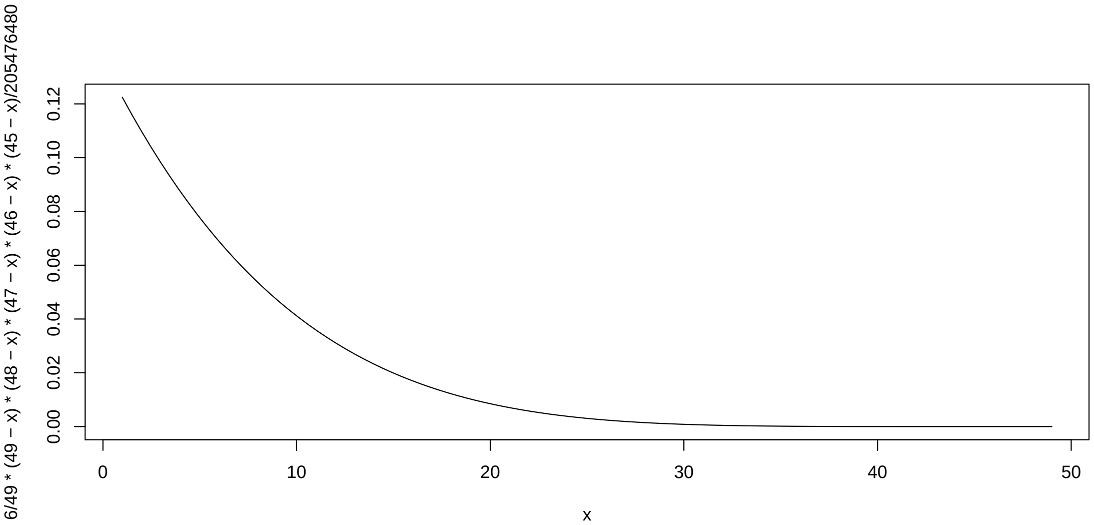
we check that $P(Y_1 = 44)$ is equal to :

$$\begin{aligned} P(Y_1 = 44) &= \frac{6}{49} \times \frac{5!}{48 \times 47 \times 46 \times 45 \times 44} && (6) \\ &= \frac{1}{C_{49}^6} && (7) \end{aligned}$$

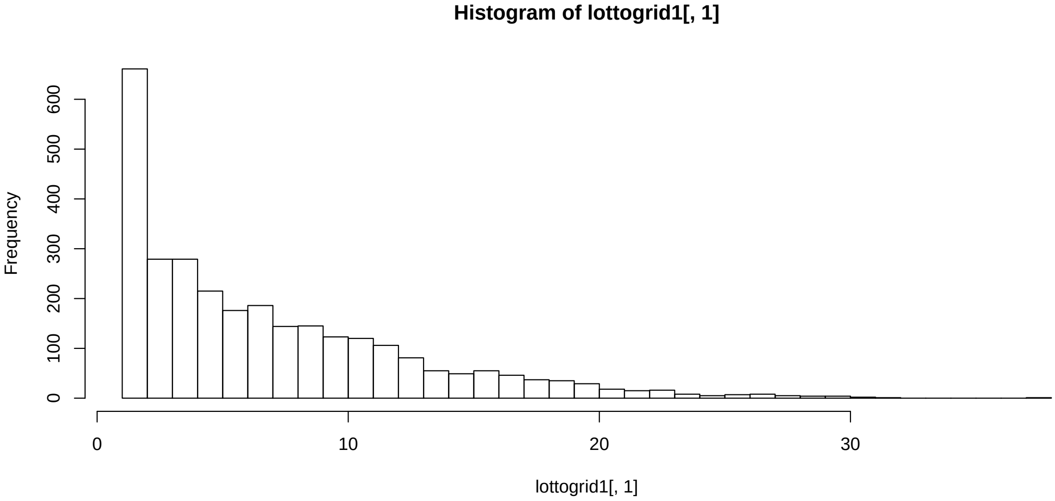
which is the probability of winning with the grid $\{44, 45, 46, 47, 48, 49\}$ since when Y_1 is equal to 44 any grid may only be $\{44, 45, 46, 47, 48, 49\}$

plotting this law we compair it with the samples at our disposal from the canadian Lottery:

Theorical distribution:



Samples (from the database):



We see that the theorical and the experimental distributions fits together.

2 Laws of Y_j , $2 \leq j \leq 6$

When considering Y_2, we need to consider both the apparition of the unique element from $A(i)$ among the prelevement (the sample) of 6 specimen among the population E , as we did before but also the fact that a unique specimen from the population $\{1, \ldots, i-1\}$ is to be found in the sample obtained from the original sample from which we exclude $A(i)$.

We have to compute an hypergeometric law with parameters $p = (i-1)/48, 1-p =$

$(49-i)/48, N=48, n=5$ and $k=1$.

We get :

$$P(Y_2=i)=u\times\frac{C_{i-1}^{1}\times C_{48-(i-1)}^{5-1}}{C_{48}^{5}}=5\times u\times(i-1)\times(49-i)(48-i)(47-i)(46-i)/48\times47\times46\times45\times44 \tag{8}$$

for the others probabilities we must simply consider that the number of specimen from the population $\{1,\ldots,i-1\}$ to be found in the 5-sample is 2 for Y_3 , 3 for Y_4 , etc...

this gives us the general formula for $j=2,\ldots 6$:

$$P(Y_j=i)=u\times\frac{C_{i-1}^{j-1}\times C_{48-(i-1)}^{5-(j-1)}}{C_{48}^{5}}=u\times C_{i-1}^{j-1}\times\frac{C_{49-i}^{6-j}}{C_{48}^{5}} \tag{9}$$

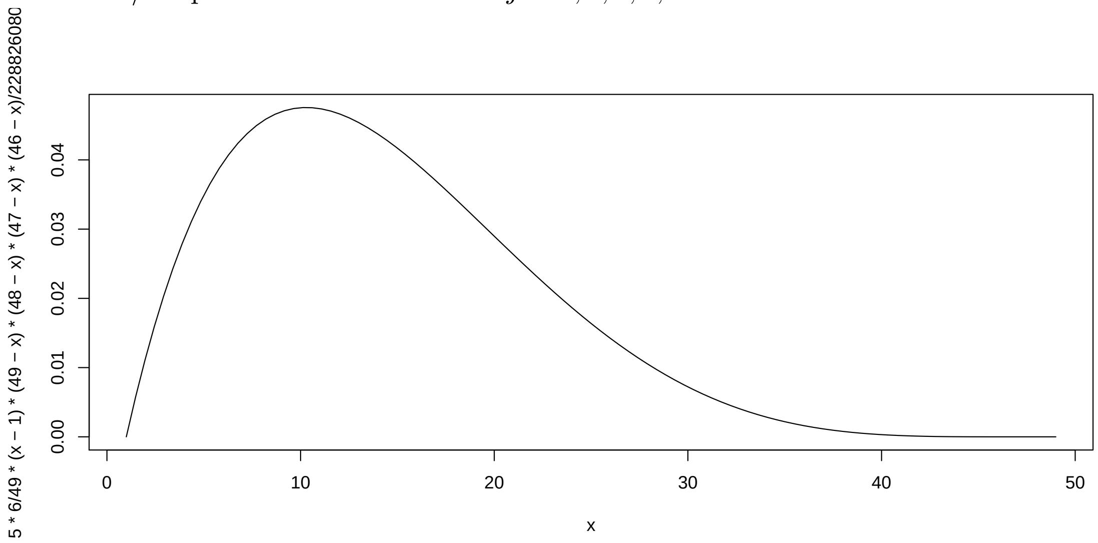
All these repartitions are represented by 5^th degree polynomnial curves who have roots among the first and last values of the set E.

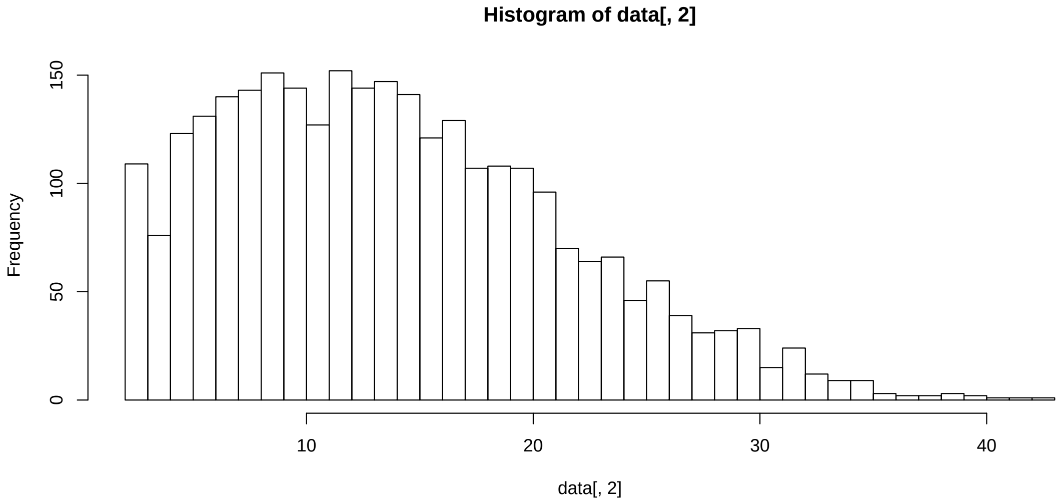
The following table details the distribution for $j\geq 2$:

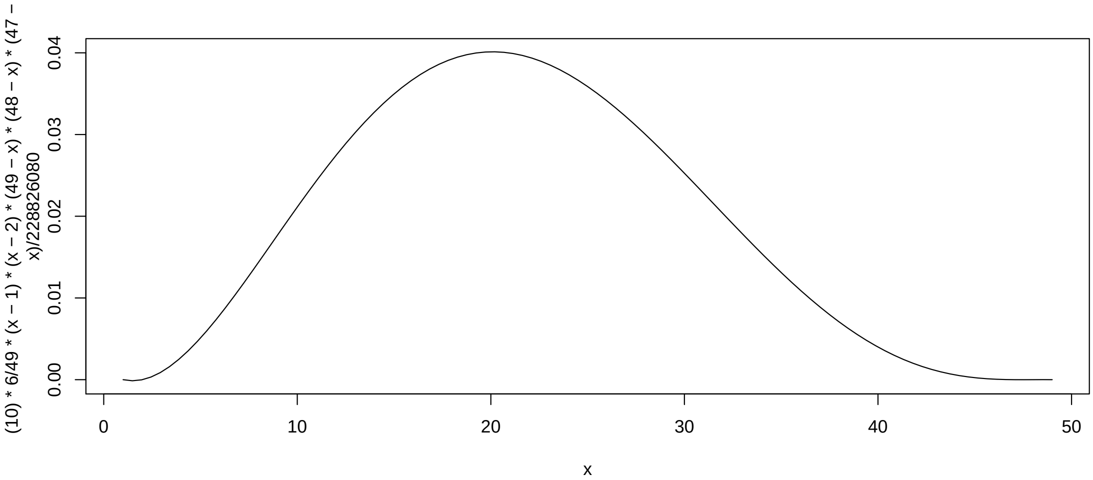
j	$P(Y_j=i)$
2	$5\times u\times(i-1)(49-i)(48-i)(47-i)(46-i)/(48\times47\times46\times45\times44)$
3	$10\times u\times(i-1)(i-2)(49-i)(48-i)(47-i)/(48\times47\times46\times45\times44)$
4	$10\times u\times(i-1)(i-2)(i-3)(49-i)(48-i)/(48\times47\times46\times45\times44)$
5	$5\times u\times(i-1)(i-2)(i-3)(i-4)(49-i)/(48\times47\times46\times45\times44)$
6	$u\times(i-1)(i-2)(i-3)(i-4)(i-5)/(48\times47\times46\times45\times44)$

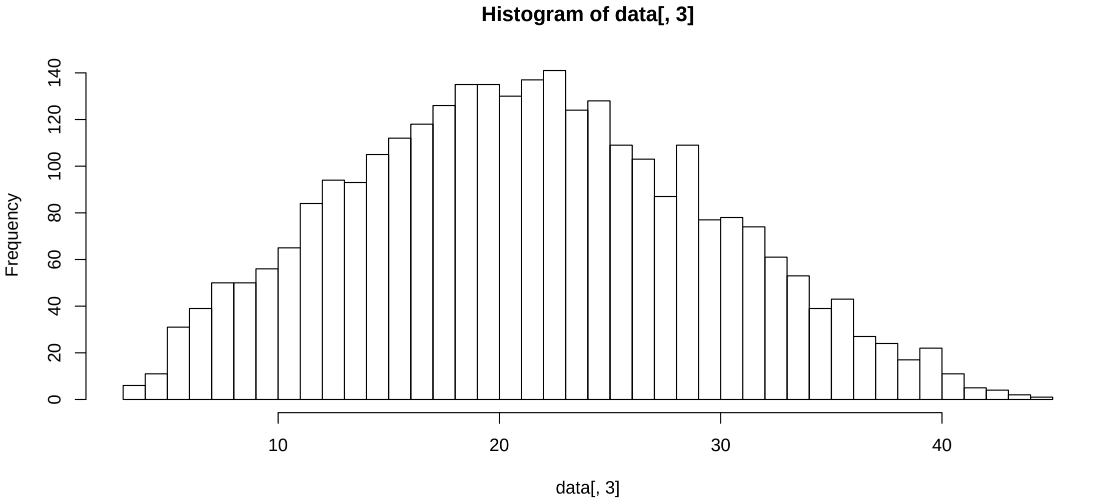
Plotting these curves we get the following results:

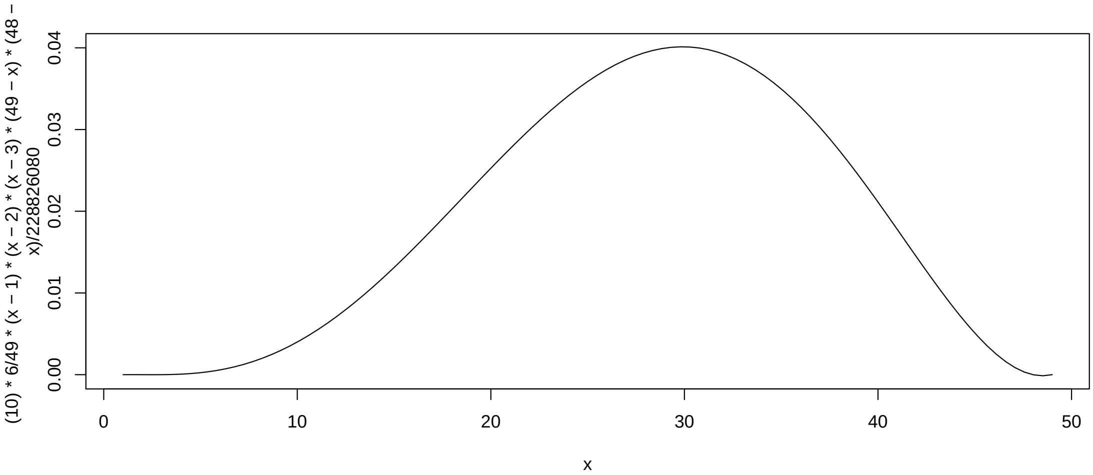
Theorical / Experimental distributions $j = 2, 3, 4, 5, 6$:

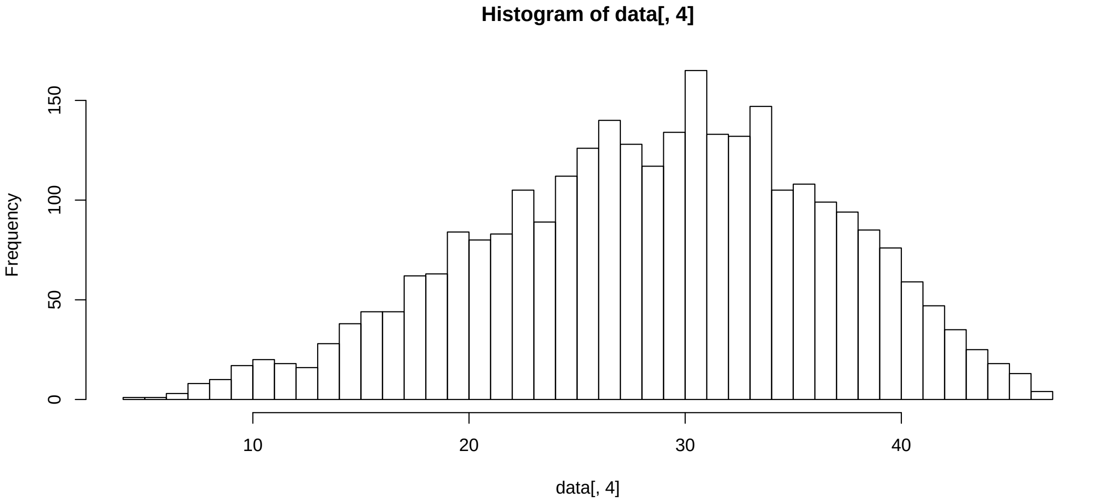


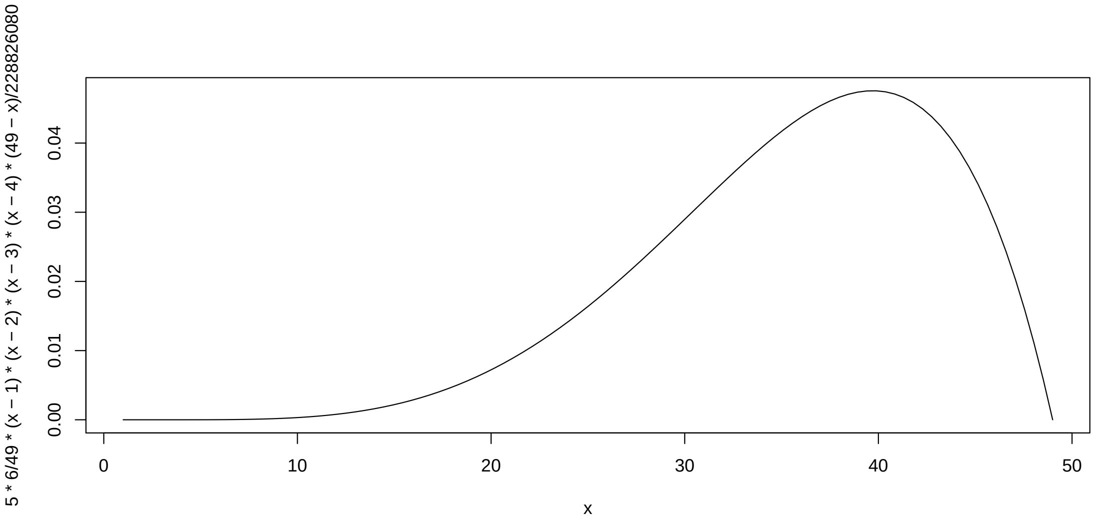



(10) * 6/49 * (x − 1) * (x − 2) * (49 − x) * (48 − x) * (47 −
x)/228826080
0.00
0.01
0.02
0.03
0.04
0
10
20
30
40
50
x

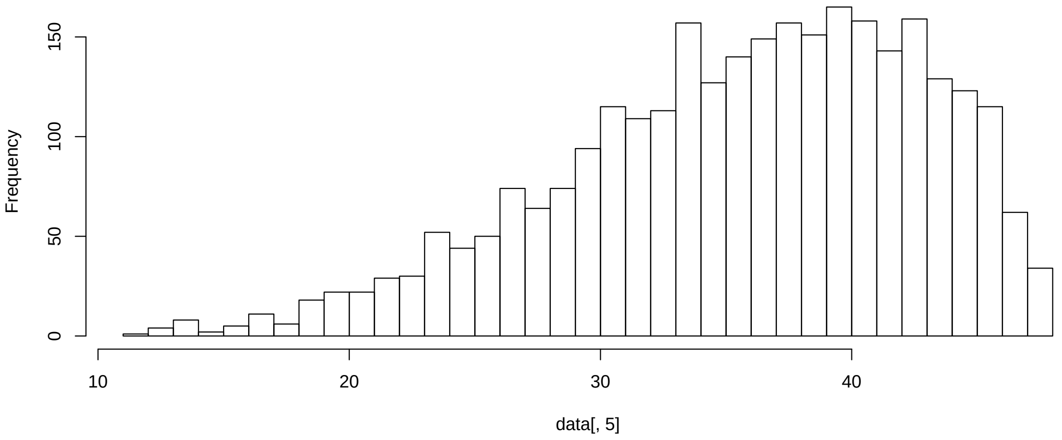

Histogram of data[, 3]
Frequency
0
20
40
60
80
100
120
140
10
20
30
40
data[, 3]

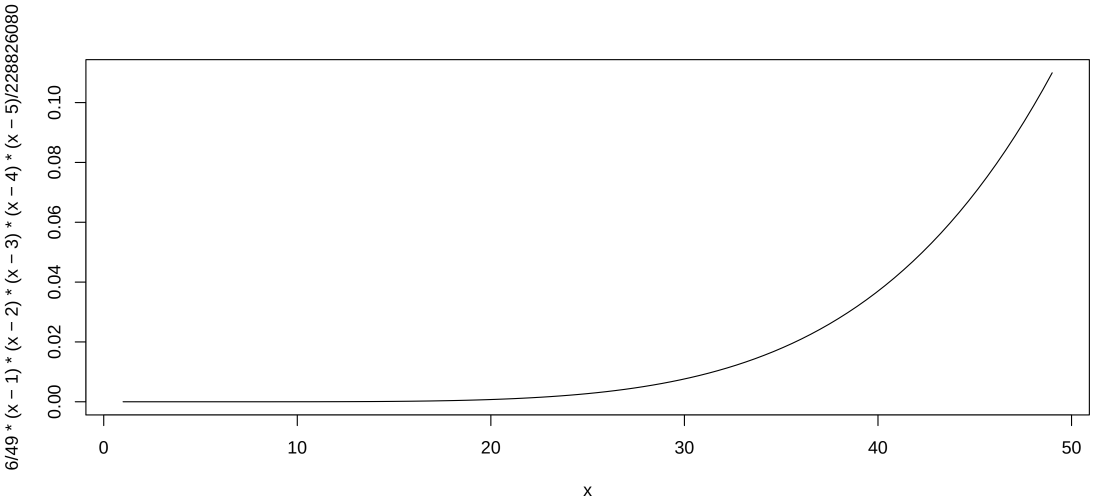

(10) * 6/49 * (x − 1) * (x − 2) * (x − 3) * (49 − x) * (48 − x)/228826080
0.00
0.01
0.02
0.03
0.04
0
10
20
30
40
50
x

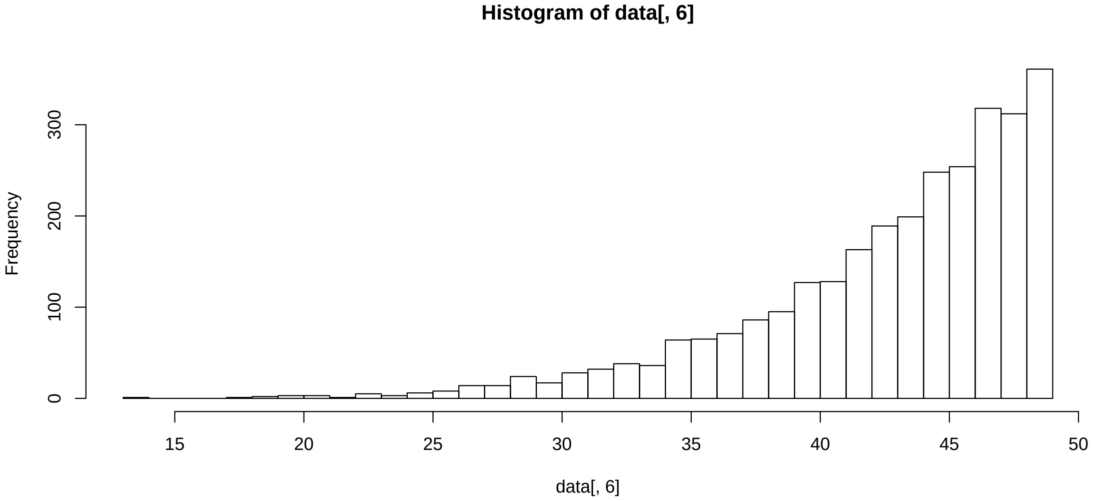

Histogram of data[, 4]
Frequency
0
50
100
150
10
20
30
40
data[, 4]



Histogram of data[, 5]







3 extremums

since we are only interested in points among the finite set we spot visually the extremums points m_j and compute the corresponding optimal probabilities P^j_{max} for $f_j, j \geq 2$:

j	m_j	P^j_{max}
1	1	≈ 0.122
2	10	≈ 0.047
3	20	≈ 0.040
4	30	≈ 0.040
5	40	≈ 0.047
6	49	≈ 0.122

4 remarks

We can see the symetric structure of the curves from $j = 1$ to 6.Also we note that Y_1/Y_6 and $Y_j, 2 \leq j \leq 5$ have different structure.

if we consider the transformation $S = \{x \to 50 - x\}$ (wich is a symmetry with respect to the y-axis x=25) , then :

$$S(f_1) = f_6, S(f_2) = f_5, S(f_3) = f_4 \tag{10}$$

All the distributions have remarkable extremum values in the range of their definition which means the existence of a value(s) which will have the greatest probability of apparition for each Y_j.

All these probabilities are polynomnial functions of degree 5 who vanishes at certain remarkable points and by looking at it , we already knew that the law f_1 of Y_1 may be defined up to a constant by the fact that $f_1(49) = f_1(48) = f_1(47) = f_1(46) = f_1(45) = 0$ since none of these values may hold for Y_1.

The same occurs for the other laws f_j of the $Y_j's$, $j \geq 2$.

Also we can see the rather 'unusual' strange shape of the curves who although they look basically shaped in the area where they have no zeros , they (have to) oscillate betweens their roots - which are outside the domain of definition for the probabilities butv this is not interesting for our concerns.

5 Analysis

At first glance we could think we have found a way to maximize the chance of finding winning numbers. We could also conclude that with our calculs the grid $\{1, 10, 20, 30, 40, 49\}$ should *always* been played.

First let us note that the probability that *any* given number appears in a winning grid is given by the hypergeometric law with parameters $p = 1/49, N = 49, n = 6, k = 1$ that is to say by the number $u \approx 0.122$, then the best probabilities we have computed for each numbers are less that this number and oscillate usually between 0.04 and 0.05.Concretely there is a chance of ≈ 0.122 that the number 10 appears in the winning grid but "only" of ≈ 0.047 that it appears at rank #2 , an this is supposed to be the best number for #2!

You have 1/49 chances that your number is picked if only one ball is picked. If 6 balls are picked you have 6 more chances that is to say 6/49 which is *exactly* the value of $u \approx 0.122$! but if you restrict that this number is to appear at a given position then you put more constraints and then dim out the chances - and so this explain this paradox.

It becomes even weirder when we look at 40: if the probability of 40 to appears at rank #6 is only ≈ 0.047 and I know I have ≈ 0.122 chances to have this number present in the winning grid, I should play it an an other rank, right? Not really in fact...

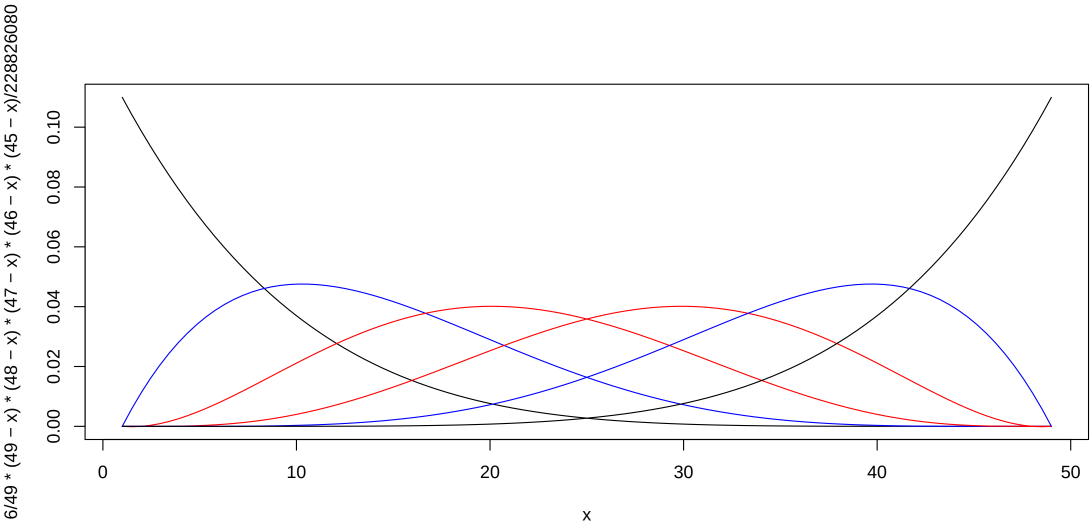
If we look at the sum $\sum_{j=1,\ldots,6} P(Y_j = 40)$ we will find that this number is equal precisely to u so in fact playing 40 at a different rank will dramatically reduce the chances of having 40 appearing! This is again a paradox and the explaination is that the probability ≈ 0.122 of having the number 40 (or *any* other numbers) appearing in the winning grid is *streamed* into the ranks .

If a player wants to have ≈ 0.122 of having a number appearing in the winning grid then he has to chose in a random way a rank for this number each times he/she plays.

This gives us incidentally and other way of using our statistics: rather by asking which number have the highest probability to be present at a given rank , we could ask ourselves where is it best to play a given number?

It seems intuitively rather obvious that playing the number 40 or any 'big' lotto number at the first ranks is not very clever since that restricts the choice of the other 5 numbers.

If we plot the 6 repartition for the ranks $\#1, 2, 3, 4, 5, 6$ alltogethers , we get the following diagram:



We can clearly use this diagram to see which rank we better use of a given number to be played: for example we clearly read that if we like to play the number '20' then it is best to play it at 3^{rd} position.

In fact we see that the set E is divided in 6 domains (intervals) where a given ranks 'dominates' over the others:

domain	best rank
$1 \leq i \leq 8$	#1
$9 \leq i \leq 16$	#2
$17 \leq i \leq 25$	#3
$25 \leq i \leq 33$	#4
$34 \leq i \leq 41$	#5
$42 \leq i \leq 49$	#6

We remark that the point of junction between each domain do not necesseraly coincide with the extremum points computed before.

If we compute the function obtained by summing the value of the six curves , e,g

$f(x) = \sum_{j=1}^{6} f_j(x)$, we will get the constant function $f(x) = u$ which is not surprising since the sum of probability for a number to be in each rank is the probability that this number is among the winning grid , which is equal to u.

6 Conclusion

We have used basic combinatorials to get few laws about how the winning numbers streamed themselves in the different ranks.The repartition among these ranks is far from being uniform which is not surprising since 'small' numbers should be present predominantly at the first ranks and 'big' numbers should be present predominantly at the last ranks but neverthless we have proved and quantified these probabilities. This allows us to give precise and exact hints if a player has already chosen a number and wants help to get the best rank for it or reversely if a player wants the best number (or the best number among a given list) to be played at a given rank.

This method does not directly increase the chance of getting a winning number, the maximum of our probabilities is always u which is the probability to get a winning number by playing randomly a number in a random rank... but if a player wants to play always the same number or a group of number or choosing a number for a given rank then this method gives him the best chances to win, although again from a theorical point of view his chances of winning are less than if this player was playing perfectly randomly.