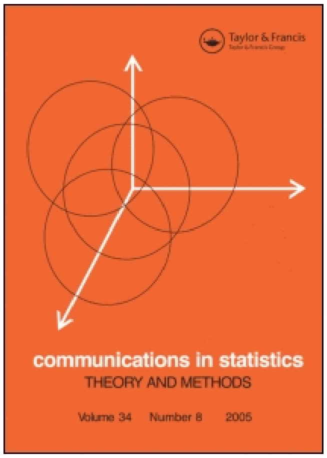

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

Waiting time to exhaust lottery numbers

José Galvão Leite [a]; Carlos Alberto de Bragança Pereira [a]; Flávio Wagner Rodrigues [a]
[a] Instituto de Matematica e Estatfstica, Universidade de Sào Paulo, CEP, SP, Brazil

Online Publication Date: 01 January 1992
To cite this Article: Leite, José Galvão, Pereira, Carlos Alberto de Bragança and Rodrigues, Flávio Wagner (1992) 'Waiting time to exhaust lottery numbers', Communications in Statistics - Theory and Methods, 22:1, 301 - 310
To link to this article: DOI: 10.1080/03610929308831019
URL: http://dx.doi.org/10.1080/03610929308831019

COMMUN. STATIST.—THEORY METH., 22(1), 301-310 (1993)

WAITING TIME TO EXHAUST LOTTERY NUMBERS

José Galvão Leite,
Carlos Alberto de Bragança Pereira,
and Flávio Wagner Rodrigues

Universidade de São Paulo - Instituto de Matemática e Estatística
C.Postal 20570. CEP 01498, SP. Brazil

Key words: Lottery, *lotto*-weeks, random selections, waiting time.

AMS classification: 60C05.

ABSTRACT

Questions related to lotteries are usually of interest to the public since people think there is a magic formula which will help them to win lottery draws. This note shows how to compute the expected waiting time to observe specific numbers in a sequence of lottery draws and show that surprising facts are expected to occur.

1. INTRODUCTION

Consider the following lottery system. called lotto: Every week, a lottery house selects randomly n distinct numbers from the first N natural numbers.

A gambler bets a fixed amount of money on a set of n numbers (among the N lottery numbers) (s)he has chosen. If these are exactly the same numbers selected by the lottery house (s)he wins a premium. This premium depends on the number of bets and on the number of winners. The lottery house profit is a fixed percentage of the money obtained from all bets. If in a certain week there is no winner, the premium accumulates for the next *lotto*-week. A gambler may be willing to bet only in weeks with accumulated premiums. (To eliminate from the study weeks whithout the game, we call *lotto*-weeks those in which the game was played.)

After many weeks, a reporter realized that r numbers, among the N, had not been selected in previous *lotto*-weeks. He then asked the authors of the present report to calculate the expected number of lotto-weeks before all r remaining numbers are selected.

The above question is related to the famous collector's problem (see Feller, 1968 and Johnson & Kotz, 1977). If a collector buys his/her collection units in boxes of n such units, one may ask "how many boxes are expected to be bought by the collector in order to have his/her collection completed?"

Another interesting situation is the capture/recapture tagging process for animal populations. Leite and Pereira (1987) discuss this model for the case where the sample size, n here, changes in each selection step (*lotto*-week here). To estimate the size, N, of the population (in the present report N is known), one needs to record, in each step, the recaptures; i.e., the number of animals that have been already captured and tagged in earlier steps. In this way, one learns how many distinct animals have been selected in the process of tagging.

2. DISTRIBUTION OF THE NUMBER OF UNSELECTED UNITS

Let R_k denote the number of units that have not been selected in k consecutive *lotto*-weeks. After recording the r numbers that have not been selected in the k consecutive *lotto*-weeks denote by T_r the number of consecutive *lotto*-weeks before all r numbers are selected. The probability distributions of these quantities are given in the following lemma:

Lemma 1.

For all $k \geq 1$, $N \geq n$, $N - \min\{kn, N\} \leq r \leq N - n$, and $p_j = \frac{\binom{N-j}{n}}{\binom{N}{n}}$,

we have

(i) $\Pr\{R_k = r\} = p_r(k,n,N) = \binom{N}{r} \sum_{i=0}^{N-r} (-1)^i \binom{N-r}{i} p_{r+i}^k$,

(ii) $\Pr\{T_r > k\} = \sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i} p_i^k$, and

(ii') $\Pr\{T_r = k\} = \sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i} p_i^{k-1}(1 - p_i)$.

Proof.

Let A_{ij} be the event "number i was not selected in the j-th *lotto*-week." Hence, $B_{ik} = A_{i1} \cap A_{i2} \cap \ldots \cap A_{ik}$ is the event "number i was not selected in k consecutive *lotto*-weeks". Since the numbers are exchangeable, without loss of generality, we consider here the first r natural numbers.

Letting $p_j = \frac{\binom{N-j}{n}}{\binom{N}{n}}$ we have that $\Pr(B_{ik}) = p_1^k$, for any $1 \leq i \leq r$, $\Pr(B_{ik} \cap B_{jk}) = p_2^k$, for $1 \leq i < j \leq r$, and so on up to $\Pr(B_{1k} \cap B_{2k} \cap \ldots \cap B_{rk}) = p_r^k$.

Now we note the following interesting facts: a) The event $\{R_k = 0\}$ is the event $\left\{\bigcup_{i=1}^{N} B_{ik}\right\}^c$, where c indicates complement, and b) the event $\{T_r > k\}$ is the event $\bigcup_{i=1}^{r} B_{ik}$. The probabilities of these events are given by

$$\begin{aligned}\Pr\{T_r > k\} &= \Pr\left\{\bigcup_{i=1}^{r} B_{ik}\right\} = \binom{r}{1} p_1^k - \binom{r}{2} p_2^k + \ldots + (-1)^{r-1} \binom{r}{r} p_r^k \\ &= \sum_{i=1}^{r} (-1)^{i-1} \binom{r}{i} p_i^k ,\end{aligned}$$

and

$$\begin{aligned}\Pr\{R_k = 0\} &= p_0(k,n,N) = 1 - \Pr\left\{\bigcup_{i=1}^{N} B_{ik}\right\} = 1 - \sum_{i=1}^{N} (-1)^{i-1} \binom{N}{i} p_i^k = \\ &= \sum_{i=0}^{N} (-1)^i \binom{N}{i} p_i^k , \qquad \text{for} N \leq kn .\end{aligned}$$

To conclude (ii') we only recall that $\Pr\{T_r = k\} = \Pr\{T_r > k-1\} - \Pr\{T_r > k\}$. Finally, notice that $p_0(k, n, N-r)$ is the probability that "exactly r specified numbers were selected in none of the k consecutive *lotto*-weeks." Since there are $\binom{N}{r}$ ways of selecting r numbers from the N lottery ones, we finally have that $p_r(k, n, N) = p_0(k, n, N-r)\binom{N}{r}$, which is positive only for $N - \min\{kn, N\} \leq r \leq N - r$. Replacing the appropriate expressions in this last equality we obtain (i), concluding the proof. □

The above result shows that the probability function of T_r is a linear combination of r geometric probability functions. Hence a moment of T_r is also a linear combination of the corresponding moments of r geometric distributions. Recall that the first and second moments of a geometric distribution with parameter q are, respectively, q^{-1} and $(2-q)q^{-2}$. The following result is then straightforward.

Lemma 2.

For $p_j = \frac{\binom{N-j}{n}}{\binom{N}{n}}$, we have

(iii) $E\{T_r\} = \sum_{i=1}^{r}(-1)^{i-1}\binom{r}{i}(1-p_i)^{-1}$ and (iv) $E\{T_r^2\} = \sum_{i=1}^{r}(-1)^{i-1}\binom{r}{i}\frac{(1+p_i)}{(1-p_i)^2}$

3. ALTERNATIVE EXPRESSIONS FOR THE DISTRIBUTION, THE MEAN, AND THE VARIANCE OF T_N

In this section we use the distruibution of R_k to obtain alternative expressions for the distribution, the mean, and the variance of $T = T_N$. Note that the event $\{T = t\}$ is equivalent to the event $\{R_{t-1} > 0\} \cap \{R_t = 0\}$. The probability function may also be expressed as follows:

$$\Pr\{T = t\} = \begin{cases} 0 & \text{if } nt < N \\ \Pr\{R_t = 0\} & \text{if } N \leq nt < N + n \\ \sum_{i=1}^{n} \Pr\{R_{t-1} = i\}\Pr\{R_t = 0 | R_{t-1} = i\} & \text{if } nt \geq N + n \end{cases}$$

Note now that $\Pr\{R_t = 0\} = p_0(t, n, N), \Pr\{R_t = 0 | R_{t-1} = i\} = \frac{\binom{N-i}{n-i}}{\binom{N}{n}}$, and $\Pr\{R_{t-1} = r\}$ is given by Lemma 1. After some simplifications, we obtain, for $nt \geq N + n$,

$$\Pr\{T = t\} = \sum_{s=1}^{n} \binom{n}{s} \sum_{i=0}^{N-s} (-1)^i \binom{N-s}{i} p_{s+i}^{t-1}$$

and

$$\Pr\{T > t\} = \sum_{s=1}^{n} \binom{n}{s} \sum_{i=0}^{N-s} (-1)^i \binom{N-s}{i} \frac{p_{s+i}^{t}}{1 - p_{s+i}} .$$

Before using this expression to compute the mean and the variance of T, we present the following result. Let $b = \left[\frac{N}{n}\right]$, the largest integer that is smaller than or equal to $\frac{N}{n}$.

Lemma 3.

(v) Mean: $\mu = E\{T\} = b + \sum_{t=b}^{\infty} \Pr\{T > t\}$ and

(vi) Variance: $\sigma^2 = (\mu - b)(b + 1 - \mu) + 2 \sum_{s=b+1}^{\infty} \left(\sum_{t=s}^{\infty} \Pr\{T > t\} \right)$.

Proof.

(v) $\mu = \sum_{t=b}^{\infty} t\Pr\{T = t\} = \sum_{i=0}^{\infty} (b + i)\Pr\{T = b + i\} = b\sum_{i=0}^{\infty} \Pr\{T = b + i\} + \sum_{i=1}^{\infty} i\Pr\{T = b + i\}$. Noticing that the first term is b and rearranging the second term, we obtain the result.

(vi) $E\{T^2\} = \sum_{t=b}^{\infty} t^2\Pr\{T = t\} = b^2 \sum_{t=b}^{\infty} \Pr\{T = t\} + [(b+1)^2 - b^2] \sum_{t=b+1}^{\infty} \Pr\{T = t\} + \ldots + [k^2 - (k-1)^2] \sum_{t=k}^{\infty} \Pr\{T = t\} + \ldots = b^2 + \sum_{t=b}^{\infty} [(t+1)^2 - t^2]\Pr\{T > t\} = b^2 + \sum_{t=b}^{\infty} (2t + 1)\Pr\{T > t\} = b^2 + \sum_{t=b}^{\infty} [2b + 1 + 2(t - b)]\Pr\{T > t\} = b^2 + (2b + 1) \sum_{t=b}^{\infty} \Pr\{T > t\} + 2\sum_{i=0}^{\infty} i\Pr\{T > b + i\}$. Finally, rearranging the last sum of this expression, we obtain

$$E\{T^2\} = b^2 + (2b + 1) \sum_{t=b}^{\infty} \Pr\{T > t\} + 2 \sum_{s=1}^{\infty} \sum_{i=s}^{\infty} \Pr\{T > b + i\} .$$

The conclusion now is straightforward. □

TABLE 1

Mean and standard deviation of T_i. p_i is the probability of drawing a number not equal to i specified numbers

i	p_i	mean	std.dev.	i	p_i	mean	std.dev.
1	0.8800	8.3333	7.8174	26	0.0085	30.8912	9.7706
2	0.7722	12.2760	8.7319	27	0.0064	31.1841	9.7725
3	0.6757	14.9117	9.0966	28	0.0047	31.4664	9.7741
4	0.5895	16.8883	9.2911	29	0.0034	31.7391	9.7756
5	0.5126	18.4696	9.4107	30	0.0024	32.0026	9.7769
6	0.4442	19.7873	9.4909	31	0.0017	32.2577	9.7780
7	0.3836	20.9168	9.5480	32	0.0012	32.5047	9.7791
8	0.3301	21.9051	9.5904	33	0.0008	32.7443	9.7800
9	0.2830	22.7836	9.6230	34	0.0005	32.9769	9.7807
10	0.2415	23.5742	9.6486	35	0.0003	33.2028	9.7814
11	0.20533	24.2930	9.6691	36	0.0002	33.4224	9.7820
12	0.1737	24.9519	9.6858	37	0.0001	33.6361	9.7825
13	0.1463	25.5600	9.6997	38	0.0001	33.8441	9.7829
14	0.1226	26.1248	9.7113	39	0.0000	34.0469	9.7834
15	0.1021	26.6519	9.7210	40	0.0000	34.2445	9.7836
16	0.0846	27.1460	9.7293	41	0.0000	34.4374	9.7839
17	0.0697	27.6111	9.7365	42	0.0000	34.6257	9.7838
18	0.0570	28.0504	9.7426	43	0.0000	34.8094	9.7845
19	0.0463	28.4665	9.7479	44	0.0000	34.9889	9.7854
20	0.0374	28.8618	9.7526	45	0.0000	35.1649	9.7843
21	0.0299	29.2383	9.7566	46	0.0000	35.3385	9.7779
22	0.0237	29.5977	9.7602	47	0.0000	35.5057	9.7812
23	0.0186	29.9414	9.7633	48	0.0000	35.6798	9.7474
24	0.0145	30.2709	9.7660	49	0.0000	35.8234	9.8109
25	0.0111	30.5871	9.7684	50	0.0000	36.0194	9.6711

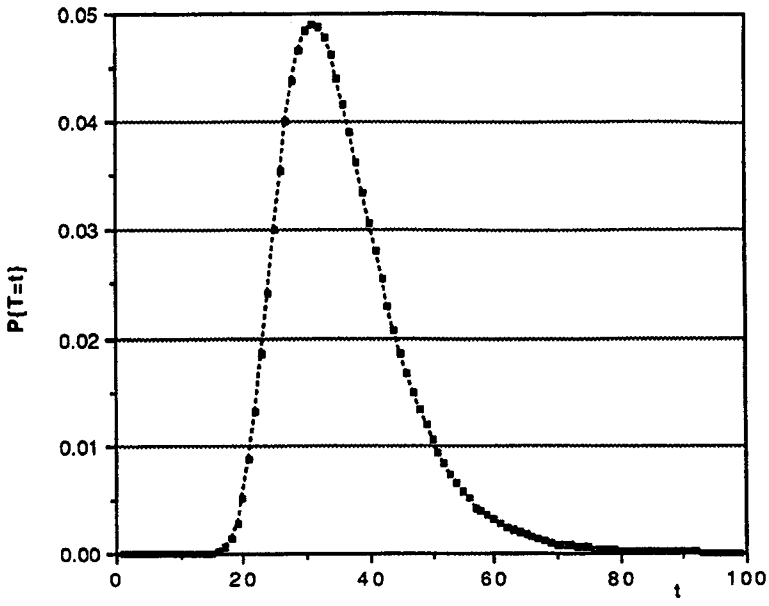

FIGURE 1
Probability function of $T = T50$

To obtain the mean and the variance of T we can apply to the above result both the formula of $\Pr\{T > t\}$ presented in this section and the one obtained from Section 2. We obtain the following surprising equalities:

$$\mu = b + \sum_{s=1}^{n} \binom{n}{s} \sum_{i=0}^{N-s} (-1)^i \binom{N-s}{i} \frac{p_{s+i}^b}{(1-p_{s+i})^2} = b + \sum_{i=1}^{N} (-1)^{i-1} \binom{N}{i} \frac{p_i^b}{1-p_i}$$

and

$$\begin{aligned} \sigma^2 &= (\mu - b)(b+1-\mu) + 2\sum_{s=1}^{n} \binom{n}{s} \sum_{i=0}^{N-s} (-1)^i \binom{N-s}{i} \frac{p_{s+i}^{b+1}}{(1-p_{s+i})^3} = \\ &= (\mu - b)(b+1-\mu) + \sum_{i=1}^{N} (-1)^{i-1} \binom{N}{i} \frac{p_i^{b+1}}{(1-p_i)^2} \,. \end{aligned}$$

In the next section we apply these formulas in the particular case of the Brazilian lottery.

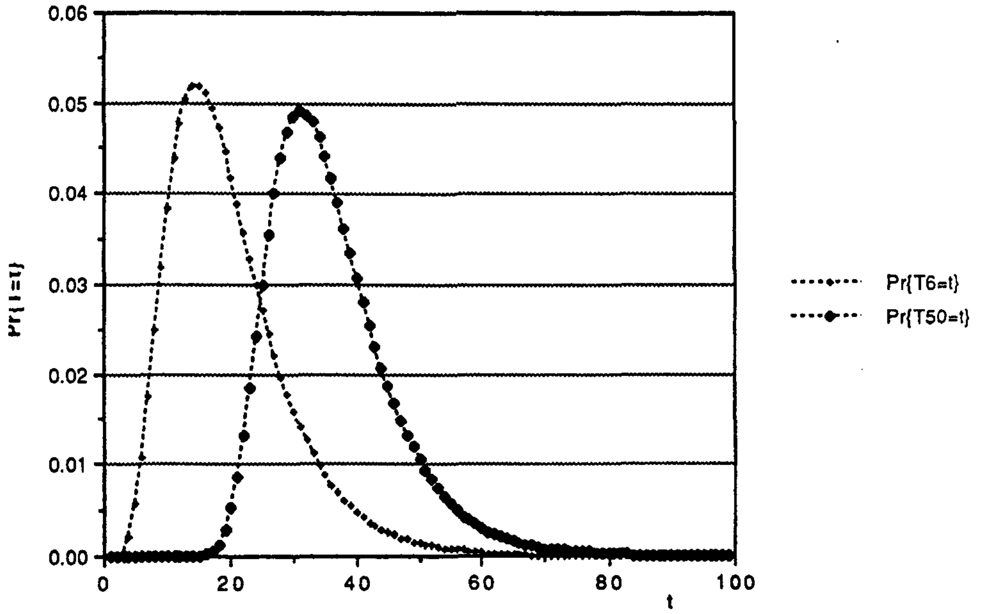

FIGURE 2
Probability functions of $T6$ and $T50$

4. CASE OF N = 50 AND n = 6

In this section we discuss the case of the Brazilian lottery called "Sena". In this case the lottery house belongs to the government and there are different kinds of winnings. However the drawing procedure is exactly as described in the present report. Here $N = 50$ and $n = 6$. The questions of interest are also related with the waiting time to special numbers to be drawn for the first time.

Table I presents the expectation and the standard deviation of the waiting time to observe $j (= 1, 2, \ldots, 50)$ specified numbers, in the "Sena". For example, if $j = 1$ then, $\mu = 8.33$ and $\sigma = 7.82$. On the other hand, if $j = 50$ then, $\mu = 36.02$ and $\sigma = 9.67$. Figure 1 shows the probability distribution of $T = T_{50}$, the waiting time to all 50 numbers being selected. Figure 2 presents

TABLE II

Number of the draw that number i has occurred

i	1	2	3	4	5	6	7	8	9	10
Draw #	6	10	3	4	4	13	7	3	5	6
i	11	12	13	14	15	16	17	18	19	20
Drau #	17	16	13	20	3	29	8	2	4	4
i	21	22	23	24	25	26	27	28	29	30
Draw #	12	3	11	13	6	16	7	1	2	2
i	31	32	33	34	35	36	37	38	39	40
Draw #	14	3	8	10	12	1	11	1	12	10
i	41	42	43	44	45	46	47	48	49	50
Drau #	1	7	1	1	5	6	4	10	2	2

both the probability distribution of T_{50} and that of T_6. Note that the mode of this probability function is in $T = 31$ where $\Pr\{T = 31\} = 0.0491$. Also, for $j = 6$, we observe that the mode of T_6 is in $T_6 = 15$ where $\Pr\{T_6 = 15\} = 0.0519$. In the simple case of $j = 1$, where we have a geometric distribution, the mode is clearly in $T_1 = 1$, where $\Pr\{T_1 = 1\} = 1 - p_1 = 0.12$. For space reasons, we did not present here the distributions of all possible j's. We believe however that the distributions of $T = T_{50}$ and of T_6 give a good idea of all those distibutions.

Observing the results of the Brazilian lottery, "Sena", we may conclude that what was observed is not surprising. For each of the possible numbers, i, Table II presents the number of the *Sena's draw* (number of the *Sena*-week) in which i occurs for the first time. It is interesting to note that number 16 was drawn for the first time in the 29th draw and, after the other 49 numbers had occurred, we had to wait 9 weeks to obtain number 16. We observe also that the number of weeks we waited to have all 50 numbers drawn was 29, which is smaller than the expectation of T, 36.01. That explains why some number has to wait long to be drawn. The reporter interested in the lottery results asked for an interview with the authors immediately after the 13th week. At that time, only 6 numbers had not been drawn before. These 6 numbers were drawn within the 14th to the 29th weeks. That is, starting the process in the

14th week, the value observed for T_6 was $T_6 = 15$ which is the mode of the distribution of T_6.

REFERENCES

Feller, W. (1968). An introduction to probability theory and its applications. 3rd ed. *New York. John Wiley.* 509pp.

Johnson, N.L. & Kotz. S. (1977). Urn models and their applications: an approach to modern discrete probability theory. *New York. John Wiley.* 402pp.

Leite, J.G. & Pereira, C.A. de B.(1987). An urn model for the multisample capture/recapture sequential tagging process. *Sequential Analysis.* 6(2): 179-86.

Received June 1992; Revised July 1992