Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2010, Article ID 279154, 16 pages
doi:10.1155/2010/279154

Research Article

A Winner's Mean Earnings in Lottery and Inverse Moments of the Binomial Distribution

Konstantinos Drakakis[1,2]

[1] *UCD CASL, University College Dublin, Belfield, Dublin 4, Ireland*
[2] *School of Electronic, Electrical & Mechanical Engineering, University College Dublin, Dublin 4, Ireland*

Correspondence should be addressed to Konstantinos Drakakis, drakakis@gmail.com

Received 5 November 2009; Revised 30 January 2010; Accepted 16 February 2010

Academic Editor: Daniel Zelterman

We study the mean earnings of a lottery winner as a function of the number n of participants in the lottery and of the success probability p. We show, in particular, that, for fixed p, there exists an optimal value of n where the mean earnings are maximized. We also establish a relation with the inverse moments of a binomial distribution and suggest new formulas (exact and approximate) for them.

1. Introduction

The game of lottery is both popular and simple. Focusing on the essentials, and leaving aside additional features of secondary importance, which vary across different lottery implementations, the rules of the game are as follows: each player submits to the lottery organizers a ticket consisting of M integers (selected by the player, without repetitions, selection order being unimportant) from the range $1, \ldots, N$; within the prespecified time period the game is set to last; upon the expiry of this period no more ticket submissions are accepted, and an M-tuple of distinct integers (the "winning set") is selected uniformly at random by the lottery organizers; each submitted ticket gets compared against the winning set, and, if they match, the corresponding player "wins." This is known as an M/N lottery system, and the winning probability is clearly $p = 1/\binom{N}{M}$. The money the winners earn depends on the number n of submitted tickets: each submitted M-tuple incurs a certain fee (which we assume, without loss of generality, to be equal to 1), and some fixed ratio of the total sum collected (which we assume, again without loss of generality, to be 100%, namely, the entire sum) is returned as prize money back to the winners and equally split among them. As long as no winner is found, earnings of earlier games accumulate until winners are found, who then split equally the entire sum. This is an important feature in the implementation of lottery systems in practice, known as *rollover*.

For a given success probability p, what is the effect of the total number of participants n on a winner's mean earnings? This is the object of study of this article. Clearly, more participants lead not only to more prize money, but also to more potential winners. Intuitively, and by the elementary properties of the binomial distribution, we expect that the area of the (n,p)-plane where $np \approx 1$ is an important borderline: as long as $np \ll 1$, existence of winners is highly improbable, so most likely the mean earnings will trivially be 0, while, as long as $np \gg 1$, the law of large numbers applies and suggests that there will be approximately np winners; each of when will receive an amount of money equal to $n/(np) = 1/p$. We first analyze the independent draws scenario (without rollover), and then use this result to deduce the corresponding result with rollover.

An additional complication rollover presents that it is not only that the number of participants in the various draws can vary, but also that this variation can potentially exhibit strong correlation, specifically be strictly increasing: as a rule, larger money prizes motivate more extensive participation [1, 2]. A reliable model for this variation can only be obtained through detailed statistical analysis and the psychology of gambling, which both lie beyond the scope of the present article (see, e.g., [1–4] for an analysis of the playing style of lottery participants). Perhaps surprisingly, though, some analysis we performed on Greek lottery data [5] did not invariably lead to the conclusion that the number of participants in the various games during a rollover round shows a clear upward trend (a similar phenomenon is observed in [3]). On many occasions, these fluctuations appear indeed to be random and, more importantly, the order of magnitude of the number of participants does not vary. Accordingly, we will derive the general formula for an arbitrary fluctuation of the number of participants during lottery games in a single rollover round, but we will then focus on the special case where this number is constant, as it is not only interesting but also mathematically tractable.

Needless to say, actual lottery implementations are actually more complicated, incorporating a plethora of secondary features. For example, lower winning levels may be introduced, corresponding to partial matches between the tickets and the winning set, and the prize money they win may either be smaller fractions of the total sum or simply a fixed sum. Alternatively, an extra "bonus number" may be introduced, acting, in a sense, like a "second chance": players are allowed to submit a guess for this number, in addition to their guess for the winning set, while the selection of the winning set is followed by the (also uniformly random amongst the remaining integers) bonus number. If a player's choice has an overlap of $M-1$ integers with the winning set, but the player also guessed correctly the bonus number, a lesser prize is offered to the player. This may, of course, be combined with all lower winning levels, whereby an overlap of $K-1$ integers ($K \leq M$) plus correct guess of the bonus number leads to a smaller prize compared to an overlap of K integers.

To the best of our knowledge, the issue of winners' mean earnings has not been considered in the lottery literature before. It is a problem of mathematical interest, relating lottery to the study of inverse moments of the positive binomial distribution (as will be seen below), and, for this reason, this article can be considered falling into the same group of publications by the present author [6–8], as well as other authors (e.g., [4, 9]), whose objective is to consider nontrivial probability issues related to the lottery. We should perhaps state that the purpose of this article is *not* to consider winning strategies or to advise a prospective player about the investive value of the lottery. For readers interested in these aspects of the lottery, we recommend the study in [1], a recent and mathematically sophisticated analysis of the lottery's viability as an investment, and also those in [10, 11].

One final point that needs to be addressed is our choice to assume throughout this article that players choose their numbers independently of each other, namely, that the probability distribution describing the choice of an M-tuple of integers by any player is uniform over all possible M-tuples, and to actually do so despite the fact that this has been extensively investigated and found, perhaps surprisingly, not to hold in practice [1, 2, 10–13]! Indeed, humans seem to favor some classes of M-tuples more than others (for several reasons), and the net effect of this is that the probability distribution of the number of winners depends on the winning set. Lottery organizers seem to encourage this phenomenon, as it leads to more frequent rollovers, which has been documented, at least, in some lotteries (e.g., US and UK), to increase participation [1, 2]. This already suggests a playing strategy: assuming that players have information on which M-tuples are "unpopular" and that a player should choose one of those. Indeed, the probability distribution of the winning set *is* uniform (assuming that the lottery organizers are honest), and, if the chosen "unpopular" M-tuple wins, the likelihood of other winners existing will be minimal. In this article, however, we choose to ignore this complication, as mentioned above, on account of three reasons: (a) there appears to be no universal and conclusive statistical evidence to support the position that the skewness present in the players' choices is strong enough to unquestionably reject the validity of the uniform players' choices model, at least as an approximation; (b) as lottery players become more mature and aware, this skewness may progressively disappear (there is already evidence that, in some lotteries, more than 70% of the choices entered are picked randomly by computers [1]); and (c) as mentioned earlier, this article focuses on the principles of the lottery rather than its actual implementations.

2. Results without Rollover

Consider the lottery game as described in the Introduction: the probability distribution for the number of winners is the binomial $B(n,p)$. Assuming that $m > 0$ winners exist, each one's earnings will be n/m, while, for $m = 0$, there is no winner to earn something, so earnings are conventionally taken to be 0. It follows that the mean earnings $G(n,p)$ are

$$G(n,p) = \sum_{m=1}^{n} \binom{n}{m} p^m (1-p)^{n-m} \frac{n}{m} = n(1-p)^n \sum_{m=1}^{n} \binom{n}{m} \frac{1}{m} \left(\frac{p}{1-p}\right)^m. \tag{2.1}$$

This expression is essentially the inverse first moment of the positive binomial distribution $B_+(n,p)$. Assuming that $X \sim B(n,p)$ and setting $q = 1-p$,

$$\begin{aligned} X_+ \sim B_+(n,p) \Longleftrightarrow \mathbb{P}[X_+ = m] &= \mathbb{P}[X = m \mid X > 0] \\ &= \frac{\mathbb{P}[X=m]}{\mathbb{P}[X>0]} = \frac{1}{1-q^n}\binom{n}{m} p^m q^{n-m}, \quad m = 1, \ldots, n, \end{aligned} \tag{2.2}$$

and the inverse kth moment is defined as

$$\mu_{-k} = \mathbb{E}\left[X_+^{-k}\right] = \frac{1}{1-q^n} \sum_{m=1}^{n} \frac{1}{m^k} \binom{n}{m} p^m q^{n-m} = \frac{q^n}{1-q^n} \sum_{m=1}^{n} \frac{1}{m^k} \binom{n}{m} \left(\frac{p}{q}\right)^m, \quad k \in \mathbb{N}^*. \tag{2.3}$$

Clearly, (2.1) and (2.3) (for $k = 1$) consist of the same sum (with different multiplicative factors):

$$G(n,p) = n(1-q^n)\mu_{-1}. \tag{2.4}$$

The problem of the inverse moments of the positive binomial distribution was first considered by Stephan in 1946 [14], and has received a fair share of attention in the literature ever since. For example, Grab and Savage [15] provided a recursive formula in n for $k = 1$, which was later generalized by Govindarajulu [16], while Chao and Strawderman [17] considered $\mathbb{E}[1/(X+c)]$, where $c \in \mathbb{R}$ such that $\mathbb{P}[X+c>0] = 1$. Recently, Wuyungaowa and Wang provided asymptotic expansions of μ_{-k}, $k \in \mathbb{N}^*$.

We now proceed to derive a more convenient formula for (2.1), hence, in view of (2.4), for μ_{-1} as well. To the best of our knowledge, this formula has not appeared in the literature before.

Theorem 2.1. *$G(n,p) = n\sum_{m=1}^{n}((q^{n-m}-q^n)/m)$. Furthermore, setting $F(x) = e^{-x}\int_0^x((e^u-1)/u)du$,*

$$\left|G(n,p) - nF(-n\ln(q))\right| \le \frac{n\ln^2(q)}{24\sqrt{q}} + 2\frac{\ln(q)}{\sqrt{q}} + O\left(n^{-1}\right). \tag{2.5}$$

For $p \ll 1$, in particular, $-n\ln(q) \approx np$, $\ln^2(q)/\sqrt{q} \approx p^2$, and $2(\ln(q)/\sqrt{q}) \approx -2p$.

Proof. We can transform (2.1) using integration:

$$\begin{aligned} G(n,p) &= n(1-p)^n \int_0^{p/(1-p)} \sum_{m=1}^{n} \binom{n}{m} u^{m-1} du \\ &= n(1-p)^n \int_0^{p/(1-p)} \frac{(1+u)^n - 1}{u} du. \end{aligned} \tag{2.6}$$

We now use the identity $x^n - 1 = (x-1)(x^{n-1} + \cdots + 1)$ to get

$$\begin{aligned} G(n,p) &= n(1-p)^n \int_0^{p/(1-p)} \sum_{m=0}^{n-1} (1+u)^m du \\ &= n(1-p)^n \sum_{m=1}^{n} \frac{\left(1+p/(1-p)\right)^m - 1}{m} \\ &= n\sum_{m=1}^{n} \frac{(1-p)^{n-m} - (1-p)^n}{m} \\ &= n\sum_{m=1}^{n} \frac{q^{n-m} - q^n}{m}, \end{aligned} \tag{2.7}$$

which proves our first claim.

We now approximate the sum in (2.7) by an integral (considering it to be a Riemann sum, or, even better, through the Euler-McLaurin formula [18]):

$$\begin{aligned} G(n,p) &= n\sum_{m=1}^{n}\frac{q^{n-m}-q^n}{m} = nq^n\sum_{m=1}^{n}\frac{q^{-m}-1}{m} = nq^n\sum_{m=1}^{n}\frac{\left(q^n\right)^{-m/n}-1}{m/n}\frac{1}{n} \\ &\longrightarrow nq^n\int_0^1\frac{q^{-nx}-1}{x}dx = nq^n\int_0^n\frac{q^{-x}-1}{x}dx = nr^{-n}\int_0^n\frac{r^x-1}{x}dx \\ &= ne^{-n\ln(r)}\int_0^{n\ln(r)}\frac{e^x-1}{x}dx = nF(n\ln(r)), \end{aligned} \tag{2.8}$$

where

$$r=\frac{1}{q},\qquad F(x)=e^{-x}\int_0^x\frac{e^u-1}{u}du. \tag{2.9}$$

But this can be further simplified as

$$\ln(r) = -\ln(q) = -\ln(1-p)\approx p \quad \text{for } 0<p\ll 1, \tag{2.10}$$

implying that

$$\frac{G(n,p)}{n}\approx F(n\ln(r))\approx F(pn) = e^{-pn}\int_0^{pn}\frac{e^x-1}{x}dx. \tag{2.11}$$

The error of this approximation is readily given by the lowest-order Newton-Cotes numerical integration formula known as the midpoint rule [19], which, applied in the function at hand, yields

$$\begin{aligned} &\left|\sum_{m=1}^{n}\frac{q^{n-m}-q^n}{m} - q^n\int_0^1\frac{q^{-nx}-1}{x}dx\right| \\ &\quad\le E(n,p) := \frac{M_2}{24n^2} + \left|q^n\int_0^{1/(2n)}\frac{q^{-nx}-1}{x}dx\right| + \left|q^n\int_1^{1+1/(2n)}\frac{q^{-nx}-1}{x}dx\right|, \end{aligned} \tag{2.12}$$

where

$$\begin{aligned} M_2 &= \max_{x\in[1/(2n),1+1/(2n)]}\left(\frac{q^{n(1-x)}-q^n}{x}\right)'' \\ &= \frac{n^2\ln^2(q)}{(1+1/2n)\sqrt{q}} + \frac{2n\ln(q)}{(1+1/2n)^2\sqrt{q}} + \frac{2\left(1/\sqrt{q}-q^n\right)}{(1+1/2n)^3}, \end{aligned} \tag{2.13}$$

as the second derivative is monotonically increasing, and, therefore, the maximum corresponds to the right endpoint of the interval, namely, $x = 1 + 1/(2n)$. For large n, the contribution of the two integral error terms becomes insignificant, while, asymptotically,

$$M_2 = n^2 \frac{\ln^2(q)}{\sqrt{q}} + 2n \frac{\ln(q)}{\sqrt{q}} + \frac{2}{\sqrt{q}} + O\left(n^{-1}\right). \tag{2.14}$$

Assuming further that $p \ll 1$, we find that $\ln^2(q)/\sqrt{q} = \ln^2(1-p)/\sqrt{1-p} \approx p^2$, and $2\ln(q)/\sqrt{q} = 2\ln(1-p)(1-p)^{-1/2} \approx -2p - p^2$; hence the total error is approximately bounded above by

$$E(n,p) = \frac{p^2}{24} - \frac{p}{12n} + O\left(n^{-2}\right). \tag{2.15}$$

This completes the proof. □

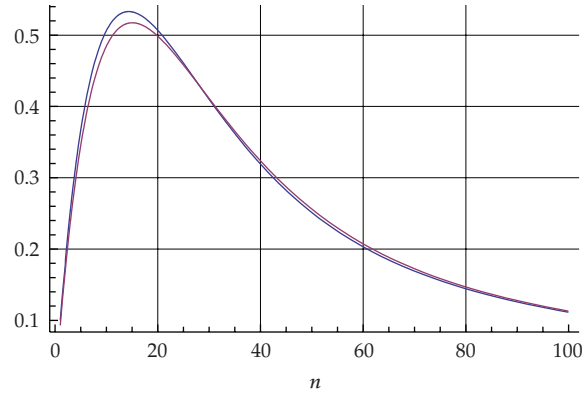
The approximation (2.11) proved in the theorem above is compared in Figure 1 against the exact (2.6), or, equivalently, (2.7). Our experiments show that for $p < 0.01$ the approximation is almost exact, while for larger values of p, where (2.15) does not yet hold, there is noticeable deviation between the two curves.

Figure 1 suggests that, for fixed p, $G(n,p)$ attains a global maximum for some n. We now prove this to be the case.

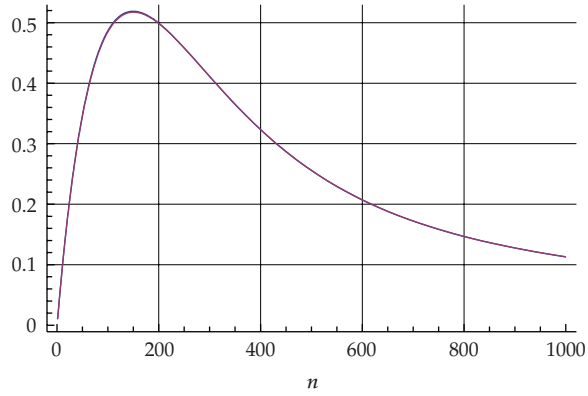
Theorem 2.2. *For fixed $p \in (0,1)$, $G(n,p)$ attains a global maximum for some n.*

Proof. To begin with, note that $G(1,p) = p$ by (2.1). Furthermore, we obtain from (2.7) that

$$\begin{aligned}
G(n,p) &= n \sum_{m=0}^{n-1} \frac{q^m}{n-m} - nq^n H(n) \\
&= \sum_{m=0}^{n-1} q^m + \sum_{m=0}^{n-1} \frac{m}{n-m} q^m - nq^n H(n) \\
&= \frac{1-q^n}{1-q} + \sum_{m=0}^{n-1} \frac{m}{n-m} q^m - nq^n H(n) \\
&= \frac{1}{p} - q^n \left(\frac{1}{p} + nH(n)\right) + \sum_{m=0}^{n-1} \frac{m}{n-m} q^m \\
&= \frac{1}{p} - q^n \left(\frac{1}{p} + nH(n)\right) + \frac{1}{n} \sum_{m=1}^{n-1} mq^m + \frac{1}{n} \sum_{m=1}^{n-1} \frac{m^2}{n-m} q^m,
\end{aligned} \tag{2.16}$$



(a)



(b)

Figure 1: Comparison of $G(n,p)/n$ as given by (2.6) versus F (in blue and red, resp.) for $p = 0.1$ (a) and $p = 0.01$ (b). For $p \leq 0.01$ the two curves are indistinguishable.

where $H(n) = \sum_{k=1}^{n}(1/k)$ denotes the nth harmonic number. We further break the last sum into two sums:

$$\begin{aligned}\frac{1}{n}\sum_{m=1}^{n-1}\frac{m^2}{n-m}q^m &= \frac{1}{n}\sum_{m=1}^{N_1}\frac{m^2}{n-m}q^m + \frac{1}{n}\sum_{m=N_1+1}^{n-1}\frac{m^2}{n-m}q^m \\ &\leq \frac{1}{n}\sum_{m=1}^{N_1}\frac{m^2}{n-m}q^m + \frac{1}{n}\sum_{m=N_1+1}^{n-1}m^2q^m,\end{aligned} \tag{2.17}$$

where $N_1 < n$ is a positive integer to be specified later. Since $\sum_{m=0}^{\infty} m^2q^m := C < \infty$ for $q < 1$, for any $\epsilon_1 > 0$ there exists a suitable N_1 such that

$$\sum_{m=N_1+1}^{n-1} m^2q^m < \sum_{m=N_1+1}^{\infty} m^2q^m < \epsilon_1. \tag{2.18}$$

On the other hand, for this specific N_1, and for any $\epsilon_2 > 0$, there exists $N_2 > N_1$ such that, for any $n > N_2$,

$$\sum_{m=1}^{N_1} \frac{m^2}{n-m} q^m < \frac{1}{n-N_1} \sum_{m=1}^{N_1} m^2 q^m < \frac{C}{n-N_1} < \epsilon_2. \tag{2.19}$$

Putting (2.17), (2.18), and (2.19) together, we obtain that, for any $\epsilon_1, \epsilon_2 > 0$, there exist $N_1 > 0$ and $N_2 > N_1$ such that, for any $n > N_2$,

$$\frac{1}{n} \sum_{m=1}^{n-1} \frac{m^2}{n-m} q^m = \frac{1}{n} \sum_{m=1}^{N_1} \frac{m^2}{n-m} q^m + \frac{1}{n} \sum_{m=N_1+1}^{n-1} \frac{m^2}{n-m} q^m < \frac{\epsilon_1 + \epsilon_2}{n}. \tag{2.20}$$

Furthermore,

$$\sum_{m=1}^{n-1} m q^m \longrightarrow \sum_{m=1}^{\infty} m q^m = \frac{q}{p^2}, \tag{2.21}$$

so that, for any $\epsilon_3 > 0$, $\exists N_3 > 0$ such that, for any $n > N_3$,

$$0 < \frac{q}{p^2} - \sum_{m=1}^{n-1} m q^m < \epsilon_3. \tag{2.22}$$

Finally, for any $n > 1/p$,

$$n q^n \left(\frac{1}{p} + nH(n) \right) < n q^n \left(n + n^2 \right) < 2n^3 q^n, \tag{2.23}$$

and we know that this decays to 0 much faster than any power: in particular, for any $\epsilon_4 > 0$, there exists $N_4 > 0$ such that, for any $n > N_4$,

$$n q^n \left(\frac{1}{p} + nH(n) \right) < \frac{\epsilon_4}{n^2}. \tag{2.24}$$

From (2.16), then, using (2.22) and (2.24), we obtain

$$\begin{aligned} G(n,p) - \frac{1}{p} &= -q^n \left(\frac{1}{p} + nH(n) \right) + \frac{1}{n} \sum_{m=1}^{n-1} m q^m + \frac{1}{n} \sum_{m=1}^{n-1} \frac{m^2}{n-m} q^m \\ &> \frac{q/p^2 - \epsilon_3}{n} - \frac{\epsilon_4}{n^2}, \end{aligned} \tag{2.25}$$

for $n > N = \max\{N_2, N_3, N_4\}$, whence, if we ensure that $q/p^2 > \epsilon_3 + \epsilon_4$, it follows that

$$G(n,p) - \frac{1}{p} > 0 \quad \text{for } n > N. \tag{2.26}$$

On the other hand,

$$\left| G(n,p) - \frac{1}{p} \right| < \frac{q/p^2 + \epsilon_1 + \epsilon_2}{n} + \frac{\epsilon_4}{n^2}, \quad n > \max\{N_2, N_4\} \Longrightarrow \lim G(n,p) = \frac{1}{p}. \tag{2.27}$$

To recapitulate, we have shown that $G(1,p) = p < 1/p$, whereas, as $n \to \infty$, $G(n,p)$ converges to $1/p$ *from above*: more specifically, we have shown that, for all n sufficiently large,

$$0 < G(n,p) = \frac{1}{p} + \frac{q}{p^2 n} + o\left(n^{-1}\right). \tag{2.28}$$

Hence, $G(n,p)$ exhibits a global maximum for some value of n, for fixed p, and this concludes the proof. □

Alternatively, the eventual positivity of $G(n,p) - 1/p$ also follows immediately from the more general asymptotic formulas presented in [20]. Having now demonstrated that $G(n,p)$ has a global maximum, the next step is to locate and calculate its value. To do this, we do not work on $G(n,p)$ directly, but rather on the approximation $G_a(n,p) = nF(np)$, derived in Theorem 2.1.

Theorem 2.3. *For fixed p, $G_a(n,p)$ is maximized for $n_{\text{opt}} \approx 4.168485/p$, and $G_a(n_{\text{opt}},p) \approx 0.310724 n_{\text{opt}} \approx 1.295248/p$. Furthermore, $G_a(n,p)/n$ is maximized for $n_{\text{opt,r}} \approx 1.502861/p$, and $G_a(n_{\text{opt,r}},p)/n_{\text{opt,r}} \approx 0.517351$.*

In particular, the maximal mean earnings possible are about 29.5% above the asymptotic value, and about 31% of the total prize money.

Proof. Considering n to be a continuous quantity, we determine the global maximum of G_a from the condition $\partial G_a(n,p)/\partial n = 0$:

$$\begin{aligned} \frac{\partial G_a(n,p)}{\partial n} &= (1-np)e^{-np} \int_0^{np} \frac{e^x - 1}{x} dx + ne^{-np} p \frac{e^{np} - 1}{np} = 0 \\ &\Longleftrightarrow \int_0^s \frac{e^x - 1}{x} dx = \frac{e^s - 1}{s-1}, \end{aligned} \tag{2.29}$$

where $s = np$. This equation can be solved numerically to yield $s \approx 4.168485$. It follows that the optimal n is asymptotically given by the formula

$$n_{\text{opt}} \approx \frac{s}{p} \approx \frac{4.168485}{p}. \tag{2.30}$$

The value of the maximum, consequently, is

$$G_a(n_{\text{opt}},p)\approx\frac{s}{p}e^{-s}\int_0^s\frac{e^x-1}{x}dx=\frac{s}{p}\frac{1-e^{-s}}{s-1}\approx\frac{1.295248}{p}\Longleftrightarrow\frac{G(n_{\text{opt}},p)}{n_{\text{opt}}}\approx 0.310724. \tag{2.31}$$

What about maximizing the relative mean earnings $G_a(n,p)/n=F(np)$? Going back to (2.11), we need to determine the roots of the derivative of F:

$$0=F'(s)=-e^{-s}\int_0^s\frac{e^x-1}{x}dx+e^{-s}\frac{e^s-1}{s}\Longleftrightarrow\int_0^s\frac{e^x-1}{x}dx=\frac{e^s-1}{s}. \tag{2.32}$$

Solving numerically we obtain $s=n_{\text{opt,r}}p\approx 1.502861$, whence

$$F(s)=e^{-s}\frac{e^s-1}{s}\approx 0.517351. \tag{2.33}$$

This completes the proof. □

As a brief numerical example, consider the Greek lottery, a 6/49 lottery system, for which $p=1/\binom{49}{6}=1/13983816$. Theorem 2.3 shows that $G_a(n,p)$ is maximized for $n_{\text{opt}}\approx 4.168485/p\approx 58.3$ million, while $G_a(n,p)/n$ is maximized for $n_{\text{opt}}\approx 1.502861/p\approx 21.0$ million. The actual number of tickets normally played, however, is significantly lower, ranging from 1 to 8 million [5], so, for the given level of participation, a 6/49 system is not optimal.

Before we conclude this section, let us investigate the variance of a winner's earnings. Our goal is to obtain a similar formula as the one presented in Theorem 2.1, and then to generalize it into a formula for any μ_{-k}, $k\in\mathbb{N}^*$. We begin by calculating the mean square earnings

$$\begin{aligned}
G^{(2)}(n,p)&=\sum_{m=1}^n\binom{n}{m}p^mq^{n-m}\left(\frac{n}{m}\right)^2=n^2q^n\sum_{m=1}^n\binom{n}{m}\frac{1}{m^2}\left(\frac{p}{q}\right)^m\\
&=n^2q^n\sum_{m=1}^n\binom{n}{m}\frac{1}{m}\int_0^{p/q}u^{m-1}du=n^2q^n\int_0^{p/q}\frac{du}{u}\sum_{m=1}^n\binom{n}{m}\int_0^u dv\,v^{m-1}\\
&=n^2q^n\int_0^{p/q}\frac{du}{u}\int_0^u\frac{dv}{v}\sum_{m=1}^n\binom{n}{m}v^m=n^2q^n\int_0^{p/q}\frac{du}{u}\int_0^u\frac{dv}{v}\left[(1+v)^m-1\right]\\
&=n^2q^n\int_0^{p/q}\frac{du}{u}\int_0^u dv\sum_{m=0}^{n-1}(1+v)^m=n^2q^n\int_0^{p/q}\frac{du}{u}\sum_{m=1}^n\frac{(1+u)^m-1}{m}\\
&=n^2q^n\int_0^{p/q}du\sum_{m=1}^n\frac{1}{m}\sum_{l=0}^{m-1}(1+u)^l=n^2q^n\sum_{m=1}^n\frac{1}{m}\sum_{l=1}^m\frac{1}{l}\left[\left(1+\frac{p}{q}\right)^l-1\right]\\
&=n^2\sum_{m=1}^n\frac{1}{m}\sum_{l=1}^m\frac{1}{l}\left[q^{n-l}-q^n\right].
\end{aligned} \tag{2.34}$$

It follows that the variance is

$$\begin{aligned} V(n,p) &= G^{(2)}(n,p) - G^2(n,p) \\ &= n^2\left[\sum_{m=1}^{n}\frac{1}{m}\sum_{l=1}^{m}\frac{1}{l}\left[q^{n-l}-q^n\right] - \left(\sum_{m=1}^{n}\frac{1}{m}\left[q^{n-m}-q^n\right]\right)^2\right]. \end{aligned} \tag{2.35}$$

Asymptotics can be found as in Theorem 2.1:

$$\begin{aligned} G^{(2)}(n,p) &= n^2q^n\sum_{m=1}^{n}\frac{1}{n}\frac{1}{m/n}\sum_{l=1}^{m}\frac{1}{n}\frac{1}{l/n}\left[e^{-(l/n)n\ln(q)}-1\right] \\ &\longrightarrow n^2e^{n\ln(q)}\int_0^1\frac{du}{u}\int_0^u\frac{dv}{v}\left[e^{-nv\ln(q)}-1\right] \\ &= n^2e^{n\ln(q)}\int_0^{-n\ln(q)}\frac{du}{u}\int_0^u\frac{dv}{v}\left[e^v-1\right]. \end{aligned} \tag{2.36}$$

The generalization is clear. For any $k \in \mathbb{N}^*$,

$$\begin{aligned} G^{(k)}(n,p) &= \sum_{m=1}^{n}\binom{n}{m}p^mq^{n-m}\left(\frac{n}{m}\right)^k \\ &= n^k\sum_{m_1=1}^{n}\frac{1}{m_1}\sum_{m_2=1}^{m_1}\frac{1}{m_2}\cdots\sum_{m_k=1}^{m_{k-1}}\frac{1}{m_k}\left(q^{n-m_k}-q^n\right), \end{aligned} \tag{2.37}$$

while the asymptotic expression as $n \to \infty$ becomes

$$G_a^{(k)}(n,p) = n^kF^{(k)}(-n\ln(q)), \quad \text{where } e^xF^{(k)}(x) = \int_0^x e^uF^{(k-1)}(u)du, \quad F^{(0)}(x) = \frac{1-e^{-x}}{x}. \tag{2.38}$$

The relation to the inverse moments follows from the generalization of (2.4)

$$G^{(k)}(n,p) = n^k(1-q^n)\mu_{-k}, \quad k \in \mathbb{N}^*. \tag{2.39}$$

3. Results with Rollover

When rollover is introduced, a sequence of lottery draws is played till a winner is found, at which point the total prize money (consisting of the prize money of the current draw plus the *jackpot*, namely, the accumulated prize money over the previous draws) is equally split among the winners. We assume that the (infinite) sequence $n \in (\mathbb{N}^*)^\infty$ of the number of participants in the various draws is known and fixed, and we denote by $G(k,n,p)$ the mean earnings in the kth lottery draw, assuming that no winner was found in the first $k-1$ draws and that a winner was found in the kth draw. It is clear that this quantity depends not on

the full n but only on its starting subsequence $(n_1, \ldots, n_k)$; hence, for example, $G(n,p)$ of the previous section equals $G(1,n',p)$ for any $n' \in (\mathbb{N}^*)^\infty$ with $n'_1 = n$.

We denote the mean earnings under rollover by $G_r(n,p)$. Letting $K(n)$ be the random variable denoting, for a fixed n, the number of draws played till a winner is found, then

$$G_r(n,p) = \sum_{k=1}^{\infty} \mathbb{P}[K(n) = k] G(k,n,p). \tag{3.1}$$

It is, in fact, possible to determine the probability distribution of $K(n)$: $K(n) = k$ if and only if no winner is found in the first $k-1$ draws and a winner is found in the kth draw. Taking into account that the probability that no winner is found in the ith draw is $(1-p)^{n_i} \approx e^{-n_i p}$ for $p \ll 1$, we immediately find

$$\mathbb{P}[K(n) = k] = e^{-p\sum_{i=1}^{k-1} n_i}\left(1 - e^{-n_k p}\right). \tag{3.2}$$

Furthermore, revisiting (2.11) in Theorem 2.1, we see that the first factor of n in the equation denotes the prize money; whereas every other occurrence of n is within the expression np and denotes the number of tickets played; whence, in the present case, the asymptotic expression $G_a(k,n,p)$ for $G(k,n,p)$ is given by

$$\frac{G_a(k,n,p)}{\sum_{i=1}^{k} n_i} = e^{-pn_k} \int_0^{pn_k} \frac{e^x - 1}{x} dx = F(pn_k). \tag{3.3}$$

To sum up, the asymptotic expression $G_{r,a}(n,p)$ for $G_r(n,p)$ is given by

$$G_{r,a}(n,p) = \sum_{k=1}^{\infty} \left(\sum_{i=1}^{k} n_i \right) e^{-p\sum_{i=1}^{k-1} n_i} \left(1 - e^{-n_k p}\right) e^{-pn_k} \int_0^{pn_k} \frac{e^x - 1}{x} dx. \tag{3.4}$$

It would be, of course, far more interesting to enrich the model by allowing n to be random, following some infinite-dimensional discrete probability distribution $\mathbb{P}[N = n]$, in which case the averaged $G_{r,a}$ would be

$$G_{r,a,\mathrm{avg}}(p) = \sum_{n \in (\mathbb{N}^*)^\infty} \mathbb{P}[N = n] G_{r,a}(n,p). \tag{3.5}$$

In general, (3.4) cannot be simplified further in a meaningful way. This, however, is possible in the special case, where, for some $n \in \mathbb{N}^*$, $n_i = n$ for all $i \in \mathbb{N}^*$, in which case we may substitute the vector $n \in (\mathbb{N}^*)^\infty$ appearing in the preceding expressions by the value $n \in \mathbb{N}$ of its first coordinate (since all coordinates are equal) and prove the following theorem, corresponding to Theorem 2.3.

Theorem 3.1. *Assuming that the number of tickets submitted remains constant throughout rollover draws, the asymptotic expression for $G_r(n,p)$ as $n \to \infty$ and $p \ll 1$ is*

$$G_{r,a}(n,p) = \frac{ne^{-np}}{1-e^{-np}} \int_0^{np} \frac{e^x - 1}{x} dx = \frac{1}{1-e^{-np}} G_a(n,p). \tag{3.6}$$

This function is maximized at $n'_{\text{opt}} = 3.750147/p$ and $G_{r,a}(n'_{\text{opt}}, p) = 1.320264/p = 0.352057 n'_{\text{opt}}$.

In particular, the maximal mean earnings possible are about 32% above the asymptotic value, and about 35% of the (first draw) total prize money.

Proof. Under the stated assumptions, (3.4) becomes

$$\begin{aligned} G_{r,a}(n,p) &= \sum_{k=1}^{\infty} \mathbb{P}(K=k) G_a(k,n,p) \\ &= \sum_{k=1}^{\infty} e^{-np(k-1)} \left(1-e^{-np}\right) kne^{-np} \int_0^{np} \frac{e^x - 1}{x} dx \\ &= n\left(1-e^{-np}\right) \int_0^{np} \frac{e^x - 1}{x} dx \sum_{k=1}^{\infty} ke^{-npk} \\ &= \frac{ne^{-np}}{1-e^{-np}} \int_0^{np} \frac{e^x - 1}{x} dx \\ &\iff \frac{G_{r,a}(n,p)}{n} = \frac{F(np)}{1-e^{-np}} = F_r(np) = \frac{e^{-np}}{1-e^{-np}} \int_0^{np} \frac{e^x - 1}{x} dx, \end{aligned} \tag{3.7}$$

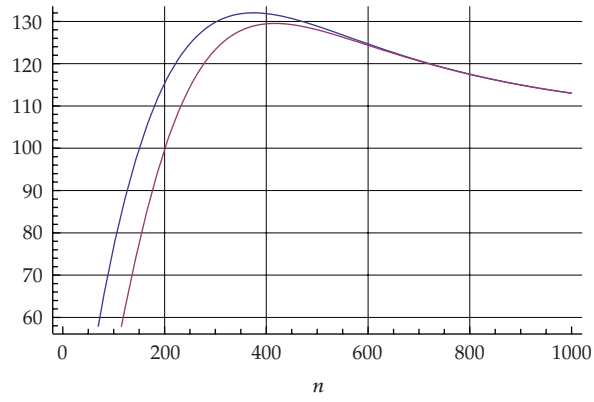
and this proves our first claim.

We now note that the asymptotic behavior of G_a and $G_{r,a}$ is identical as $n \to \infty$, while $\lim_{n\to 0} G_{r,a}(n,p) = 0$ (n is treated here as a continuous quantity). As in Theorem 2.2, it follows that $G_{r,a}$ has a global maximum, which, as before, we locate through the condition $\partial G_{r,a}(n,p)/\partial n = 0$:

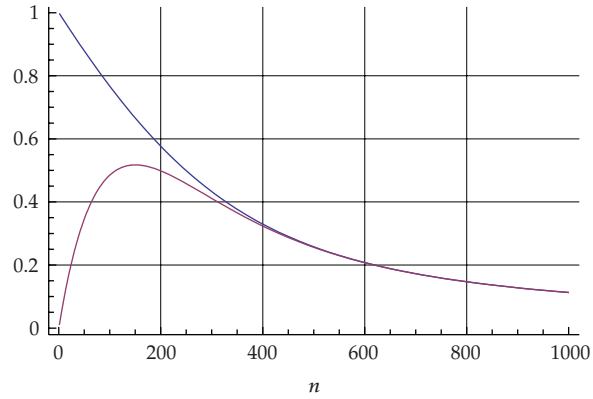
$$\begin{aligned} 0 &= \frac{(1-np)e^{-np}(1-e^{-np}) - npe^{-2np}}{(1-e^{-np})^2} \int_0^{np} \frac{e^x - 1}{x} dx + \frac{ne^{-np}}{1-e^{-np}} p \frac{e^{np} - 1}{np} \\ &= 1 + \frac{e^{-np}(1-np-e^{-np})}{(1-e^{-np})^2} \int_0^{np} \frac{e^x - 1}{x} dx \\ &\iff \int_0^s \frac{e^x - 1}{x} dx = \frac{(1-e^{-s})^2}{e^{-s}(s+e^{-s}-1)}. \end{aligned} \tag{3.8}$$

This equation can be solved numerically to yield $s \approx 3.750147$, whence

$$n'_{\text{opt}} \approx \frac{s}{p} \approx \frac{3.750147}{p}. \tag{3.9}$$



(a)



(b)

Figure 2: (a) G_a (in red) and $G_{r,a}$ (in blue) for $p = 0.01$, as given by (2.11) and (3.7), respectively. (b) F_r (in blue) and F (in red) for $p = 0.01$.

The value of the maximum, consequently, is

$$\begin{aligned} G_{r,a}\left(n'_{\text{opt}}, p\right) &\approx \frac{s}{p}\frac{e^{-s}}{1-e^{-s}}\int_0^s \frac{e^x-1}{x}dx \\ &= \frac{s}{p}\frac{e^{-s}}{1-e^{-s}}\frac{(1-e^{-s})^2}{e^{-s}(s+e^{-s}-1)} \\ &= \frac{s}{p}\frac{(1-e^{-s})}{s+e^{-s}-1} \approx \frac{1.320264}{p} \\ \Longleftrightarrow \frac{G_{r,a}\left(n'_{\text{opt}}, p\right)}{n'_{\text{opt}}} &\approx 0.352057. \end{aligned} \tag{3.10}$$

This completes the proof. □

Continuing the numerical example for the Greek lottery given right below Theorem 2.3, $G_{r,a}(n,p)$ is maximized for $n'_{\text{opt}} \approx 3.750147/p \approx 52.4$ million, which is also significantly higher than the actual number of tickets normally played.

We may attempt, as in Theorem 2.3, to maximize the relative (to the prize money accumulated in the first draw) mean earnings. To do this, we refer back to (3.7) and ask for the roots of the derivative of F_r:

$$0 = F'_r(s) = -\frac{e^s}{(e^s-1)^2}\int_0^s \frac{e^x-1}{x}dx + \frac{1}{e^s-1}\frac{e^s-1}{s} \Longleftrightarrow \int_0^s \frac{e^x-1}{x}dx = \frac{(e^s-1)^2}{se^s}. \tag{3.11}$$

This equation, however, has no root! It turns out that F_r is strictly decreasing. A more appropriate quantity to consider is the mean earnings relative to the mean accumulated prize money

$$g_r = \sum_{k=1}^{\infty} kne^{-np(k-1)}\left(1-e^{-pn}\right) = \frac{n}{1-e^{-pn}}, \tag{3.12}$$

whence it follows that $G_{r,a}/g_r = F$.

Figure 2 plots G_a, $G_{r,a}$, F_r, and F for $p = 0.01$.

4. Conclusion

We investigated the mean earnings of a lottery winner as a function of the number n of players participating in a lottery with success probability p, under the assumption that the total amount of money offered as a prize to the winners is equal (or, in general, linearly proportional) to n. We considered two versions of the lottery game, namely, with and without rollover. We concluded that both with and without rollover, both the absolute mean earnings and the fraction of the mean earnings over the total accumulated sum for an individual winner are maximized when np equals a certain constant (different, in general, for the four different stated cases), which we determined numerically. In the course of our investigation, we linked the mean earnings formula to the classical problem of determining the inverse moments of a binomial distribution, offering for them novel formulas, both exact and approximate.

Acknowledgments

The author would like to thank the two anonymous referees for their detailed and insightful comments, which greatly improved this paper; in particular, for bringing to the author's attention that the problem studied here is related to the classical problem of determining the inverse moments of the binomial distribution, and for providing an extensive list of additional references.

References

[1] A. Abrams and S. Garibaldi, "Finding good bets in the lottery, and why you shouldn't take them," *American Mathematical Monthly*, vol. 117, no. 1, pp. 3–26, 2010.

[2] J. Simon, "An analysis of the distribution of combinations chosen by UK National Lottery players," *Journal of Risk and Uncertainty*, vol. 17, no. 3, pp. 243–276, 1998.

[3] L. DeBoer, "Lotto sales stagnation: product maturity or small jackpots?" *Growth and Change*, vol. 21, no. 1, pp. 73–77, 2006.
[4] N. Henze and H. Riedwyl, *How to Win More: Strategies for Increasing a Lottery Win*, A.K. Peters, Natick, Mass, USA, 1998.
[5] "Greek Lottery past draws data," http://www.opap.gr/lottoGame.fds?langid=1.
[6] K. Drakakis, "A note on the appearance of consecutive numbers amongst the set of winning numbers in Lottery," *Facta Universitatis: Mathematics and Informatics*, vol. 22, no. 1, pp. 1–10, 2007.
[7] K. Drakakis, "On the maximal distance between consecutive choices in the set of winning numbers in Lottery," *Applied Mathematical Sciences*, vol. 3, no. 55, pp. 2725–2738, 2009.
[8] K. Drakakis and K. Taylor, "A statistical test to detect tampering with lottery results," in *Proceedings of the 2nd International Conference on Mathematics in Sport*, Groningen, The Netherlands, 2009.
[9] N. Henze, "The distribution of spaces on lottery tickets," *Fibonacci Quarterly*, vol. 33, pp. 426–431, 1995.
[10] H. Joe, "A winning strategy for lotto games?" *The Canadian Journal of Statistics*, vol. 18, pp. 233–244, 1990.
[11] P. Roger and M.-H. Broihanne, "Efficiency of betting markets and rationality of players: evidence from the French 6/49 lotto," *Journal of Applied Statistics*, vol. 34, no. 6, pp. 645–662, 2007.
[12] H. Joe, "An ordering of dependence for distribution of k-tuples, with applications to lotto games," *The Canadian Journal of Statistics*, vol. 15, pp. 227–238, 1987.
[13] H. Joe, "Tests of uniformity for sets of lotto numbers," *Statistics and Probability Letters*, vol. 16, no. 3, pp. 181–188, 1993.
[14] F. F. Stephan, "The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate," *Annals of Mathematical Statistics*, vol. 16, pp. 50–61, 1946.
[15] E. L. Grab and I. R. Savage, "Tables of the expected value of $1/X$ for positive Bernoulli and Poisson variables," *Journal of the American Statistical Association*, vol. 49, pp. 169–177, 1954.
[16] Z. Govindarajulu, "Recurrence relations for the inverse moments of the positive binomial variable," *Journal of the American Statistical Association*, vol. 58, pp. 463–473, 1963.
[17] M. Y. Chao and W. E. Strawderman, "Negative moments of positive random variables," *Journal of the American Statistical Association*, vol. 67, pp. 429–431, 1972.
[18] R. Graham, D. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, Addison-Wesley, Harlow, UK, 2nd edition, 1994.
[19] E. Isaakson and H. Keller, *Analysis of Numerical Methods*, Dover, New York, NY, USA, 1994.
[20] Wuyungaowa and T. Wang, "Asymptotic expansions for inverse moments of binomial and negative binomial," *Statistics and Probability Letters*, vol. 78, no. 17, pp. 3018–3022, 2008.