Applied Mathematical Sciences, Vol. 3, 2009, no. 55, 2725 - 2738

On the Maximal Distance between Consecutive Choices in the Set of Winning Numbers in Lottery

Konstantinos Drakakis[1]

UCD CASL, University College Dublin
Belfield, Dublin 4, Ireland
Konstantinos.Drakakis@ucd.ie

Abstract

We study the probability distribution of the biggest gap between two consecutive choices in an m-tuple of distinct integers, assuming the m-tuple is chosen uniformly from within the range $1, \dots, n > m$, or, in other words, in the winning set of an m/n lottery game. We further study the asymptotic behavior of its mean and variance.

Mathematics Subject Classification: 91A60, 97A20, 00A08

Keywords: Lottery, maximal gap, random choices

1 Introduction

The game of lottery is popular the world over. In the m/n lottery game, the player chooses m integers from among the integers $1, \dots, n > m$, the order of choice being unimportant; the lottery organizers choose publicly m numbers uniformly at random in the same way, and if they turn out to be the same numbers as the ones the player chose, the player wins. The specific values of m and n vary from lottery to lottery, but $m = 6$ and $n = 49$ appears to be a popular choice for many national lotteries. In practice, the lottery is hardly ever a win/lose game, as various lesser winning prizes are commonly awarded to the player according to how many chosen numbers the player and the organizers have in common.

The media usually publish the winning set of numbers, along with (simple) statistics on the number of times each particular number from 1 to n has appeared in the winning set. The lottery is, however, an opportunity to carry out much more sophisticated statistical/probabilistic studies, such as the ones

[1] The author is also affiliated with the School of Electronic, Electrical & Mechanical Engineering, University College Dublin, Ireland.

already presented in various papers in the mathematical literature [2, 5, 6, 7]. In a previous work of ours [2], in particular, we studied the following question: *"What is the probability that, out of $m > 0$ numbers drawn uniformly randomly from the range $1, \dots, n$, where $n \geq m$, at least two are consecutive, or, more generally, that no two numbers are closer than k integers apart?"*

The answer turned out to be a non-trivial exercise in constrained combinatorics, and rather surprising as well, to the extent of being almost paradoxical, as it showed that the chosen numbers tend to form clusters much more often than we would intuitively expect: for example, in the case $m = 6$ and $n = 49$, almost half of the time the winning sets contain consecutive integers! In this work, we propose to study the related question of how far apart two consecutive choices are likely to be:
"What is the probability that, out of $m > 0$ numbers drawn uniformly randomly from the range $1, \dots, n$, where $n > m$ (and subsequently sorted), two consecutive such choices lie k integers apart or further?"

In other words, in our previous work we studied the smallest gap between two consecutive choices; we now study the largest gap between two consecutive choices. Both of these problems are actually problems in discrete probability only motivated by the game of lottery, in line with the tradition of the development of probability theory out of problems originating in games of chance.

2 The result

Let us choose an m-tuple $(X_1, \dots, X_m)$ uniformly randomly out of the m-tuples in the range $1, \dots, n > m$; without loss of generality we assume $X_i < X_j$ whenever $i < j$. For a given integer k, we define the event $X_{i+1} - X_i \geq k$ by A_i, $i = 1, \dots, m-1$. We define the new random variable $Y = \max\limits_{i=1,\dots,m-1}(X_{i+1} - X_i)$, and we seek $p(k, m, n) = \mathbb{P}(Y = k)$.

2.1 The probability distribution

Let us start by computing

$$P(k,m,n) = \mathbb{P}(Y \geq k) = \mathbb{P}(\cup_{i=1}^{m-1} A_i) = \frac{N(\cup_{i=1}^{m-1} A_i)}{\binom{n}{m}}. \quad (1)$$

The Inclusion-Exclusion Principle [1] gives:

$$N(\cup_{i=1}^{m-1} A_i) = \sum_{i=1}^{m-1} N(A_i) - \sum_{1 \leq i_1 < i_2 \leq m-1} N(A_{i_1} \cap A_{i_2}) + \sum_{1 \leq i_1 < i_2 < i_3 \leq m-1} N(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots \quad (2)$$

We first observe that there are $\binom{m-1}{l}$ ways to choose l out of $m-1$ indices. In order to compute $N(A_{i_1} \cap \ldots \cap A_{i_l})$, we note that, for any set of choices satisfying the event $A_{i_1} \cap \ldots \cap A_{i_l}$, choices $i_1, \ldots, i_l$ are each followed by at least $k-1$ integers, none of which lies among our choices. An equivalent way then for making such a choice is to consider $n - l(k-1)$ "white balls" ordered on a line, choose m among them at random, insert $k-1$ "black balls" in the line after each one of the choices $i_1, \ldots, i_l$, and finally drop the colors and order the balls consecutively from 1 to n. It follows that

$$N(A_{i_1} \cap \ldots \cap A_{i_l}) = \binom{n - l(k-1)}{m}, \tag{3}$$

hence finally that

$$N(\cup_{i=1}^{m-1} A_i) = \sum_{l=1}^{m-1} (-1)^{l-1} \binom{m-1}{l} \binom{n - l(k-1)}{m} \tag{4}$$

and that

$$P(k,m,n) = \mathbb{P}(\cup_{i=1}^{m-1} A_i) = \\ = \left(\sum_{l=1}^{m-1} (-1)^{l-1} \binom{m-1}{l} \binom{n - l(k-1)}{m} \right) \Big/ \binom{n}{m}. \tag{5}$$

In the expressions above we follow the usual convention that $\binom{n}{m} = 0$ if $n < m$. The following facts are easily verified:

- $P(1,m,n) = 1$, $P(k,m,n) > 0$ for $1 \leq k \leq n-m+1$, and $P(k,m,n) = 0$ for $k \geq n-m+2$.
- The nonzero terms in the sum for a given $k > 1$ correspond to $1 \leq l \leq \min\left(m-1, \left\lfloor \frac{n-m}{k-1} \right\rfloor\right)$ and to $1 \leq l \leq m-1$ for $k = 1$.

Consequently, the probability $p(k,m,n) = \mathbb{P}(Y = k)$ can be computed as

$$p(k,m,n) = P(k,m,n) - P(k+1,m,n) = \\ = \left(\sum_{l=1}^{m-1} (-1)^{l-1} \binom{m-1}{l} \left[\binom{n - l(k-1)}{m} - \binom{n - lk}{m} \right] \right) \Big/ \binom{n}{m} \tag{6}$$

for $k \geq 1$. Note that $p(k,m,n) = 0$ for $k \leq 0$.

2.2 Asymptotics, mean value, and variance

For large n, (6) can be simplified, as, setting $x = n - lk$, it follows that

$$\binom{x+l}{m} - \binom{x}{m} = \binom{x}{m}\left[\frac{(x+1)\ldots(x+l)}{(x-m+1)\ldots(x-m+l)} - 1\right] \approx \binom{x}{m}\frac{lm}{x}, \quad (7)$$

and hence that

$$p(k,m,n) \approx \left(\sum_{l=1}^{m-1}(-1)^{l-1}\binom{m-1}{l}\binom{n-lk}{m}\frac{lm}{n-lk}\right)\Bigg/\binom{n}{m}. \quad (8)$$

Expanding the binomial coefficients using their factorial representation and using the approximation

$$\frac{(n-lk-1)!}{(n-lk-m)!} \approx (n-lk)_+^{m-1}, \quad (9)$$

valid for $n \gg m$, yields the further simplification that

$$p(k,m,n) \approx \left(\sum_{l=1}^{m-1}(-1)^{l-1}\frac{(n-lk)_+^{m-1}}{(l-1)!(m-1-l)!}\right)\Bigg/\binom{n}{m}, \quad (10)$$

where $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ if $x \leq 0$.

In order to compute the mean value $\mu = \mathbb{E}(X) = \sum_{k\geq 1} kp(k,m,n)$ of this probability distribution, we first consider the sum

$$\sum_{k=0}^{n/l} k(n-lk)^{m-1} \approx \int_0^{n/l} x(n-lx)^{m-1}dx = \frac{1}{l^2}\int_0^n (n-y)y^{m-1}dy = \frac{n^{m+1}}{l^2 m(m+1)}, \quad (11)$$

whereby, using the approximation

$$\frac{n^{m+1}(n-m)!}{n!} \approx n, \quad (12)$$

valid for $n \gg m$, we finally obtain

$$\mu \approx \frac{n}{m+1}\sum_{l=1}^{m-1}\frac{(-1)^{l-1}}{l}\binom{m-1}{l} = \frac{n}{m+1}\sum_{l=1}^{m-1}\frac{1}{l}. \quad (13)$$

The last equality is by no means evident, though it is a known combinatorial identity. To prove it, start with the well known binomial identity:

$$(x+y)^m = \sum_{l=0}^{m} \binom{m}{l} x^l y^{n-l} = y^m + \sum_{l=1}^{m} \binom{m}{l} x^l y^{n-l} \Leftrightarrow$$

$$\frac{(x+y)^m - y^m}{x} = \sum_{l=1}^{m} \binom{m}{l} x^{l-1} y^{n-l}. \quad (14)$$

Then take the anti-derivative with respect to x and set $y = 1$ to obtain

$$\int_0^x \frac{(u+1)^m - 1}{u} du = \sum_{l=1}^{m} \binom{m}{l} \frac{x^l}{l}. \quad (15)$$

Set further $x = -1$ to obtain

$$\sum_{l=1}^{m} \binom{m}{l} \frac{(-1)^{l-1}}{l} = \int_{-1}^{0} \frac{(u+1)^m - 1}{u} du =$$

$$= \int_0^1 \frac{u^m - 1}{u-1} du = \int_0^1 \sum_{i=0}^{m-1} u^i du = \sum_{i=0}^{m-1} \int_0^1 u^i du = \sum_{i=1}^{m} \frac{1}{i}. \quad (16)$$

This proves the identity.

Our asymptotic analysis shows that the mean value of X is asymptotically linear in n (assuming $n \gg m$); this result agrees with our observations in numerical simulations. In particular, note that this result is valid for all m, not just large m.

In order to compute the variance $\sigma^2 = \mathbb{E}(X^2) - \mu^2$, we follow a similar procedure. We first consider the sum

$$\sum_{k=0}^{n/l} k^2 (n-lk)^{m-1} \approx \int_0^{n/l} x^2 (n-lx)^{m-1} dx =$$

$$= \frac{1}{l^3} \int_0^n (n-y)^2 y^{m-1} dy = \frac{2n^{m+2}}{l^3 m(m+1)(m+2)}, \quad (17)$$

and, using this with (10), we obtain that

$$\mathbb{E}(X^2) \approx \frac{2n^2}{(m+1)(m+2)} \sum_{l=1}^{m-1} \binom{m-1}{l} \frac{(-1)^{l-1}}{l^2} =$$

$$= \frac{2n^2}{(m+1)(m+2)} \sum_{i=1}^{m-1} \frac{1}{i} \sum_{j=1}^{i} \frac{1}{j}. \quad (18)$$

Hence the asymptotic variance is

$$\sigma^2 = \mathbb{E}(X^2) - \mu^2 \approx \frac{n^2}{m+1}\left[\frac{2}{m+2}\sum_{i=1}^{m-1}\frac{1}{i}\sum_{j=1}^{i}\frac{1}{j} - \frac{1}{m+1}\left(\sum_{i=1}^{m-1}\frac{1}{i}\right)^2\right], \quad (19)$$

and therefore the standard deviation is also asymptotically linear in n, assuming $n \gg m$. Again, we stress that this result is valid for all m, not just large m.

In both (11) and (17), in order to approximate the sum by an integral, we made use of the Euler summation formula [4].

The asymptotic forms for μ and σ we have obtained are still rather complicated, but simpler forms can be obtained if we focus on large values of m (namely assuming that $m \gg 1$), by simplifying the discrete sums appearing in the two expressions (note that the current expressions are accurate for all m, as long as $n \gg m$). We will make use of the well known asymptotic expansion [4]

$$H(n) = \sum_{i=1}^{n}\frac{1}{i} = \ln(n) + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + O(n^{-6}), \quad (20)$$

where $\gamma = 0.5772156649\ldots$ is the Euler-Mascheroni constant. It follows immediately that

$$\mu = \frac{n}{m+1}(\gamma + \ln(m-1)), \quad (21)$$

but, in order to simplify σ^2, we also need an asymptotic expression for the sum $\sum\limits_{i=1}^{m-1}\frac{H(i)}{i}$:

$$\sum_{i=1}^{m-1}\frac{H(i)}{i} = \sum_{i=1}^{m-1}\left(\frac{\ln(i)}{i} + \frac{\gamma}{i} + \frac{1}{2i^2} - \frac{1}{12i^3} + \frac{1}{120i^5} + \ldots\right) \approx$$
$$\approx \sum_{i=1}^{m-1}\frac{\ln(i)}{i} + \gamma H(m-1) + \alpha \approx \sum_{i=2}^{m-1}\frac{\ln(i)}{i} + \gamma(\gamma + \ln(m-1)) + \frac{1}{2}\alpha, \quad (22)$$

where

$$\alpha = \zeta(2) - \frac{1}{6}\zeta(3) + \frac{1}{60}\zeta(5) + \ldots \quad (23)$$

is a constant and ζ denotes Riemann's ζ-function. Applying further Euler's summation formula on the remaining sum we get

$$\sum_{i=2}^{m-1} \frac{\ln(i)}{i} = \sum_{i=1}^{m-2} \frac{\ln(i+1)}{i+1} \approx \int_0^{m-1} \frac{\ln(x+1)}{x+1} dx - \frac{1}{2}\frac{\ln(m)}{m} = \\ = \frac{1}{2}\left(\ln^2(m) - \frac{\ln(m)}{m}\right), \quad (24)$$

whence, only keeping terms asymptotically larger than 1,

$$\sigma^2 \approx \frac{n^2}{(m+1)^2}\left[\frac{m+1}{m+2}\ln^2(m) + 2\frac{m+1}{m+2}(\gamma^2 + \gamma\ln(m-1) + \alpha/2) \right. \\ \left. - \gamma^2 - 2\gamma\ln(m-1) - \ln^2(m-1)\right]. \quad (25)$$

We now observe that

$$\frac{m+1}{m+2}\ln^2(m) - \ln^2(m-1) \approx -\frac{\ln^2(m)}{m+2} - \frac{1}{m^2} \text{ and } 2\gamma\frac{\ln(m-1)}{m+2} \quad (26)$$

are asymptotically negligible compared to 1, whence

$$\sigma^2 \approx \frac{n^2}{(m+1)^2}(\alpha + \gamma^2) \approx 1.795\frac{n^2}{(m+1)^2}. \quad (27)$$

To sum up,

- the probability distribution of Y is given by (6);
- the asymptotic behavior of its mean and variance is given by (13) and (19), respectively, assuming $n \gg m$;
- assuming further that $n \gg m \gg 1$, the asymptotic behavior of its mean and variance can be further simplified and is given by (21) and (27), respectively.

Figure 1 shows how the mean and the standard deviation of Y, along with their asymptotic estimates (21) and (27), vary with m and n. For fixed m and varying n, the estimates are linear in n and are very good approximations of the actual quantities, only off by a small constant. For fixed n and varying m, however, we verify that the estimates are accurate only as long as $m \ll n$; beyond this, the curves diverge.

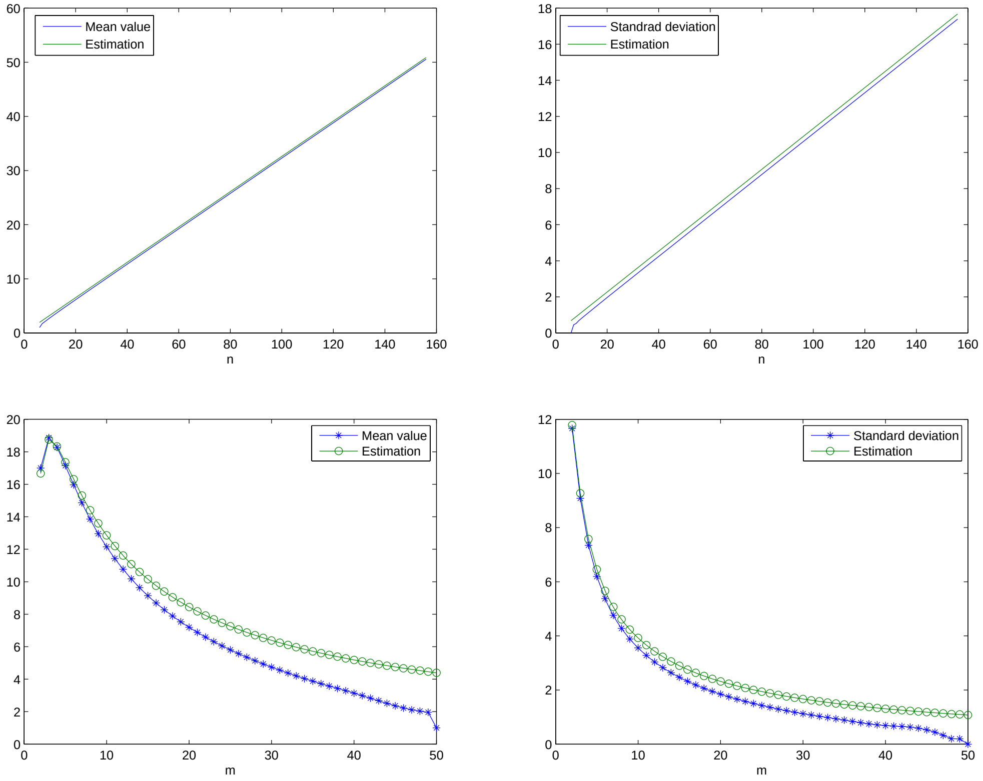

Figure 1: Mean value (left) and variance (right) of the probability distribution of Y along with the estimates (13) and (19) for $m = 6$ and $6 \leq n \leq 150$ (top row) and for $n = 50$ and $2 \leq m \leq 50$ (bottom row).

3 Attempts to approximate the probability distribution

3.1 Simplification of the original formulas

Can we get a simpler closed form solution for (6) or (10), in particular not involving a sum? Revisiting (10), expanding the binomial coefficients, and using the approximation that

$$(n-lk)_+^{m-1}\frac{(n-m)!}{n!} \approx \frac{1}{n}\left(1-\frac{lk}{n}\right)_+^{m-1} \approx \frac{1}{n}e^{-lk(m-1)/n}\mathbf{1}_{l<n/k}(l), \tag{28}$$

which grows less and less accurate as the product kl approaches n, we obtain

$$p(k,m,n) \approx \frac{m}{n}\sum_{l=0}^{\min(m-1,n/k)}(-1)^{l-1}l\binom{m-1}{l}e^{-lk(m-1)/n}. \tag{29}$$

In order to sum this in closed form, we use $m-1$ as the upper limit in the sum independently of k (which is a rather extreme approximation step):

$$\min(m-1, n/k) \approx m-1 \text{ for all } k. \tag{30}$$

Revisiting the binomial identity in (14), using $m-1$ in place of m and differentiating with respect to x yields

$$(m-1)(x+y)^{m-2} = \sum_{l=1}^{m-1} \binom{m-1}{l} l x^{l-1} y^{m-1-l}. \tag{31}$$

Use $y=1$ and $x=-e^{-k(m-1)/n}$ to obtain

$$(m-1)(1-e^{-k(m-1)/n})^{m-2} = \sum_{l=1}^{m-1} \binom{m-1}{l} l(-1)^{l-1} e^{-lk(m-1)/n} e^{k(m-1)/n}, \tag{32}$$

whence

$$p(k,m,n) \approx (m-1)\frac{m}{n} e^{-k(m-1)/n}(1-e^{-k(m-1)/n})^{m-2}. \tag{33}$$

In order to find the most likely value k_{ml}, we set $x=e^{-k(m-1)/n}$, then set the derivative of $x(1-x)^{m-1}$ with respect to x equal to 0, finding $x=1/(m-1)$ as the only root, and then set

$$e^{-k_{ml}(m-1)/n} = \frac{1}{m-1} \Leftrightarrow k_{ml} = n\frac{\ln(m-1)}{m-1}. \tag{34}$$

Hence, the (approximate) most likely value of k is close to μ and also asymptotically linear in n; this result also agrees with our observations in numerical simulations that the peak of the probability distribution is very close to the mean. Unfortunately, a direct comparison of the graphs of (6) and (33) (see Figure 2) reveals that the latter is not a very good approximation of the former (it may not even be a proper probability distribution, as its terms may not be summable to 1), as it exhibits a much heavier tail, and hence that (30) is not a valid approximation, though both curves have the same basic features: they both start with zero values, decay to zero values, and have a single local (hence global) maximum.

3.2 Negative binomial fit

Getting back to (33), let us perform the approximation

$$e^{-k(m-1)/n} \approx \left(1-\frac{m-1}{n}\right)^k \text{ and } 1-e^{-k(m-1)/n} \approx$$
$$\approx 1-\left(1-\frac{m-1}{n}\right)^k \approx k\frac{m-1}{n}, \tag{35}$$

valid when $m \ll n$, which leads to the approximation

$$p(k,m,n) \approx k^{m-2}(m-1)\frac{m}{n}\left(1-\frac{m-1}{n}\right)^k \left(\frac{m-1}{n}\right)^{m-2} \approx$$
$$\approx k^{m-2}(m-1)\left(1-\frac{m}{n}\right)^k \left(\frac{m}{n}\right)^{m-1}. \quad (36)$$

Setting

$$p = m/n, \quad (37)$$

p is the probability of choosing an integer out of the range $1, \dots, n$, when m such integers are chosen in total. Therefore,

$$p(k,m,n) \sim (1-p)^k p^{m-1}, \quad (38)$$

which is the probability mass function of the negative binomial distribution. This implies that, as long as $m \ll n$, $p(k,m,n)$ follows approximately a negative binomial distribution, whose full mass function is

$$p_{nb}(k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \quad (39)$$

and expresses the probability that, in $k+r$ independent and identically distributed trials of an experiment with success probability p, exactly r successes occur with the last success occurring at trial $k+r$; the formula remains valid, however, for non-integer values of r. In order to fit the best possible negative binomial distribution to our original distribution, we use the mean and variance of the former

$$\mu_{nb} = r\frac{1-p}{p},\ \sigma_{nb}^2 = r\frac{1-p}{p^2} \quad (40)$$

and match them to those of the latter:

$$\mu_{nb} = \mu,\ \sigma_{nb}^2 = \sigma^2. \quad (41)$$

It follows that

$$p = \frac{\mu}{\sigma^2} \approx \frac{m+1}{n} \frac{H(m-1)}{2\frac{m+1}{m+2}\sum_{i=1}^{m-1}\frac{H(i)}{i} - H^2(m-1)} \text{ and } r = \mu\frac{p}{1-p}. \quad (42)$$

The estimate for p agrees with the cruder previous estimate (37), with the exception of the factor involving the harmonic numbers.

Figure 2 compares the various distributions for Y we have proposed so far, namely the exact distribution (6), the approximation (33), and the negative binomial fit (39), for two sets of parameters: for the first set $n = 150$ and $m = 6$ we see that the binomial fit is almost identical to the exact distribution, while for the second set $n = 50$ and $m = 10$ we see that the binomial fit is a good but not perfect approximation. In both cases, the approximation (33) is very far from the exact distribution, though its peak (most likely value) is quite accurate, as we saw above.

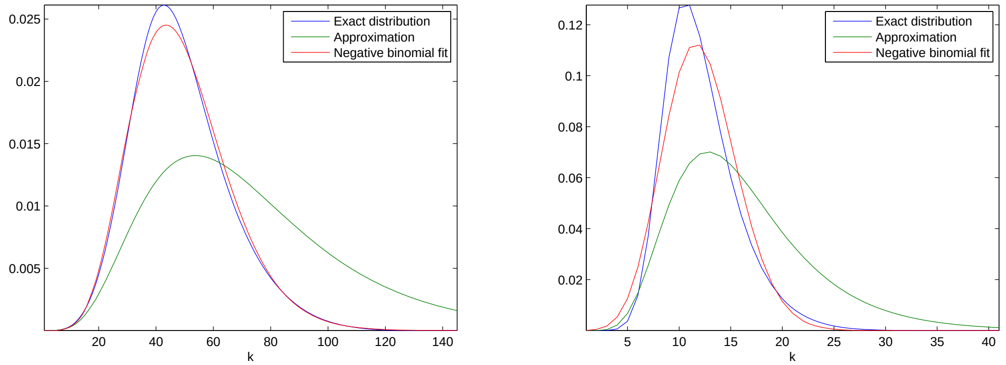

Figure 2: Comparison of the various distributions proposed for Y, for $n = 150$ and $m = 6$ (left), and for $n = 50$ and $m = 10$ (right): the exact distribution (6), the approximation (33), and the negative binomial fit (39).

4 Comparison with actual lottery data

In order to verify the correctness of our derivation of (6), we compare it against actual lottery data collected from the website of the Italian national lottery http://www.lottomatica.it, which uses $n = 90$, $m = 5$. The advantage of the Italian national lottery over other national lotteries is that it has been running under the same rules since 1939: the archives contained 45221 winning sets till March 2009. Figure 3 shows the distribution of Y against the histogram of the data: the match is virtually exact.

5 Some joint probability distributions

In addition to the random variable Y defined in the beginning of Section 2, and in the same context, let us define $Z = \min_{i=1,\dots,m-1}(X_{i+1} - X_i)$. We shall now endeavor to define the joint probability distribution of Y and Z, and more specifically $\mathbb{P}(Y \geq K, Z \geq k)$ with $K \geq k \geq 1$. To begin with, any choice of m integers in the range $1, \dots, n$ that satisfies both conditions has the property that between any two consecutive choices lie at least $k-1$ non-chosen integers, and there are, of course, exactly $m-1$ such pairs of consecutive choices. We can then remove $(m-1)(k-1)$ integers, and by renumbering we reach an m-tuple chosen within the range $1, \dots, n-(m-1)(k-1)$ with a maximal distance between consecutive choices of at least $K-k+1$. Conversely, starting with such an m-tuple, adding $k-1$ "spaces" after each one of the first $m-1$ choices, and renumbering, including the spaces, consecutively, we obtain an m-tuple in the the range $1, \dots, n$ with the two original properties. It follows from (5) that:

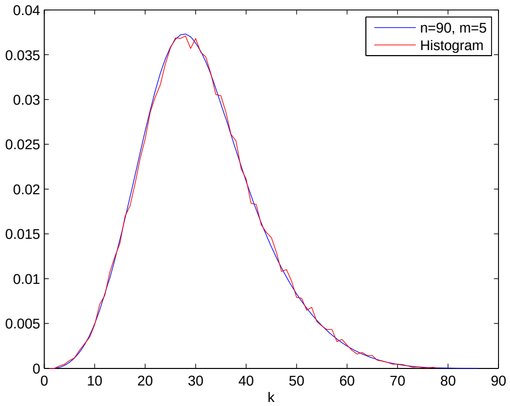

Figure 3: Comparison of the histogram of Y over the winning sets of the Italian lottery, which uses $n = 90$ and $m = 5$, versus the theoretical distribution of Y for this parameters as given by (6).

$$\mathbb{P}(Y \geq K, Z \geq k) = \\ = \left(\sum_{l=1}^{m-1} (-1)^{l-1} \binom{m-1}{l} \binom{n-(m-1)(k-1)-l(K-k)}{m}\right) \Big/ \binom{n}{m}, \quad (43)$$

while the probability mass function can be obtained by the formula:

$$\mathbb{P}(Y \geq K, Z \geq k) = \mathbb{P}(Y \geq K, Z \geq k) - \mathbb{P}(Y \geq K+1, Z \geq k) \\ - \mathbb{P}(Y \geq K, Z \geq k+1) + \mathbb{P}(Y \geq K+1, Z \geq k+1), \quad (44)$$

a result of the Inclusion-Exclusion Principle.

Incidentally, by setting $K = k$, (43) also proves that

$$\mathbb{P}(Z \geq k) = \mathbb{P}(Y \geq k, Z \geq k) = \binom{n-(m-1)(k-1)}{m} \Big/ \binom{n}{m}, \quad (45)$$

which was the subject of study of a previous work of ours [2].

We can now find the joint probability distribution of Y, Z, X_1, and X_m using the formula:

$$\mathbb{P}(Y \geq K, Z \geq k, X_1 \geq s, X_m \leq S) = \\ = \mathbb{P}(Y \geq K, Z \geq k | X_1 \geq s, X_m \leq S) \mathbb{P}(X_1 \geq s, X_m \leq S). \quad (46)$$

Using (44), we see that the effect of setting $X_1 \geq s$ and $X_m \leq S$ is effectively to use $S - s + 1$ in place of n in the numerator, whence

$$\mathbb{P}(Y \geq K, Z \geq k | X_1 \geq s, X_m \leq S) = \\ = \left(\sum_{l=1}^{m-1} (-1)^{l-1} \binom{m-1}{l} \binom{S-s+1-(m-1)(k-1)-l(K-k)}{m} \right) \Big/ \binom{n}{m}. \quad (47)$$

Applying the same substitution, we see immediately that

$$\mathbb{P}(X_1 \geq s, X_m \leq S) = \binom{S-s+1}{m} \Big/ \binom{n}{m}. \quad (48)$$

Combining the three formulas,

$$\binom{n}{m}^2 \mathbb{P}(Y \geq K, Z \geq k, X_1 \geq s, X_m \leq S) \Big/ \binom{S-s+1}{m} = \\ = \left(\sum_{l=1}^{m-1} (-1)^{l-1} \binom{m-1}{l} \binom{S-s+1-(m-1)(k-1)-l(K-k)}{m} \right) \quad (49)$$

6 Conclusion

Assuming $(X_1, \ldots, X_m)$ is an integer m-tuple without repeated entries chosen uniformly from within the range $1, \ldots, n$, so that $n \geq m$, and so that $X_i < X_j$ whenever $i < j$, what is the probability distribution of $Y = \max_{i=1,\ldots,m-1} (X_{i+1} - X_i)$? We found the exact form of this distribution in the form of a sum involving binomial coefficients, which we were unable to simplify further. In terms of the popular game of chance known as the lottery, we seek the maximal distance between two consecutive choices in the winning set of numbers of the lottery (assuming the choices are sorted increasingly and that the winning set is chosen uniformly randomly).

We then turned our attention to the calculation/estimation of the mean value and standard deviation of this probability distribution. We were successful in finding asymptotic formulas valid for $n \gg m$, both linear in n, which we simplified even further under the assumption that $n \gg m \gg 1$. We also attempted to find a simplified expression approximating the probability distribution, and some asymptotic manipulation suggested that the standard negative binomial distribution might be a good candidate: fitting the negative binomial with the correct mean and standard deviation, we observed that the fit is indeed good as long as $n \gg m$.

We tested our result against real lottery data collected from the website of the Italian national lottery: the resulting histogram matched perfectly the distribution we derived. This result can be used to test for tampering with lottery results, based on the law of large numbers, along the lines we followed in [3]. We finally formulated some joint probability distributions involving Y and $Z = \min_{i=1,\dots,m-1}(X_{i+1} - X_i)$.

References

[1] Ch. Charalambides. “Enumerative combinatorics.” Chapman & Hall/CRC, 2002.

[2] K. Drakakis. “A note on the appearance of consecutive numbers amongst the set of winning numbers in Lottery.” Facta Universitatis: Mathematics and Informatics 22(1), 2007, pp. 1–10.

[3] K. Drakakis, K. Taylor: “A statistical test to detect tampering with lottery results.” 2nd International Conference on Mathematics in Sport, 2009.

[4] R. Graham, D. Knuth, and O. Patashnik. “Concrete Mathematics: A Foundation for Computer Science (2nd edition).” Addison-Wesley, 1994.

[5] N. Henze. “The distribution of spaces on lottery tickets.” Fibonacci Quarterly 33, 1995, pp. 426–431.

[6] N. Henze and H. Riedwyl. “How to win more: Strategies for increasing a lottery win.” A.K. Peters, Natick, Massachusetts 1998.

[7] L. Holst. “On discrete spacings and the Bose-Einstein distribution.” Contributions to probability and statistics, Hon. G. Blom, 1985, pp. 169–177.

Received: May, 2009