The North Carolina Lottery Coincidence

Leonard A. STEFANSKI

The sets of five numbers picked in the North Carolina Cash-5 Lottery game were identical on July 9th and 11th, 2007. This coincidence was the topic of a local television station news story on July 12 in which I played a minor role. This article documents the coincidence, my interactions with the television reporter seeking to understand how likely, or unlikely, the coincidence was, and some afterthoughts, including an analysis of the likelihood of matching sets of numbers.

KEY WORDS: Birthday problem; Matching probability; Recursion formula; Television reporter.

1. WHAT ARE THE CHANCES?

Late in the morning of Thursday, July 12, 2007, a reporter with WTVD (Channel 11) in Raleigh, NC, telephoned the Department of Statistics and asked if he could interview someone about the lottery coincidence that occurred earlier that week. He was referring to the fact that the North Carolina Lottery Cash-5 numbers came up the same on Monday and Wednesday. In the Cash-5 game, five distinct numbers between 1 and 39 are picked; order does not matter. On July 9 and 11, the same set of five numbers, \{4, 21, 23, 34, 39\}, bubbled up in the number-selection machines. The reporter wanted to know the probability of such a coincidence, or so I assumed. In hindsight, what he really wanted was someone with a modicum of authority to quote a very small probability on camera, and he got it eventually—read on.

The reporter’s call was forwarded to me. Being unfamiliar with the North Carolina lottery, and having little experience in the statistical analysis of coincidences (I wish I had read Diaconis and Mosteller (1989) prior to July 12), I was hesitant. However, in light of recent involvement with the ASA’s efforts in public relations, and the benefits to the Department of maintaining good relations with the media, my sense of obligation won out over my hesitancy, and I agreed to an interview. A meeting was set for 2:30 pm, and the reporter and a cameraman showed up 15 minutes early.

In the interim, I met with colleagues and decided what I should tell him. The problem here is that the calculation of probabilities of events after they have occurred is fraught with difficulties, akin to assessing significance of observed clusters of cancer cases; see Hanley (1992) for a discussion of this issue in the context of lottery coincidences. Among other problems, there is no fixed or well-defined sample space with which to assess the likelihood of the observed coincidence. However, not wanting to miss the chance of being on television, I forged ahead. I was on television, but it was the 5:00 p.m. news, not the 6:00 p.m. news that I told friends to watch—Doh! I wrote a few formulas on my office white board, prepared notes, and arranged for my colleague John Monahan to be present to correct me in the event I mis spoke. I presented the probabilities in Table 1 calculated via

\[
p_k = 1 - \frac{(n-1)(n-2)\ldots(n-k+1)}{n^{k-1}}, \tag{1}
\]

where \(n = 39C5 = 39!/(34! 5!) = 575,757\) is the number of possible sets of five numbers chosen from \{1, 2, \ldots, 39\}, and \(k = \) the number of drawings. These are probabilities of at least one matching set of five numbers in a given set of \(k\) drawings. So the chances of getting at least one match in a given \(k = 3\) days is approximately 1 in 191,919.

Formula (1) is like that for the Birthday Problem in which one calculates the probability of at least one birthday match in a group of \(k\) people,

\[
\Pr(\text{Birthday match}) = 1 - \frac{(365 - 1)(365 - 2)\ldots(365 - k + 1)}{365^{k-1}}.
\]

Instead of 365 possible days to match, there are \(n = 575,757\) possible sets of five numbers to match; and instead of \(k\) people, there are drawings on \(k\) days.

The reporter was noticeably disappointed that the three-day probability, \(p_3 \approx 1/191,919\), was not “astronomically” small. He expected a probability on the order of \(1/(575,757)^2\), and was surprised to learn that \(p_3\) is approximately three times greater than the probability of winning on a given day (\(= p_2\) in Table 1).

I explained off-camera, and again on camera, that these calculations assumed a given three-day period, and did not tell the whole story in that a refinement of the calculations that addressed the encountered-coincidence nature of the problem.

| # Days | Probability \((p_k)\) | Approximately one in ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.737e-006</td>
<td>575,757</td>
</tr>
<tr>
<td>3</td>
<td>5.211e-006</td>
<td>191,919</td>
</tr>
<tr>
<td>7</td>
<td>3.647e-005</td>
<td>27,417</td>
</tr>
<tr>
<td>31</td>
<td>0.0008073</td>
<td>1,239</td>
</tr>
<tr>
<td>258</td>
<td>0.05596</td>
<td>18</td>
</tr>
<tr>
<td>365</td>
<td>0.1090</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 1. Probabilities of at least one matching set of five numbers in the North Carolina Cash-5 lottery game in any given set of \(k\) drawings.

Leonard A. Stefanski is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (E-mail: steffanski@stat.ncsu.edu). The author acknowledges helpful discussions with Dennis Boon, David Dickey, Sujit Ghosh, Donald Martin, and John Monahan.
would result in larger probabilities. With help from John Mon-
ahan I argued that what we thought was interesting, and what
the reporter should be interested in, is the probability that the
same set of numbers would appear within a short-enough pe-
riod of time so that people would notice it, anytime since the
start of the lottery. In other words, the relevant calculation is
the probability of at least one match in some short period of time
since the inception of the Cash-5 game. We also pointed out that
this more-relevant probability (more relevant for the purpose of
assessing evidence of tampering, equipment malfunctioning, rarer-
ness, etc.) would be greater than the “1 in 191,919” chance in
Table 1. The reporter then asked “So having the same set of
numbers appear within a small number of days is not really all
that extremely unlikely?” (paraphrasing him). And I replied
“Correct” (paraphrasing myself).
The reporter’s recognition of the relative magnitudes of the
various probabilities seemed like a success, just as when a per-
son realizes that birthday matches are more common than intu-
tion suggests. The reporter then asked whether I thought that
the Monday–Wednesday lottery match was just a coincidence,
and with some qualifying phrases about reviewing the proce-

\[
p = \left( \frac{3}{2} \right) \left( \frac{1}{n} \right)^2 \left( 1 - \frac{1}{n} \right)^1 + \left( \frac{3}{2} \right) \left( \frac{1}{n} \right)^3 \left( 1 - \frac{1}{n} \right)^0
\]

\[
= \frac{3}{n^2} \left( \frac{n - 1}{n} \right) + \frac{1}{n^3}
\]

\[
= \frac{3n - 2}{n^3} \approx 9 \times 10^{-12}
\]

The reporter then asked if I would go back on camera and an-
swer his question about winning twice in three days by playing a
favorite set of numbers and respond with “9 out of a trillion.” He
asked the question on camera. I did not simply respond “9 out of
a trillion,” but rather I restated his question as part of my answer,
saying “We’re talking about one set of favorite numbers that one
person is playing over and over again. The chance of those com-
ing up two times out of three days is much, much smaller. It’s
on the order of nine in a trillion” (quote taken from news story).
The reporter then asked something like “So the probability of
someone in North Carolina winning twice in three days is as-
tronomically small?” And I said “No, there are a lot of people
playing favorite numbers each day, so that the probability of at
least one of them winning two times in three days would be
higher than 9 in a trillion.” Neither the reporter’s original ques-
tion or the latter statement about multiple players made it on
television.

On the 5:00 p.m. version of the WTVD news you could see
my hands writing out the binomial probabilities that result in the
9-in-a-trillion figure, and then on camera you see and hear me
respond with the quote in the previous paragraph. The full news
story was available on the WTVD web page http://abclocal.go.
com/wtvd/story?section=triangle&id=3475916 at the time of
writing this article. It is evident that I am not very telegenic, so
maybe the editor did me a favor by cutting much of the inter-
view filmed in my office!

2. AFTERTHOUGHTS

Whew, what an ordeal! I liked the reporter, and I had a good
time. I was frustrated at being selectively quoted, but reck-
oned that his interest is entertainment, not statistics. Coinci-
dences capture the attention of statisticians (e.g., Diaconis and
Mosteller 1989) and nonstatisticians alike. That there are a lot
more of the latter explains in part the frustration that statisticians
experience with the media’s coverage of coincidences. Most
people are not interested in the same aspects of coincidences
that statisticians are, and the media responds accordingly.

I now think (wishfully perhaps) that the reporter left my of-


ice understanding that lottery matches were more common than
he thought when he entered. Also that he realized (or advan-
tageously decided on the fly) during our meeting, that for his
story he needed an assessment of the individual lottery player’s
musing “What are the chances that my favorite numbers would
come up twice in three days?” and 9 × 10^{−12} is not too inap-
propriate for that question (although, perhaps the individual lot-
tery player should really be wondering “What are the chances
that my numbers would come up twice in three days sometime
during my adult, lottery-playing career?”). It is also possible that
the reporter just kept fishing until he got me to state very long
odds.

Both probabilities presented to the reporter, 1/191,919 and
9 × 10^{−12}, are correct answers, but correct for different ques-
tions. There is value in both. One from the “wide” perspective
of trying to understand whether latent causes are behind the
coincidence or whether something truly rare has occurred. The
other, from the “narrow” perspective of a lottery player wonder-
ing what is the chance that such a coincidence could happen to
him/her. The latter perspective is not without merit. I have never
played the NC lottery, but I have wondered what it would be like
to win—hmmm, maybe the revision of this article would never
be completed . . . . The fact that the probability is close to 1 that
someone, somewhere, sometime, will win mega-millions isn’t
very relevant during those reveries. However, that the proba-

Statisticians should point out when seemingly rare events are
not really that rare. But in doing so we should not lose sight of
the fact that for some human interest stories, a probability calcu-

lation from the “narrow perspective” is appropriate. My hunch
is that we sometimes do lose sight of the human-interest angle
because we are geared toward the “wide” perspective. My own
experiences supports this conclusion. Before my meeting with the reporter I was planning to tell him that the question that he should be interested in is “How likely is it that duplicate sets of numbers will be drawn within some reasonably narrow time frame, anytime since the start of the NC lottery?” The fact that he might be interested in a probability from the individual player’s perspective did not occur to me until he asked it. My reading of other statisticians’ comments on lottery coincidences suggests that they too sometimes emphasize the “wide” perspective at the expense of possibly overlooking the relevance of the “narrow” perspective. This is evident in a letter to the Editor of the New York Times by Samuels and McCabe (1986) refuting the odds quoted in a February 14, 1986 Times news story (McFadden 1986) on a repeat lottery winner; and in another Times article on coincidences (Kolata 1990) citing sources P. Diaconis and F. Mosteller that mentions the same repeat lottery event that inspired the letter written by Samuels and McCabe (1986). In both cases the statisticians’ point of view is that the wide sense probability is correct and the narrow sense probability is wrong. Yet the Times article focused on Mrs. Adams’ (the lucky double winner) good fortune. The article talked about the odds that she was “up against,” and also quoted her assessment of the repeat win, “Shocking—definitely shocking.” Was Mrs. Adams 30-to-1 shocked (the wide-sense probability of a repeat winner calculated by Samuels and McCabe)? Or was she closer to 17.3 trillion-to-1 shocked (the “wrong” narrow-sense probability to which the statisticians objected)? Neither, in fact. In this case there is yet another perspective that needs to be considered, the “wide-narrow” perspective—the 17.3 trillion-to-1 odds applies to just two ticket purchases, yet Mrs. Adams had been playing for several years purchasing multiple tickets per drawing, thus her chance of winning twice was better than 17.3 trillion-to-1. Hanley (1992) approximated it at between 1-in-a-million and 1-in-50 million under various simplifying assumptions. The point is, that although odds of 30-to-1 is a fair assessment of the likelihood of some repeat winners, somewhere, sometime, it is not appropriate for assessing Mrs. Adams’ sense of surprise at her good fortune.

Evidently, statisticians (including myself) think that “wide” sense lottery probabilities are more relevant than “narrow” sense ones. They are for the scientific objectives we usually deal with, but not always for the task of quantifying how “lucky” a single player is, or how “shocked” he/she is. Perhaps if we are to avoid being “grossly misunderstood or misquoted—the common fate of our profession” in the words of Samuels and McCabe (1986), we should acknowledge more forcefully that there are different questions that have different answers, and rather than passing judgment on the questions, we should point out the relevant purposes of each. If I am ever again approached by a reporter about a lottery coincidence I plan to ask him/her two questions: “How surprised would you be if someone wins the NC Cash-5 lottery next week?” and “How surprised would you be if you won the NC Cash-5 lottery next week?” These opening questions set the stage for explaining the difference, and relevance, of both the wide-sense and narrow-sense probability calculations. With a little luck I might be able to get the reporter to include both types of probabilities in his/her story thereby satisfying the media’s appetite for “astronomical” odds and the statistician’s desire to assess coincidences from a broader perspective.

3. THE CHANCES ARE . . .

We now present what Diaconis and Mosteller (1989) called a special-purpose model for analyzing the NC lottery coincidence. It captures both the matching feature and the multiple-drawings feature of the observed coincidence. The probability modeling is at the level of Ross (2005) and Casella and Berger (2002) and is suitable for using as an exercise in advanced undergraduate and graduate courses. Consider the event,

$$E(k, m) = \{ \text{at least one occurrence of at least one match within any } k \text{ consecutive days, over a contiguous } m\text{-day period} \}.$$ 

For example, with $k = 3$, and $m = 258 (= \text{number of days since the Cash-5 game started until July 11}$), $E(k, m)$ is the event that a duplicate (or triplicate) set of numbers occurred during days (1,2,3), or days (2,3,4), or days (3,4,5), or . . . , or days (256,257,258).

Let $p(k, m) = \Pr(E(k, m))$. Note that $k \leq m$ by definition. From (1) we have

$$p(k, k) = 1 - \frac{n!}{(n-k)!(n^k)}.$$ 

Define $T_j$ equal to the event that a match occurs during the $k$ consecutive days $\{j, \ldots, k+j-1\}$, for $j = 1, \ldots, m+1-k$; and define $T^*_j$ to be the event that a match occurs for the first time in the set of days $\{j, \ldots, k+j-1\}$. Then because a match occurs in one of the $k$-day periods if and only if it occurs for the first time in one of the periods, and it can occur for the first time only once,

$$p(k, m) = \Pr(E(k, m)) = \Pr(T_1 \cup \cdots \cup T_{m+1-k})$$

$$= \Pr(T^*_1 \cup \cdots \cup T^*_{m+1-k})$$

$$= \Pr(T^*_1) + \cdots + \Pr(T^*_m + 1 - k).$$

But $\Pr(T^*_j) = p(k, m)$ and for $j > 1$, $T^*_j$ occurs if and only if: a match has not occurred in the previous $j-1$ periods, having probability $1 - p(k, k+j-2)$; and the match occurs on day $k+j-1$, with probability $(k-1)/n$, because it must be that on day $k+j-1$ the set of numbers matches one of the necessarily unique sets on the previous $k-1$ days. Thus

$$p(k, m) = p(k, k) + \sum_{j=2}^{m+1-k} \frac{k-1}{n} \left[ 1 - p(k, k+j-2) \right]$$

$$= p(k, k) + \sum_{j=0}^{m-k-1} \frac{k-1}{n} \left[ 1 - p(k, k+j) \right].$$

For $m > k$, $p(k, m)$ can be found from $p(k, k+j)$ for $j = 0, \ldots, m-k-1$. Alternatively, direct calculation of the probability using counting methods results in

$$p(k, k+j) = \Pr(E(k, k+j)) = 1 - \frac{n!(n-k+1)^j}{(n-k)!n^{(k+j)}}.$$ 

$$j = 0, 1, 2, \ldots$$
which is shown, by way of an instructional exercise, to satisfy (4).

Table 2 displays probabilities \( p(k, m) \) calculated using (5) for \( m = 258 \) and select values of \( k \). The Cash-5 lottery had 258 drawings between inception and July 11. I think that having the same set of numbers appear within about a two-week period is noteworthy, and would be recognized. For \( k = 14 \) (two weeks) and \( m = 258 \) (= how long the Cash-5 game had been played as of July 11), \( \Pr(E(14, 258)) \approx 0.0057 \). So it is rare to have the same set of numbers appear at least twice within a two-week span in the first 258 days of operation, but not amazingly so.

Figure 1 displays plots of \( p(k, m) = \Pr(E(k, m)) \) for a selection of values \( k \) and \( m \). The left panel is most relevant for assessing the likelihood of lottery coincidences. The curves in this panel are \( p(k, m) \) for \( k = 3, 7, 14, \) and 31, and \( m = k + 1, \ldots, 258 \). The horizontal line is drawn at 0.01. Thus only for \( k = 31 \) (one month) does the probability of at least one match since the start of the Cash-5 game exceed 0.01 for \( m = 258 \).

The near linearity of the curves in the left panel of Figure 1 is due to the fact that when \( (k - 1)/n \) is small

\[
\left( \frac{n - k + 1}{n} \right)^{(m-k)} \approx 1 -(m-k) \left( \frac{k-1}{n} \right). 
\]

Rearranging (5), results in the approximation that is linear in \( m \)

\[
\frac{1-p(k,m)}{a_{n,k}} \approx 1 + \frac{k(k-1)}{n} - \frac{(k-1)}{n} m, \tag{6}
\]

where \( a_{n,k} = n!/(n-k)!n\).

It is customary in explanations of the Birthday Problem to show that the probability of at least one match first exceeds 0.5 at \( k = 23 \) people. The right panel of Figure 1 displays \( p(k, m) \) for large \( k \) and \( m \). The panel spans a total playing time of four years (\( m = 4 \times 365 = 1460 \) days), and shows plots of \( p(k, m) \) for \( k \) corresponding to 2, 4, \ldots, 12 months and \( m = k, \ldots, 1460 \). The horizontal line is drawn at 0.5.

Starting with (5), the equation \( p(k, m) = \alpha \) has solution \( m \), in terms of \( k \) and \( \alpha \),

\[
m_{k,\alpha} = k + \frac{\ln \left( n^k (1-\alpha) (n-k)!/n! \right)}{\ln \left( 1- (k-1)/n \right)}. \tag{7}
\]

Taking \( k = 304 \) (10 months) and \( \alpha = 0.5 \) results in \( m_{k,\alpha} = 1468.7767 \). Using (5), one finds \( p(304, 1468) = 0.4998 \) and \( p(304, 1469) = 0.5001 \). Thus, for a matching window of 10 months \((k = 304)\), the probability of a matching set of lottery numbers first exceeds 0.5 at \( m = 1469 \) drawings (\( \approx 4.02 \) years).

4. SUMMARY

The television segment on the lottery coincidence was a human interest story and not dealing with serious scientific issues. During the interview I never had the impression that the reporter was looking for evidence of fraud, incompetence, or equipment malfunction in the conduct of the lottery, although the news story touched on this angle. So, because of the entertainment focus of the story, the reporter’s quest for a sound bite quoting an “astronomically” small probability is understandable. However, my experience with the reporter was a poignant reminder of the power of selective editing. If there is a next time that I am asked to be interviewed on short notice, I will try to weave qualifying remarks into the likely sound bites, if for no other reason than to make the editing more apparent should it occur.

One complication that I avoided in my discussion with the reporter is the fact that the Cash-5 game is not the only game in town. North Carolina also runs a Pick-3 game and is part of a multistate Power Ball Lotto. Thus, if the intrepid WTVD reporter who interviewed me has been keeping his eye on the lottery system looking for potential stories since the start of the lottery in North Carolina (March–October, 2006 for various games), then a more appropriate analysis would be to calculate the probability of at least one coincidence (comparably rare as the Cash-5 matching coincidence) in at least one North Carolina lottery game since the start of the North Carolina lottery.

Finally, the television segment also featured the lottery director. He stated on camera that different machines with different sets of balls were used on Monday and Wednesday, thus all but eliminating the possibility of equipment malfunction or tampering, and supporting the opinion that I expressed to the reporter that the matching numbers was just a coincidence.

5. EPILOGUE: COINCIDENTIALLY...

Shortly after the first submission of this article yet another lottery coincidence hit the news in the form of a repeat winner, Mr. Eugene Angelo Sr., in the New York State Lottery on August 18, 2007. Another repeater! Kudos to Samuels and McCabe (1986); see also Hanley (1992). The lottery post, http://www.lotterypost.com/news/161939, contains the statement: “The chances of winning once are 22 million to 1—so the odds of doing it two times are ‘galactically astronomical,’ said New York Lottery spokesman John Charlson.” We could debate the meaning of the sentence—someone in all of New York, or maybe the entire northeast, “doing it two times,” or Mr. Eugene Angelo Sr. “doing it two times,” but instead let us simply adopt the wide-narrow perspective and calculate whether Mr. Angelo Sr. really was galactically astronomically lucky.

In the NY 6-number lottery, six numbers from 1 to 59 are selected per card. One dollar buys two cards. The game is played twice a week, Wednesday and Saturday. The news story indicated that the two-time winner was spending $42 per week on lottery tickets for the past 30 years. Assume that the $42 is evenly split between Wednesday and Saturday, so that on each day the player has 42 chances to win ($21 \times 2 \text{ cards/dollar}$). Thus the probability of winning on a single day is $p = \frac{42}{n}$ where $n = \binom{59}{6} = \frac{59!}{6!(53)!} = 45,057,474$ is the number of sets of five numbers chosen from $\{1, 2, \ldots, 59\}$. There are about 104 drawings per year, and Mr. Angelo has been playing for 30 years for a total of about 3,120 drawings. So $X =$ number of wins in 30 years for Mr. Angelo is distributed $\text{Bin}(3120, p)$. The probability that he wins at least twice is $p_2 = \Pr(X \geq 2) = 1 - (1 - p)^{3120} - 3120p(1 - p)^{3119} \approx 4.220e-006$. The diameter of Pluto's orbit is approximately 3.666e+009 kilometers and provides a lower bound on the diameter of the solar system. Thus the odds of Mr. Angelo's two wins, $(1 - p_2)/p_2 \approx 2.370e+005$, are not even "solar-systemly astronomical," never mind "galactically astronomical." Alternatively, the ratio of $p_2$ to the probability of winning with a single $1$ investment is $p_2/(2/n) \approx 95$. Thus a person with Mr. Angelo's playing habits is 95 times more likely to win twice, than a person playing once ($$2$ for two numbers) is likely to win on that one occasion.

[Received September 2007. Revised December 2007.]

REFERENCES