A Random Mapping Statistics and a Related Identity

Stanislav Sykora, Extra Byte, www.ebyte.it
First published in June 2014

Investigating elementary properties of random selections from a finite, discrete set, an identity popped up which can be cast both as a finite probability distribution and as a decomposition of n^n. They are reported here, together with a few notes on the original object of this study.

Keywords: math, random mapping, discrete distribution, identity, pigeonhole principle

Introduction

This tiny essay started as a reflection on how soon/late should/should not a random numbers generator repeat a value[1]. This is a classical problem in the art of random number generators [1-4], but here we will pursue a point of view which is a bit off-stream.

Terminology: Let us start with a random mapping $R(i) \rightarrow S$ of natural numbers i into a set S of n elements $\{s_1, s_2, \ldots, s_n\}$. The result of an application R(i), henceforth called a "shot", is one of the elements of S, with all of them having the same probability to come up. We will say that the selected element was "hit" and we will be interested in how many times each element was hit.

Notes: The only purpose of the argument i of R(i) is to label the i-*th* shot. Upon every shot an element of S gets certainly hit[2], but there is no way to know which one, and the outcomes of consecutive shots are independent of each other. Hitting an element does not remove it from the set S. We actually implicitly assume that each element carries a "counter" telling how many times it was hit.

There are many approximate practical models of this type of processes, such as:

- Running a discrete random numbers generator (well, all RND's are discrete).
- Bombing an area divided into a discrete number of squares.
- Bombardment of a celestial body by meteorites (divide the body surface into finite elements).
- Winning a raffle round [5] in a group of habitual players (each game representing one shot).

Since the *probability* of hitting any particular element is known and equal to 1/n, the most obvious queries that arise in this context are of this kind[3]:

1. How many hits does it take for the probability of hitting a particular element to become 1/2?
2. How many hits does it take, on the average, for half of the elements to get hit?
3. How many hits it takes before the probability of having hit all elements reaches a certain level.
4. How many hits is it likely to take before a *particular* element will be hit twice?
5. How many hits is it likely to take before *some* element will be hit twice?

Most of these are easy questions leading to the binomial statistics [6] with probability $p = 1/n$. After K shots there will be elements of S that were hit j times, with $j = 0, 1, 2, \ldots, K$. The probabilities a(K,j) of each such outcome are of course the terms in the expansion of $(p+(1-p))^K$:

$$(p+(1-p))^K = \sum_{j=0,K} a(K,j) = 1, \text{ where } a(K,j) = ((n-1)^j/n^K)C(K,j), \qquad (1)$$

C(k,l) being the usual binomial coefficient [7].

[1] Which does not mean that the generator will start repeating the whole sequence, unless it is really primitive.
[2] In processes where "shots" may fail to hit any element, one can resort to the derived conditional variety whereby a shot is counted "valid" only after it has resulted in a hit.
[3] Both lists are illustrative, but not at all exhaustive.

Stanislav Sykora, *A Random Mapping Statistics and a Related Identity*, DOI: 10.3247/SL5Math14.003

Note that in this context K is unlimited. Even if it were incomparably larger than n, there might still be elements which did not get hit at all (lucky survivors) and, vice versa, it might potentially happen that every shot would hit the same element (a bad-luck attractor). All this is interesting but conventional and we will not pursue it further in this Note. Let me only point out, for reference purposes, that when putting K = n in (1), one obtains the following trivial decomposition of n^n:

$$\Sigma_{j=0,n}\,(n-1)^j\,C(n,j) = n^n. \qquad (2)$$

What will interest us more are identities that come up when statistical considerations are combined with the pigeonhole principle [8,9].

A reflection on multiple hits

If, after i shots, no element was hit twice, what is the probability that the next, (i+1)-*st*, shot will also hit an element which was not yet hit. Since there are exactly (n-i) such unscathed elements, the answer is (n-i)/n. The statement that *prior to the k-th shot, no element was hit twice* is equivalent to saying that no element was hit twice in any of the shots up to *k-th*. The probability of this happening is therefore

$$\Pi_{i=0,k-1}(n-i)/n = (1/n^k)[n!/(n-k)!] = (k!/n^k)C(n,k), \qquad (3)$$

Note that the shot index i starts formally at 0 (which is inconsequential since the 0-*th* factor is 1) and stops at (k-1), just prior to k.

Complementing (1), one obtains the *cumulative distribution function* [10] for the event of hitting some element more than once in the k shots:

$$D(n,k) = 1 - (k!/n^k)C(n,k). \qquad (4)$$

The k in this formula ranges from k=0 with D(n,0) = 0 (no shot, no hit), through 1 with D(n,1)=1 (one shot cannot produce multiple hits), up to k=n with $D(n,n) = 1 - n!/n^n$.

The validity of D(n,k) can be extended up to k=n+1 by setting D(n,n+1)=1. The rationale behind this is provided by the **pigeonhole principle** [8]: if the n elements of S get hit a total of (n+1) times, at least one element *must* be hit more than once, so that the probability of that event must be 1.

The corresponding *probability mass function* [11] that some element will be hit twice in the shot number (j+1), but not before, is then, for j = 1 through n,

$$P(n,j) = D(n,j+1) - D(n,j) = (j/n)(j!/n^j)C(n,j), \qquad (5)$$

Equations (4) and (5) define a discrete probability distribution on the finite set of $j \in \{1,2,...,n\}$ shots which looks rather novel, since I could not find a category [12] where to put it.

However, it is essential that it be understood properly and this is best done using as model the random number generator (RNG) which selects randomly one element out of n. Let us run the RNG until it selects an element which was already selected before, an event which will certainly happen in at most (n+1) shots. Let's say it happened in shot (j+1), i.e., the last shot which did not lead to any repetition was the j-th. Let us consider this a single "run", the result of which is the number j, $1 \le j \le n$. Then the probability density (5) describes the distribution of the values of j *when one repeats whole runs* for an unlimited number of times.

It should be now understandable why, for n>2, P(n,j) has a maximum. When n is large, the runs when the first double-hit occurred very early (say after just two shots) are bound to be very rare. Likewise, runs

Stanislav Sykora, *A Random Mapping Statistics and a Related Identity*, DOI: 10.3247/SL5Math14.003

where after almost n shots there was still no double hit will also be extremely rare (it is unlikely that almost every element would get hit, but none more than once).

The location of the maximum is best estimated by writing P(n,j) in a recursive form, P(n,j+1) = f.P(n,j), where f = [(j+1)/j][(n-j)/n]. Clearly, f is a decreasing factor starting at 2(n-1)/n which, for n>2, is greater than 1, and decreasing towards 0. The maximum of P(n,j) occurs just before f crosses 1. Setting f = 1 leads to a quadratic equation for j whose solution is

$$j_{max} = \left[\sqrt{4n+1} - 1\right]/2. \qquad (6)$$

The value of j_{max} equals approximately $\sqrt{n}$, with a discrepancy smaller than 1 for any $n > 2$.

One of the consequences of this is that, for example, a "correctly random" RNG should lead to a duplicate hit in about $\sqrt{n}$ shots (which turns out to be quite early). Trying to achieve as long "cycles" as possible is unnatural and causes deviations from randomness.

An interesting identity

The distribution (5) certainly merits to be studied in more detail but that will take another essay. For now, let us just notice a direct corollary which looks quite interesting on its own:

Thanks to the pigeonhole principle, P(n,j) is a true probability density on a discrete set of n elements (shots, not those of S), and the sum of all P(n,j), for j = 1 to n, must be 1. Hence

$$\Sigma_{j=1,n}\,(j/n)(j!/n^j)C(n,j) = 1. \qquad (7)$$

Using simple manipulations, this becomes another integer decomposition of n^n,

$$\Sigma_{j=0,n}\,(n^{n-j-1})(j)(j!)C(n,j) = n^n. \qquad (8)$$

Alternatively, it can be also written as an intriguing decomposition of unity:

$$\sum_{j=1}^{n} \frac{j}{n} \prod_{k=1}^{j-1}\left(1-\frac{k}{n}\right) \equiv \sum_{j=1}^{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{j-1}{n}\right)\frac{j}{n} = 1 \qquad (9)$$

Let us compare the n^n expansions (2) and (8). Both have the form

$$\Sigma_{j=0,n}\,w(n,j)C(n,j) = n^n, \qquad (10)$$

but the 'weight' coefficients w(n,j) are quite different:

$$\text{in (2)}, \qquad w(n,j) = (n-1)^j = b(n,j) \qquad (11)$$
$$\text{in (8)}, \qquad w(n,j) = (n^{n-j-1})(j)(j!) = c(n,j) \qquad (12)$$

A numeric example of both for n= 7, j = 0,1,2,...,7,

b(7,j) = {1, 6, 36, 216, 1296, 7776, 46656, 279936},
c(7,j) = {0, 16807, 9604, 6174, 4704, 4200, 4320, 5040},

reveals, apart from the first element, a much more 'balanced' distribution of the values for c(n,j) with even a flat minimum in the upper half. A consequence of this is that the values w(n,j)C(n,j) are in identity (8) much more peaked around the central values of j, rather than close to the end as in identity (2):

b(7,j)C(7,j) = {1, 42, 756, 7560, 45360, 163296, 326592, 279936},
c(7,j)C(7,j) = {0, 117649, 201684, 216090, 164640, 88200, 30240, 5040}.

Stanislav Sykora, *A Random Mapping Statistics and a Related Identity*, DOI: 10.3247/SL5Math14.003

Conclusion

Analyzing the process of random statistical selection (or 'shooting') of elements in a finite set typical, for example, of random number generators, we have derived a statistics on the shot number when the first multiple hit occurs. This appears to be a novel type of a discrete distribution function.

It also leads to an interesting identity which can be cast in two ways: either as an integer decomposition of the number n^n in terms of weighed binomial coefficients C(n,j), or as a rational decomposition of unity. The triangular matrices of the n^n decomposition (both the integer 'weights' and the full terms) were registered[4] in OEIS [13] (numbers A243202 and A243203).

The analysis also leads to an insight regarding discrete random number generators on sets with n elements. It turns out that trying forcibly to make an RNG exhibit a "cycle" much longer than about $\sqrt{n}$ is statistically unnatural. Rather, a single-value repetition after about $\sqrt{n}$ steps should be allowed, but avoiding, obviously, a complete repetition of the subsequent sequence.

References and links

[1] Knuth D.E., *Seminumerical Algorithms*, 3rd edition, Vol.2. of *The Art of Computer Programming*, Addison-Wesley 1997, Chapter 3. ISBN 978-0201896848.

[2] Bratley P., Fox B.L., Schrage E.L., *A Guide to Simulation*, Softcover reprint of 1983 edition. Springer-Verlag 2012. ISBN 978-1468401691.

[3] Goldreich O., *Modern Cryptography, Probabilistic Proofs and Pseudorandomness*, Springer 1998. ISBN 978-3540647669. Available online.

[4] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., *Numerical Recipes in C. The Art of Scientific Computing*, Cambridge University Press 1992, Section 7. ISBN 978-0521431088. Partially available online.

[5] Wikipedia, Raffle.

[6] Wikipedia, Binomial distribution.

[7] Wikipedia, Binomial coefficient.

[8] Wikipedia, Pigeonhole principle.

[9] Grimaldi R.P., *Discrete and Combinatorial Mathematics: An Applied Introduction*, 5th edition, Pearson 2003. ISBN 978-0201726343.

[10] Wikipedia, Cumulative distribution function.9

[11] Wikipedia, Probability mass function.

[12] Wikipedia, List of probability distributions.

[13] OEIS Foundation Inc. (2011), The On-Line Encyclopedia of Integer Sequences.

History of this document

1 June 2014: Assigned a DOI (10.3247/SL5Math14.003) and uploaded online.

[4] While registering the sequence of the full terms, it turned out that it is not quite novel; it arises also in a different context and was already registered in A066324, but with the starting zero terms omitted. Conceptually, considering the summations over the binomial coefficients C(n,j) I believe that maintaining the j=0 terms is preferable, even if they are identically zero. At least in the current context (no double-hit upon the first shot).