
Sequential Searches: Proofreading, Russian Roulette, and the Incomplete q -Eulerian Polynomials Revisited

Don Rawlings

1. INTRODUCTION. So, you’ve read your fifty page manuscript a half dozen times and you continue to stumble across annoying typos. Is there any end in sight? Just how many more times do you need to read it before you’re satisfied?

For purposes of discussion, let’s say that your manuscript originally contained 10 errors and that, because of fatigue and a certain level of impatience, the probability that you notice an error when staring right at it is only $1/10$. Then the expected number of errors found is one after the first reading, 1.9 after the second reading, and so on. Several such results rounded to the nearest hundreth are shown in Table 1.

TABLE 1.

# of readings	1	2	5	15	25	29	30
expected # of errors found	1.00	1.90	4.10	7.94	9.28	9.53	9.58

Although easily computed and certainly informative, Table 1 is not entirely satisfying. If you decide that 90 percent of the errors must be found and corrected, then it suggests that fewer than 25 readings will do. To estimate how much less, one could of course just extend the table.

What’s really called for, though, is the inverse expectation, that is, the expected number of searches needed to find a specified percentage of the errors. Under the assumptions of the preceding paragraph, formula (13) in Section 4 generates the following table.

TABLE 2.

# of errors to be found	1	2	5	9	10
expected # of readings needed	1.54	2.51	6.63	18.81	28.30

So, if finding 90 percent of the errors is desired, then 18.81 readings is recommended. Should you demand perfection, then brace yourself for 28.3 readings!

The act of proofreading a manuscript is but one natural context in which a sequence of searches is conducted for lost objects. One can easily imagine other sequential search scenarios: Ranging from the commonplace to the exhilarating to the risky, they include easter egg hunts, treasure hunts, the clearing of dangerously littered live munitions from a region following either a war or wargames, and even Russian roulette with several players.

As in proofreading, the same fundamental question arises in all such scenarios: What is the expected number of searches needed to find all, or some acceptable percentage, of the lost objects? Time and resources are only finite after all!

Two technical matters must first be settled. In most situations, both the set of lost objects and the probability of finding a given object on a particular search must be estimated. The methods needed to resolve these issues will of course depend on the context. In the proofreading scenario, a typist's error-per-page average could be estimated through sampling. For treasure hunts, historical records and legend will be of importance. To approximate the find probabilities, the methods and resources available for the search as well as the various properties of the objects, such as size and general location (thick rain forest, high seas, etc.), should all be taken into account.

Leaving these issues to the applied statistician, let's assume that such estimates have been made. Denote the lost objects by $\theta_1, \theta_2, \dots, \theta_n$, and let q_{ij} signify the probability of θ_j remaining lost under the conditions and strategies of the i^{th} search. Also, q_{ij} is assumed to be independent of $q_{i'j'}$ for $(i, j) \neq (i', j')$. So, if l_{sj} denotes the probability of θ_j remaining lost during the initial s searches, then $l_{sj} = q_{1j}q_{2j} \cdots q_{sj}$.

The relevant distributions associated with a sequential search are:

The sequential search distribution. For a fixed nonnegative integer s , the probability that exactly k objects are found in the initial s searches is denoted by $M_{n,s}(k)$.

The inverse sequential search distribution. For a fixed integer k from 1 to n , the probability that s searches are needed to find k objects is denoted by $P_{n,k}(s)$.

In the latter, each search is to be conducted in an order consistent with the labeling of the objects (if not previously found, θ_1 is sought first, then θ_2, \dots) and the final search is terminated with the k^{th} find. Such an order of search is inherent in the proofreading of a manuscript: Each reading constitutes a search for errors, typically proceeding from the first page to the last. In the case $k = n$, order is actually irrelevant.

The inverse sequential search distribution has some amusing sidelights. Herbranson and Rawlings determined $P_{n,k}(s)$ under the assumption that each find probability depends on the previous number of finds, which is reasonable if either resources are depleted or knowledge is gained in making finds [6]. Comparison of the results in this article with those in [6] leads to a completely probabilistic proof of an identity for the q -Eulerian polynomials studied by Carlitz [3] and to the discovery of a new formula for the incomplete q -Eulerian polynomials introduced in [6]. Also, a certain issue of machismo resolved in [6] may be extended to Sandell's variation of n -player Russian roulette [10].

2. THE SEQUENTIAL SEARCH DISTRIBUTION. The assumption of independence makes the formula for $M_{n,s}(k)$ transparent. By identifying a find with success, the process associated with the sequential search distribution may be viewed as little more than Bernoulli trials with variable probabilities of success: For $1 \leq j \leq n$, the probability of θ_j being found during the initial s searches is $1 - q_{1j}q_{2j} \cdots q_{sj}$. Thus,

Theorem 1. *The probability of finding k of n lost objects in the initial s searches is*

$$M_{n,s}(k) = \sum_J \prod_{j \in J} (1 - l_{sj}) \prod_{j \in \{1,2,\dots,n\} \setminus J} l_{sj}, \tag{1}$$

where the sum is over all $J \subseteq \{1, 2, \dots, n\}$ of cardinality k and $l_{sj} = q_{1j}q_{2j} \cdots q_{sj}$. The expected number of finds made in the initial s searches is the sum of the expectations of finding the individual objects:

$$\sum_{k=0}^n k M_{n,s}(k) = \sum_{j=1}^n (1 - l_{sj}) = n - \sum_{j=1}^n l_{sj}. \tag{2}$$

When $q_{ij} = q < 1$ for all i, j , (1) and (2) respectively reduce to

$$M_{n,s}(k) = \binom{n}{k} (1 - q^s)^k q^{s(n-k)} \quad \text{and} \quad \sum_{k=0}^n k M_{n,s}(k) = n(1 - q^s). \quad (3)$$

The assumption $q_{ij} = q$ for all i, j is reasonable if the method of search is fixed and if the objects are identical and lost under similar circumstances. The distribution $M_{n,s}(k)$ in this case is exposed by (3) as the binomial distribution with success probability $1 - q^s$. The second part of (3) may of course be used to generate the entries in the second row of Table 1.

3. THE INVERSE SEQUENTIAL SEARCH DISTRIBUTION: $k = n$. In some cases, such as the recent one at a Los Alamos laboratory involving missing computer disks containing sensitive nuclear information, all lost objects must be found. To get at the associated distribution, we need a few preliminaries. First,

$$\prod_{j=1}^n (1 + tx_j) = \sum_{k=0}^n t^k e_k(x_1, x_2, \dots, x_n) \quad (4)$$

where

$$e_k(x_1, x_2, \dots, x_n) = \sum_{J \subseteq \{1, 2, \dots, n\}, |J|=k} \prod_{j \in J} x_j$$

is the *elementary symmetric polynomial* of degree k in the indeterminates x_1, x_2, \dots, x_n . For $n = 3$, we have $e_0 = 1$, $e_1 = x_1 + x_2 + x_3$, $e_2 = x_1x_2 + x_1x_3 + x_2x_3$, and $e_3 = x_1x_2x_3$. We also need

Lemma 1. *If $\{a_k\}$ is a monotonically decreasing sequence of nonnegative real numbers and $\sum a_k$ converges, then $\lim_{m \rightarrow \infty} ma_m = 0$.*

The inverse sequential distribution for $k = n$ is deduced from Theorem 1 as follows. Formula (8) was established in [6].

Corollary 1. *The probability that s searches are needed to find all n lost objects is*

$$P_{n,n}(s) = \prod_{j=1}^n (1 - l_{sj}) - \prod_{j=1}^n (1 - l_{(s-1)j}) \quad (5)$$

where $l_{sj} = q_{1j}q_{2j} \cdots q_{sj}$. If

$$\sum_{s \geq 1} l_{sj} \text{ converges for all } j \in \{1, 2, \dots, n\}, \quad (6)$$

then the expected number of searches needed to find all n objects is finite and given by

$$\sum_{s \geq 1} s P_{n,n}(s) = 1 + \sum_{k=1}^n (-1)^{k-1} \sum_{s \geq 1} e_k(l_{s1}, l_{s2}, \dots, l_{sn}). \quad (7)$$

When $q_{ij} = q < 1$ for all i, j , (5) and (7) imply $P_{n,n}(s) = (1 - q^s)^n - (1 - q^{s-1})^n$ and

$$\sum_{s \geq 1} s P_{n,n}(s) = 1 + \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} \frac{q^k}{1 - q^k}. \quad (8)$$

Proof. The probability that s searches are required to find the entire set of lost objects is equal to the probability that n objects are found in s searches minus the probability that n objects are found in $s - 1$ searches:

$$P_{n,n}(s) = M_{n,s}(n) - M_{n,s-1}(n).$$

Thus, (1) implies (5).

Now suppose (6) holds. To get (7), first note that (5) and (4) together lead to the fact that

$$\sum_{s \geq 1} s P_{n,n}(s) = \sum_{s \geq 1} \sum_{k=1}^n (-1)^{k-1} \sum_{|J|=k} s \left(\prod_{j \in J} l_{(s-1)j} - \prod_{j \in J} l_{sj} \right). \quad (9)$$

For a fixed $J \subseteq \{1, 2, \dots, n\}$, consider the partial sum

$$\sum_{s=1}^m s \left(\prod_{j \in J} l_{(s-1)j} - \prod_{j \in J} l_{sj} \right) = 1 - m \prod_{j \in J} l_{mj} + \sum_{s=1}^{m-1} \prod_{j \in J} l_{sj}.$$

From (6) and the comparison test, we may conclude that all sums of the form

$$\sum_{s \geq 1} \prod_{j \in J} l_{sj}$$

converge. By invoking Lemma 1, we therefore have

$$\sum_{s \geq 1} s \left(\prod_{j \in J} l_{(s-1)j} - \prod_{j \in J} l_{sj} \right) = 1 + \sum_{s \geq 1} \prod_{j \in J} l_{sj}. \quad (10)$$

Then (10) and (9) imply (7). Finally, as $e_k(q^s, q^s, \dots, q^s) = \binom{n}{k} q^{ks}$, (8) follows from (7). ■

Besides (8), there is another notable special case for which (7) is reasonably tractable: For $1 \leq j \leq n$, suppose that (i) finding θ_j is search independent so that q_{ij} may be replaced by q_j and that (ii) $q_j < 1$, which guarantees that (6) holds. Under these assumptions, (7) gives the expected number of searches needed to find all of $n = 3$ objects as

$$\begin{aligned} \sum_{s \geq 1} s P_{3,3}(s) &= 1 + \sum_{k=1}^3 (-1)^{k-1} \sum_{s \geq 1} e_k(q_1^s, q_2^s, q_3^s) \\ &= 1 + \sum_{s \geq 1} (q_1^s + q_2^s + q_3^s - q_1^s q_2^s - q_1^s q_3^s - q_2^s q_3^s + q_1^s q_2^s q_3^s) \quad (11) \\ &= 1 + \sum_{j=1}^3 \frac{q_j}{1 - q_j} + \sum_{1 \leq j < k \leq 3} \frac{q_j q_k}{1 - q_j q_k} + \frac{q_1 q_2 q_3}{1 - q_1 q_2 q_3}. \end{aligned}$$

Incidentally, if $q_j = 1$ for some j , then the probability of θ_j remaining lost for s searches is $l_{sj} = 1$, $\sum_{s \geq 1} l_{sj}$ diverges, and the expected number of searches needed to find θ_j (and therefore to find all n objects) is infinite.

4. THE INVERSE SEQUENTIAL SEARCH DISTRIBUTION: $k \leq n$. There are certainly situations in which recovery of only some of the lost set may be reasonable. The discussion here is limited to the case when the find probability is constant throughout the process.

Corollary 2. *If $q_{ij} = q < 1$ for all i, j , then the probability that s searches are needed to find k of n objects, $1 \leq k \leq n$, is*

$$P_{n,k}(s) = (1-q)^k q^{(s-1)n} \sum_{i=0}^{k-1} \sum_{j=0}^{n-k} \binom{n}{i} \binom{k-i-1+j}{j} \left(\frac{q^{1-s}-1}{1-q} \right)^i q^j. \quad (12)$$

The expected number of searches needed to find k of n objects is

$$\sum_{s \geq 1} s P_{n,k}(s) = \sum_{i=0}^{k-1} \sum_{j=0}^{n-k} \sum_{m=0}^i \binom{n}{i} \binom{k-i-1+j}{j} \binom{i}{m} \frac{(-1)^m (1-q)^{k-i} q^j}{(1-q^{n-i+m})^2}. \quad (13)$$

Proof. If s searches are required to make k finds, then the first $s-1$ searches must result in i finds for some i from 0 to $k-1$, leaving the s^{th} search to account for $k-i$ finds made from the remaining $n-i$ lost objects. Hence,

$$P_{n,k}(s) = \sum_{i=0}^{k-1} M_{n,s-1}(i) P_{n-i,k-i}(1). \quad (14)$$

To compute $P_{n,k}(1)$, consider the event that one search is needed to find k of n objects. As the probability of θ_{k+j} being the k^{th} find is $\binom{k-1+j}{j} (1-q)^k q^j$, we have

$$P_{n,k}(1) = (1-q)^k \sum_{j=0}^{n-k} \binom{k-1+j}{j} q^j. \quad (15)$$

Then (14), (15), and (3) imply (12).

To obtain (13), the probability generating function for $P_{n,k}(s)$ is useful. From the calculation

$$\begin{aligned} \sum_{s \geq 1} (q^{1-s} - 1)^i q^{(s-1)n} z^s &= \sum_{m=0}^i (-1)^m \binom{i}{m} q^{i-m-n} \sum_{s \geq 1} (q^{n+m-i} z)^s \\ &= \sum_{m=0}^i (-1)^m \binom{i}{m} \frac{z}{1 - q^{n-i+m} z} \end{aligned}$$

and (12), we readily obtain

$$\sum_{s \geq 1} P_{n,k}(s) z^s = \sum_{i=0}^{k-1} \sum_{j=0}^{n-k} \sum_{m=0}^i \binom{n}{i} \binom{k-i-1+j}{j} \binom{i}{m} \frac{(-1)^m (1-q)^{k-i} q^j z}{1 - q^{n-i+m} z}. \quad (16)$$

The derivative of (16) evaluated at $z = 1$ gives (13). ■

5. THE INCOMPLETE q -EULERIAN POLYNOMIALS. Herbranson and Rawlings used the incomplete q -Eulerian polynomials as a combinatorial means of computing probabilities [6]. The tables can now be turned. In particular, (16) leads directly to a new identity for the incomplete q -Eulerian polynomials.

For $1 \leq k \leq n$, let $\mathcal{I}_{n,k}$ denote the set of injections from $\{1, 2, \dots, k\}$ to the set $\{1, 2, \dots, n\}$. We express an injection $f \in \mathcal{I}_{n,k}$ as the list $f(1)f(2)\dots f(k)$ of its range values. The *descent set*, *descent number*, and *comajor index* of $f \in \mathcal{I}_{n,k}$ are, respectively, defined by

$$\text{Des } f = \{j : 1 \leq j < k, f(j) > f(j+1)\}, \quad \text{des } f = |\text{Des } f|, \quad \text{and} \\ \text{cmj } f = |\{1, 2, \dots, f(k)\} \setminus \{f(1), f(2), \dots, f(k)\}| + \sum_{j \in \text{Des } f} (n - j).$$

For $f = 27436 \in \mathcal{I}_{8,5}$, note that $\text{Des } f = \{2, 3\}$, $\text{des } f = 2$, and $\text{cmj } f = |\{1, 5\}| + (8 - 2) + (8 - 3) = 13$. For $k = n$, the set $\mathcal{I}_{n,n}$ of course coincides with the set of permutations on $\{1, 2, \dots, n\}$ and the comajor index is a close relative of the major index, initially known as the greater index, first considered by MacMahon [7, Vol. 1, p. 135].

As defined in [6], the $(n, k)^{\text{th}}$ incomplete q -Eulerian polynomial is

$$E_{n,k}(z) = \sum_{f \in \mathcal{I}_{n,k}} q^{\text{cmj } f} z^{\text{des } f}.$$

For example, $E_{4,3}(z) = (1 + 3q) + q^2(5 + 6q + 5q^2)z + q^5(3 + q)z^2$. The function $E_{n,n}(z)$ is the q -Eulerian polynomial considered by Carlitz [3]. A specialization of Theorem 1 in [6] gives

Theorem 2. (Herbranson and Rawlings) *If $q_{ij} = q < 1$ for all i, j , then the probability generating function for the inverse sequential search distribution is*

$$\sum_{s \geq 1} P_{n,k}(s)z^s = \frac{z(1 - q)^k E_{n,k}(z)}{(zq^{n-k+1}; q)_k} \quad (17)$$

where $(z; q)_m = (1 - z)(1 - zq) \cdots (1 - zq^{m-1})$ is the q -shifted factorial.

The following are now consequences of probability. Equivalent to identity (1) in [3], Corollary 3 is a special case of a theorem due to MacMahon [7, Vol. 2, p. 211]. Corollary 4 is immediately implied by (16) and (17).

Corollary 3. (MacMahon) *The q -Eulerian polynomials are generated by*

$$E_{n,n}(z) = (z; q)_{n+1} \sum_{m \geq 0} [m + 1]^n z^m$$

where $[j] = (1 - q^j)/(1 - q)$ denotes the q -analog of j .

Proof. By Corollary 1, $P_{n,n}(s) = (1 - q^s)^n - (1 - q^{s-1})^n$ when $q_{ij} = q < 1$. It follows that

$$\sum_{s \geq 1} P_{n,n}(s)z^s = z(1 - z) \sum_{m \geq 0} (1 - q^{m+1})^n z^m. \quad (18)$$

Then (18) and (17) with $k = n$ complete the proof. ■

Corollary 4. *The incomplete q -Eulerian polynomials are generated by*

$$E_{n,k}(z) = (zq^{n-k+1}; q)_k \sum_{i=0}^{k-1} \sum_{j=0}^{n-k} \sum_{m=0}^i (-1)^m \binom{n}{i} \binom{k-i-1+j}{j} \binom{i}{m} \frac{(1-q)^{-i} q^j}{1-zq^{n-i+m}}.$$

6. RUSSIAN ROULETTE AND A QUESTION OF MACHISMO. The rules for n -player Russian roulette are as follows. Players $\theta_1, \theta_2, \dots, \theta_n$ are seated around a table. Beginning with player θ_1 and proceeding in order, a partially loaded revolver is passed from hand to hand. Upon receiving the gun, a player spins its chamber, points it to his head, and pulls the trigger. Needless to say, any player who receives a head wound is removed from the game. Play terminates when a single player (the survivor) remains.

Several interesting questions pertaining to n -player Russian roulette were considered and resolved in [2], [5], [6], and [10]. We now revisit and extend a certain issue of machismo addressed in [6].

Prospective players should ask whether or not they have the guts and foolhardiness required of Russian roulette. To help decide, the number \mathcal{E}_n of times the survivor expects to pull the trigger is key. Herbranson and Rawlings [6] computed \mathcal{E}_n for the case when the gun is reloaded so that its discharge probability remains constant and the case when the gun, initially containing at least $n-1$ bullets, is not reloaded. Corollary 1 allows the computation of \mathcal{E}_n for a further variation introduced by Sandell [10].

Concerned that Russian roulette as just described is unfair (particularly to θ_1), Sandell assigns each player a different revolver. Assuming θ_j 's gun fires with probability $(1 - q_j)$, Sandell shows that θ_j wins with probability

$$R_{nj} = (1 - q_j) \sum_{k \geq 1} q_j^k \prod_{m=1}^{j-1} (1 - q_m^{k+1}) \prod_{m=j+1}^n (1 - q_m^k). \quad (19)$$

He then specifies q_1, q_2, \dots, q_n so that Russian roulette becomes fair in that each player enjoys the same probability of surviving ($R_{n1} = \dots = R_{nn}$).

To determine the machismo factor for Sandell's variation, we frame n -player Russian roulette as an inverse sequential search process (with $k = n-1$) by viewing each sweep through the playing order as a search. Each discharge of the gun is to be interpreted as a bullet finding a lost soul.

The computation of \mathcal{E}_n in [6] readily extends to Sandell's variation with $q_1, q_2, \dots, q_n < 1$, as follows. First, suppose θ_j plays Russian roulette alone and to the death. By (7), the expected number $\mathcal{E}(\theta_j)$ of times θ_j pulls the trigger in this suicidal game of solitaire is

$$\mathcal{E}(\theta_j) = 1 + \sum_{s \geq 1} e_1(l_{sj}) = 1 + \sum_{s \geq 1} q_j^s = \frac{1}{1 - q_j}.$$

Next note that \mathcal{E}_n is equal to the expected number of times the winner pulls the trigger in a game of Russian roulette played until all players have head wounds minus the expected number of times the winner pulls the trigger in a game of suicidal solitaire. To illustrate the case $n = 3$, our analysis and (11) give

$$\mathcal{E}_3 = 1 + \sum_{j=1}^3 \frac{q_j}{1 - q_j} + \sum_{1 \leq j < k \leq 3} \frac{q_j q_k}{1 - q_j q_k} + \frac{q_1 q_2 q_3}{1 - q_1 q_2 q_3} - \sum_{j=1}^3 R_{3j} \mathcal{E}(\theta_j).$$

The case $q_j = q$ for all j of (19) is due to Blom, Englund, and Sandell [2]. An alternate formula was deduced in [9], namely,

$$R_{nj} = \frac{(1-q)^n}{(q;q)_n} \sum_{\sigma \in S_{nj}} q^{\text{cmj } \sigma},$$

where S_{nj} denotes the set of permutations σ of $\{1, 2, \dots, n\}$ satisfying $\sigma(n) = j$. Equating the two expressions gives the curious identity

$$\sum_{\sigma \in S_{nj}} q^{\text{cmj } \sigma} = \frac{(q;q)_n}{(1-q)^{n-1}} \sum_{k \geq 0} q^k (1-q^{k+1})^{j-1} (1-q^k)^{n-j}.$$

To place this in perspective, a special case of a result due to MacMahon [7, Vol. 2, p. 189] is equivalent to

$$\sum_{\sigma \in S_n} q^{\text{cmj } \sigma} = \frac{(q;q)_n}{(1-q)^n}, \quad (20)$$

where S_n is the set of all permutations of $\{1, 2, \dots, n\}$. Moritz and Williams [8] re-discovered (20) while studying a process that in essence is n -player Russian roulette played until no survivor remains.

AN ACKNOWLEDGEMENT AND CLOSING REMARKS. Donald Knuth deserves special mention. His entertaining and enlightening letters were inspirational. Besides several valuable suggestions, he also deduced (8) and (11) independently using an inclusion-exclusion approach.

Other sequential search schemes have been considered by Benkerouf and Bather [1] and Dunkl [4]. In fact, Dunkl studied a variation of proofreading wherein the reader returns to the beginning of the manuscript each time an error is discovered (so that a maximum of one find is allowed per search).

REFERENCES

1. L. Benkerouf and J. A. Bather, Oil exploration: sequential decisions in the face of uncertainty, *J. Appl. Probab.* **25** (1988) 529–543.
2. G. Blom, J. E. Englund, and D. Sandell, General Russian roulette, *Math. Mag.* **69** (1996) 293–297.
3. L. Carlitz, A combinatorial property of q -Eulerian numbers, *Amer. Math. Monthly* **82** (1975) 51–54.
4. C. F. Dunkl, The absorption distribution and the q -binomial theorem, *Comm. Statist. Theory Methods* **10** (1981) 1915–1920.
5. D. E. Knuth, *The Art of Computer Programming*, Vol. 3, 2nd ed., Addison-Wesley, Reading, MA, 1997.
6. T. Herbranson and D. P. Rawlings, A sequential search distribution: Proofreading, Russian roulette, and the incomplete q -Eulerian polynomials, *Discrete Math. Theor. Comput. Sci.* to appear.
7. P. A. MacMahon, *Combinatory Analysis*, Vols. 1, 2, Cambridge University Press, London/New York, 1915; reprinted by Chelsea Publishing Co., New York, 1960.
8. R. H. Moritz and R. C. Williams, A coin-tossing problem and some related combinatorics, *Math. Mag.* **61** (1988) 24–29.
9. D. P. Rawlings, Absorption processes: Models for q -identities, *Adv. in Appl. Math.* **18** (1997) 133–148.
10. D. Sandell, Fair Russian roulette, *Math. Sci.* **22** (1997) 52–57.

DON RAWLINGS received his Ph.D. from University of California at San Diego under the direction of Adriano Garsia and Dominique Foata. He has incredible difficulty in spotting his own typos and absolutely hates to look for lost keys. To relax, Don enjoys playing his guitar and kicking up his heels doing a form of tap dance known as clogging.

California Polytechnic State University, San Luis Obispo, CA 93407
 drawling@math.calpoly.edu