[1]This research was supported in part by the U.S. Army Research Offic, Durham, under Contract No. DAHCO4-71-0042.

A RETURN TO REPETITIONS

N.L. Johnson

Department of Statistics
University of North Carolina at Chapel Hill

Institute of Statistics Memeo Sereis No. 871

June, 1973

A RETURN TO REPETITIONS

by

N.L. Johnson
University of North Carolina at Chapel Hill

1. Introduction

In [1] it was shown that in a series of independent trials with m mutually exclusive possible outcomes $E_1, E_2, \ldots, E_m$, and $\Pr[E_j] = p_j$ $(j=1,\ldots,m;$ $p_j > 0, \sum_{j=1}^{m} p_j = 1)$ at each trial, then denoting by N_L the number of trials needed to obtain a randomly chosen sequence of L outcomes (an *L-sequence*)

$$E[N_L] = m^L + L - 1 , \tag{1}$$

whatever be the p_j's. (The random choice of the L-sequence is supposed to be such that the probability of obtaining a sequence with b_j E_j's $(j=1,\ldots,m)$ in a specified order is $\prod_{j=1}^{m} p_j^{b_j}$.)

In the present paper we first give formulae for the variance of N_L and then exploit (1) to obtain some further results.

2. Variance of N_L

For a specified L-sequence with k critical points (see [1]) and a_{ji} E_j's occuring up to the i-th critical point, the conditional distribution of N_L has probability generating function

$$[1 + (1-t)g(t)]^{-1}$$

where

$$g(t) = \sum_{i=1}^{k} \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right) t^{-a_i}$$

with $a_i = \sum_{j=1}^{m} a_{ji}$. Note that $a_k = L$, since the last (k-th) critical point is always at the end of the sequence.

The conditional expected value of N_L is

$$g(1) = \sum_{i=1}^{k} \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right)$$

and the conditional second factorial moment is

$$2[\{g(1)\}^2 + g'(1)] = 2[\{ \sum_{i=1}^{k} \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right) \}^2 - \sum_{i=1}^{k} a_i \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right)] .$$

The unconditional moments of N_L are obtained by taking expected values over the distribution of L-sequences. Using the symbol $\sum_S$ to denote summation over all L-squences we have

$$E[N_L] = \sum_S \left(\prod_{j=1}^{m} p_j^{a_{jk}} \right) \sum_{i=1}^{k} \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right) = m^2 + L-1 \tag{2}$$

(shown in [1]), and

$$E[N_L(N_L-1)] = 2\sum_S \left(\prod_{j=1}^{m} p_j^{a_{jk}} \right) \left[\{ \sum_{i=1}^{k} \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right) \}^2 - \sum_{i=1}^{k} a_i \left(\prod_{j=1}^{m} p_j^{-a_{ji}} \right) \right] \tag{3}$$

Note that k, as well as the a_{ji}'s will depend on the particular L-sequence in S.

We use the formula

$$Var(N_L) = E[N_L(N_L-1)] - E[N_L]\{E[N_L] - 1\} \tag{4}$$

to obtain $\mathrm{var}(N_L)$.

We first evaluate

$$T = \sum_S \Big(\prod_{j=1}^{m} p_j^{a_{jk}} \Big) \sum_{i=1}^{k} a_i \Big(\prod_{j=1}^{m} p_j^{-a_{ji}} \Big) .$$

Remembering that each L-sequence has $a_k = L$, and using (2) we have

$$T = Lm^L + \sum_S \Big(\prod_{j=1}^{m} p_j^{a_{jk}} \Big) \sum_{i=1}^{k-1} a_i \Big(\prod_{j=1}^{m} p_j^{-a_{ji}} \Big)$$

(with $\sum_{i=1}^{0} a_i = 0$).

Following an argument similar to that in Sections 2 and 3 of [1], the second term in T can be shown to be equal to

$$1 + 2 + \ldots + (L-1) = \frac{1}{2}L(L-1) .$$

Hence from (4)

$$\text{(5)} \qquad \mathrm{Var}(N_L) = 2\sum_S \Big(\prod_{j=1}^{m} p_j^{a_{jk}} \Big) \Big\{ \sum_{i=1}^{k} \Big(\prod_{j=1}^{m} p_j^{-a_{ji}} \Big) \Big\}^2 - 2Lm^L - L(L-1) - (m^L+L-1)(m^L+L-2)$$

$$= 2\sum_S \Big(\prod_{j=1}^{m} p_j^{a_{jk}} \Big) \Big\{ \sum_{i=1}^{k} \Big(\prod_{j=1}^{m} p_j^{-a_{ji}} \Big) \Big\}^2 - m^{2L} - (4L-3)m^L - 2(L-1)^2 .$$

The first term has to be evaluated by direct enumeration, though some partial simplification is possible. Table 1 shows the results obtained for $L = 1,2,3,4,5$. Note that, in contradistinction to $E[N_L]$, the variance of N_L does depend on the p_j's.

Table 1. Variance of N_L (Note: $\phi_h = \sum_{j=1}^{m} p_j^{-h}$)

L	$Var(N_L)$	$Var(N_L)$ with $p_1=\ldots=p_m = m^{-1}$
1	$2\phi_1-m(m+1)$	m^2-m
2	$2[\phi_1^2+2\phi_1]-m^4-5m^2+2m-2$	m^4-m^2+2m-2
3	$2[\phi_1^3+2\phi_2+(2m+1)\phi_1]-m^6-9m^3+6m-8$	$m^6-m^3+2m^2+6m-8$
4	$2[\phi_1^4+2\phi_3+2\phi_1^2+\phi_2+2(m^2+1)\phi_1]$ $-m^8-13m^4+2m^2+6m-14$	$m^8-m^4+2m^3+6m^2+6m-14$
5	$2[\phi_1^5+2\phi_4+\phi_3+2\phi_1\phi_2+2m\phi_1^2+(2m^3+m+2)\phi_1]$ $-m^{10}-17m^5+2m^2+14m-28$	$m^{10}-m^5+2m^4+6m^3+6m^2+14m-28$

Since the minimum value of ϕ_L is attained when $p_1=\ldots=p_m = m^{-1}$, these values of the p_j's minimize $Var(N_L)$. The minimum values of $Var(N_L)$ are shown in the last column of Table 1. Note that (m-1) is a factor in each expression, as must be the case since $N_L = L$ and $Var(N_L) = 0$ when $m = 1$.

3. Effects of Partial Repetition

Suppose that, in the situation described in Section 1, we wait until the first L' (<L) terms of the required L-sequence appear. What is the expected number of further trials needed in order to obtain a completion of the L-sequence?

Let $N_{L|L'}$ denote the needed number of trials. Since the L' term initial subsequence must be obtained prior to obtaining the L-sequence, we have

$$N_L = N_{L'} + N_{L|L'}$$

where $N_{L'}$ has the same distribution as N_L, with L' replacing L. Hence

$$(6) \qquad \begin{aligned} E[N_{L|L'}] &= E[N_L] - E[N_{L'}] \\ &= (m^L + L-1) - (m^{L'} + L'-1) \\ &= m^L - m^{L'} + L - L' \end{aligned}$$

As with $E[N_L]$ this is an integer, independent of the p_j's. At each trial the probability that there is E_j at that position of the L-sequence is p_j, and the probability that E_j is observed in the trial is also p_j, so the probability of correct matching is $\sum_{j=1}^{m} p_j^2$.

So, in the next s trials ($s \le L-L'$) the probability of obtaining correctly the next s terms (i.e. the $(L'+1)$-th,...,$(L'+s)$-th terms) in the L-sequence is on average $(\sum_{j=1}^{m} p_j^2)^s$. If these terms are obtained correctly, the conditional expected number of further terms needed to complete the L-sequence is $E[N_{L|L'+s}]$ since we now already have the initial $(L'+s)$ terms of the L-sequence.

Hence

$$(7) \qquad E[N_{L|L'}] = s + (\sum_{j=1}^{m} p_j^2)^s E[N_{L|L'+s}] + \{1 - (\sum_{j=1}^{m} p_j^2)^s\} \varepsilon_{L|L',s}$$

where $\varepsilon_{L|L',s}$ = expected number of further trials needed to complete the L-sequence when the first L' terms have been obtained correctly but the correct sequence has been lost during the next s trials, ($s \le L - L'$).

From (6) and (7), putting $\theta = \sum_{j=1}^{m} p_j^2$ we have

$$\varepsilon_{L|L',s} = m^L - m^{L'} + L - L' - s + \theta^s (1-\theta^s)^{-1} m^{L'} (m^s - 1) \tag{8}$$

or, equivalently

$$\varepsilon_{L|L',s} = m^L - m^{L'+s} + L - L' - s + (1-\theta^s)^{-1} m^{L'} (m^s - 1) \tag{8}'$$

It is of interest to enquire whether an initial correct L'-subsequence followed by s trials which are not all correct (with respect to the relevant L-sequence) is on balance favorable to early completion of the L-sequence, or not. We compare $\varepsilon_{L|L',s}$ with the expected number of trials needed to complete a random L-sequence, after the first $s(\leq L)$ trials. The latter is

$$E[N_L] - s = m^L + L - 1 - s$$

(taking the needed number of trials to be zero when $s = L = N_L$).

Now

$$\varepsilon_{L|L',s} - \{E[N_L] - s\} = (1-\theta^s)^{-1}(m^s\theta^s - 1)m^{L'} - L' + 1 .$$

(Note that this does not depend on L.)

The difference is positive if

$$(m^s\theta^s - 1)m^{L'} > (1-\theta^s)(L'-1) .$$

That is

$$\theta^s > \frac{1}{m^s} \cdot \frac{1 + (L'-1)m^{-L'}}{1 + m^{-s}(L'-1)m^{-L'}} \tag{9}$$

The right hand side of this inequality is greater than or equal to m^{-s}, and less than or equal to

$$m^{-s}(1+G)/(1+m^{-s}G)$$

where $G = \max_{L'} (L'-1)m^{-L'}$.

Remembering that L' must be a positive integer, and $m > 1$, we find $G = (L'-1)m^{-L'}\big|_{L'=2} = m^{-2}$.

Hence the right hand side of (9) lies between

$$m^{-s} \text{ and } m^{-s}(1+m^{-2})/(1+m^{-s-2}) = (1+m^{-2})/(m^{s}+m^{-2}) .$$

Since $m^{-1} \leq \theta = \sum_{j=1}^{m} p_j^2 < 1$, it follows that

$$m^{-s} \leq \theta^{s} < 1 .$$

Further, it is possible to choose the p_j's so that θ is arbitrarily close to 1. In particular, the p_j's can be chosen so that

(10) $$\theta > [(1+m^{-2})/(m^{s}+m^{-2})]^{1/s}$$

i.e. so that $$\varepsilon_{L|L',s} > E[N_L] - s$$

for all L' and $L > L'$.

On the other hand, for all $L' > 1$, it is possible to choose the p_j's so that

(11) $$\theta < m^{-1}[\{1 + (L'-1)m^{-L'}\}/\{1 + (L'-1)m^{-L'-s}\}]^{1/s}$$

i.e. so that

$$\varepsilon_{L|L',s} < E[N_L]-s$$

for this particular (but not necessarily all) L' and $L > L'$.

Note that to get low values of θ we must make the p_j's nearly equal (to m^{-1}). If this is the case then there is still on average, some residual advantage in having had the first L'-subsequence in correct order, even though the next s terms were not in correct order.

Table 2 shows values of the right hand side of (10) and Table 3 the ratio of this quantity to m^{-1} (the minimum possible value of θ). It can be seen that only for a relatively small range of values of θ is it possible to have $\varepsilon_{L|L',s} < E[N_L]-s$ for any L' at all, and this range decreases as m and/or s increases.

Table 2. Values of $\left(\frac{1+m^{-2}}{m^s+m^{-2}}\right)^{1/s} = \theta_B$.

m\s	1	2	3	4	5	6
2	0.556	0.543	0.533	0.527	0.522	0.519
3	0.357	0.349	0.345	0.342	0.340	0.339
4	0.262	0.257	0.255	0.254	0.253	0.253
5	0.206	0.204	0.203	0.202	0.202	0.201
6	0.171	0.169	0.168	0.168	0.168	0.167

Table 3. Values of θ_B/m^{-1}

m\s	1	2	3	4	5	6
2	1.111	1.085	1.066	1.053	1.044	1.037
3	1.071	1.048	1.034	1.026	1.021	1.018
4	1.046	1.029	1.020	1.015	1.012	1.010
5	1.032	1.019	1.013	1.010	1.008	1.007
6	1.023	1.013	1.009	1.007	1.005	1.005

4. Derived Sequences

From any L-sequence based on m possible outcomes $E_1,\ldots,E_m$, an L-sequence based on m' ($<m$) possible outcomes $E'_1,\ldots,E'_{m'}$ can be derived

by combining certain of the first set of outcomes to form single outcomes in the second set. Thus, from an L-sequence based on $E_1,\ldots,E_{10}$ we might derive an L-sequence based on

$E'_1 \equiv E_1$; $E'_2 \equiv E_2 \cup E_3$; $E'_3 \equiv E_4$; $E'_4 \equiv E_5 \cup E_6 \cup E_7$; $E'_5 \equiv E_8 \cup E_9$; $E'_6 \equiv E_{10}$. (Here $m = 10$, $m' = 6$).

For example, the 8-sequence $E_2E_2E_7E_1E_5E_1E_{10}E_3$ becomes the 8-sequence $E'_2\ E'_2\ E'_4\ E'_1\ E'_4\ E'_1\ E'_6\ E'_2$.

Of course there are many ways in which this can be done.

Let us denote the class of sequences based on $\{E_j\}$ by S, and the class based on $\{E'_j\}$ by S'. Any L-sequence in S corresponds to a single L-sequence in S' (but not conversely). If the original sequence is randomly chosen (in the sense described in Section 1) in S, so will be the derived sequence in S', The expected number of trials needed to produce the latter is

$$m'^L + L - 1 \quad .$$

This is, of course, less than $E[N_L] = m^L + L - 1$, because the L-sequence in S' which is first obtained may not actually be constructed from the same L-sequence in S as that required. Given that the L-sequence in S' has just been completed, the expected number of further trials needed to complete the original L-sequence in S is, by an argument similar to that used in Section 2 to obtain $E[N_{L|L'}]$,

$$(m^L + L - 1) - (m'^L + L - 1) = m^L - m'^L \quad . \tag{12}$$

The probability that a realized L-sequence in S' is indeed constructed from the original L-sequence in S is

$$\theta^{L}_{(S')} = \{\sum_{u=1}^{m'} [\sum^{(u)} p_j^2 / \sum^{(u)} p_j]\}^{L} \tag{13}$$

where $\sum^{(u)}$ denotes summation over the values of j for which E_j belongs to E'_u $(u = 1,\ldots,m')$. This is, in particular, the probability that the *first* realization of the derived L-sequence in S' is a realization of the original L-sequence in S.

So if ε' denotes the expected number of further trials needed to obtain a realization of the original L-sequence following a realization of the derived L-sequence in S' which is not also of the original L-sequence in S, then

$$m'^{L} + L - 1 + (1-\theta^{L}_{(S')})\varepsilon' = m^{L} + L - 1$$

whence

$$\varepsilon = (m^{L}-m^{L'})/(1-\theta^{L}_{(S')}) \; . \tag{14}$$

(Note that $\theta_{(S')} = 1$ if and only if $S' \equiv S$, which is excluded by the condition $m' < m$.)

The expected value of the number of additional trials needed given only that the first L have not produced the original L-sequence is

$$E[N_L | N_L > L] = (m^{L}-1)/\{1 - (\sum_{j=1}^{m} p_j^2)^{L}\} \tag{15}$$

($1 - (\sum_{j=1}^{m} p_j^2)^{L}$ = probability that first L trial does not give the randomly chosen L-sequence.)

Now $E[N_L | N_L > L] > \varepsilon'$ if

$$(m^{L}-1)(1-\theta^{L}_{(S')} > (m^{L}-m'^{L})\{1 - (\sum_{j=1}^{m} p_j^2)^{L}\}$$

that is

$$(16) \qquad m'^{L}-1 > (m^{L}-1)\theta^{L}_{(S')} - (m^{L}-m'^{L})\Big(\sum_{j=1}^{m} p_j^2\Big)^{L} .$$

In the special case $p_1 = p_2 = \ldots = p_m = m^{-1}$

$$\theta_{(S')} = \sum_{u=1}^{m} \frac{n_u m^{-2}}{n_u m^{-1}} = m'/m$$

(where n_u = number of E_j's belonging to E'_u.)

Hence (14) and (15) become

$$(14') \qquad \varepsilon' = (m^{L} - m'^{L})/\{1 - (m'/m)^{L}\}$$

$$= m^{L} \quad \text{(whatever the value of } m')$$

and

$$(15') \qquad E[N_L | N_L > L] = (m^{L}-1)/(1-m^{-L}) = m^{L}$$

respectively.

So in this case $\varepsilon' = E[N_L | N_L > L]$. It makes no difference, on average, to the expected number of additional trials needed, whether or not the unsuccessful first L trials constitute the derived L-sequence in S'. This is not, of course, true in the general case.

REFERENCE

[1] Johnson, N.L. (1960) "Repetitions," *American Mathematical Monthly*, *75*, 382-3.