SOME LIMIT THEOREMS ASSOCIATED WITH MULTINOMIAL TRIALS

D. A. DARLING
UNIVERSITY OF MICHIGAN

1. Introduction

Let $X_1, X_2, \cdots$ be independent random variables, each having the same distribution $\Pr\{X_i = k\} = p_k$, $k = 1, 2, \cdots$. We assume without loss of generality that $p_i > 0$ and $p_1 \geq p_2 \geq p_3 \geq \cdots$.

Let $N_n(k)$ be the number of those X_j which equal k, $j = 1, 2, \cdots, n$. In this paper we are going to study certain limiting properties of the random variables

$$R_n = \sum_{N_n(k)>0} 1, \tag{1.1}$$

$$L_n = \sum_{N_n(k) \equiv 1 \pmod 2} 1. \tag{1.2}$$

Thus R_n is the number of distinct values assumed by the sequence

$$\{X_1, X_2, \cdots, X_n\}, \tag{1.3}$$

or the "range" of this sequence, while L_n is the number of values assumed an odd number of times. In principle, other random variables of the form $\sum_{k=1}^{\infty} \phi(N_n(k))$, where ϕ has a finite range, could be studied by the methods of this paper. But the important case of the "coverage" C_n,

$$C_n = \sum_{N_n(k)>0} p_k, \tag{1.4}$$

cannot apparently be so studied.

The random variable R_n is related to the "coupon collector's problem" (cf. Feller [1], p. 102) and has been studied in the case of finitely many equal $p_i > 0$ by Békéssy [2] among others. The random variable L_n is related to a random walk on a simple Abelian group, as described in section 3. It turns out that the studies of the random variables R_n and L_n are almost identical.

The main results of this paper are given in (2.9), (2.11), (3.10), (3.11), and (4.2).

2. The generating functions

As is well known and easily proved, if in the definition of $N_n(k)$ of section 1 we replace n by a random variable Λ which is independent of the $\{X_i\}$ and has a Poisson distribution with parameter λ, the random variables $N_\Lambda(k) = \Lambda_k$ are independent Poisson random variables, $k = 1, 2, \cdots$, Λ_k having a parameter λp_k,

$$\Pr\{\Lambda_k = j\} = e^{-\lambda p_k}\frac{(\lambda p_k)^j}{j!}, \qquad j = 0, 1, \cdots. \tag{2.1}$$

Thus from the equalities

$$\Pr\{\Lambda_k > 0\} = 1 - e^{-\lambda p_k}, \tag{2.2}$$

$$\Pr\{\Lambda_k \equiv 1 \ (\mathrm{mod}\ 2)\} = e^{-\lambda p_k}\sinh \lambda p_k, \tag{2.3}$$

we conclude that for $|t| < 1$,

$$\begin{aligned} E(t^{R_\Lambda}) &= \prod_1^\infty (e^{-\lambda p_k} + t(1 - e^{-\lambda p_k})) \\ &= e^{-\lambda}\prod_1^\infty ((1 - t) + te^{\lambda p_k}), \end{aligned} \tag{2.4}$$

$$\begin{aligned} E(t^{L_\Lambda}) &= e^{-\lambda}\prod_1^\infty (\cosh \lambda p_k + t\sinh \lambda p_k) \\ &= e^{-\lambda}\prod_1^\infty \left(e^{\lambda p_k}\left(\frac{1+t}{2}\right) + e^{-\lambda p_k}\left(\frac{1-t}{2}\right)\right). \end{aligned} \tag{2.5}$$

Let now $0 < t < 1$, and let $W_1, W_2, \cdots$ be independent Bernoulli random variables $\Pr\{W_i = 1\} = t$, $\Pr\{W_i = 0\} = 1 - t$, and put

$$S = \sum_{t=1}^\infty p_i W_i. \tag{2.6}$$

Let also $Y_1, Y_2, \cdots$ be independent random variables of the same character $\Pr\{Y_i = 1\} = (1 + t)/2$, $\Pr\{Y_i = -1\} = (1 - t)/2$, and put $T = \sum p_i Y_i$. We then conclude from (2.4) and (2.5) that

$$\sum_{n=0}^\infty E(t^{R_n})\frac{\lambda^n}{n!} = E(e^{\lambda S}), \tag{2.7}$$

$$\sum_{n=0}^\infty E(t^{L_n})\frac{\lambda^n}{n!} = E(e^{\lambda T}), \tag{2.8}$$

or, equating powers of λ, that

$$\begin{aligned} E(t^{R_n}) &= E(S^n), \\ E(t^{L_n}) &= E(T^n). \end{aligned} \tag{2.9}$$

It is convenient for later purposes to transform T linearly as follows. With the random variables Y_i as above, define $Z_i = (1 - Y_i)/2$. Then Z_i has the distribution $\Pr\{Z_i = 0\} = (1 + t)/2$, $\Pr\{Z_i = 1\} = (1 - t)/2$. Setting

$$U = \sum p_i Z_i, \tag{2.10}$$

we have $T = 1 - 2U$, and

$$E(t^{L_n}) = E((1 - 2U)^n). \tag{2.11}$$

Thus the random variables R_n and L_n have generating functions which are the n-th moments of fixed random variables S and $1 - 2U$ respectively, S and U

being weighted sums (with weights p_i) of simple Bernoulli random variables. This representation is useful in studying limiting properties of R_n and L_n.

3. Recurrence properties of L_n

The random variables L_n (and R_n) do not form a Markov chain, but the events $\{L_n = 0\}$ clearly form a sequence of "recurrent events" in the sense of Feller ([1], p. 282), and it is easily seen that $\Pr\{L_n = j \text{ infinitely often}\}$ is independent of j, and hence, specializing to $j = 0$, it is either 0 or 1, according as $\sum \Pr\{L_n = 0\}$ converges or diverges.

Spitzer ([3], p. 91) has analyzed the recurrence of L_n by considering a random walk on an Abelian group $\mathcal{G}$ in the following way: let the group elements g be all infinite sequences of 0's or 1's ultimately terminating in zeros $g = \{\omega_1, \omega_2, \cdots\}$, $\omega_i = 0$, or 1, $\sum \omega_i < \infty$, with multiplication defined as component-wise addition mod 2. A set of generators for $\mathcal{G}$ is $\{g_1, g_2, \cdots\}$ where g_n has all zeros, except at the n-th coordinate where it has a one. A "random walk" Q_n on $\mathcal{G}$ is defined by setting $Q_0 = e = (0, 0, \cdots,)$, $Q_{n+1} = Q_n G_{r+1}$ where the G_n are independent and $\Pr\{G_n = g_j\} = p_j$. If we set L_n as the sum of the components of Q_n, this conforms distribution-wise to the L_n defined in (1.2), and the realization of infinitely many $L_n = 0$ is equivalent to the recurrence of Q_n.

If we set $t = 0$ in (2.10) and (2.11), we have the Z_i as fair $\{0, 1\}$ random variables $\Pr\{Z = 1\} = \Pr\{Z = 0\} = \frac{1}{2}$, Z_i independent and

$$\begin{aligned} U &= \sum p_i Z_i, \\ \Pr\{L_n = 0\} &= E((1 - 2U)^n). \end{aligned} \tag{3.1}$$

It follows from the general theory (cf. Feller [1], p. 285) that a necessary and sufficient condition for recurrence is that $\sum \Pr\{L_n = 0\} z^n$ be infinite at $z = 1$, or that

$$E\left(\sum_0^\infty (1 - 2U)^n\right) = \frac{1}{2} E\left(\frac{1}{U}\right) = \infty. \tag{3.2}$$

In other words, a necessary and sufficient condition for recurrence is that the function whose Rademacher coefficients are the p_j have a nonintegrable reciprocal. (Here the Rademacher functions $r_n(x)$ are defined as the n-th term in the binary expansion of x, $0 \leq x < 1$, using, say, the expansion terminating in 0's if x is of the form $k/2^n$). Thus, a necessary and sufficient condition is that

$$\int_0^1 \frac{dx}{\sum p_n r_n(x)} = \infty. \tag{3.3}$$

Also, using (2.5) with $t = 0$, we have

$$e^{-\lambda} \sum \Pr\{L_n = 0\} \frac{\lambda^n}{n!} = e^{-\lambda} \prod_1^\infty \cosh \lambda p_n, \tag{3.4}$$

and the terms on the left being nonnegative, an easy Tauberian theorem for

Borel summability shows that a necessary and sufficient condition for recurrence is that

$$\int_0^\infty e^{-\lambda} \prod_0^\infty \cosh \lambda p_n \, d\lambda = \infty. \tag{3.5}$$

Neither of the above two results is very informative, and indeed it would appear that, from the result to be given next, the necessary and sufficient conditions on $\{p_n\}$ to ensure recurrence are rather delicate and not to be exhibited in a neat form.

In the series (3.1) for U, let J_k be the number of terms separating the $(k-1)$-st and k-th occurrence of the event $Z_i = 1$, so that $J_1, J_2, \cdots$ are independent, identically distributed random variables with $\Pr\{J_k = n\} = 1/2^n$, $n = 1, 2, \cdots$. Let also $S_k = J_1 + J_2 + \cdots + J_k$ be the index at which $Z_i = 1$ for the k-th time. Then $U = p_{S_1} + p_{S_2} + \cdots$. If we set $f_k = p_k + p_{k+1} + \cdots$ and define

$$U_k = p_{S_1} + p_{S_2} + \cdots + p_{S_k}, \tag{3.6}$$

$$V_k = p_{S_1} + p_{S_2} + \cdots + p_{S_{k-1}} + f_{S_k}, \tag{3.7}$$

we have $U_k \leq U \leq V_k$, and U_k monotonically increases, V_k monotonically decreases to U, $k \to \infty$.

If we next define

$$\alpha_k(r) = \sum (p_{r_1} + p_{r_2} + \cdots + p_{r_k})^{-1}, \qquad 1 \leq r_1 < r_2 < \cdots < r_k = r, \tag{3.8}$$

$$\beta_k(r) = \sum (p_{r_1} + p_{r_2} + \cdots + p_{r_{k-1}} + f_{r_k})^{-1}, \qquad 1 \leq r_1 < r_2 < \cdots < r_k = r, \tag{3.9}$$

a straightforward calculation gives

$$E\left(\frac{1}{U}\right) < E\left(\frac{1}{U_k}\right) = \sum_{r=k}^\infty \frac{1}{2^r \alpha_k(r)}, \tag{3.10}$$

$$E\left(\frac{1}{U}\right) > E\left(\frac{1}{V_k}\right) = \sum_{r=k}^\infty \frac{1}{2^r \beta_k(r)}, \tag{3.11}$$

for $k = 1, 2, \cdots$. Consequently, a necessary and sufficient condition for recurrence is that for all k the series on the right of (3.10) diverge. Equivalently, a necessary and sufficient condition for recurrence is that the series on the right of (3.11) diverge for some k.

For any fixed $k \geq 1$ the divergence of (3.10) is necessary, and the divergence of (3.11) is sufficient, but there is a gap (which vanishes as $k \to \infty$) which seems difficult to bridge. For $k = 1$, this criterion was given by Spitzer using different methods, based on determining a set of group characters for $\mathcal{G}$ (which are simply related to the Rademacher functions $r_n(k)$ given above).

4. Limiting results

The limiting behavior of R_n and L_n are essentially identical, since the generating functions are given as moments of essentially identical random variables S and $1 - 2U$ (S and U are defined in (2.6) and (2.10)). We thus consider only R_n and make the following assumption about the sequence $\{p_n\}$. Define $g(\xi) = \max\{j|p_j > 1/\xi\}$, $0 < \xi < \infty$, and assume that $g(\xi) = \xi^\alpha L(\xi)$ where $L(\xi)$ is a slowly varying function; $L(a\xi)/L(\xi) \to 1$, $\xi \to \infty$, $a > 0$.

We necessarily have $0 \le \alpha < 1$, and it seems indispensable that some such regularity condition be satisfied in order for limiting distributions to exist. It is interesting that if this condition is slightly strengthened, the series in (3.10) and (3.11) converge or diverge together for all k, and one obtains, for this class of $\{p_n\}$, necessary and sufficient conditions for recurrence.

In (2.9) we set $t = e^{-\epsilon}$ and $\delta = 1 - e^{-\epsilon}$, and let $J_1, J_2, \cdots$ be independent random variables with the common distribution $\Pr\{J = k\} = \delta(1 - \delta)^{k-1}$, $k = 1, 2, \cdots$, $S_k = J_1 + J_2 + \cdots + J_k$. We then have

$$E(e^{-\epsilon R_n}) = E((1 - p_{S_1} - p_{S_2} - \cdots)^n), \tag{4.1}$$

where S_k represents the index of the k-th occurrence of $W_i = 1$ in the series (2.6).

Let us define the stochastic process $S_\delta(t)$ as $S_\delta(t) = \sum_{S_i < t} 1$; it is then easy to verify that $S_\delta(t/\delta)$ converges in distribution to $X(t)$, the Poisson process with rate 1. From (4.1) we have

$$\begin{aligned} E(e^{-\epsilon R_n}) &= E\left(\left(1 - \int_0^\infty p_t\, dS_\delta(t)\right)^n\right) \\ &= E\left(\left(1 - \int_0^\infty p_{t/\delta}\, dS_\delta(t/\delta)\right)^n\right) \\ &= E\left(e^{-n\int_0^\infty p_{t/\delta}\, dS_\delta(t/\delta) + nZ_\delta}\right), \end{aligned} \tag{4.2}$$

where the quotient of Z_δ by the integral in the last exponent converges to zero in distribution as $\delta \to 0$.

Thus we can express limiting distributions in terms of the distributions of functionals of the form $\int_0^\infty f(t)\, dX(t)$, where $X(t)$ is the Poisson process. As an example, consider the case when $\alpha > 0$ where, because of the form of $g(\xi)$ presumed above, we have

$$np_{[tg(n)]} \to t^{-1/\alpha}, \qquad n \to \infty. \tag{4.3}$$

Now since $\epsilon = \delta + 0(\delta)$, $\delta \to 0$, an application of the J-convergence theorem for functionals of additive processes of Skorohod ([4], p. 221), enables one to conclude that, setting $\delta = h/g(n)$,

$$E\left(e^{-h\frac{R_n}{g(n)}}\right) \to E\left(e^{-h^{1/\alpha}\int_0^\infty \frac{dx(t)}{t^{1/\alpha}}}\right) = e^{-h\Gamma(1-\alpha)} \tag{4.4}$$

This last remark follows from the fact that $\int_0^\infty (dX(t))/t^{1/\alpha}$ has a positive stable distribution of index α, as is readily established.

5. Concluding remark

At the time of presenting these results at the Symposium, a central limit theorem was given for the random variables R_n, by refining the above calculations. The author learned at the Symposium that this had been established independently by S. Karlin (unpublished), using different methods. It was decided that we would publish our results jointly elsewhere.

ADDED IN PROOF. In the presentation of the above paper, the author was unaware of the work of R. R. Bahadur, "On the number of distinct values in a large sample from an infinite discrete distribution," *Proc. Nat. Inst. Sci. India*, Vol. 26 (1960), pp. 67–75. In this paper Bahadur obtains estimates for $E(R_n)$ in a number of interesting cases, partly overlapping section 4 above.

REFERENCES

[1] W. FELLER, *An Introduction to Probability Theory and Its Applications*, New York, John Wiley, 1957 (2d ed.).

[2] A. BÉKÉSSY, "A new proof of a theorem concerning a distribution problem," *Magyar Tud. Akad. Mat. Fiz. Oszt. Kozl.*, Vol. 12 (1962), pp. 329–334.

[3] R. SPITZER, *Principles of Random Walk*, New York, Van Nostrand, 1964.

[4] A. V. SKOROHOD, *Random Processes with Independent Increments*, Moscow, Izdatel'stvo "Nauka," 1964.