ON THE SERIAL TEST FOR RANDOM SEQUENCES

By I. J. Good

Let $a_1, a_2, \cdots, a_N$ be a finite random sequence, $\mathfrak{S}$, of independent random variables with $P(a_j = r) = p_r (j = 1, 2, \cdots, N; r = 0, 1, \cdots, t-1)$, where $\sum p_r = 1$. We call each of $0, 1, \cdots, t-1$ "digits." Associated with $\mathfrak{S}$ is the corresponding cyclic sequence $\bar{\mathfrak{S}}$, defined by regarding the first digit of $\mathfrak{S}$ as immediately following the last one. We shall always denote properties of $\bar{\mathfrak{S}}$ by placing a bar over the corresponding algebraic symbol relating to $\mathfrak{S}$.

A sequence of ν digits is called a ν-sequence. A ν-sequence is said to belong to $\mathfrak{S}$ if it is of the form $a_j, a_{j+1}, \cdots, a_{j+\nu-1} (j = 1, 2, \cdots, N-\nu+1)$ and to belong to $\bar{\mathfrak{S}}$ if j is also allowed to take the values $N-\nu+2, N-\nu+3, \cdots N$, where a_{N+k} is identified with a_k for all integers k. Let $n_{r_1, r_2, \cdots, r_\nu}$, or $n_\mathfrak{r}$ for short, be the number of ν-sequences in $\mathfrak{S}$ which are the ν-sequence $(r_1, r_2, \cdots, r_\nu) = \mathfrak{r}$ (where $r_1, r_2, \cdots, r_{\nu-1}$ and r_ν are digits). Let

$$p_\mathfrak{r} = p_{r_1} \cdots p_{r_\nu}, \tag{1}$$

$$\psi_\nu^2 = \sum_\mathfrak{r} \frac{[n_\mathfrak{r} - (N-\nu+1)p_\mathfrak{r}]^2}{(N-\nu+1)p_\mathfrak{r}}, \tag{2}$$

$$\bar{\psi}_\nu^2 = \sum_\mathfrak{r} \frac{(\bar{n}_\mathfrak{r} - Np_\mathfrak{r})^2}{Np_\mathfrak{r}}, \tag{3}$$

where $\mathfrak{r}$ runs through all its t^ν possible values. Let

$$\psi_0^2 = \bar{\psi}_0^2 = 1. \tag{4}$$

We shall prove that if[1]

$$\nu \leqq \tfrac{1}{2}(N+1), \tag{5}$$

then

$$\mathcal{E}(\psi_\nu^2) = t^\nu - 1, \tag{6}$$

and

$$\mathcal{E}(\bar{\psi}_\nu^2) = t^\nu - 1. \tag{7}$$

The special case $p_0 = p_1 = \cdots = p_{t-1} = t^{-1}$ was dealt with by Good [1], who also proved some asymptotic results that have been generalized to arbitrary random sequences by Billingsley [2]. When we apply these asymptotic results to actual (finite) values of N, our confidence in their approximate validity is increased on finding that equations (6) and (7) are exact.

We mention in passing that the variances of ψ_ν^2 and $\bar{\psi}_\nu^2$ are asymptotically

$$\frac{2}{t-1}(t^{\nu+1} + t^\nu - (2\nu+1)t + 2\nu - 1), \tag{8}$$

but we do not know how good the approximation is for actual values of N.

Received April 4, 1956.

[1] The condition (5) was also required by Good [1], but in error it was not mentioned.

The results can be generalized to random arrays (rectangles or "blocks") in two or more dimensions, with circularization in any number of these dimensions. (The word "dimension" is here being used in its obvious Euclidean sense and not in the sense of Good [1], p. 280.) The length, breadth, height, ... of the whole array may be denoted by $N_1, N_2, N_3, \cdots$. A ν-sequence or ν-block is then defined as an array of digits forming a block of length, breadth, height, ... $\nu_1, \nu_2, \nu_3, \cdots$, digits, and with the obvious definitions of ψ_ν^2 and $\bar{\psi}_\nu^2$ it can be proved that if

$$\nu_1 \leqq \tfrac{1}{2}(N_1 + 1),\ \nu_2 \leqq \tfrac{1}{2}(N_2 + 1), \cdots, \tag{9}$$

then

$$\mathcal{E}(\psi_\nu^2) = t^{\nu_1\nu_2\cdots} - 1, \tag{10}$$

and

$$\mathcal{E}(\bar{\psi}_\nu^2) = t^{\nu_1\nu_2\cdots} - 1. \tag{11}$$

The proofs of these multidimensional results are essentially the same as for the one-dimensional case, but involve extra notation that tends to obscure the issue. We shall therefore content ourselves with giving the proof for the one-dimensional case only.

If $1 \leqq j \leqq j + \nu - 1 \leqq N$ (and continuing to regard $a_1, a_2, \cdots, a_N$ as random variables), let

$$P(a_j = r_1, \cdots, a_{j+\nu-1} = r_\nu) = P(\mathfrak{r}),$$

and if $1 \leqq j$, $1 \leqq j + m$ (where m may be negative), $j + m + \nu - 1 \leqq N$, let

$$\begin{aligned} P(a_{j+m} = r_1, a_{j+m+1} = r_2, \cdots, a_{j+m+\nu-1} = r_\nu \mid a_j = r_1, \cdots, a_{j+\nu-1} = r_\nu) \\ = P_m(r_1, \cdots, r_\nu) = P_m(\mathfrak{r}), \end{aligned}$$

the probability that a ν-sequence agrees with an assigned one m places to its left. Everything will depend on the following lemma proved at the end of the paper.

LEMMA. *If $\nu + |m| \leqq N$, then*

$$\sum_{\mathfrak{r}} P_m(\mathfrak{r}) = \begin{cases} 1 & \text{if } m \neq 0, \\ t^\nu & \text{if } m = 0, \end{cases}$$

where $\mathfrak{r}$ runs through all of its t^ν possible values.

We shall now prove equation (6).

By (12) and (14) of Good [1], we have

$$\mathcal{E}(n_{\mathfrak{r}}) = (N - \nu + 1)P(\mathfrak{r}),$$

$$V(n_{\mathfrak{r}}) = P(\mathfrak{r}) \sum_{m}^{|m|<\nu} (N - \nu + 1 - |m|)(P_m(\mathfrak{r}) - P(\mathfrak{r})).$$

Therefore

$$\mathcal{E}(\psi_\nu^2) = \frac{1}{N-\nu+1}\sum_{\mathfrak{r}}\sum_{m}^{|m|<\nu}(N-\nu+1-|m|)(P_m(\mathfrak{r})-P(\mathfrak{r})),$$

$$= \frac{1}{N-\nu+1}\sum_{m}^{|m|<\nu}(N-\nu+1-|m|)\sum_{\mathfrak{r}}(P_m(\mathfrak{r})-P(\mathfrak{r})).$$

By the lemma, the inner sum vanishes except when $m = 0$, and equation (6) follows immediately. Equation (7) may be proved in a similar but slightly simpler manner, and the proof will be omitted. We now prove the lemma.

PROOF. Negative values of m may be treated by the same method as positive values, and the case $m = 0$ is trivial since $P_0(\mathfrak{r}) = 1$, so we shall suppose m to be positive. The lemma is also trivial if $|m| \geqq \nu$, since $P_m(\mathfrak{r})$ is then equal to $P(\mathfrak{r})$, but in the application $|m| < \nu$, so we suppose $1 \leqq m < \nu$. We may suppose $j = 1$ without loss of generality.

By the multiplicative axiom of probability, we have

$$P_m(\mathfrak{r}) = P(a_{1+m} = r_1, a_{2+m} = r_2, \cdots, a_\nu = r_{\nu-m} \mid a_1 = r_1, \cdots, a_\nu = r_\nu)$$
$$\times P(a_{\nu+1} = r_{\nu+1-m}, \cdots, a_{\nu+m} = r_\nu$$
$$\mid a_1 = r_1, \cdots, a_\nu = r_\nu ; a_{1+m} = r_1, \cdots, a_\nu = r_{\nu-m}).$$

The first factor is either 0 or 1, depending on whether the following conditions are not or are satisfied:

$$r_{1+m} = r_1, r_{2+m} = r_2, \cdots, r_\nu = r_{\nu-m}.$$

When these conditions are satisfied, the second factor is simply equal to $P(a_{\nu+1} = r_{\nu+1-m}, \cdots, a_{\nu+m} = r_\nu)$, since $\mathfrak{S}$ is a random sequence. Therefore

$$\sum_{\mathfrak{r}} P_n(\mathfrak{r}) = \sum_{r_1,\cdots r_\nu}^{r_{1+m}=r_1,\cdots,r_\nu=r_{\nu-m}} P(a_{\nu+1} = r_{\nu+1-m}, \cdots, a_{\nu+m} = r_\nu)$$

$$= \sum_{r_{\nu+1-m},\cdots,r_\nu}^{0,1,\cdots,t-1} P(a_{\nu+1} = r_{\nu+1-m}, \cdots, a_{\nu+m} = r_\nu),$$

since the conditions above the sign of summation in the previous line determine $r_1, r_2, \cdots, r_{\nu-m}$ uniquely in terms of $r_{\nu+1-m}, \cdots, r_\nu$. Our sum is now merely the total probability of all t^m possible m-sequences and is therefore unity.

REFERENCES

[1] I. J. GOOD, "The serial test for sampling numbers and other tests for randomness," *Proc. Cambridge Philos. Soc.*, Vol. 49 (1953), pp. 276–284.

[2] P. BILLINGSLEY, "Asymptotic distributions of two goodness of fit criteria," *Ann. Math. Stat.*, Vol. 27, (1956) pp.1123–1129.