# NOTE ON A MATCHING PROBLEM

By Solomon Kullback

1. **Introduction.** There is to be found in the literature [1] a number of discussions of the matching problem i.e., the problem of deriving the distribution of the number of correct matchings when two sequences of elements are placed in correspondence. However, the formulation of the matching problem discussed and illustrated herein is somewhat different from those problems already discussed in the literature [1], and may be of interest. A rather general statement of the problem follows.

2. **The Problem.** Consider urns $U_i$, $i = 1, 2, \cdots, n$ each of which contains some or all of the $r$ different elements $E_1, E_2, \cdots, E_r$. The relative proportions of the $r$ elements in the $i$-th urn are $p_{i1}, p_{i2}, \cdots, p_{ir}$ $(i = 1, 2, \cdots, n)$ such that

$$p_{i1} + p_{i2} + \cdots + p_{ir} = 1 \qquad i = 1, 2, \cdots, n \tag{1}$$

$$p_{i1}^2 + p_{i2}^2 + \cdots + p_{ir}^2 = p_i \qquad i = 1, 2, \cdots, n \tag{2}$$

(Some $p_{ij}$ $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, r$ may be zero).

Assuming each urn to be an infinite source, consider two sequences made by drawing, at random, a single element from each urn in turn. If the two sequences are placed in correspondence there will be a number of correct matchings. What is the distribution of the number of correct matchings if the foregoing process be indefinitely repeated?

3. **Solution of the Problem.** The probability that the elements in the $k$-th position of the two sequences match may be derived by the following simple considerations. Since all the drawings are independent, the probability that both elements in the $k$-th position are $E_m$ is $p_{km}^2$. Accordingly, the probability that both elements are the same, irrespective of their particular identity is $p_{k1}^2 + p_{k2}^2 + \cdots + p_{kr}^2 = p_k$.

The theory for the number of correct matchings in this case thus corresponds to that for the Poisson series, which is well known [2]. For the special case in which $p_k = p$, $k = 1, 2, \cdots, n$ the distribution of the number of correct matchings is in accordance with the binomial $(q + p)^n$ where $q = 1 - p$.

4. **Numerical Illustration and Verification.** The following illustration corresponds to the special case in which the urns are taken to be identical with equal proportions of each of the $r$ elements.

Random sequences of 300 digits each were matched and the number of correct matchings recorded. The result of 457 such observations is given in Table 1.

TABLE 1

*Observed distribution of number of correct matchings per sequences of 300 random digits each*

| Number of correct matchings | Observed frequency | Number of correct matchings | Observed frequency |
|---|---|---|---|
| 18 | 1 | 32 | 35 |
| 19 | 2 | 33 | 25 |
| 20 | 3 | 34 | 15 |
| 21 | 5 | 35 | 20 |
| 22 | 9 | 36 | 20 |
| 23 | 22 | 37 | 17 |
| 24 | 18 | 38 | 6 |
| 25 | 21 | 39 | 10 |
| 26 | 41 | 40 | 7 |
| 27 | 28 | 41 | 1 |
| 28 | 30 | 42 | 3 |
| 29 | 31 | 43 | 1 |
| 30 | 42 | 44 | 0 |
| 31 | 42 | 45 | 2 |
| | | Total | 457 |

| Average number of correct matchings | Standard deviation |
|---|---|
| 29.9934 | 4.8484 |

TABLE 2

*Values of* $P_x = (300!/x!(300 - x)!)(0.1)^x(0.9)^{300-x}$

| $x$ | $P_x$ | $x$ | $P_x$ | $x$ | $P_x$ | $x$ | $P_x$ |
|---|---|---|---|---|---|---|---|
| 14 | 0.00033 | 23 | 0.03240 | 32 | 0.06920 | 41 | 0.00875 |
| 15 | .00070 | 24 | .04156 | 33 | .06245 | 42 | .00599 |
| 16 | .00139 | 25 | .05099 | 34 | .05499 | 43 | .00400 |
| 17 | .00257 | 26 | .05992 | 35 | .04601 | 44 | .00259 |
| 18 | .00449 | 27 | .06756 | 36 | .03763 | 45 | .00164 |
| 19 | .00741 | 28 | .07319 | 37 | .02984 | 46 | .00101 |
| 20 | .01156 | 29 | .07628 | 38 | .02294 | 47 | .00061 |
| 21 | .01713 | 30 | .07656 | 39 | .01713 | 48 | .00036 |
| 22 | .02413 | 31 | .07409 | 40 | .01242 | 49 | .00020 |

In accordance with paragraph 3, the distribution in Table 1 should correspond to the binomial distribution with $n = 300$ and $p = 10(1/10^2) = 1/10$. For the

TABLE 3

*Comparison of observed distribution with the theoretical distribution*
*457 (0.9 + 0.1)*$^{300}$

| Number of correct matchings | Frequency | | |
|---|---|---|---|
| | Observed | Theoretical | |
| | $F_0$ | $f$ | $F = 457f$ |
| 14–16 | 0 | 0.00242 | 1.1 |
| 17–19 | 3 | .01447 | 6.6 |
| 20–22 | 17 | .05282 | 24.1 |
| 23–25 | 61 | .12495 | 57.1 |
| 26–28 | 99 | .20067 | 91.7 |
| 29–31 | 115 | .22693 | 103.7 |
| 32–34 | 75 | .18614 | 85.1 |
| 35–37 | 57 | .11348 | 51.9 |
| 38–40 | 23 | .05249 | 24.0 |
| 41–43 | 5 | .01874 | 8.6 |
| 44–46 | 2 | .00524 | 2.3 |
| 47–49 | 0 | .00117 | .5 |
| | 457 | | 456.7 |

TABLE 4

| $F_0$ | $F$ | $(F_0 - F)^2/F$ | |
|---|---|---|---|
| 0, 3 | 1.1, 6.6 | 2.87 | |
| 17 | 24.1 | 2.09 | |
| 61 | 57.1 | .27 | |
| 99 | 91.7 | .58 | $\chi_0^2 = 10.48$ |
| 115 | 103.7 | 1.23 | |
| 75 | 85.1 | 1.20 | |
| 57 | 51.9 | .50 | $n = 8$ |
| 23 | 24.0 | .04 | |
| 5, 2, 0 | 8.6, 2.3, .5 | 1.70 | $P(\chi^2 > \chi_0^2) = .235$ |
| | | 10.48 | |

binomial distribution we have $m = np = 30$, $\sigma = \sqrt{npq} = \sqrt{27} = 5.1962$. To compare the observed distribution with the expected distribution we calculated the values of $P_x = (300!/x!(300 - x)!)(0.1)^x(0.9)^{300-x}$ for values of $x$ from 14 to 49 inclusive which are given in Table 2.

To compare the observed and the theoretical distributions, and test the "Goodness of Fit," the distributions were grouped in classes of three. The results are shown in Tables 3 and 4.

**5. Conclusion.** The agreement between the observed distribution and the theoretical distribution derived on the basis of the argument in paragraph 3 is quite satisfactory.

We have shown herein, that if two sequences be matched under certain conditions, the distribution of the number of correct matchings will, in general, be that of a Poisson series and in special cases the binomial distribution. The theory was illustrated by an experiment which yielded results in satisfactory agreement with the theory.

## REFERENCES

[1] An indication of, and some of the problems discussed, will be found in: E. G. OLDS, "A moment generating function which is useful in solving certain matching problems," *Bull. Am. Math. Soc.*, Vol. 44, June 1938, pp. 407–413.

[2] Some references to discussions of the theory, moments, and distribution connected with the Poisson series follow:

(a) H. L. RIETZ, *Mathematical Statistics*, Carus Mathematical Monograph No. 3, Open Court Pub. Co. (1927), pp. 148–152.

(b) CHARLES JORDAN, *Statistique Mathematique*, pp. 109–110.

(c) T. KAMEDA, "Theory of generating functions and its application to the theory of probability," *Journal of the Faculty of Science*, Imperial Univ. Tokyo, Section I, Vol. I, Part 1, (1925), Theorem XXIV, p. 48.

(d) P. R. RIDER, "The third and fourth moments of the generalized Lexis Theory," *Metron*, Vol. 12, No. 1, (1934) p. 195.

(e) S. KULLBACK, On the Bernoulli Distribution, *Bull. Am. Math. Soc.*, Vol. 41, (1935) p. 861.

THE GEORGE WASHINGTON UNIVERSITY.