ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVI, 2 1978)

ADRIENNE W. KEMP (Bradford)

ON PROBABILITY GENERATING FUNCTIONS FOR MATCHING AND OCCUPANCY DISTRIBUTIONS

The three distributions which Irwin [11] derived by Whitworth's theorem, namely the classical matching, the classical occupancy and a line-division distribution, are re-examined in this paper. It is shown that these distributions, together with certain important generalizations, all belong to the class of distributions which Kemp and Kemp [13] called *generalized hypergeometric factorial-moment distributions*, i.e., they all have probability generating functions of the form ${}_pF_q[\lambda(s-1)]$.

1. Introduction. Almost twenty years ago Irwin [11], in an expository paper, attempted to unify certain areas of discrete distribution theory by the use of Whitworth's theorem. The three discrete distributions discussed by Irwin were the following:

(i) the classical matching distribution,

(ii) the distribution of the number of occupied classes when N objects are assigned at random to k different classes,

(iii) the distribution of the number of intervals greater than u/k when a line of fixed length u is divided into n intervals by $n-1$ randomly placed points.

Barton's [2] paper contains a short historical survey of the matching distribution. Generalizations include Gumbel's matching distribution and the Laplace-Haag matching distribution, Fréchet [8], Anderson's [1] distribution of matching K-tuples for K identical packs of N cards (see also [2]). The classical occupancy distribution is known also by the names Arfwedson, coupon-collecting, cigarette card, and Stevens-Craig. Harkness [10] reviews previous work on this distribution, together with a more general form of the distribution, where each object has a probability p of remaining in the randomly chosen class and $1-p$ of escaping (see also [14]). Irwin's third distribution is closely related to Fisher's modification of Schuster's criterion in harmonic analysis, and to a result by Garwood [9] concerning the operation of vehicular-controlled traffic signals.

Whitworth's theorem is a consequence of the principle of inclusion and exclusion (see [5], p. 102); hence methods utilizing Whitworth's theorem are closely related to Fréchet's (see [7] and [8]) treatment for matching distributions (which was based on earlier work by Haag and de Finetti) and to David and Barton's use of characteristic random variables. The principle of inclusion and exclusion was used as long ago as 1714 by Montmort in his study of the game of Treize (Rencontre). Czuber's [4] proof assumed the independence of events; Broderick [3] appears to have given the first proof not assuming independence.

Suppose now that the events of interest can be defined in terms of more elementary events $A_1, A_2, \ldots, A_n$ one, two, or more of which can occur simultaneously with probabilities

$$P\{A_i\}, P\{A_iA_j\}, \ldots, P\{A_iA_j \ldots A_r\}, \ldots,$$

and that

$$S_1 = \sum_i P\{A_i\}, \quad S_2 = \sum_{\substack{i,j \\ i<j}} P\{A_iA_j\}, \quad \ldots,$$

$$S_r = \sum_{\substack{i,j,\ldots,r \\ i<j<\ldots<r}} P\{A_iA_j \ldots A_r\}, \quad \ldots,$$

where S_r is the sum of ${}_NC_r$ probabilities (see [6], p. 88 ff.). Using the inclusion-exclusion principle both Fréchet and also David and Barton show that $\mu'_{[r]} = S_r \times r!$, where $\mu'_{[r]}$ is the r-th factorial moment, and that

$$P_{[r]} = S_r - {}_{r+1}C_1 S_{r+1} + {}_{r+2}C_2 S_{r+2} - \ldots + (-1)^{N-r} {}_NC_{N-r} S_N$$

$$= \sum_{t=r}^{N} \frac{(-1)^{t-r} {}_tC_r \mu'_{[t]}}{t!},$$

where $P_{[r]}$ is the probability that exactly r among the N events $A_1, A_2, \ldots$ $\ldots, A_N$ occur simultaneously (this "inversion" formula was given by Steffensen [16]). Hence the probability generating function (p.g.f.) for the required distribution is

$$G(s) = \sum_{r=0}^{N} S_r (s-1)^r = \sum_{r=0}^{N} \frac{\mu'_{[r]} (s-1)^r}{r!}.$$

2. Matching distributions. We begin by considering the Laplace-Haag matching distribution (see [8], p. 148 ff.). Suppose that there are N objects na of which carry the labels $1, 2, \ldots, n$, each label occurring a times, and that $N - na$ are unlabelled; a coincidence (match) occurs if an object with the label j occurs at the j-th (random) draw. In Barton's [2] notation the compositions of the target and matching packs are (1^N)

and $((N-an), a^n)$, respectively. Fréchet shows that the distribution of the number of matches has the factorial moments

$$\mu'_{[r]} = \frac{a^r n!(N-r)!}{(n-r)!N!}$$

and that the probabilities are

$$P_{[r]} = \frac{1}{r!}\sum_{t=r}^{n} \frac{(-1)^{t-r}a^t n!(N-t)!}{(t-r)!(n-t)!N!}.$$

Hence we deduce that the p.g.f. may be stated in terms of the confluent hypergeometric function as

$$G(s) = {}_1F_1[-n;\ -N;\ a(s-1)].$$

For the Laplace matching distribution, $N = an$ (see [8], p. 150); for the Gumbel distribution, $N = n$ (see [8], p. 192); and for the classical matching distribution, $N = n$, $a = 1$.

The p.g.f. for the Laplace-Haag distribution bears a striking resemblance to that for the Poisson-beta distribution, i.e., to

$${}_1F_1[a;\ a+b;\ \lambda(s-1)], \qquad a, b, \lambda > 0.$$

The method given by Kemp and Kemp [13] for obtaining a recurrence formula for the Poisson-beta probabilities yields immediately, for the Laplace-Haag probabilities,

$$(r+2)(r+1)P_{[r+2]} = (r+1)(a+r-N)P_{[r+1]} + a(n-r)P_{[r]},$$

where

$$P_{[0]} = {}_1F_1[-n,\ -N;\ -a] \quad \text{and} \quad P_{[1]} = na\,{}_1F_1[1-n;\ 1-N;\ -a]/N.$$

This result may usefully be recast in the form

$$a(n-r)P_{[r]} = (r+2)(r+1)P_{[r+2]} + (r+1)(N-a-r)P_{[r+1]},$$

where

$$P_{[n]} = a^n(N-n)!/N! \quad \text{and} \quad P_{[n-1]} = (1+N-n-a)(N-n)!na^{n-1}/N!.$$

The class of distributions with p.g.f.'s of the form

$${}_pF_q[(a);\ (b);\ \lambda(s-1)]$$

form the subject matter of Kemp and Kemp [13], where they are called *generalized hypergeometric factorial-moment* (g.h.f.) *distributions* (since they have factorial-moment generating functions of the form ${}_pF_q[\lambda t]$).

Three limiting formulae for generalized hypergeometric functions were stated in [12]. Application of these formulae to g.h.f. distributions leads to the following general results:

The distribution with p.g.f. ${}_pF_q[(a);(b);\lambda(s-1)]$ is the limiting form, $d\to\pm\infty$, of that with p.g.f.

$${}_{p+1}F_q[(a),c+d;(b);\lambda(s-1)/d].$$

It is also the limiting form, $d\to\pm\infty$, of the distribution with p.g.f.

$${}_pF_{q+1}[(a);(b),c+d;d\lambda(s-1)].$$

It is furthermore the limiting form, $d\to\pm\infty$, of the distribution with p.g.f.

$${}_{p+1}F_{q+1}[(a),d;(b),c+dk;k\lambda(s-1)].$$

The penultimate limiting forms of these three distributions are

$${}_pF_q[(a);(b);\lambda(1+c/d)(s-1)],\quad {}_pF_q[(a);(b);\lambda(s-1)/(1+c/d)],$$

and

$${}_pF_q[(a);(b);\lambda(s-1)/(1+c/dk)],$$

respectively.

For the Laplace-Haag distribution, we find that as both N and a become large, the penultimate limiting form of the p.g.f. is ${}_1F_0[-n;;a(1-s)/N]$, i.e., it is binomial and, as is well known, when both n and N become large, the ultimate limiting distribution is Poisson with parameter an/N.

Consider now the Levene [15] type matching scheme, where one pack of composition (2^N) is randomly split into two equally sized packs. Barton [2] shows that the factorial moments for the distribution of the number of matches are

$$\mu'_{[r]}=\frac{N!(N-\frac{1}{2}-r)!}{(N-r)!(N-\frac{1}{2})!2^r},$$

whence the p.g.f. can be obtained as ${}_1F_1[-N;\frac{1}{2}-N;(s-1)/2]$; as N becomes large, this tends to $\exp[N(s-1)/(2N-1)]$, and so to $\exp[(s-1)/2]$. Barton proceeds to give the generalization to a pack of composition (K^N), randomly split into K equally sized packs; he shows that if a match is defined as existing when all K cards in the same position are of the same kind, then in this case the factorial moments are

$$\mu'_{[r]}=\frac{N!N!(K!)^r(KN-Kr)!}{(N-r)!(N-r)!(KN)!}.$$

Hence, using the Gauss-Legendre multiplication formula for the γ-function, we deduce that the p.g.f. is

$${}_1F_{K-1}\left[-N;\frac{1}{K}-N,\frac{2}{K}-N,\ldots,\frac{K-1}{K}-N;\frac{K!(s-1)}{(-K)^K}\right].$$

As N becomes large, this tends, *via* the Poisson distribution with parameter $(N^2/{}_{NK}C_K)$, to the degenerate distribution with all probability concentrated at zero.

Suppose now that the initial pack with composition (K^N) is split into K packs each of composition (1^N); Anderson [1] shows that the probabilities are now

$$P_{[r]} = \sum_{j=0}^{N-r} \frac{(-1)^j}{r!j!}\left[\frac{(N-r-j)!}{N!}\right]^{K-2},$$

whence we deduce that the new p.g.f. is

$${}_1F_{K-1}[-N;\ -N, \ldots, -N; (-1)^K(s-1)].$$

As N becomes large, this p.g.f. tends, *via* the Poisson distribution with parameter $(1/N^{K-2})$, to the degenerate distribution.

3. Occupancy distributions. David and Barton ([5], p. 243), using the method of indicator functions, derived the factorial moments for the number of classes unoccupied in the classical occupancy problem as

$$\mu'_{[r]} = \frac{k!}{(k-r)!}\left(\frac{k-r}{k}\right)^N,$$

where N is the number of objects and k is the total number of classes (cells). We can deduce from these factorial moments that the p.g.f. for the number of unoccupied classes is

$${}_NF_{N-1}[1-k, \ldots, 1-k,;\ -k, \ldots, -k; 1-s].$$

Note that when $N < k$, the first $k-N$ probabilities, i.e., $P_{[0]}, \ldots$ $\ldots, P_{[k-N-1]}$, are zero.

David and Barton ([5], p. 251) examine the problem of specified occupancy. Let N objects be assigned to k classes l of which (less than k) are specified. David and Barton show that the distribution of the number of empty classes amongst the l specified classes has the factorial moments

$$\mu'_{[r]} = \frac{l!}{(l-r)!}\left(\frac{k-r}{k}\right)^N.$$

The corresponding p.g.f. is

$${}_{N+1}F_N[-l, 1-k, \ldots, 1-k;\ -k, \ldots, -k; 1-s],$$

which reduces, of course, to the previous p.g.f. for $l = k$; for k large (and $N \geqslant l$) the p.g.f. tends, *via* the binomial distribution with p.g.f.

$${}_1F_0[-l;\ ;(k-1)^N(1-s)/k^N],$$

to the degenerate distr/bution with probability that all l specified classes are empty equal to unity.

This distribution is treated again in [5], p. 269 ff., under the title "golliwog problem"; it is also equivalent to the randomized occupancy distribution of Harkness [10] and Kotz and Srinivasan [14] in which N objects are assigned at random to k classes, and each is then allowed to stay in its class with probability p and to escape with probability $1-p$. In this notation the p.g.f. becomes

$$ {}_{N+1}F_N[-k, 1-k/p, \ldots, 1-k/p; -k/p, \ldots, -k/p; 1-s]. $$

4. Irwin's line division problem. Irwin's [11] line division problem concerns the distribution of the number of intervals greater than u/k when a line of fixed length u is divided into n intervals by $n-1$ randomly placed points. He shows, using the Whitworth theorem form of the inclusion-exclusion principle, that the probability that exactly r of the intervals exceed u/k is

$$ P_{[r]} = {}_nC_r \sum_{j=0}^{[k]-r} {}_{n-r}C_j(-1)^j\left[1-\frac{r+j}{k}\right]^{n-1}. $$

This expression for the probabilities yields the p.g.f.

$$ {}_{n+1}F_n\big[-[k], -n, 1-k, \ldots, 1-k; -[k], -k, \ldots, -k; 1-s\big], $$

where $[k]$ denotes the integer part of k. (It is necessary to introduce $-[k]$ into the p.g.f. since $P_{[r]} = 0$ when $r > [k]$.) When k is an integer, the p.g.f. reduces to

$$ {}_nF_{n-1}[-n, 1-k, \ldots, 1-k; -k, \ldots, -k; 1-s], $$

which closely resembles the p.g.f. for the problem of specified (randomized) occupancy. Note that $P_{[0]} = 0$ unless $n > k$.

References

[1] T. W. Anderson, *On card matching*, Ann. Math. Statist. 14 (1943), p. 426-435.

[2] D. E. Barton, *The matching distributions: Poisson limiting forms and derived methods of approximation*, J. Roy. Stat. Soc. B20 (1958), p. 73-92.

[3] T. S. Broderick, *On some symbolic formulae in probability theory*, Proc. Roy. Irish Acad. 44 (1937), p. 19-28.

[4] E. Czuber, *Wahrscheinlichkeitsrechnung*, Vols. 1 and 2, Teubner, Leipzig 1908.

[5] F. N. David and D. E. Barton, *Combinatorial chance*, Griffin, London 1962.

[6] W. Feller, *An introduction to probability theory and its applications*, Vol. 1, Wiley, New York 1957.

[7] M. Fréchet, *Les probabilités associées à un système d'événements compatibles et dépendants, 1. Evénements en nombre fini fixe*, Actualités Scientifiques et Industrielles 859, Hermann, Paris 1940.

[8] — *Les probabilités associées à un système d'événements compatibles et dépendants, 2. Cas particuliers et applications*, ibidem 942, Hermann, Paris 1943.

[9] F. Garwood, *An application of the theory of probability to the operation of vehicular-controlled traffic signals*, J. Roy. Stat. Soc. Suppl. 7 (1940), p. 65-77.

[10] W. L. Harkness, *The classical occupancy problem revisited* in *Random counts in scientific work*, Vol. 3, p. 107-126, The Pennsylvania State University Press 1970.

[11] J. O. Irwin, *A unified derivation of some well-known frequency distributions of interest in biometry and statistics*, J. Roy. Stat. Soc. A118 (1955), p. 389-398.

[12] A. W. Kemp, *A wide class of discrete distributions and the associated differential equations*, Sankhyā A30 (1968), p 401-410.

[13] — and C. D. Kemp, *A family of discrete distributions defined via their factorial moments*, Commun. Statist. 3 (1974), p. 1187-1196.

[14] S. Kotz and R. Srinivasan, *Randomized occupancy models*, paper presented at NATO Advanced Study Institute, University of Calgary, August 1974.

[15] H. Levene, *On a matching problem arising in genetics*, Ann. Math. Statist. 20 (1949), p. 91-94.

[16] J. F. Steffensen, *Factorial moments and discontinuous frequency-functions*, Skand. Aktuarietidskr. 6 (1923), p. 73-89.

SCHOOL OF MATHEMATICS
UNIVERSITY OF BRADFORD
BRADFORD BD7 1DP, ENGLAND

Received on 5. 5. 1976

ADRIENNE W. KEMP (Bradford)

O FUNKCJACH TWORZĄCYCH PRAWDOPODOBIEŃSTWA DLA ROZKŁADÓW SKOJARZENIA I ROZMIESZCZENIA

STRESZCZENIE

W pracy rozpatruje się trzy rozkłady, które Irwin [11] wyprowadził z twierdzenia Whitwortha: klasyczny rozkład skojarzenia, klasyczny rozkład rozmieszczenia oraz pewien rozkład podziału prostej. Pokazuje się, że rozkłady te oraz pewne ich ważne uogólnienia należą do klasy rozkładów, mających funkcje tworzące prawdopodobieństwa postaci ${}_pF_q[\lambda(s-1)]$, a nazywanych w [13] *generalized hypergeometric factorial-moment distributions*.
