The Annals of Statistics
1974, Vol. 2, No. 5, 977-987

SPARSE AND CROWDED CELLS AND DIRICHLET DISTRIBUTIONS

BY MILTON SOBEL[1] AND V. R. R. UPPULURI

University of Minnesota and Oak Ridge National Laboratory[2]

1. Introduction. In recent years a number of applications have been found for the Dirichlet distributions; another application is considered in this paper. A multinomial distribution with k cells is given with b cells ($1 \leqq b \leqq k$) having common cell probability p ($0 < p \leqq 1/b$); these are called blue cells. Dual concepts of sparseness and crowdedness are introduced for these b blue cells based on a fixed number n of observations. The (Type 1) Dirichlet distribution is used to evaluate the probability laws, the cumulative distribution functions (cdf's), the moments, the joint probability law and the joint moments of the number S of sparse blue cells and the number C of crowded blue cells. The results are put in the form of moment generating functions. Applications of some of these results are also considered in Sections 7 and 8.

2. The distribution of S. A sparse blue cell is one with at most u observations in it. A crowded blue cell is one with at least v observations in it. Let $S_{b,p}^{(u,n)} = S$ denote the random number of sparse blue cells when there are b blue cells with common probability p, n observations, and u defines sparseness; similarly, let $C_{b,p}^{(v,n)} = C$ denote the random number of crowded blue cells with v defining crowdedness. We use the symbolism max $(j, n) \leqq u$ (for integers u) to denote the event that the maximum frequency (based on n observations) in a specified set of j blue cells is at most u; similarly, min $(j, n) \geqq v$ (for integers v) denotes the event that the minimum frequency (based on n observations) in a specified set of j blue cells is at least v. It has already been noted elsewhere (cf. e.g., [2] and [4]) that

$$
\begin{aligned}
(2.1) \qquad & P\{\min(j, n) \geqq v \mid p\} \\
&= I_p^{(j)}(v, n) \\
&= \frac{\Gamma(n+1)}{\Gamma^j(v)\Gamma(n+1-jv)} \int_0^p \cdots \int_0^p \Big(1 - \sum_{\alpha=1}^{j} x_\alpha\Big)^{n-jv} \prod_{\alpha=1}^{j} x_\alpha^{v-1}\, dx_\alpha ,
\end{aligned}
$$

where $0 \leqq p \leqq 1/b \leqq 1/j$. For $j = 1$ it is easily seen that $I_p^{(1)}(v, n) = I_p(v, n - v + 1)$, where the latter is the usual incomplete beta function.

It is clear that $P\{S = s \mid b, p, u, n\}$ is the probability that in exactly s (out of b) cells the frequency (based on n observations) is at most u. Using the method

Received October 1972; revised September 1973.

[1] Research supported by NSF Grant GP-28922X at the University of Minnesota and by subcontract No. 3325 with the Oak Ridge National Laboratory.

[2] Supported by the U.S. Atomic Energy Commission under contract with the Union Carbide Corporation.