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

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

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.

of inclusion-exclusion, we obtain for $0 \leqq s \leqq b$

$$(2.2)\qquad \begin{aligned} P\{S = s \mid b, p, u, n\} &= \binom{b}{s} \sum_{\gamma=0}^{s} (-1)^{\gamma} \binom{s}{\gamma} P\{\min(b - s + \gamma, n) \geqq u + 1 \mid p\} \\ &= \binom{b}{s} \sum_{\gamma=0}^{s} (-1)^{\gamma} \binom{s}{\gamma} I_p^{(b-s+\gamma)}(u + 1, n), \end{aligned}$$

where $I_p^{(0)}(u + 1, n) \equiv 1$ by definition for all $u \geqq 0$, $p \geqq 0$ and $n \geqq 0$.

For the special case $s = b$ and $p = 1/b$ (so that $k = b$), it is clear that the value in (2.2) equals zero when $n > bu$ and that it equals one when $n \leqq u$. In this case we are dealing with the probability that the maximum frequency in a homogeneous multinomial is at most u and this probability was tabulated by Steck [6]; thus the concept of sperse blue cells is a direct generalization of the maximum frequency in a homogeneous multinomial. For $p < 1/b$ we can assume that $k = b + 1$.

Another use of the sparse concept is to generalize the so-called "empty-cell test." By considering the number of sparse cells as our statistic instead of the number of empty cells, we can improve the power of the test of homogeneity, (i.e., the test that all the cells have the same probability p) against certain alternatives. This application will be discussed below.

From (2.2) we can also get a fairly simple expression for the cdf of S. Replacing s by t in (2.2) and summing t from 0 to s, we obtain by straightforward algebra

$$(2.3)\qquad \begin{aligned} P\{S \leqq s \mid b, p, u, n\} &= \sum_{\alpha=0}^{s} (-1)^{\alpha} \binom{b}{\alpha} \sum_{t=\alpha}^{s} \binom{b-\alpha}{t-\alpha} I_p^{(b-(t-\alpha))}(u + 1, n) \\ &= (b - s)\binom{b}{s} \sum_{\alpha=0}^{s} \frac{(-1)^{\alpha}\binom{s}{\alpha}}{b - s + \alpha} I_p^{(b-s+\alpha)}(u + 1, n); \end{aligned}$$

for $s = b$ the result is one. Thus we find that both the individual probabilities and the cdf of S are expressible through the (Type 1) Dirichlet functions, $I_p^{(j)}(v, n)$, for $p \leqq 1/b$ and $n \geqq jv$.

3. The distribution of C. As a dual to the concept of sparseness, we now consider the concept of crowdedness; the results are quite similar and we omit the intermediate steps. It is clear that $P\{C = c \mid b, p, v, n\}$ is the probability that in exactly c (out of b) cells the frequency (based on n observations) is at least v. Using the method of inclusion-exclusion, we obtain for $0 \leqq c \leqq b$

$$(3.1)\qquad \begin{aligned} P\{C = c \mid b, p, v, n\} &= \binom{b}{c} \sum_{\alpha=0}^{b-c} (-1)^{\alpha} \binom{b-c}{\alpha} P\{\min(c + \alpha, n) \geqq v \mid p\} \\ &= \binom{b}{c} \sum_{\alpha=0}^{b-c} (-1)^{\alpha} \binom{b-c}{\alpha} I_p^{(c+\alpha)}(v, n). \end{aligned}$$

For the special case $c = b$ this reduces to $I_p^{(b)}(v, n)$ and for any $p \leqq 1/b$ its value is zero if $n < bv$. In this case we are dealing with the minimum frequency among b cells with common cell probability p in a multinomial distribution with $b + 1$ cells.

For the cdf of C we first note that for $c \geqq b$ the result is clearly unity. Then

for $c < b$

$$P\{C \leqq c \mid b, p, v, n\}$$

$$= 1 - \sum_{t=c+1}^{b} \binom{b}{t} \sum_{\alpha=0}^{b-t} (-1)^{\alpha} \binom{b-t}{\alpha} I_p^{(t+\alpha)}(v, n) \tag{3.2}$$

$$= 1 - (b - c)\binom{b}{c} \sum_{\alpha=0}^{b-c-1} \frac{(-1)^{\alpha}}{c + 1 + \alpha} \binom{b - 1 - c}{\alpha} I_p^{(c+1+\alpha)}(v, n) .$$

If $c = b - 1$ then (3.2) reduces to $1 - I_p^{(b)}(v, n)$.

4. Joint distribution results. To conserve space we give without derivation an expression for the joint distribution of S and C (see [5] for details).

Let $(F(s, c) = P\{\max(s, n) \leqq u, \min(c, n) \geqq v\}$. Then

$$F(s, c) = \sum_{\alpha=0}^{n} b_{\alpha}(n, sp) Q(s, \alpha) I_{p/(1-sp)}^{(c)}(v, n - \alpha) \tag{4.1}$$

where $Q(s, \alpha)$, which does not depend on p, is given for $\alpha \geqq n$ by

$$Q(s, \alpha) = P\left\{\max(s, \alpha) \leqq u \,\middle|\, \frac{1}{s}\right\} = \sum_{\gamma=0}^{s} (-1)^{\gamma} \binom{s}{\gamma} I_{1/s}^{(\gamma)}(u + 1, \alpha) ; \tag{4.2}$$

for $\alpha < u$ the result in (4.2) is equal to 1. The joint pdf of S and C can be written in terms of $F(s, c)$ in (4.1) as

$$P\{S = s, C = c\}$$

$$= \sum_{\beta=0}^{b-s-c} (-1)^{\beta} \sum_{\alpha=0}^{\beta} \binom{s+\alpha}{\alpha} \binom{c+\beta-\alpha}{\beta-\alpha} \left[{}_{s+\alpha, c+\beta-\alpha}^{\ b}\right] F(S + \alpha, c + \beta - \alpha) \tag{4.3}$$

where $\left[{}_{s,c}^{\ b}\right]$ denotes the usual multinomial coefficient.

5. Moment of C and S. It has been previously pointed out in an unpublished technical report by Sobel [3] that the factorial moments of C can be simply expressed in terms of the Type 1 Dirichlet function (see also Barton and David [1]).

Let $E\{C^{[m]}\}$ denote the mth factorial moment of C, let $b^{[m]} = b(b - 1) \cdots (b - m + 1)$ and let M denote the largest integer contained in n/m. For $0 \leqq m \leqq M$ and $n \geqq mv$ the result is

$$E\{C^{[m]}\} = b^{[m]} I_p^{(m)}(v, n) , \tag{5.1}$$

where the second factor does not depend on b. As a corollary we obtain the first moment and variance of C. For $m = 1$, $n \geqq v$ and $p \leqq 1/b$, we have

$$E\{C\} = bI_p^{(1)}(v, n) = bI_q(v, n - v + 1) . \tag{5.2}$$

For $m = 2$, $n \geqq 2v$ and $p \leqq 1/b$, we have

$$E\{C(C - 1)\} = b(b - 1)I_p^{(2)}(v, n) , \tag{5.3}$$

$$\sigma^2(C) = b(b - 1)I_p^{(2)}(v, n) + bI_p^{(1)}(v, n) - \{bI_p^{(1)}(v, n)\}^2 . \tag{5.4}$$

For the special case $v = 1$ and $u = 0$ this gives for the number C of occupied cells and the number S of empty cells

$$E\{C\} = b(1 - q^n) = b - E\{S\} \tag{5.5}$$

$$\sigma^2(C) = b(b-1)[1-2q^n+(q-p)^n] + b(1-q^n) - b^2(1-q^n)^2 \tag{5.6}$$
$$= bq^n(1-bq^n) + b(b-1)(q-p)^n$$
$$= \sigma^2(S);$$

these also give correct answers for $n = 0$ and $n = 1$.

The factorial moments of S can be obtained in two different ways, both of which are useful and make use of I-functions. One uses the idea that $b - S$ is the number of crowded cells if crowdedness is defined by having a frequency of at least $u + 1$. Hence from (5.1)

$$E\{(b-S)^{[m]}\} = b^{[m]} I_p^{(m)}(u+1, n). \tag{5.7}$$

Another method is to use the 'binomial theorem for factorial powers', namely the identity for any b, c

$$(b-c)^{[m]} = \sum_{\alpha=0}^{m} (-1)^\alpha \binom{m}{\alpha} (b-\alpha)^{[m-\alpha]} c^{[\alpha]}. \tag{5.8}$$

(This identity is easily proved by induction; we omit the proof.) Putting C for c and then S for $b - C$, we obtain from (5.8) and (5.1)

$$E\{S^{[m]}\} = b^{[m]} \sum_{\alpha=0}^{m} (-1)^\alpha \binom{m}{\alpha} I_p^{(\alpha)}(u+1, n). \tag{5.9}$$

In (5.9) we used the fact that if $v = u + 1$ then $b - C = S$ identically, but we note that this was not needed in (5.7).

Some special cases of these results are included for completeness. For $m = 1$, $p = 1 - q \leqq 1/b$ and $n \geqq u + 1$, we have from (5.7)

$$E\{S\} = b[1 - I_p^{(1)}(u+1, n)] \tag{5.10}$$
$$= b[1 - I_p(u+1, n-u)] = bI_q(n-u, u+1);$$

for $n \leqq u$ all cells are sparse and hence $E\{S\} = b$. For $m = 2$, $p \leqq 1/b$ and $n \geqq 2(1+u)$, we have

$$E\{S(S-1)\} = b^{[2]}[1 - 2I_p^{(1)}(u+1, n) + I_p^{(2)}(u+1, n)]. \tag{5.11}$$

The variance of S is identical with the result in (5.4) for $\sigma^2(C)$ if we replace v by $u + 1$. The results, after integrations, are generally found to hold also for $n < 2(u+1)$. Results for $u = 0$ are given in (5.5) and (5.6). A more explicit expression for $E\{S(S-1)\}$ for $n \geqq u + 1 > 0$ and (only) for $p = 1/b$ is given by David and Barton ([1], page 279).

6. Joint moment results. It can also be shown (see [5] for details) using (4.3) and (5.8) that

$$E\{S^{[g]}C^{[h]}\} = b^{[g+h]} F(g, h). \tag{6.1}$$

$$E\{C^{[h]}(b-h-S)^{[g]}\} = b^{[g+h]} \sum_{\alpha=0}^{g} (-1)^\alpha \binom{g}{\alpha} F(\alpha, h) \tag{6.2}$$
$$= b^{[g+h]} I_p^{(g+h)}((u+1)_g, (v)_h, n)$$

where $(v)_h$ denotes a repetition h times of the argument v and

$$
\begin{aligned}
&I_p^{(\alpha+\beta)}((t)_\alpha, v_\beta, n) \\
&\quad = \frac{\Gamma(n+1)}{\Gamma^\alpha(t)\Gamma^\beta(v)\Gamma(n+1-\alpha t-\beta v)} \\
&\qquad \times \int_0^p \cdots \int_0^p (1-\textstyle\sum_{i=1}^{\alpha+\beta} x_i)^{n-\alpha t-\beta v} \prod_{i=1}^{\alpha} x_i^{t-1}\, dx_i \prod_{j=1}^{\beta} x_{\alpha+j}^{v-1}\, dx_{\alpha+j}\,;
\end{aligned} \tag{6.3}
$$

this is a generalization of the Dirichlet integral in (2.1).

From (6.1) and (6.2) with $g = h = 1$ and (5.1) we can write

$$
E\{SC\} = b^{[2]}F(1,1) = b^{[2]}[I_p^{(1)}(v,n) - I_p^{(2)}(u+1,v,n)]\,. \tag{6.4}
$$

For $u = 0$ and any $v \geqq 1$ we get further simplification here; by straightforward integration we easily obtain

$$
E\{SC \mid u = 0,\ v \geqq 1\} = b^{[2]}q^n I_{p/q}^{(1)}(v,n) = b^{[2]}q^n I_{p/q}(v, n-v+1)\,. \tag{6.5}
$$

For example, if $b = 3$, $p = 1/b = \frac{1}{3}$, $u = 0$ and $v = 2$ then for any $n > 2$ the result is by (6.5)

$$
E\{SC \mid u = 0,\ v = 2\} = 6(\tfrac{2}{3})^n I_{\frac{1}{2}}(2, n-1) = \frac{2(2^n - n - 1)}{3^{n-1}}\,. \tag{6.6}
$$

Another way of combining these moment results is to write them in the form of decreasing factorial moment generating functions (df mgf). For C from (3.1) we easily obtain

$$
\begin{aligned}
E\{(1+t)^{[C]}\} &= \textstyle\sum_{c=0}^{b} \binom{b}{c}(1+t)^c \sum_{\alpha=0}^{b-c}(-1)^\alpha \binom{b-c}{\alpha} I_p^{(c+\alpha)}(v,n) \\
&= \textstyle\sum_{\beta=0}^{b}(-1)^\beta I_p^{(\beta)}(v,n)\binom{b}{\beta}\sum_{c=0}^{\beta}(-1)^c\binom{\beta}{c}(1+t)^c \\
&= \textstyle\sum_{\beta=0}^{b}\binom{b}{\beta}t^\beta I_p^{(\beta)}(v,n)\,.
\end{aligned} \tag{6.7}
$$

The df mgf that gives rise to the moments in (6.2) is

$$
E\left\{\left(1+\frac{t_2}{1+t_1}\right)^C (1+t_1)^{b-S}\right\} = \sum_{g,h}\frac{t_1^g}{g!}\frac{t_2^h}{h!}E\{C^{[h]}(b-S-h)^{[g]}\} \tag{6.8}
$$

and hence by our results in (6.2) we must have

$$
\begin{aligned}
&E\{(1+t_1+t_2)^C(1+t_1)^{b-S-C}\} \\
&\qquad = \sum_{g,h}\frac{t_1^g}{g!}\frac{t_2^h}{h!}\,b^{[g+h]}I_p^{(g+h)}((u+1)_g, (v)_h, n)\,.
\end{aligned} \tag{6.9}
$$

REMARK 1. In the special case $v = u + 1$ we have $F(s,c) = 0$ unless $s + c = b$. Starting with the probability that c cells are crowded, we subtract the probability that $c+1$ cells (including these c) are crowded etc., obtaining for a specified partition of the b cells (with $b = s + c \geqq 1$ and $v = u + 1$)

$$
F(s,c) = I_p^{(c)}(v,n) - \binom{s}{1}I_p^{(c+1)}(v,n) + \cdots + (-1)^s\binom{s}{s}I_p^{(b)}(v,n)\,. \tag{6.10}
$$

This agrees with (2.2) for $c = 0$ and with (3.1) for $s = 0$; it also agrees with (4.1) for $c = s = 0$, if we define the probabilities in (4.2) involving $\frac{1}{0}$ as being equal to one.

REMARK 2. The above method can also be used for the general case, $u < v$ and $s + c \leqq b$, if we use the more general Dirichlet integral introduced in (6.3). We use the result proved in [2] that for a single set of n observations

$$I_p^{(c+\alpha)}((v)_c, (u+1)_\alpha, n) = P\{\min(c, n) \geqq v, \min(\alpha, n) \geqq u+1\} \tag{6.11}$$

is the probability that a specified set of c cells have frequency at least v and a disjoint specified set of α cells have frequency at least $u + 1$ (where $c + \alpha \leqq b$). Hence by inclusion-exclusion we obtain for the general case

$$F(s, c) = \sum_{\alpha=1}^{s} (-1)^\alpha \binom{s}{\alpha} I_p^{(c+\alpha)}((v)_c, (u+1)_\alpha, n)\,, \tag{6.12}$$

which is equal to (4.1) and somewhat simpler than (4.1) for small values of s. For $v = u + 1$ and $s + c = b$ this clearly reduces to (6.10).

REMARK 3. One advantage of (4.1) is that it gives us a limit theorem as $n \to \infty$, $n/b \to \alpha > 0$ and $p \leqq 1/b \to 0$ like $1/b$, i.e., $\lim pb > 0$. We take $c = 0$ in (4.1) so that the final I-function in (4.1) is logically replaced by one. Then the binomial $b_\alpha(n, sp)$ for fixed s tends to a Poisson distribution with parameter $\lambda = s$. Hence from (4.2), letting $\max f(s)$ denote the maximum frequency for a specified set of s cells, we obtain for $u \geqq 1$

$$\begin{aligned} \lim P\{\max f(s) \leqq u \mid p\} &= \sum_{x=0}^{\infty} e^{-s} \frac{s^x}{x!} P\left\{\max(s, x) \leqq u \,\middle|\, \frac{1}{s}\right\} \\ &= e^{-s} \sum_{\gamma=0}^{s} \sum_{x=x_0}^{\infty} \frac{s^x}{x!} (-1)^\gamma \binom{s}{\gamma} I_{1/s}^{(\gamma)}(u+1, x) \end{aligned} \tag{6.13}$$

where $x_0 = \gamma(u + 1)$ and the I-function is clearly equal to one for $x = 0$. If we sum on x under the integral sign, we obtain

$$\begin{aligned} \lim F(s, 0) &= \sum_{\gamma=0}^{s} (-1)^\gamma \binom{s}{\gamma} \left[\frac{1}{u!} \int_0^1 t^u e^{-t}\, dt\right]^\gamma = \left[1 - \frac{1}{u!} \int_0^1 t^u e^{-t}\, dt\right]^s \\ &= \left[\frac{1}{e} \sum_{j=0}^{u} \frac{1}{j!}\right]^s = P\{MP_\lambda(s) \leqq u \mid \lambda = 1\} = G(u) \quad \text{(say)}\,, \end{aligned} \tag{6.14}$$

where $MP_\lambda(s)$ is the maximum of s independent Poisson chance variables all with parameter λ.

Hence the asymptotic expected maximum frequency for the above limiting process is the same as the expectation of $MP_\lambda(s)$ given $\lambda = 1$. For $s = 1$ in (6.14) it is clear that $E\{MP_\lambda(1)\} = \lambda = 1$ and for $s \geqq 2$ an approximate solution for $E\{MP_\lambda(s)\}$ is obtained by finding the decimal value of u that solves the equation

$$1 - \frac{1}{\Gamma(u+1)} \int_0^1 t^u e^{-t}\, dt = s/(s+1)\,. \tag{6.15}$$

Since the left side varies from $1/e$ to 1 and $s/(s + 1) \geqq \frac{1}{2} > 1/e$ for $s \geqq 1$, the solution in u will be unique. A good starting value in searching for this root is the decimal solution for u of the equation $\Gamma(u + 2) = s/e$ and this describes the rate at which $u \to \infty$ as $s \to \infty$, i.e., it shows that u grows like $\log s/(\log \log s)$

as $s \to \infty$. In the absence of exact tables for $E\{MP_\lambda(s)\}$ for $\lambda = 1$, the approximation (6.15) should be useful for a wide range of s-values.

A more simple-minded approximation for $E\{MP_1(s)\}$ is to solve for u and θ (u an integer and $0 \leqq \theta < 1$) the equation

$$e^{-1} \sum_{j=0}^{u} \frac{1}{j!} + \theta e^{-1} \frac{1}{(u+1)!} = \frac{s}{s+1}\,; \tag{6.16}$$

this can be done with a table of the Poisson cdf for $\lambda = 1$. Then $u + \theta$ is the required approximation.

REMARK 4. If we use (4.1) with $s = 0$ (or (3.1) with $b = c$) to find the expected minimum frequency in c specified cells, we first obtain the exact results for $p \leqq 1/c$ and n observations

$$E\{\min f(c) \mid n, p\} = \sum_{j=1}^{J} I_p^{(c)}(j, n)\,, \tag{6.17}$$

where J is the integer part of n/c; the sum clearly vanishes for $J = 0$. Using normal approximations to $I_p^{(c)}(v, n)$ as $n \to \infty$, $n/b \to \alpha > 0$ and $p \leqq 1/b \to 0$ like $1/b$ with c and v fixed, we find that

$$\lim I_p^{(c)}(v, n) = \left[1 - \Phi\left(\frac{\alpha - v}{v^{\frac{1}{2}}}\right)\right]^c, \tag{6.18}$$

where $\Phi(x)$ is the standard normal cdf.

7. An application to the empty-cell test. In this section we illustrate the changes in power if we replace the empty-cell test (ECT) by the sparse-cell test (SCT); both of these tests are facilitated by the use of a table of the Type 1 Dirichlet distribution. The changes can take place in both directions, depending on the alternative being considered.

Suppose we have a multinomial with (say) $k = b = 10$ cells and $n = 40$ observations; we wish to test the hypothesis $H_0\colon p_1 = p_2 = \cdots = p_{10} = \frac{1}{10}$. One alternative of interest is $H_1\colon p_1 = p_2 = \cdots = p_9 = p < \frac{1}{10}$ and $p_{10} = 1 - 9p$; another alternative of interest is $H_2\colon p_1 = p_2 = \cdots = p_9 = \frac{1}{9}$ and $p_{10} = 0$. We shall not attempt to get the best sparse-cell test but merely use $u = 1$ to define sparseness, as opposed to the empty-cell test where we have to use $u = 0$; in this latter case S becomes the number of empty cells.

For the empty-cell test we use (2.2) and (2.3) to find the smallest integer s_0 such that

$$P\{S \leqq s_0 \mid 10, \tfrac{1}{10}, 0, 40\} \geqq P^* \tag{7.1}$$

for preassigned P^*; we will use $P^* = .95$. From an unpublished table of the Dirichlet distribution we obtain

$$\begin{aligned} P\{S = 0\} &= I_{1/10}^{(10)}(1, 40) = .8581 \\ P\{S \leqq 1\} &= 10 I_{1/10}^{(9)}(1, 40) - 9 I_{1/10}^{(10)}(1, 40) = .9942\,. \end{aligned} \tag{7.2}$$

In order to attain a test size of exactly .05 we reject H_0 if $S \geqq 2$ and also with

probability p_0 when $S = 1$; then p_0 is found by setting

$$p_0(.9941 - .8581) + (1 - .9941) = .05 \tag{7.3}$$

and we easily find that $p_0 = .324$. To write the power of this test against H_1 we let S_9 denote the value of S when $b = 9$ as in H_1 and use $f(10)$ to denote the frequency in the tenth cell. Then

$$\begin{aligned} P_1 = \text{Power (ECT vs. } H_1) &= P\{S_9 \geqq 2\} + .324P\{S_9 = 1, f(10) > 0\} \\ &\quad + P\{S = 1, f(10) = 0\} + .324P\{S = 0, f(10) = 0\} . \end{aligned} \tag{7.4}$$

The last two terms are less than $(\frac{9}{20})^{40} < 10^{-12}$ and do not affect our calculations. Using (2.2) we obtain

$$\begin{aligned} P_1 &= 1 - P\{S = 0 \mid 9, \tfrac{1}{20}, 0, 40\} - .676P\{S = 1 \mid 9, \tfrac{1}{20}, 0, 40\} \\ &= 1 - I^{(9)}_{1/20}(1, 40) - (.676)9[I^{(8)}_{1/20}(1, 40) - I^{(9)}_{1/20}(1, 40)] = .4580 . \end{aligned} \tag{7.5}$$

For the corresponding sparse-cell test we take $u = 1$ and again use (2.2) and (2.3), obtaining for H_0

$$\begin{aligned} P\{S = 0\} &= I^{(10)}_{1/10}(2, 40) = .3858 \\ P\{S \leqq 1\} &= 10I^{(9)}_{1/10}(2, 40) - 9I^{(10)}_{1/10}(2, 40) = .8296 \\ P\{S \leqq 2\} &= 45I^{(8)}_{1/10}(2, 40) - 80I^{(9)}_{1/10}(2, 40) + 36I^{(10)}_{1/10}(2, 40) = .9800 . \end{aligned} \tag{7.6}$$

Hence we reject H_0 if $S \geqq 3$ and also with probability p_0' if $S = 2$; it is easily seen that $p_0' = .200$. The power calculation against H_1 is

$$\begin{aligned} P_1' = \text{Power (SCT vs. } H_1) &= P\{S_9 \geqq 3\} + .2\mathrm{P}\{S_9 = 2, f(10) > 1\} \\ &\quad + P\{S = 2, f(10) \leqq 1\} + .2P\{S = 1, f(10) \leqq 1\} . \end{aligned} \tag{7.7}$$

As in the previous case we can omit the last two terms; an additional reason here is that they can only improve our result. Using (2.2) and (2.3) we obtain

$$\begin{aligned} P_1' &= 1 - P\{S \leqq 1 \mid 9, \tfrac{1}{20}, 1, 40\} - .8P\{S = 2 \mid 9, \tfrac{1}{20}, 1, 40\} \\ &= 1 - [9I^{(8)}_{1/20}(2, 40) - 8I^{(8)}_{1/20}(2, 40)] - (.8)36[I^{(7)}_{1/20}(2, 40) \\ &\quad - 2I^{(8)}_{1/20}(2, 40) + I^{(9)}_{1/20}(2, 40)] = .8372 . \end{aligned} \tag{7.8}$$

Thus the sparse-cell with $u = 1$ already gives a better power against H_1; calculations for $u \geqq 2$ have not been carried out.

It should also be pointed out that under the alternative H_2 (or others like it with $b' < b$ cells having common probability $1/b'$ and $b - b'$ cells with probability zero) the empty-cell test is preferable, i.e., $u = 0$ gives a better power against H_2 than $u \geqq 1$; it suffices to consider the case $u = 1$. The power calculations against H_2 for the empty-cell test are

$$\begin{aligned} P_2 = \text{Power (ECT vs. } H_2) &= P\{S \geqq 2 \mid H_2\} + .324P\{S = 1 \mid H_2\} \\ &= 1 - P\{S = 0 \mid H_2\} - .676P\{S = 1 \mid H_2\} \\ &= 1 - .676P\{S = 0 \mid 9, \tfrac{1}{9}, 0, 40\} = 1 - .676I^{(9)}_{1/9}(1, 40) \\ &= .3777 . \end{aligned} \tag{7.9}$$

For the sparse-cell test we again use (2.2) and obtain for the power against H_2

$$
\begin{aligned}
P_2' &= \text{Power (SCT vs. } H_2) = P\{S \geqq 3 \mid H_2\} + .2P\{S = 2 \mid H_2\} \\
&= 1 - P\{S \leqq 1 \mid H_2\} - .8P\{S = 2 \mid H_2\} \\
&= 1 - P\{S = 0 \mid 9, \tfrac{1}{9}, 1, 40\} - .8P\{S = 1 \mid 9, \tfrac{1}{9}, 1, 40\} \\
&= 1 - I_{1/9}^{(9)}(1, 40) - (.8)9[I_{1/9}^{(8)}(1, 40) - I_{1/9}^{(9)}(1, 40)] \\
&= .0171 .
\end{aligned} \tag{7.10}
$$

Thus for the extreme alternatives like H_2 the empty-cell test is much better than the sparse-cell test with $u = 1$, but for alternatives that leave some probability in the "odd" cells (like H_1) the sparse-cell test with some $u \geqq 1$ has much better power.

8. An application to clustering. Another application shows that the concepts of sparseness and crowdedness are related to the notion of clustering. Since we use 'cluster' for a set of cells in close proximity, we define the term 'crowded cluster' for a set of closely-grouped crowded cells (with crowded cell being defined in terms of v as above). Suppose we have a square T of size $t \times t$ (t an integer) marked off into unit cells, so that t^2 is our original total k. For a fixed positive integer $d \leqq t$ we define a cluster as any $d \times d$ square matrix of contiguous cells, so that there are $D = (t - d + 1)^2$ clusters in all. *A cluster is called crowded if each of the d^2 ($= k'$, say) cells in the cluster is crowded.* For some purposes we may also want to impose the additional condition that the cells bordering a crowded cluster are not crowded, but for our problem this condition does not affect the result and is omitted. Our problem is to compute the probability of having at least one crowded cluster (among the D), if all the $k = t^2$ cells have common probability $p \leqq 1/k$ and n is the total number of observations taken.

This problem was suggested by a model dealing with the formation of tumors (cancer cells) in animal tissue. Here the multinomial cell corresponds to the biological cell. The observation is a radiation 'hit' and a 'crowded' cell is one in which the number of hits is above some threshold value v. If too many cells in close proximity are crowded then the chances of forming a tumor at that location are close to certainty. The cells in close proximity are our cell clusters and a crowded cluster is the origin of the tumor. One interesting quantity for this application is the probability A of at least one crowded cluster, since one crowded cluster is sufficient to start the formation of a tumor.

To get the answer, A, to this, we apply inclusion-exclusion methods to the crowdedness of the individual clusters $L_1, L_2, \cdots, L_D$ and let the respective b-values (all equal to d^2 in our application) be denoted by $b_1, b_2, \cdots, b_D$. We let $P(b_i, b_j)$ denote the probability that the ith and jth cluster are both crowded. By (3.1) this event occurs if every cell in the clusters L_i and L_j is crowded. Thus

$$
P(b_i, b_j) = P\{L_i \cup L_j\} = I_p^{(|i \cup j|)}(v, n) \tag{8.1}
$$

where $L_i \cup L_j$ denotes the union of L_i and L_j and $|i \cup j|$ is the total number of unit cells in this union. By inclusion-exclusion

$$A = \sum_{i=1}^{D} P(b_i) - \sum_{i<j} P(b_i, b_j) + \cdots + (-1)^{D-1} P(b_1, b_2, \cdots, b_D) \tag{8.2}$$

and we need to know the frequency of the various overlaps if we select (say, 2) smaller squares from the larger square. These can be computed for certain pairs (t, d) and the formula (8.2) then simplifies somewhat further. For example, if $t = 4$ and $d = 2$ then $D = 9$, and we obtain after a careful geometric analysis

$$\sum_{i=1}^{D} P(b_i) = 9I_p^{(4)}(v, n), \tag{8.3}$$

$$\sum_{i<j} P(b_i, b_j) = 12I_p^{(6)}(v, n) + 8I_p^{(7)}(v, n) + 16I_p^{(8)}(v, n), \tag{8.4}$$

$$\begin{aligned}\sum_{i,j,\alpha} P(b_i, b_j, b_\alpha) = {} & 22I_p^{(8)}(v, n) + 16I_p^{(9)}(v, n) + 34I_p^{(10)}(v, n) \\ & + 4I_p^{(11)}(v, n) + 8I_p^{(12)}(v, n),\end{aligned} \tag{8.5}$$

$$\begin{aligned}\sum_{i,j,\alpha,\beta} P(b_i, b_j, b_\alpha, b_\beta) = {} & 4I_p^{(9)}(v, n) + 32I_p^{(10)}(v, n) + 32I_p^{(11)}(v, n) \\ & + 37I_p^{(12)}(v, n) + 12I_p^{(13)}(v, n) \\ & + 8I_p^{(14)}(v, n) + I_p^{(16)}(v, n),\end{aligned} \tag{8.6}$$

$$\begin{aligned}\sum_{i,j,\alpha,\beta,\gamma} P(b_i, b_j, b_\alpha, b_\beta, b_\gamma) = {} & 16I_p^{(11)}(v, n) + 37I_p^{(12)}(v, n) \\ & + 36I_p^{(13)}(v, n) + 28I_p^{(14)}(v, n) \\ & + 4I_p^{(15)}(v, n) + 5I_p^{(16)}(v, n),\end{aligned} \tag{8.7}$$

$$\begin{aligned}\sum p(b_{i_1}, b_{i_2}, \cdots, b_{i_6}) = {} & 4I_p^{(12)}(v, n) + 24I_p^{(13)}(v, n) + 34I_p^{(14)}(v, n) \\ & + 12I_p^{(15)}(v, n) + 10I_p^{(16)}(v, n),\end{aligned} \tag{8.8}$$

$$\sum P(b_{i_1}, b_{i_2}, \cdots, b_{i_7}) = 14I_p^{(14)}(v, n) + 12I_p^{(15)}(v, n) + 10I_p^{(16)}(v, n), \tag{8.9}$$

$$\sum P(b_{i_1}, b_{i_2}, \cdots, b_{i_8}) = 4I_p^{(15)}(v, n) + 5I_p^{(16)}(v, n), \tag{8.10}$$

$$P(b_1, \cdots, b_9) = I_p^{(16)}(v, n). \tag{8.11}$$

Using (8.2) to combine these, we obtain the answer A as a linear combination of eight I-functions, all with the same arguments p, v, n and only the superscript varying, namely

$$\begin{aligned}A = {} & 9I_p^{(4)}(v, n) - 12I_p^{(6)}(v, n) - 8I_p^{(7)}(v, n) + 6I_p^{(8)}(v, n) \\ & + 12I_p^{(9)}(v, n) + 2I_p^{(10)}(v, n) - 12I_p^{(11)}(v, n) + 4I_p^{(12)}(v, n).\end{aligned} \tag{8.12}$$

Note that the sum of the coefficients in equation (8.2) is $\binom{9}{i}$ $(i = 1, 2, \cdots, 9)$ and hence it should be one in (8.12); this is a partial check on (8.12).

If we had defined a crowded cluster to mean that it has at least one crowded cell, than the result is much simpler. The probability of at least one crowded cluster is then equal to the probability of at least one crowded cell in T and this is simply the complement of (2.2) with $s = b = 16$ and $u + 1 = v$.

9. Acknowledgment. The authors wish to thank Dr. David Hoel of the

National Institute of Environmental Health Sciences at Research Triangle Park, North Calolina, for suggesting the application discussed in Section 8.

REFERENCES

[1] DAVID, F. N. and BARTON, D. E. (1962). *Combinatorial Chance*. Griffin, London.

[2] OLKIN, I. and SOBEL, M. (1965). Integral expressions for tail probabilities of the multinomial and negative multinomial distributions. *Biometrika* **52** 167–179.

[3] SOBEL, M. (1967). Notes on a multiple occupancy problem. Technical Report No. 98, Dept. of Statistics, Univ. of Minnesota.

[4] SOBEL, M. and UPPULURI, V. R. R. (1972). On Bonferroni-type inequalities of the same degree for the probability of unions and intersections. *Ann. Math. Statist.* **43** 1549–1558.

[5] SOBEL, M. and UPPULURI, V. R. R. (1972). Sparse and crowded cells and Dirichlet distributions. Technical Report No. 183. Dept. of Statistics, Univ. of Minnesota.

[6] STECK, G. (*ca.* 1965). Table of the distribution of the maximum frequency in a homogeneous multinomial. (Unpublished table—personal communication.)

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455

OAK RIDGE NATIONAL LABORATORY
OAK RIDGE, TENNESSEE 37830