SPARSE AND CROWDED CELLS AND DIRICHLET DISTRIBUTIONS

by

Milton Sobel[1] and V. R. R. Uppuluri
University of Minnesota Oak Ridge National Laboratory[2]

Technical Report No. 183

September 1972

University of Minnesota
Minneapolis, Minnesota

[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.

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 $(1 \leq b \leq k)$ having common cell probability p $(0 < p \leq 1/b)$; 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 (c.d.f.'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 at the end of Section 6. Applications of some of these results are considered in Sections 7 and 8. Corresponding sequential or waiting-time problems will be treated in a separate paper.

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 $\text{Max}(j, n) \leq u$ (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, $\text{Min}(j, n) \geq v$ (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

$$(2.1) \qquad P\{\mathrm{Min}(j, n) \geq v|p\} = I_p^{(j)}(v,n)$$
$$= \frac{\Gamma(n+1)}{\Gamma^j(v)\Gamma(n+1-jv)} \int_0^p \cdots \int_0^p (1-\sum_{\alpha=1}^{j} x_\alpha)^{n-jv} \prod_{\alpha=1}^{j} x_\alpha^{v-1} dx_\alpha,$$

where $0 \leq p \leq 1/b \leq 1/j$, since $j \leq b$. A generalization of this form is in Section 6, equation (6.12). 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 the $P\{S = s|b, p, u, n\}$ is the probability that in exactly s (out of b) cells the frequency (based on n observations) is at most u. Using the method of inclusion-exclusion, we obtain for $0 \leq s \leq b$

$$(2.2) \qquad P\{S = s|b, p, u, n\} = \binom{b}{s} \sum_{\gamma=0}^{s} (-1)^\gamma \binom{s}{\gamma} P\{\mathrm{Min}(b-s+\gamma, n) \geq u+1|p\}$$
$$= \binom{b}{s} \sum_{\gamma=0}^{s} (-1)^\gamma \binom{s}{\gamma} I_p^{(b-s+\gamma)}(u+1, n),$$

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

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 $n \leq u$. 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 [5]; thus the concept of sparse blue cells is a direct generalization of the maximum frequency in a homogeneous multinomial. For $p < 1/b$ we can assume that $k = b + 1$ and the value of k does not enter into any of the formulas in this paper.

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 in Section 7 below.

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

$$(2.3)\qquad P\{S \le s|b, p, u, n\} = \sum_{\alpha=0}^{s} (-1)^{\alpha}\binom{b}{\alpha} \sum_{t=\alpha}^{s} \binom{b-\alpha}{t-\alpha} I_p^{(b-(t-\alpha))}(u+1, n)$$

$$= \sum_{j=0}^{s} \binom{b}{j} I_p^{(b-j)}(u+1, n) \sum_{\alpha=0}^{s-j} (-1)^{\alpha}\binom{b-j}{\alpha}$$

$$= (b-s)\binom{b}{s} \sum_{\alpha=0}^{s} \frac{(-1)^{\alpha}\binom{s}{\alpha}}{b-s+\alpha} I_p^{(b-s+\alpha)}(u+1, n);$$

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

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 $P\{C = c|b, p, v, n\}$ 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 $0 \le c \le b$

$$(3.1)\qquad P\{C = c|b, p, v, n\} = \binom{b}{c} \sum_{\alpha=0}^{b-c} (-1)^{\alpha}\binom{b-c}{\alpha} P\{\mathrm{Min}(c+\alpha, n) \ge v|p\}$$

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

For the special case $c = b$ this reduces to $I_p^{(b)}(v, n)$ and for any $p \leq 1/b$. 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. Thus the type 1 Dirichlet integral is equal to this probability; this is only one but perhaps one of the more important uses of the Dirichlet distribution.

For the c.d.f. of C we first note that for $c \geq b$ the result is clearly unity. Hence for $c < b$

$$(3.2) \qquad P\{C \leq c | b, p, v, n\} = 1 - \sum_{t=c+1}^{b} \binom{b}{t} \sum_{\alpha=0}^{b-t} (-1)^{\alpha} \binom{b-t}{\alpha} I_p^{(t+\alpha)}(v, n)$$

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

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

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

If $c = b - 1$ then (3.2) reduces to $1 - I_p^{(b)}(v, n)$; this is correct since the complement of this event states that all blue cells are crowded or that $\text{Min}(b, n) \geq v$, which was shown in Section 2 to have probability $I_p^{(b)}(v, n)$.

4. The Joint Probability Law of S and C.

In this section we consider an inclusion-exclusion argument that operates simultaneously on the number S of sparse blue cells and the number C of crowded blue cells. It is not known whether this type of "2-dimensional" inclusion-exclusion operation has been used heretofore. We assume $v > u + 1$ $s \geq 0$, $c \geq 0$, $s + c \leq b$ and $0 < p \leq 1/b$; we use $\left[\begin{smallmatrix} b \\ s,c \end{smallmatrix}\right]$ for the multinomial

coefficient $b!/s!c!(b-s-c)!$. It should be pointed out that in the notation $P\{\text{Max}(s, n) \leq u, \text{Min}(c, n) \geq v\}$ there is only one set of n observations even though n is repeated twice. By inclusion-exclusion, we obtain for $v > u + 1$

$$
\begin{aligned}
(4.1)\qquad P\{S = s, C = c \mid b, p, u, v, n\} &= \left[{}_{s,c}^{\,b}\right]P\{\text{Max}(s, n) \leq u, \text{Min}(c, n) \geq v\} \\
&\quad - \left[{}_{s+1,\,c}^{\,b}\right]\binom{s+1}{s}P\{\text{Max}(s+1, n) \leq u, \text{Min}(c, n) \geq v\} \\
&\quad - \left[{}_{s,\,c+1}^{\,b}\right]\binom{c+1}{c}P\{\text{Max}(s, n) \leq u, \text{Min}(c+1, n) \geq v\} \\
&\quad + \left[{}_{s+2,\,c}^{\,b}\right]\binom{s+2}{s}P\{\text{Max}(s+2, n) \leq u, \text{Min}(c, n) \geq v\} \\
&\quad + \left[{}_{s+1,\,c+1}^{\,b}\right]\binom{s+1}{s}\binom{c+1}{c}P\{\text{Max}(s+1, n) \leq u, \text{Min}(c+1, n) \geq v\} \\
&\quad + \left[{}_{s,\,c+2}^{\,b}\right]\binom{c+2}{c}P\{\text{Max}(s, n) \leq u, \text{Min}(c+2, n) \geq v\} \\
&\quad + \ldots + (-1)^{b-s-c}\left[{}_{s,\,b-s}^{\,b}\right]\binom{b-s}{c}P\{\text{Max}(s, n) \leq u, \text{Min}(b-s, n) \geq v\} \\
&= \left[{}_{s,\,c}^{\,b}\right]\{1 - \binom{b-s-c}{1}(E_s + E_c) + \ldots + (-1)^{b-s-c}\binom{b-s-c}{b-s-c}(E_s + E_c)^{b-s-c}\}F(s,c) \\
&= \left[{}_{s,\,c}^{\,b}\right](1-\theta)^{b-s-c}F(s, c)
\end{aligned}
$$

where $F(s, c) = P\{\text{Max}(s, n) \leq u, \text{Min}(c, n) \geq v\}$, E_s (resp., E_c) is the finite difference operator that adds one to the s (resp., c) argument, and $\theta = E_s + E_c$.

If $v = u + 1$ then $C = b - S$, $E\{SC\} = E\{S(b-S)\}$, etc.; the one-dimensional distribution of S (or C) gives all the moments. However the above discussion and the subsequent one both hold for $v = u + 1$ and this case need not be treated separately.

To evaluate each of the terms above we need an expression for $F(s, c)$ and again we utilize the Dirichlet distribution and its identities. Conditioning on the numbers α and β of observations in the s sparse and c crowded cells of $F(s, c)$, respectively, we can then express $F(s, c)$ as a product

of probabilities for homogeneous multinomials. For $s = 0$ the first inequality in $F(s, c)$ is logically removed, for $c = 0$ the second inequality is logically removed and for $s = c = 0$ the entire probability, $F(s, c)$, is one. For $s \geq 1$ and $c \geq 1$ we obtain

$$(4.2) \qquad F(s, c) = \sum_{\substack{\alpha\geq 0,\beta\geq 0 \\ \alpha+\beta\leq n}} [{}_{\alpha,\ \beta}^{\ n}](sp)^{\alpha}(cp)^{\beta}P\{\mathrm{Max}(s, \alpha) \leq u|\tfrac{1}{s}\}P\{\mathrm{Min}(c, \beta) \geq v|\tfrac{1}{c}\}$$

$$= \sum_{\alpha=0}^{n} \binom{n}{\alpha}(sp)^{\alpha}(1-sp)^{n-\alpha}P\{\mathrm{Max}(s, \alpha) \leq u|\tfrac{1}{s}\}$$

$$\cdot \sum_{\beta=0}^{n-\alpha} \binom{n-\alpha}{\beta}\left(\frac{cp}{1-sp}\right)^{\beta}\left(1 - \frac{cp}{1-sp}\right)^{n-\alpha-\beta}P\{\mathrm{Min}(c, \beta) \geq v|\tfrac{1}{c}\}$$

$$= \sum_{\alpha=0}^{n} b_{\alpha}(n, sp)P\{\mathrm{Max}(s, \alpha) \leq u|\tfrac{1}{s}\}I^{(c)}_{\frac{p}{1-sp}}(v, n-\alpha),$$

where $b_{\alpha}(n, sp)$ is the binomial probability element $\binom{n}{\alpha}(sp)^{\alpha}(1-sp)^{n-\alpha}$ and we have yet to show that the third line of (4.2) can be replaced by an I-function. To show the latter, we consider a multinomial with N observations and $b + 1$ cells, b of which have common probability p, and we expand with respect to the last cell. This gives us the identity

$$(4.3) \qquad I^{(b)}_{p}(v, N) = \sum_{i=0}^{N} \binom{N}{i}(bp)^{i}(1-bp)^{N-i}I^{(b)}_{1/b}(v, i),$$

which was used in the last step of (4.2) with b, p, and N replaced by c, $p/(1-sp)$ and $n-\alpha$, respectively. The middle factor in the last line of (4.2) is equal to (2.2) with $b = s$, n replaced by α, and $p = 1/s$, i.e., for $\alpha \geq u$

$$(4.4) \qquad P\{\mathrm{Max}(s, \alpha) \leq u|\tfrac{1}{s}\} = \sum_{\gamma=0}^{s} (-1)^{\gamma}\binom{s}{\gamma}I^{(\gamma)}_{1/s}(u+1, \alpha),$$

which does not depend on p; for $\alpha < u$ the result is clearly equal to one. The final result in (4.2) is a polynomial in p of degree n. The expression (4.4) is the c.d.f. of the maximum frequency in a homogeneous multinomial

with α observations and, as noted before, was tabulated by Steck [5]. It is also the probability that all s cells are sparse and hence is zero for $\alpha > us$ and is equal to one for $\alpha \leq u$.

An important special case arises when $s = c = 0$ and (4.1) then becomes

$$(4.5) \qquad P\{S = C = 0 | b, p, u, v, n\} = \{1 - \binom{b}{1}\theta + \ldots + (-1)^b \theta^b\} F(s, c)\Big|_{s=c=0}$$

$$= (1-\theta)^b F(s, c)\Big|_{s=c=0},$$

where $\theta = E_s + E_c$ is the same operator as in (4.1). Although no further simplification arises in (4.5), this does appear to be a good method of calculating the probability that all frequencies in a homogeneous multinomial are strictly between u and v. In addition, (4.1) is used in Section 6 below.

5. Moments 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 all be simply expressed in terms of the type 1 Dirichlet function; this result is generalized in Section 6. Another (rather complicated) exact expression for factorial moments is given by Barton and David [1]. It would be desirable to include some of these expressions here (without derivation) because of their relevance to this paper.

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

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

where the second factor does not depend on b, the integer v defines crowdedness, and n is the number of observations. As a corollary we obtain the first moment and variance of C. For $m = 1$, $n \geq v$ and $p \leq 1/b$, we have

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

where the last symbol is the standard beta function notation. For $m = 2$, $n \geq 2v$ and $p \leq 1/b$, we have

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

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

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

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

(5.6) $\quad \sigma^2(C) = b(b-1)[1-2q^n + (q-p)^n] + b(1-q^n) - b^2(1-q^n)^2$

$$= bq^n(1-bq^n) + b(b-1)(q-p)^n = \sigma^2(S);$$

these also give the correct answer for $n = 0$ and $n = 1$, e.g., the common $\sigma^2 = 0$ for $n = 0$ and $\sigma^2 = bp(1-bp)$ for $n = 1$.

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 $\geq u + 1$. Hence from (5.1)

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

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

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

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

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

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

Some special cases of these results are included for completeness. For $m = 1$, $p = 1 - q \leq 1/b$ and $n \geq u + 1$, we have from (5.7) or (5.9)

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

for $n \leq u$ all cells are sparse and hence $E\{S\} = 0$. For $m = 2$, $p \leq 1/b$ and $n \geq 2(1+u)$, we have

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

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 integration, 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 $n \geq u + 1 > 0$ and (only) for $p = 1/b$ given by David and Barton [1 - page 279] is in our notation

$$(5.12) \qquad E\{S(S-1)\} = \frac{(b-1)}{b^{n-1}} \sum_{\alpha=0}^{2u} \binom{n}{\alpha} (b-2)^{n-\alpha} \sum_{j=\alpha-u}^{n-\alpha+u} \binom{\alpha}{j},$$

where the terms are zero for $\alpha > \min(n,\ u + n/2)$. For $n \leq u$, the result is $b^{[2]}$ since $S = b$ identically. This can also be used to derive an explicit expression for $E\{C(C-1)\}$ but the result is even more complicated and we omit it.

6. Joint Moments of S and C.

Joint moments of S and C can also be conveniently expressed in terms of I-functions. We consider joint factorial moments in the two forms

(6.1) $\quad E\{S^{[g]}C^{[h]}\}$ and $E\{C^{[h]}(b-S-h)^{[g]}\}$,

for nonnegative integers g and h; special results are then obtained for $E\{SC\}$. Our main results in (6.6) and (6.10) hold for $v > u + 1$ and also for $v = u + 1$. For $v = u + 1$ (and also for $g = 0$) the second form in (6.1) reduces to $E\{C^{[g+h]}\}$, which was treated in [3]; for $h = 0$ the result is in (5.7).

From (4.1) for $g \geq 0$ and $h \geq 0$

$$(6.2) \qquad P\{S = s, C = c\} = \sum_{d=0}^{b-s-c} (-1)^d \sum_{\alpha=0}^{d} \binom{s+\alpha}{\alpha}\binom{c+d-\alpha}{d-\alpha}\begin{bmatrix} b \\ s+\alpha,\ c+d-\alpha \end{bmatrix} F(s+\alpha,\ c+d-\alpha),$$

where $F(s, c)$ is given by (4.2). Summing on s and c ($s \geq 1$, $c \geq 1$, $s+c \leq b-d$), we obtain for the first form in (6.1)

$$(6.3) \qquad E\{S^{[g]}C^{[h]}\} = \sum_{c=0}^{b-g-h} (-1)^d \sum_{c=h}^{b-d-g} \sum_{s=g}^{b-d-c} \sum_{\alpha=0}^{d} s^{[g]}c^{[h]}\binom{s+\alpha}{\alpha}\binom{c+d-\alpha}{d-\alpha} \cdot \begin{bmatrix} b \\ s+\alpha,\ c+d-\alpha \end{bmatrix} F(s+\alpha,\ c+d-\alpha).$$

Let $y = s + \alpha$ and $z + c + d - \alpha$; then (6.3) reduces to

$$(6.4) \qquad E\{S^{[g]}C^{[h]}\} = \sum_{z=h}^{b-g} \sum_{y=g}^{b-z} y^{[g]}z^{[h]}\begin{bmatrix} b \\ y,z \end{bmatrix} F(y,z) \sum_{d=0}^{b-g-h} (-1)^d \sum_{\alpha=0}^{d} \binom{y-g}{\alpha}\binom{z-h}{d-\alpha}.$$

Using the hypergeometric identity and the well-known identity

$$(6.5) \qquad \sum_{d=0}^{b-g-h} (-1)^d \binom{y+z-g-h}{d} = (-1)^{b-g-h}\binom{y+z-g-h-1}{b-g-h},$$

the result for (6.4) reduces to zero for $y > g$ and $z > h$ since $y + z - 1 < b$. For $y = g$ and $z > h$ we set $\alpha = 0$ to get a non-zero term in (6.4) and use the same identity (6.5) to obtain zero again; similarly for $z = h$ and $y > g$. Finally for $y = g$ and $z = h$ we obtain the simple result that

$$(6.6) \qquad E\{S^{[g]}C^{[h]}\} = g!h!\begin{bmatrix} b \\ g,\ h \end{bmatrix} F(g, h) = b^{[g+h]}F(g, h);$$

here $F(g, h)$ is the probability that a specified subset of g cells are all sparse and another disjoint specified subset of h cells are all crowded. The latter interpretation assumes that $g + h \leq b$; if this is not the case then both sides of (6.6) are clearly zero. Since $F(0, 0)$ was noted in Section 4 to be equal to 1, the case $g = h = 0$ in (6.6) shows that the probabilities (4.1) sum to one.

For the second form in (6.1) we use (5.8) and (6.6) to write

$$(6.7) \qquad E\{C^{[h]}(b-h-S)^{[g]}\} = \sum_{\alpha=0}^{g} (-1)^{\alpha}\binom{g}{\alpha}(b-h-\alpha)^{[g-\alpha]}E\{S^{[\alpha]}C^{[h]}\}$$

$$= \sum_{\alpha=0}^{g} (-1)^{\alpha}\binom{g}{\alpha}(b-h-\alpha)^{[g-\alpha]}b^{[\alpha+h]}F(\alpha, h)$$

$$= b^{[g+h]} \sum_{\alpha=0}^{g} (-1)^{\alpha}\binom{g}{\alpha}F(\alpha, h).$$

As a generalization of (2.1), we define the I-function

$$(6.8) \qquad I_p^{(\alpha+\beta)}((t)_\alpha, (v)_\beta, n) = \frac{\Gamma(n+1)}{\Gamma^{\alpha}(t)\Gamma^{\beta}(v)\Gamma(n+1-\alpha t-\beta v)}$$

$$\cdot \int_0^p \cdots \int_0^p (1 - \sum_{i=1}^{\alpha+\beta} x_i)^{n-\alpha t-\beta v} \prod_{i=1}^{\alpha} x_i^{t-1}dx_i \prod_{j=1}^{\beta} x_{\alpha+j}^{v-1}dx_{\alpha+j},$$

where $(t)_\alpha$ and $(v)_\beta$ stand for $t,\ldots, t$ repeated α times and $v,\ldots, v$ repeated β times respectively. By Lemma 2.2 of [2] this represents the probability that a specified set of α blue cells have frequency $\geq t$ and another disjoint specified set of β blue cells have frequency $\geq v$, when there are n observations in all, and all the blue cells have common probability p, with $p \leq 1/(\alpha+\beta)$.

Using an inclusion-exclusion argument on the first argument of $F(\alpha, h)$ in (6.7) we obtain

$$(6.9) \qquad F(\alpha, h) = \sum_{\gamma=0}^{\alpha} (-1)^{\gamma}\binom{\alpha}{\gamma}I_p^{(\gamma+h)}((u+1)_\gamma, (v)_h, n).$$

Substituting this in (6.7) and summing through on α, we obtain our second main result

$$(6.10)\qquad E\{C^{[h]}(b-h-S)^{[g]}\} = b^{[g+h]} \sum_{\gamma=0}^{g} I_p^{(\gamma+h)}((u+1)_\gamma,(v)_h,n)\binom{g}{\gamma} \sum_{\alpha=\gamma}^{g}(-1)^{\alpha-\gamma}\binom{g-\gamma}{\alpha-\gamma}$$

$$= b^{[g+h]} I_p^{(g+h)}((u+1)_g,\ (v)_h,\ n),$$

since the summation on α in (6.10) is one for $g = \gamma$ and zero otherwise.

From (6.6), (6.10) with $g = h = 1$ and (5.1) we can write

$$(6.11)\qquad E\{SC\} = b^{[2]}F(1,1) = b^{[2]}[I_p^{(1)}(v,n) - I_p^{(2)}(u+1,\ v,\ n)].$$

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

$$(6.12)\qquad E\{SC|u=0,\ v\geq 1\} = b^{[2]}q^n I_{p/q}^{(1)}(v,n) = b^{[2]}q^n I_{p/q}(v,n-v+1).$$

For example, if $b = 3$, $p = 1/b = 1/3$, $u = 0$ and $v = 2$ then for any $n > 2$ the result is by (6.12)

$$(6.13)\qquad E\{SC|u=0,\ v=2\} = 6(\tfrac{2}{3})^n I_{1/2}(2,\ n-1) = \frac{2(2^n-n-1)}{3^{n-1}}.$$

This result (which also holds for $n = 0$, 1, and 2) can also be obtained 'ab initio' by considering the 3 cases for which $SC \neq 0$ or by computing $F(1, 1)$ and using (6.6).

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

$$(6.14)\qquad E\{(1+t)^C\} = \sum_{c=0}^{b}\binom{b}{c}(1+t)^c \sum_{\alpha=0}^{b-c}(-1)^\alpha\binom{b-c}{\alpha}I_p^{(c+\alpha)}(v,n)$$

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

$$= \sum_{\beta=0}^{b}\binom{b}{\beta}t^\beta I_p^{(\beta)}(v,n).$$

The dfmgf that gives rise to the moments in the second form in (6.1) is

$$(6.15)\qquad E\{(1+\frac{t_2}{1+t_1})^C(1+t_1)^{b-S}\} = \sum_{g,h}\frac{t_1^g}{g!}\frac{t_2^h}{h!}E\{C^{[h]}(b-S-h)^{[g]}\}$$

and hence by our result in (6.10) we must have

$$(6.16)\qquad E\{(1+t_1+t_2)^C(1+t_1)^{b-S-C}\} = \sum_{g,h}\frac{t_1^g}{g!}\frac{t_2^h}{h!}\,b^{[g+h]}I_p^{(g+h)}((u+1)_g,\ (v)_h,\ n).$$

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: $p_1 = p_2 = \ldots = p_{10} = 1/10$. One alternative of interest is H_1: $p_1 = p_2 = \ldots = p_9 = p < 1/10$ and $p_{10} = 1 - 9p$; another alternative of interest is H_2: $p_1 = p_2 = \ldots = p_9 = 1/9$ 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

$$(7.1)\qquad P\{S \le s_0 | 10,\ 1/10,\ 0,\ 40\} \ge P^*$$

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

$$P\{S = 0\} = I^{(10)}_{1/10}(1, 40) = .8581$$

(7.2)

$$P\{S \leq 1\} = 10\, I^{(9)}_{1/10}(1, 40) - 9\, I^{(10)}_{1/10}(1, 40) = .9942.$$

In order to attain a test size of exactly .05 we reject H_0 if $S \geq 2$ 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

$$P_1 = \text{Power(ECT vs. } H_1) = P\{S_9 \geq 2\} + .324\, P\{S_9 = 1,\ f(10) > 0\} \tag{7.4}$$

$$+ P\{S = 1,\ f(10) = 0\} + .324\, P\{S = 0,\ f(10) = 0\}.$$

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

$$P_1 = 1 - P\{S = 0 \mid 9,\ 1/20,\ 0,\ 40\} - .676\, P\{S = 1 \mid 9,\ 1/20,\ 0,\ 40\} \tag{7.5}$$

$$= 1 - I^{(9)}_{1/20}(1,40) - (.676)9[I^{(8)}_{1/20}(1,40) - I^{(9)}_{1/20}(1,40)] = .4580.$$

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

$$P\{S = 0\} = I^{(10)}_{1/10}(2, 40) = .3858$$

$$P\{S \leq 1\} = 10\, I^{(9)}_{1/10}(2,40) - 9I^{(10)}_{1/10}(2,40) = .8296 \tag{7.6}$$

$$P\{S \leq 2\} = 45I^{(8)}_{1/10}(2,40) - 80I^{(9)}_{1/10}(2,40) + 36I^{(10)}_{1/10}(2,40) = .9800\ .$$

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

$$P_1' = \text{Power(SCT vs. } H_1) = P\{S_9 \geq 3\} + .2P\{S_9 = 2,\ f(10) > 1\} \tag{7.7}$$

$$+ P\{S = 2,\ f(10) \leq 1\} + .2P\{S = 1,\ f(10) \leq 1\}.$$

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

$$(7.8)\qquad P_1' = 1 - P\{S \le 1|9,\ 1/20,\ 1,\ 40\} - .8P\{S = 2|9,\ 1/20,\ 1,\ 40\}$$

$$= 1 - [9I^{(8)}_{1/20}(2,40) - 8I^{(8)}_{1/20}(2,40)]$$

$$-(.8)36[I^{(7)}_{1/20}(2,40) - 2I^{(8)}_{1/20}(2,40) + I^{(9)}_{1/20}(2,40)] = .8372.$$

Thus the sparse-cell with $u = 1$ already gives a better power against H_1; calculations for $u \ge 2$ 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 $u \ge 1$; it suffices to consider the case $u = 1$. The power calculations against H_2 for the empty-cell test are

$$(7.9)\qquad P_2 = \text{Power(ECT vs. } H_2) = P\{S \ge 2|H_2\} + .324\ P\{S = 1|H_2\}$$

$$= 1 - P\{S = 0|H_2\} - .676\ P\{S = 1|H_2\}$$

$$= 1 - .676\ P\{S = 0|9,\ 1/9,\ 0,\ 40\} = 1 - .676\ I^{(9)}_{1/9}(1,\ 40)$$

$$= .3777.$$

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

$$(7.10)\qquad P_2' = \text{Power(SCT vs. } H_2) = P\{S \ge 3|H_2\} + .2\ P\{S = 2|H_2\}$$

$$= 1 - P\{S \le 1|H_2\} - .8\ P\{S = 2|H_2\}$$

$$= 1 - P\{S = 0|9,\ 1/9,\ 1,\ 40\} - .8\ P\{S = 1|9,\ 1/9,\ 1,\ 40\}$$

$$= 1 - I^{(9)}_{1/9}(1,\ 40) - (.8)9[I^{(8)}_{1/9}(1,\ 40) - I^{(9)}_{1/9}(1,\ 40)]$$

$$= .0171.$$

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 $u \geq 1$ 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 $d \leq t$ we call any $d \times d$ square a cluster (of 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 can be 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 $p \leq 1/k$ 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. If too many cells in close proximity are crowded then the chances of forming a cancerous tumor at that location are very high. The cells in close proximity are our cell clusters and a crowded cluster is the origin of

the cancerous tumor. One interesting quantity for this application is the probability of at least one crowded cluster, since one crowded cluster is sufficient to start the formation of the cancerous tumor.

To get the answer, A , to this we apply inclusion-exclusion methods to the individual clusters $L_1, L_2, \ldots, L_D$ and let the respective b-values (all equal to d^2 in our application) be denoted by $b_1, b_2, \ldots, b_D$. If we let $P(b_i, b_j)$ denote the probability that the i^{th} and j^{th} cluster are both crowded, then by (3.1)

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

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

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

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

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

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

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

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

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

$$(8.8)\qquad \Sigma P(b_{i_1}, b_{i_2}, \ldots, b_{i_6}) = 4I_p^{(12)}(v, n) + 24I_p^{(13)}(v, n) + 34I_p^{(14)}(v, n) + 12I_p^{(15)}(v, n) + 10I_p^{(16)}(v, n),$$

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

$$(8.10)\qquad \Sigma P(b_{i_1}, b_{i_2}, \ldots, b_{i_8}) = 4I_p^{(15)}(v, n) + 5I_p^{(16)}(v, n),$$

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

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

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

Note that the sum of the coefficients in equation (8.2) is $\binom{9}{i}(i = 1,2,\ldots, 9)$ 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, then 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. Acknowledgement.

The authors wish to thank Dr. David Hoel of the National Institute of Environmental Health Sciences at Research Triangle Park, North Carolina for suggesting the application discussed in Section 8.

REFERENCES

[1] David, F. N. and Barton, D. E. (1962). Combinatorial Chance. London, Charles Griffin and Co.

[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. Department of Statistics Technical Report No. 98. University of Minnesota, Minneapolis.

[4] Sobel, M. and Uppuluri, V. R. R. (1972). On Bonferroni-type inequalities of the same degree for the probability of unions and intersections. (To appear in Ann. Math. Statist. October 1972.)

[5] Steck, G. (ca. 1965). Table of the distribution of the maximum frequency in a homogeneous multinomial. (Unpublished table--personal communication.)