[*] This research was sponsored by U.S. Army Research Office - Durham, under Grant Number DAH-C04-74-C-0030.

[†] This research was sponsored by U.S. Air Force Office of Scientific Research, under Grant Number 74-2701.

ON A MULTIVARIATE GENERALIZED OCCUPANCY MODEL

N.L. Johnson[*]

University of North Carolina at Chapel Hill

Samuel Kotz[†]

Temple University, Philadelphia

Institute of Statistics Mimeo Series No. 979

January, 1975

On A Multivariate Generalized Occupancy Model

by

N.L. Johnson*

University of North Carolina at Chapel Hill

and

Samuel Kotz**

Temple University, Philadelphia

0. Introduction and Summary

Uppuluri and Carpenter (1971) (referred to below as [U-C]), have discussed a generalized occupancy model related to the classical occupancy model discussed by David and Barton (1962). We first obtain Uppuluri and Carpenter's results by elementary methods, and then discuss a natural multivariate generalization.

1. Univariate Case

Consider a population containing just m categories, of which $b(\leq m)$ appear with equal frequency, each being a proportion p of the total population. We call these categories of Class I. The remaining categories (a proportion $(1 - bp)$ of the total population) we will call categories of Class II.

We first evaluate the probability that in r independent trials, in each of which an individual is chosen at random from the population, observed and returned to the population, there are observed exactly j of the b Class I categories.

We will denote by V_r the number of different Class I categories observed in r independent trials, and by R_1 the number of trials in which a Class I category (any one among the b in this Class) is observed. Then R_1 has

*This research was sponsored by U.S. Army Research Office - Durham, under Grant Number DAH-C04-74-C-0030.

**This research was sponsored by U.S. Air Force Office of Scientific Research, under Grant Number 74-2701.

a binomial distribution with parameters r, bp.

The distribution of V_r, given R_1 is the same as that of the number of different categories observed when there are just b equally likely categories (i.e. each in proportion b^{-1}) and R_1 trials. This is the classical occupancy distribution (e.g. David and Barton (1962)). Hence

$$\Pr[V_r = j|R_1] = \binom{b}{j}\Delta^j 0^{R_1}/b^{R_1} . \tag{1}$$

Averaging over the distribution of R_1

$$\Pr[V_r = j] = \binom{b}{j}E[\Delta^j 0^{R_1}/b^{R_1}] .$$

Using the formula (for A an arbitrary constant)

$$E[A^{R_1}] = (1 - bp + bpA)^r \tag{2}$$

we obtain

$$\Pr[V_r = j] = \binom{b}{j}\Delta^j(1 - bp + p\circ 0)^r . \tag{3}$$

Since $\Delta^j f(x) = \sum_{h=0}^{j}(-1)^{j-h}\binom{j}{h}f(x+h) = \sum_{h=0}^{j}(-1)^h\binom{j}{h}f(x+j-h)$, we can express (3) as

$$\Pr[V_r = j] = \binom{b}{j}\sum_{h=0}^{j}(-1)^{j-h}\binom{j}{h}\{1 - (b-h)p\}^r \tag{3'}$$

or

$$\Pr[V_r = j] = \binom{b}{j}\sum_{h=0}^{j}(-1)^h\binom{j}{h}\{1 - (b-j+h)p\}^r . \tag{3''}$$

[U-C] derive the *conditional* probability $p_{ij}^{(r)}$, given that i (<j) categories of Class I have already been observed, that after r further independent trials are carried out, the total number of different Class I categories observed (including the i already observed) will be j. To evaluate this we regard the (b-i) so far unobserved Class I categories as Class I' categories (each, of course, still with proportion p) and all others (including Class II categories) as Class II'. We then need the probability of observing (j-i) Class I'

categories. So we replace b by (b-i) and j by (j-i) in (3), obtaining

$$p_{ij}^{(r)} = \binom{b-i}{j-i}\Delta^{j-i}\{1 - (b-i)p + p\circ 0\}^{r} . \tag{4}$$

If we use the form (3)", we obtain

$$p_{ij}^{(r)} = \binom{b-i}{j-i}\sum_{h=0}^{j-i}(-1)^{h}\binom{j}{h}\{1 - (b-j+h)p\}^{r} \tag*{(4)'}$$

which agrees with the formula on page 320 of [U-C]. The methods we use are much simpler than those used in [U-C], which include the theory of absorbing Markov chains and "bidiagonal' matrices.

Results obtained in [U-C] concerning a certain waiting-time distribution can also be obtained quite simply. Let T denote the number of trials needed to observe *all* b categories of Class I, given that i have already been observed. We again introduce the Classes I', II' defined above, and note that

$$T = \sum_{h=0}^{b-i} T_h \tag{5}$$

where T_h is the number of trials needed to observe some one of (b-i-h+1) categories of Class I (each with proportion p). The T_h's are independent and T_h has a geometric distribution with parameter (b-i-h+1)p. Hence

$$E[T_h] = (b-i-h+1)^{-1}p^{-1} \tag{6}$$

and

$$\operatorname{var}(T_h) = \{1 - (b-i-h+1)p\}(b-i-h+1)^{-2}p^{-2}. \tag{7}$$

From (5), (6) and (7)

$$E[T] = p^{-1}\sum_{h=1}^{b-i}(b-i-h+1)^{-1} \tag{8}$$

$$\operatorname{var}(T) = p^{-2}\sum_{h=1}^{b-i}(b-i-h+1)^{-2} - p^{-1}\sum_{h=1}^{b-i}(b-i-h+1)^{-1} . \tag{9}$$

These formulae agree with those on page 323 of [U-C].

Of course, the expected value of the waiting time could be obtained as

$$\sum_{r=b}^{\infty} r\{\Pr[V_r=b] - \Pr[V_{r-1}=b]\}.$$

Feller (1962) and Harkness (1969) consider an extension of the classical occupancy problem in which there is a probability $(1-\omega)$ that an observation falling in any Class I category will not be recorded. (Tables of critical points for the "extended" classical occupancy distribution are available in Johnson, Kotz and Srinivasan (1974)). In the situation we have discussed here this simply means that R has a binomial distribution with parameters r, $b\omega p$ (in place of r, bp). The formulae (3) - (9) will still apply, with p replaced everywhere by ωp.

2. Multivariate Generalization

The simple techniques employed in Section 1 can be applied to obtain some results for a natural generalization.

Suppose we have b_g categories of Class I(g) for $g = 1,2,\ldots,k$ and each category of Class I(g) has proportion p_g. Class II, as before contains all the remaining categories, and has total proportion $1 - \sum_{g=1}^{k} b_g p_g = p_0$, say.

Suppose, now, that in $r = \sum_{g=0}^{k} R_g$ independent trials, R_g trials yield individuals in Class I(g) (for $g = 1,2,\ldots,k$) and R_0 yield Class II individuals. Then the conditional probability, given $R_1,R_2,\ldots,R_g$, that *exactly* j_g of the b_g Class I(g) categories have been observed $(g = 1,2,\ldots,k)$ is

$$\Pr[\underset{\sim}{V}_r = \underset{\sim}{j}|\underset{\sim}{R}] = \prod_{g=1}^{k} [V_{r,g} = j_g|R_g]$$

$$= \prod_{g=1}^{k} \binom{b_g}{j_g} \Delta^{j_g} 0^{R_g} / b_g^{R_g} \quad . \qquad (10)$$

Here $\underset{\sim}{V}'_r = (V_{r,1},\dots,V_{r,k})$ with $V_{r,g}$ being the number of Class I(g) categories observed, and $\underset{\sim}{j}' = (j_1,j_2,\dots,j_k)$. (Note that *conditionally* on $\underset{\sim}{R}$, the $V_{r,g}$'s are mutually independent.)

Now $R_0,R_1,\dots,R_k$ have a joint multinomial distribution, with parameters r; $(1 - \sum_{g=1}^{k} b_g\, p_g)$, $b_1p_1,\dots,b_kp_k$. Proceeding as in the univariate case (cf (2) - (3)) and noting that for arbitrary constants $A_1,A_2,\dots,A_k$

$$E\left[\prod_{g=1}^{k} A_g^{R_g}\right] = \left(1 - \sum_{g=1}^{k} b_g p_g + \sum_{g=1}^{k} b_g p_g A_g\right)^r \qquad (11)$$

$$= \left(p_0 + \sum_{g=1}^{k} b_g p_g A_g\right)^r \qquad (11)'$$

we obtain

$$\Pr[\underset{\sim}{V}_r = \underset{\sim}{j}] = E[\Pr[\underset{\sim}{V}_r = \underset{\sim}{j} \mid \underset{\sim}{R}]]$$

$$= \prod_{g=1}^{k} \binom{b_g}{j_g} \Delta_g^{j_g} \left(p_0 + \sum_{g=1}^{k} p_g \cdot 0_g\right)^r \qquad (12)$$

where the difference operator Δ_g operates only on 0_g.

Noting that

$$\Delta_1^{j_1}\Delta_2^{j_2}\dots\Delta_k^{j_k} 0_1^{\alpha_1} 0_2^{\alpha_2}\dots 0_k^{\alpha_k} = \prod_{g=1}^{k} (\Delta^{j_g} 0^{\alpha_g})$$

we find that (12) can be put in the form

$$\Pr[\underset{\sim}{V}_r = \underset{\sim}{j}] = \prod_{g=1}^{k} \binom{b_g}{j_g} \sum_{\underset{\sim}{h}=\underset{\sim}{0}}^{\underset{\sim}{j}} (-1)^{\sum_1^k (j_g - h_g)} \left\{ \prod_{g=1}^{k} \binom{j_g}{h_g} \right\} \left(p_0 + \sum_{g=1}^{k} h_g p_g\right)^r . \qquad (12)'$$

Noting further that

$$\binom{b_g}{j_g}\binom{j_g}{h_g} = \frac{b_g!}{(b_g - j_g)!\,h_g!\,(j_g - h_g)!} = \binom{b_g}{b_g - j_g,\, h_g,\, j_g - h_g}$$

we see that (12) or (12)' can also be expressed as

$$\Pr[\underset{\sim}{V}_r=\underset{\sim}{j}] = \sum_{\underset{\sim}{h}=\underset{\sim}{0}}^{\underset{\sim}{j}} (-1)^{\sum_1^k (j_g-h_g)} \prod_{g=1}^{k} \binom{b_g}{b_g-j_g,h_g,j_g-h_g} \left(p_0 + \sum_{g=1}^{k} h_g p_g\right)^r . \qquad (12)''$$

Yet another form is

$$\Pr[\underset{\sim}{V}_r=\underset{\sim}{j}] = \sum_{\underset{\sim}{h}=\underset{\sim}{0}}^{\underset{\sim}{j}} (-1)^{\sum_1^k h_g} \prod_{g=1}^{k} \binom{b_g}{b_g-j_g,h_g,j_g-h_g} \left\{p_0 + \sum_{g=1}^{k} (j_g-h_g) p_g\right\}^r . \qquad (12)'''$$

Remembering that $p_0=1-\sum_{g=1}^{k} b_g p_g$, the parallelism between (12)'' and (3)', and between (12)''' and (3)'' is apparent.

If we suppose that $i_1,i_2,\ldots,i_k$ categories of Classes I(1),I(2),...,I(k) respectively have already been observed, then as in Section 1, we obtain the conditional probability, that $j_1,j_2,\ldots,j_k$ categories,*in all*,of Classes I(1),I(2),...,I(k) respectively will have been observed after r further independent trials, by replacing b_g by (b_g-i_g) and j_g by (j_g-i_g) in (12), remembering the "hidden" b_g's in p_0. We obtain from (12)''

$$p^{(r)}_{\underset{\sim}{i}\underset{\sim}{j}} = \sum_{\underset{\sim}{h}=\underset{\sim}{0}}^{\underset{\sim}{j}-\underset{\sim}{i}} (-1)^{\sum_1^k (j_g-i_g-h_g)} \prod_{g=1}^{k} \binom{b_g-i_g}{b_g-j_g,h_g,j_g-i_g-h_g} \left(p_0 + \sum_{g=1}^{k} (h_g-i_g) p_g\right)^r . \qquad (13)$$

An extension, analogous to that described at the end of Section 1, is obtained by supposing that (for g = 1,2,...,k) there is a probability, $(1-\omega_g)$, that an observation of any category in Class I(g) is not recorded. The formulae (12)-(13) will still apply, with p_g replaced by $\omega_g p_g (g=1,2,\ldots,k)$, remembering that now $p_0 = 1 - \sum_{g=1}^{k} b_g \omega_g p_g$.

3. Expected Values

In the univariate case (Section 1) the probability that each of n *specified* categories of Class I are observed (at least once) in r independent trials is

$$1 - \binom{n}{1}(1-p)^r + \binom{n}{2}(1-2p)^r - \ldots + (-1)^n(1-np)^r = (-1)^n\Delta^n(1-p\circ 0)^r \qquad (14)$$

(Note that $np < 1$.)

We define

$$Z_t = \begin{cases} 1 & \text{if the t-th category in Class I is observed,} \\ 0 & \text{if not.} \end{cases} \qquad (15)$$

Then, for any subset $(a_1, a_2, \ldots, a_n)$ of $(1,2,\ldots,b)$

$$E[\prod_{t=1}^{n} Z_{\alpha_t}^{\alpha_t}] = (-1)^n\Delta^n(1-p\circ 0)^r \qquad (16)$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are positive integers.

In particular

$$E[Z_t^{\alpha_t}] = -\Delta(1-p\circ 0)^r = 1 - (1-p)^r \qquad (17)$$

whence

$$E[Z_t] = 1 - (1-p)^r \qquad (18.1)$$

$$\text{var}(Z_t) = (1-p)^r\{1-(1-p)^r\} \qquad (18.2)$$

$$\text{cov}(Z_t Z_{t'}) = \Delta^2(1-p\circ 0)^r - \{-\Delta(1-p\circ 0)^r\}^2$$

$$= (1-2p)^r - (1-p)^{2r}. \qquad (18.3)$$

Now $V_r = \sum_{t=1}^{b} Z_t$ and so

$$E[V_r] = b\{1-(1-p)^r\} \qquad (19.1)$$

$$\text{var}(V_r) = b(1-p)^r\{1-(1-p)^r\} + b(b-1)\{(1-2p)^r - (1-p)^{2r}\}$$

$$= b\{(1-p)^r - (1-2p)^r\} - b^2\{(1-p)^{2r} - (1-2p)^r\}. \qquad (19.2)$$

Note that if there is no Class II, then $p = b^{-1}$ and we have the classical occupancy problem. p cannot exceed b^{-1}.

In the multivariate case we define

$$Z_{gt} = \begin{cases} 1 \text{ if the t-th category in Class I(g) is observed,} \\ 0 \text{ if not.} \end{cases} \qquad (20)$$

so that $V_{r,g} = \sum_{t=1}^{b_g} Z_{gt}$.

Then (cf. (19.1), (19.2))

$$E[V_{r,g}] = b_g\{1 - (1-p_g)^r\} \qquad (21.1)$$

$$\text{var}(V_{r,g}) = b_g\{(1-p_g)^r - (1-2p_g)^r\} - b_g^2\{(1-p_g)^{2r} - (1-2p_g)^r\} \qquad (21.2)$$

Also, since (for $(g,t) \not\equiv (g',t')$)

$$\begin{aligned} E[Z_{gt}Z_{g't'}] &= \Pr[(Z_{gt}=1) \cap (Z_{g't'} = 1)] \\ &= 1 - (1-p_g)^r - (1-p_{g'})^r + (1-p_g-p_{g'})^r \end{aligned}$$

we have (for $g \neq g'$)

$$\begin{aligned} \text{cov}(V_{r.g},V_{r,g'}) &= b_g b_{g'} \text{ cov}(Z_{gt},Z_{g't'}) \\ &= b_g b_{g'}\{(1-p_g-p_{g'})^r - (1-p_g)^r(1-p_{g'})^r\} \end{aligned} \qquad (21.3)$$

We now find the regression of $V_{r,g'}$ on $V_{r,g}$. Since

$$\Pr[V_{r,g}= j_g|R_g] = \binom{b_g}{j_g}\Delta^{j_g}0^{R_g}/b_g^{R_g}$$

we have, using Bayes' theorem

$$\Pr[R_g=r_g|V_{r,g}=j_g]= \frac{(\Delta^{j_g}0^{r_g})\binom{r}{r_g}p_g^{r_g}(1-b_gp_g)^{r-r_g}}{\Delta^{j_g}\{1-b_gp_g+p_g\circ 0\}^r} \qquad (22)$$

Given $R_g=r_g$, $V_{r,g'}$ will be distributed as $V_{r-r_g,g'}$, with $p_{g'}$ replaced by $p_{g'}(1-b_gp_g)^{-1}$. Hence

$$E[V_{r,g'}|R_g=r_g] = b_{g'}[1 - \{1-p_{g'}(1-b_gp_g)^{-1}\}^{r-r_g}] \qquad (23)$$

and $E[V_{r,g'}|V_{r,g}=j_g]$ is the expected value of (23) taken over the distribution (22).

This is

$$b_{g'}\left[1 - [\Delta^{j_g}\{1-b_gp_g+p_g\circ 0\}^r]^{-1}\Delta^{j_g}\sum_{r_g=0}^{r}\binom{r}{r_g}(p_g\circ 0)^{r_g}(1-b_gp_g-p_{g'})^{r-r_g}\right]$$

$$= b_{g'}\left[1 - \frac{\Delta^{j_g}(1-b_gp_g-p_{g'}+p_g\circ 0)^r}{\Delta^{j_g}(1-b_gp_g+p_g\circ 0)^r}\right]$$

Thus the regression of $V_{r,g}$ on $V_{r,g'}$ is

$$E[V_{r,g'}|V_{r,g}] = b_{g'}\left[1 - \frac{\Delta^{V_{r,g}}(1-b_gp_g-p_{g'}+p_g\circ 0)^r}{\Delta^{V_{r,g}}(1-b_gp_g+p_g\circ 0)^r}\right] \qquad (24)$$

REFERENCES

David, F.N. and Barton, D.E. (1962). *Combinatorial Chance*, New York: Hafner.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol. 1, New York: Wiley.

Harkness, W.L. (1969). *The Classical Occupancy Problem Revisited*, Technical Report No. 11, Department of Statistics, Pennsylvania State University.

Johnson, N.L., Kotz, S. and Srinivasan, R. (1974). *Extended Occupancy Probability Distribution Critical Points*, Mimeo Series No. 934, Institute of Statistics, University of North Carolina.

Uppuluri, V.R.R. and Carpenter, J.A. (1971). A generalization of the classical occupancy problem, *J. Math. Anal. Appl.* *34*, 316-324.