The mode of a multinomial distribution

BY H. M. FINUCAN

*University College London**

INTRODUCTION

1. A single trial may have r different results E_i with respective probabilities $p_i\{\Sigma p = 1\}$, and the outcome of n such independent trials in which the result E_i occurs k_i times $\{i = 1, 2, \ldots, r; \Sigma k = n\}$ could be denoted by $(k_1, k_2, \ldots, k_r)$ or briefly by (k_i); this notation is based upon that of Feller (1957, p. 157). In the present paper the outcome (k_i) $\{i = 1, 2, \ldots, r\}$, for which the probability is a maximum, is sought; if unique it will be referred to as the mode, but if two or more distinct outcomes have equal maximum probability they will be referred to as joint modes. The current best procedure for finding the mode(s) is derived from Feller's problem 28 (1957, p. 161) which records a fairly sharp inequality for each k_i, due to Moran; these, in association with $\Sigma k = n$, restrict the possible modes to a few outcomes and evaluation of actual probabilities (or their ratios) is required to select the true (joint) mode(s). Moran's inequalities are very useful in practice for small r since often they restrict the possible modes to just one which is thus known to be *the* mode. The results obtained here 'tidy up' theory and practice, reveal the possible configurations of joint modes, and are more economical for larger values of r.

SUMMARY

2. The notation $I(z)$ meaning 'the integral part of z' is widely used in connexion with the modes of binomial, Poisson and other distributions (though it is not used by Feller). If z is a fraction $u+f$ with u an integer and $0 < f < 1$, then $I(z)$ denotes u; here we use $I(z)$ further, when $z = i$, an integer, to denote 'either i or $i-1$'. With this notation the result to be proved is briefly that if N is such that the r quantities $u = I(Np)$ satisfy $\Sigma u = n$ then this outcome (u) is a mode. From now on the word 'outcome' will be replaced by 'point' (u) with 'co-ordinates' u. Subscripts will be omitted when the meaning is clear without them.

3. Initially a study is made of properties **A** of our alleged modes and properties **R** of real modes; then a formal proof that $\mathbf{A} \Leftrightarrow \mathbf{R}$ is given. This includes a demonstration that the modal region is unique in the senses that there is no local maximum remote from *the* maximum and that joint modes occur all in one cluster. Next the theorem is translated into a practical procedure for locating the mode in any particular case so as to avoid prolonged search for the suitable value(s) of N. The corresponding problem for sampling from a finite population is then reduced to dependence upon the multinomial (infinite population) case; also, the inverse problem of finding the most likely population corresponding to an observed sample is treated. Finally, a few numerical examples are given.

PROPERTIES OF $\Sigma I(Np)$

4. In this section we find the properties (**A**) of points satisfying our condition

$$u_i = I(Np_i) \quad \text{and} \quad \sum^r u_i = n. \tag{1}$$

5. As N increases, each $I(Np)$ increases by 1 at certain values of N but is constant in the intervals between; our definition of this symbol endows it with both the old and the new value at such an N. Thus $y = \Sigma I(Np)$, regarded as a function of N, increases from time to time by jumps of 1 and is two-valued at a jump. Should s jumps coincide at the one N, y is $(s+1)$-valued at this N (say $y = a, a+1, a+2, \ldots, a+s$) and an intermediate one of its values $a+g$ is attributable to $\binom{s}{g}$ different combinations of g high and $(s-g)$ low alternatives of the s ambiguous $I(Np)$.

6. In solving for N, and hence for the u, the equation $\Sigma I(Np) = n$, two cases may arise according as (**U**) the value n is an ordinate of the (N, y) graph over one of its flat stretches, or (**J**) the value n is an intermediate ordinate at a multiple jump. The case **U** ('**U**' for uniform or level portion of the graph) offers an infinity of solutions for N, all of which define the same (u); while the case **J** determines a unique N which, however, gives rise to several acceptable (u).

* Permanent address: University of Queensland, Australia.

7. Thus the characteristic of a unique (**U**) alleged mode obtained by the condition (1) above is

AU: there exists a range of values N satisfying $\Sigma I(Np) = n$; each u is uniquely defined by its $I(Np)$ for any N in this range. (2)

Further the characteristic of a set of (u), i.e. an alleged set of joint (**J**) modes, obtained from (1), is:

AJ: there exists a single value of N for which s ($\geqslant 2$) of the Np are integers and, with $m = I(Np)$, several points (m) exist with $\Sigma m = n$. (3)

In fact there are $\binom{s}{g}$ points if $n = a+g$, where a is as in § 5.

Properties of a strict local mode

8. We say here that (s) is the $[a,b]$ neighbour of (k) if $s_a = k_a+1$, $s_b = k_b-1$ and $s_i = k_i$ for $i \neq a, b$. In general (k) has $r(r-1)$ neighbours, an exception occurring when any $k_i = 0$. We say also that (u) is a strict local mode if its probability exceeds that of any (i.e. all) of its neighbours.

9. If (u) is a strict local mode, the condition that its probability exceeds that of its $[i,j]$ neighbour is

$$(u_i+1)/p_i > u_j/p_j. \tag{4}$$

In general there are $r(r-1)$ such inequalities. There is no $[i,j]$ neighbour if $u_j = 0$ but obviously (4) is still true in this case; furthermore (4) is obviously true when $i = j$. Thus there is no need for qualifications (of the form $u \neq 0$, $u \neq u'$) and the real conditions for a strict local (say unique local) mode are

$$\textbf{RU}: \text{the } r^2 \text{ inequalities (4) and } \Sigma u = n. \tag{5}$$

Proof that AU ⇔ RU

10. We cannot round off the list of conditions by giving **RJ** (the real conditions for joint modes) at this stage as we must first learn more about possible types of joint modes. We therefore deal with the case of a unique mode and show that **AU** ⇔ **RU**.

11. First, **AU** ⇒ **RU**.
For consider any N given by (2); since the u are unique for this N, all Np's are fractional and for each

$$u_a < Np_a < u_a+1,$$

i.e.
$$(u_a+1)/p_a > N, \quad u_a/p_a < N \quad (a=1,2,\ldots,r),$$

i.e.
$$(u_i+1)/p_i > N > u_j/p_j \quad (i,j=1,2,\ldots,r). \tag{6}$$

As $\Sigma u = N$ is part of **AU** and since these equations (6) imply (4), we see that (5) has been established, as required.

12. Next **RU** ⇒ **AU**.
Consider the r intervals of values V

$$u_b/p_b < V < (u_b+1)/p_b \quad (b=1,\ldots,r).$$

These intervals have a common stretch 'in the middle' because (4), which is given, shows that the right boundary of any interval is to the right of the left boundary of any other. Thus for the range of values N in this common stretch

$$\left.\begin{array}{l} u_b/p_b < N < (u_b+1)/p_b, \\ u_b < Np_b < u_b+1 \end{array}\right\} \quad (b=1,\ldots,r),$$
i.e.

i.e.
$$Np_b \text{ is a fraction with } u_b = I(Np_b);$$

and $\Sigma u = n$ is part of **RU**. Thus (2) has been established, as required.

13. Thus **AU** are established as conditions for a strict local mode, from which it follows that there can be no more than one strict local mode since **AU**, as was seen in § 6, defines a unique point (u). More than this, if there is one strict local mode, there can be no modes, strict or joint, elsewhere.

Properties of a joint mode

14. From the remark immediately preceding it follows that two non-neighbours, e.g. $(e+1, f, g+1, h)$ and $(e, f+1, g, h+1)$ in a quadrinomial, could not be *the only* joint modes of a distribution. For in such a case each would have greater probability than any of its neighbours, so that each would be a strict local mode which is not possible. Of course, two non-neighbours could well be members of a larger set of joint modes. Summarizing this argument—a joint mode must have an equiprobable neighbour.

15. Conditions **RJ** for a point to be in reality a joint mode can now be derived. For let (m) be a joint mode and suppose that one of its equally probable neighbours is the $[a, b]$ neighbour, so that

$$(m_a+1)/p_a = m_b/p_b, \quad \text{which implies} \quad (m_b+1)/p_b > m_a/p_a; \tag{7}$$

also, no neighbour is more probable than (m), so that (cf. (4) above),

$$(m_i+1)/p_i \geqslant m_j/p_j \quad (i, j = 1, 2, \ldots, r). \tag{8}$$

RJ consists of these equations (7), (8) and $\Sigma m = n$.

Proof that AJ ⇔ RJ

16. First: **RJ** ⇒ **AJ**.
For, defining N as

$$(m_a+1)/p_a = m_b/p_b, \quad \text{from (7)}, \tag{9}$$

we have

$$N \geqslant m_k/p_k \quad \text{by applying (8) to } (a, k),$$

$$N \leqslant (m_k+1)/p_k \quad \text{by applying (8) to } (k, b),$$

i.e.

$$m_k \leqslant Np_k \leqslant m_k+1 \quad (k = 1, 2, \ldots, r) \tag{10}$$

and $\Sigma m = n$. But (9) and (10) are simply **AJ**: N is unique; Np_a, Np_b, and perhaps other Np's, are integers; (m) and its $[a, b]$ neighbour are points with co-ordinates of the form $m_i = I(Np_i)$ and with $\Sigma m = n$.

17. Next: **AJ** ⇒ **RJ**.
Here take one point (m) satisfying **AJ** (§ 7, with the notation of § 5). Then there are s integer values of Np and g of the m_i are these Np_i while another $(s-g)$ of the m_i are the corresponding (Np_i-1). That is

$$m_a = Np_a \quad \text{for } g \text{ co-ordinates},$$

$$m_b+1 = Np_b \quad \text{for } (s-g) \text{ co-ordinates},$$

and

$$m_f < Np_f < m_f+1 \quad \text{for } (n-s) \text{ co-ordinates},$$

so that

$$m_e/p_e \leqslant N \leqslant (m_e+1)/p_e \quad \text{in every case};$$

therefore

$$(m_i+1)/p_i \geqslant m_j/p_j \quad \text{for all pairs } (i, j), \tag{11}$$

including

$$(m_b+1)/p_b = m_a/p_a \quad \text{in certain of these}, \tag{12}$$

with, of course, $\Sigma m = n$. But these conclusions (11), (12) are only a rearrangement of **RJ** in (7), (8) above.

Possible configurations of joint modes

18. Having established that **AJ** ⇔ **RJ**, we deduce that joint modes can only occur as described in § 5. That is, a constellation of joint modes comprises all $\binom{s}{g}$ points obtainable by increasing by 1 each member of any selection of g out of a specified s co-ordinates of a given (reference, non-modal) point. Cf. example E in § 25 below, where the 'reference' point is (0, 1, 1, 3, 3, 6) and modes are obtained by unit increases in any 3 of its first 5 co-ordinates.

A practical search procedure

19. To solve the 'equation' $\Sigma u = n$, where $u = I(Np)$, for N and thus for the u, a method of successive approximation with a first trial value $T = n + \frac{1}{2}r$ is often practicable. If, for this T, all Tp are fractions and, with $v = I(Tp)$, $\Sigma v = n$, then the mode is (v); and this comment applies also to the case where some of the Tp are integers provided that certain selections of the now ambiguous v satisfy $\Sigma v = n$. Cf. examples A and B in § 25 below. If none of the (v) satisfy $\Sigma v = n$ then this T is abandoned.

20. Suppose now that $\Sigma v = n-1$, then clearly the correct values of N is greater than T and a suitable next trial value of N is $T' = T+1$, since on the average $\Sigma I(Tp)$ increases at the same rate as T. But an even better value of T' can be found if we detect that particular one of the $Tp\,(\equiv v+f)$ which is proportionately closest to its $v+1$, i.e. for which $(1-f)/(v+1)$ is least; for clearly as T increases, all Tp increase proportionally, and this one will be the first to pass its upper integer neighbour, whereupon its 'v' (and no other) increases by 1 and Σv takes the correct value. Note now the very useful feature that the actual value of T' (i.e. of N) need not enter into this method; we merely find the particular co-ordinate (of the first approximation) which must be increased by one. If, still in the case $\Sigma v = n-1$, there are $d\,(>1)$ equal minimum test fractions $(1-f)/(v+1)$ then there are d joint modes. Similar remarks apply to the case $\Sigma v = n+1$ where of course the test fractions are now f/v and the least of them indicate(s) the co-ordinate(s) to be reduced. Thus in example C a value $T' = 19$ will succeed after $T = 20$ has failed or alternatively the fourth test fraction (i.e. f_4/v_4) is clearly least.

21. If $\Sigma v = n+2$, we may of course take $T' = T-2$ and proceed. But here, too, the 'test-fraction' method is superior despite a slight new complication; the complication lies in the fact that if one p and therefore one v is rather larger than the others it is conceivable that proportional reduction of all the Tp may take this $v+f$ down below its v and then below its $v-1$ before any other $v+f$ has even fallen below its v. So the 'test-fraction' method proceeds in series rather than in parallel. That is, the two least f/v do not necessarily point to two co-ordinates to be reduced by one each. Rather, the least f/v is designated by a * (say) and replaced by $(f+1)/(v-1)$, i.e. the proportional gap to the currently contemplated 'v'; then the least of the new set (comprising this one amendment) attracts a *. Now every v is reduced by its corresponding number of *'s to give the mode (u).

22. The generalization to the case $|\Sigma v - n| = h$ will be obvious; (L) locate minimum fraction, (S) star it, and (A) amend it; then **L, S, A, L, ..., A, L, S**—stopping after h stars, and finally amend the v's as indicated by the stars; joint modes fit very naturally into this process for if, with g stars still to be placed, $s\,(>g)$ equal minimum test fractions are found, then there are $\binom{s}{g}$ joint modes obtained in the obvious way. Example D has $g = 1$, $s = 3$, with the second, fourth, and fifth fractions being equal minima in the second stage. Example E has $g = 2$, $s = 5$, in the first stage.

23. Of course, a computer whom (or which) we wish to spare a long search for N (e.g. in a case of joint modes) can very easily be instructed in the 'test-fraction' method. For a human computer there is usually no question of evaluating all these fractions precisely—it is merely a matter of finding the least, and the necessary evaluations are simplified by the fact that the denominators are always integers.

The case of a finite population

24. If there are M_i elements of type $i\,\{i=1,2,\ldots,r;\ \Sigma M = G\}$ in the population and a sample of n, drawn without replacement, contains k_i of type $i\,(i=1,2,\ldots,r;\ \Sigma k = n)$ the probability of this outcome (k_i) is proportional to

$$\prod_{i=1}^{r} \binom{M_i}{k_i}. \tag{13}$$

The condition, corresponding to (4) above, that a point (u) has greater probability than its $[a,b]$ neighbour is

$$\frac{u_a+1}{M_a+1} > \frac{u_b}{M_b+1}.$$

Thus an infinite population with p_i proportional to corresponding (M_i+1) of the finite population gives the same modes (u) as this finite population. It follows immediately that the mode(s) may be found by 'adjusting' c so that the u defined by $u = I\{c(M+1)\}$ satisfy $\Sigma u = n$. An appropriate initial value of c is $(n+\frac{1}{2}r)/(G+r)$.

'Maximum likelihood' constitution of a finite population

25. Here we suppose that the M_i of § 24 are unknown but that $G = \Sigma M$ is known and that one sample of n has been drawn and found to have constitution (k). The population constitution (M) having maximum probability of yielding such a sample is sought. In contemplating the various conceivable (M), the $[a,b]$ neighbour of a given (M) may be defined as in § 8. From (13) it is easily found that the condition for (M) to be a more likely source (of a given (k)) than its $[i,j]$ neighbour is

$$\frac{M_i+1}{k_i} > \frac{M_j}{k_j}. \tag{14}$$

The correspondence between (14) and (4) is complete and hence the present problem is reduced to the previous. The result may be stated as follows: if λ is such that the r quantities $M = I(\lambda k)$ satisfy $\Sigma M = G$, then (M) is the most likely constitution of the population corresponding to the sample constitution (k). A first approximation to a suitable value of λ is $(G+\frac{1}{2}r)/n$. Families of jointly most likely (M) form complete constellations as described in § 18.

Examples

26. The outcomes of 17 throws of a biased die which has probabilities p_i for the various faces, in a single throw, form an example with $n = 17$, $r = 6$. The table gives the modes of such distributions for five selected single-throw probabilities $(p_1, \ldots, p_6)$.

Ex.	§	$100 \times (p_1, \ldots, p_6)$	$(u_1, \ldots, u_6)$
A	19	(7, 12, 13, 19, 21, 28)	(1, 2, 2, 3, 4, 5)
B	19	(6, 11, 12, 20, 23, 28)	(1, 2, 2, 3, 4, 5)
C	20	(7, 11, 12, 21, 22, 27)	(1, 2, 2, 3, 4, 5)
D	22	(6, $10\frac{1}{2}$, 11, 21, 21, $30\frac{1}{2}$)	(1, 1, 2, 4, 4, 5) (1, 2, 2, 4, 3, 5)
			(1, 2, 2, 3, 4, 5)
E	18	the $p_i \times 1000$ are	(0, 1, 2, 4, 4, 6) (1, 1, 2, 3, 4, 6)
		(52, 104, 104, 208, 208, 324)	(0, 2, 1, 4, 4, 6) (1, 1, 2, 4, 3, 6)
			(0, 2, 2, 3, 4, 6) (1, 2, 1, 3, 4, 6)
			(0, 2, 2, 4, 3, 6) (1, 2, 1, 4, 3, 6)
			(1, 1, 1, 4, 4, 6) (1, 2, 2, 3, 3, 6)

The current best procedure of § 1 above would succeed in isolating four, ten, four, ten, and twenty-five possible modes for further testing in cases A, B, C, D and E, respectively.

27. If 17 counters are drawn from an urn containing 13, 23, 25, 37, 41, 55 counters marked A, B, C, D, E, F respectively, the most probable composition of the sample is (1, 2, 2, 3, 4, 5) of the respective letters.

Reference

Feller, W. (1957). *An Introduction to Probability Theory and its Applications*. New York: John Wiley and Sons.

The analysis of Poisson regression with an application in virology

By JOHN J. GART

*The Johns Hopkins University**

1. Introduction and summary

Consider a set of independent Poisson variates y_{ij} with means βx_i, where $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, m$; the x's are known constants, and β is an unknown parameter. This paper considers point and interval estimation for β and tests of goodness of fit of the linear model. Techniques proposed by Cox (1953) and Neyman (1959) prove useful in deriving these results. The results are applied to data from a pock-counting experiment in virology. Finally, extensions to the comparison of regression coefficients are indicated.

2. A review of Cox's chi-squared approximation for Poisson variates

Since the Poisson distribution enjoys the reproductive property we have, for all i, that $Y_{i.} = \sum_j y_{ij}$ is a Poisson variate with mean $\beta n_i x_i$. Recalling the relationship between the partial sum of the Poisson distribution and the incomplete integral of the chi-squared distribution, we may write

$$\Pr[Y_{i.} \geqslant k] = \Pr[\chi^2_{2k} < 2\beta n_i x_i],$$

and

$$\Pr[Y_{i.} > k] = \Pr[\chi^2_{2k+2} < 2\beta n_i x_i].$$

* Paper no. 347 from the Department of Biostatistics.