Iranian Journal of Science and Technology, Vol. 4, pp. 77–83. Pergamon Press 1975. Printed in Great Britain.

THE NUMBER OF HYPOTHESES OF INDEPENDENCE FOR A RANDOM VECTOR OR FOR A MULTIDIMENSIONAL CONTINGENCY TABLE, AND THE BELL NUMBERS*

I. J. GOOD

Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, U.S.A.

(*Received* 20 *April* 1975)

Abstract—Given an m-dimensional random vector or contingency table, the number of distinct hypotheses of independence is discussed and asymptotic formulae are obtained by using the saddlepoint method. There is one 'exhaustive' hypothesis corresponding to each class partition (with one exception) of the m dimensions. Both the exhaustive hypotheses and mixed conditional/marginal independence hypotheses are enumerated.

1. INTRODUCTION

Before reading this introduction it is advisable to read the definitions given in Section 2.

Multidimensional contingency tables occur frequently in medical and sociological statistics up to high dimensionalities and random vectors are even more widely encountered. For the sake of definiteness I shall word the discussion entirely in terms of contingency tables.

To simplify its analysis it is important to break down a given table into statistically independent parts, or approximately independent parts, if possible (although in the past breakdown into dependent parts has been customary). The problem of designing an efficient search for such a breakdown is unsolved and it is not the purpose of this paper to attack this problem, but one method would be to consider every possible breakdown and to compute a significance criterion for each possibility. (A Bayesian approach to this problem has been given [8].) It may not be sufficient to break the dimensions of the table into pairs of sets, and then to break up the sets further, as in a dichotomous dendroidal search strategy, because, although the independence of three tables does imply pairwise independence, it is more informative to have a statistically independent breakdown into three tables than to have only a partial breakdown. Hence an exhaustive search might be necessary although there may be more economical methods that have only a small chance of missing an 'optimal' breakdown. As a preliminary guide to deciding what kind of search to make it should be useful to know the number of possibilities because the running time of the program, and the amount of output, can then be better estimated. This rather obvious point seems to have been ignored or overlooked in the literature, presumably because the techniques for searching for independence have not yet become systematic, and have been as much intuitive as scientific. It has however, been known for some time that there are many hypotheses of independence and conditional independence for multidimensional contingency tables of high dimensionality: see, e.g. [9, 12].

The combinatorial problem of enumerating the possible hypotheses of independence for an m-dimensional table is the topic of the present paper, and asymptotic formulae are also given, but we do not deal here with statistical tests of significance. The immediate relationship to exponential or Bell numbers is pointed out. The number of possible hypotheses grows rapidly with m and the asymptotic formulae provided give results of considerable accuracy.

We shall *not* consider the even larger number of hypotheses of the vanishing of 'interactions' of the second and higher orders, although these can be considered as a kind of generalized independence especially in virtue of their connections with maximum entropy. (See e.g. [7, 13, 17].)

2. DEFINITIONS

Consider an m-dimensional (m-way) 'population contingency table' with cell probabilities $p(i_1, i_2, \ldots, i_m)$ $(1 \leq i_j \leq I_j)$. When one or more 'coordinates' is 'summed out' we write an asterisk

*This work was supported in part by the U.S. Department of Health, Education and Welfare, N.I.H. Grant No. 1 RO1 GM18770.

in the corresponding place in the notation, e.g.

$$\sum_{i_2=1}^{I_2}\sum_{i_4=1}^{I_4} p(i_1, i_2, \ldots, i_m) = p(i_1, *, i_3, *, i_5, \ldots, i_m).$$

The various hypotheses of independence that we shall consider will not involve the breaking up of the table along any of its m directions, so that there will be no further mention of the symbols $I_1, I_2, \ldots, I_m$ in this paper.

Two or more mutually exclusive subsets of the m *directions* are said to be *marginally independent* if the population probabilities factorize in the obvious way; e.g. if $m = 10$, then the subsets (2, 5), (3, 6, 10), and (4, 7) are marginally independent if

$$\begin{aligned} p(*, i_2, i_3, i_4, i_5, i_6, i_7, *, *, i_{10}) = {} & p(*, i_2, *, *, i_5, *, *, *, *, *) \\ & \times p(*, *, i_3, *, *, i_6, *, *, *, i_{10}) \\ & \times p(*, *, *, i_4, *, *, i_7, *, *, *) \end{aligned}$$

for all values of i_2, i_3, i_4, i_5, i_6, i_7, and i_{10}.

On the other hand the mutually exclusive subsets will be said to be *conditionally* independent if, for all values of the coordinates we have the natural factorization; e.g. if

$$\begin{aligned} p(i_1, i_2, i_3, i_4, \ldots, i_{10}) = {} & p(i_1, i_2, *, *, i_5, *, *, i_8, i_9, *) \\ & \times p(i_1, *, i_3, *, *, i_6, *, i_8, i_9, i_{10}) \\ & \times p(i_1, *, *, i_4, *, *, i_7, i_8, i_9, *), \\ & \div [p(i_1, *, *, *, *, *, *, i_8, i_9, *)]^2, \end{aligned}$$

for all values of $i_1, i_2, \ldots, i_{10}$, then the subsets (2, 5), (3, 6, 10) and (4, 7) are conditionally independent, given the components or directions number 1, 8, and 9.

If the mutually exclusive subsets are exhaustive of the m directions, so that the remaining directions constitute the null set, then both the kinds of independence just defined reduce to the same thing, which we may call *exhaustive independence.* Hypotheses of this kind may be thought of as a special case of marginal and conditional independence in the degenerate case where the number of 'omitted' directions happens to reduce to zero.

Finally there are *mixed (conditional/marginal) independence hypotheses,* conditional with respect to s' of the omitted directions and marginal with respect to the s remaining ones. Here again it will be convenient to think of the degenerate cases, where s' or s or both are zero, as special cases.

For an m-dimensional table, let the number of exhaustive independence hypotheses be denoted by β_m. Clearly, if we put $\beta_1 = 0$, we have

$$\beta_m = b_m - 1 \tag{1}$$

where b_m is the number of ways in which m different things can be distributed into m or fewer 'indifferent [indistinguishable] parcels' (where a parcel must contain at least one object), to use the terminology of [19]. A more classy description of b_m is given, e.g. by [3, p. 103] as the number of 'class partitions' of a set of m members. The numbers b_m are known as *exponential numbers* or *Bell numbers.* The former name presumably arises because of the generating function (e.g. [19, Proposition XXIV], or [3, p. 103]),

$$\sum_{m=0}^{\infty} b_m x^m / m! = \exp(e^x)/e \tag{2}$$

if b_0 is defined as 1. For completeness we draw a few deductions from (2), in (3)–(6), though none of these is new. (For an extensive bibliography, see [10].) Of the three obvious ways of expanding (2), as

$$\prod_{n=1}^{\infty} \exp(x^n/n!), \quad \sum_{n=0}^{\infty} \left(x + \frac{x^2}{2!} + \cdots\right)^n \Big/ n!, \text{ and } \frac{1}{e}\sum_{n=0}^{\infty} e^{nx}/n!,$$

the third is the most interesting, and leads at once to

$$b_m = \frac{1}{e}\sum_{r=0}^{\infty}\frac{r^m}{r!}. \tag{3}$$

Therefore (e.g. [1])

$$b_n e = 1^{m-1} + \frac{2^{m-1}}{1!} + \frac{3^{m-1}}{2!} + \cdots$$
$$= \sum_{\mu=0}^{\infty}\frac{1}{\mu!}\sum_{s=0}^{m-1}\binom{m-1}{s}\mu^{m-1-s}$$

so that, by reversing the order of summation, we see that[16]

$$b_m = b_{m-1} + \binom{m-1}{1}b_{m-2} + \binom{m-1}{2}b_{m-3} + \cdots + b_0. \tag{4}$$

Now one of the definitions of the Stirling numbers of the second kind [11, p. 168] is that they satisfy the identity

$$r^m = \mathcal{S}_m^{(1)}r + \mathcal{S}_m^{(2)}r(r-1) + \mathcal{S}_m^{(3)}r(r-1)(r-2) + \cdots. \tag{5}$$

Therefore, from (3) we have[2]

$$b_m = \mathcal{S}_m^{(1)} + \mathcal{S}_m^{(2)} + \cdots + \mathcal{S}_m^{(m)}. \tag{6}$$

The numbers $\mathcal{S}_m^{(n)} = (\Delta^n 0^m)/n!$ are tabulated in [11, p. 170] up to $m = 12$ ($n \le m$), and in [5, Table XXII] up to $m = 25$. The values of b_m up to $m = 20$ are given in our Table 1, up to $m = 51$ in [15], and up to $m = 74$ in [14].

Table 1. Values of b_m and of the first three terms of the asymptotic expansion

The last column gives the ratio of the sum of the 3 terms divided by the exact value.

m	$b_m = \beta_m + 1$	First term	2nd term	3rd term	Sum	Ratio
1	1	0.934	0.088	0.016	1.037	1.037
2	2	1.870	0.1206	0.0116	2.0023	1.0011
3	5	4.7583	0.2301	0.0129	5.0013	1.00026
4	15	14.440	0.5531	0.0184	15.011	1.00076
5	52	50.412	1.586	0.030	52.028	1.00054
6	203	197.772	5.250	0.046	203.067	1.00033
7	877	857.590	19.589	0.031	877.210	1.00024
8	4140	4060.020	81.050	-0.300	4140.770	1.00019
9	21147	20785.719	367.198	-2.848	21150.069	1.00015
10	1 15975	114204.254	1803.8	-19.7	115988.3	1.00011
11	6 78570	669229.55	9531.6	-128.3	678632.9	1.000093
12	42 13597	41 60921	53831	-835	42 13918	1.000076
13	276 44437	273 28564	3 23177	-5555	276 46186	1.000063
14	1908 99322	1888 94654	20 52988	-38159	1909 09438	1.000053
15	13829 58545	13695 47122	137 45573	-271615	13830 21079	1.000045
20	5172 41582 35372	5.13680(13)	3.658(11)	-8.60(9)	5.17252(13)	1.000020
30	8.46749 01451(23)	8.43259(23)	3.59226(21)	-9.522(19)	8.46756(23)	1.0000083
50	1.85724 26877(47)	1.853462(47)	3.8948(44)	-1.109(43)	1.857246(47)	1.0000017

b_{16} = 104801 42147 b_{17} = 828648 69804 b_{18} = 68 20768 06159 b_{19} = 583 27422 05057

The combinatorial interpretation of (6) is that $\mathcal{S}_m^{(n)}$ is the number of partitions of m objects into precisely n indistinguishable parcels (cf. [18, p. 91]). An asymptotic expansion for b_m is discussed in Section 5.

For an m-dimensional table, let γ_m denote the total number of mixed conditional/marginal independence hypotheses, including the degenerate ones. The number of these having r 'omitted' directions of which s are 'summed out' is clearly

$$\binom{m}{r}\binom{r}{s}\beta_{m-r}. \tag{7}$$

Therefore the number with r 'omitted' directions is, in all

$$\sum_{s=0}^{r}\binom{m}{r}\binom{r}{s}\beta_{m-r} = \binom{m}{r}2^r\beta_{m-r}. \tag{8}$$

Therefore (remembering (4) and that $\beta_1 = 0$), we see that the total number of *purely* marginal independence hypotheses, which is also the number of purely conditional independence hypotheses, is

$$\begin{aligned}\beta_m + \binom{m}{1}\beta_{m-1} + \cdots + \binom{m}{m-1}\beta_1 &= b_m + \binom{m}{1}b_{m-1} + \cdots + \binom{m}{m-1}b_1 - (2^m - 1)\\ &= b_{m+1} - 2^m = \beta_{m+1} - (2^m - 1).\end{aligned} \tag{9}$$

Therefore

$$\begin{aligned}\gamma_m &= \beta_m + \binom{m}{1}2\beta_{m-1} + \binom{m}{2}2^2\beta_{m-2} + \cdots + \binom{m}{m-2}2^{m-2}\beta_2\\ &= b_m + \binom{m}{1}2b_{m-1} + \binom{m}{2}2^2b_{m-2} + \cdots + \binom{m}{m-2}2^{m-2}b_2 + \binom{m}{m-1}2^{m-1}b_1 + 2^m b_0\\ &\quad - \left[1 + \binom{m}{1}2 + \binom{m}{2}2^2 + \cdots + 2^m\right]\\ &= b_m + \binom{m}{1}2b_{m-1} + \cdots + 2^m b_0 - 3^m.\end{aligned} \tag{10}$$

Therefore

$$\begin{aligned}\gamma_m + 3^m = \Big\{&m!\,\mathscr{C}(z^m) + \binom{m}{1}2(m-1)!\,\mathscr{C}(z^{m-1})\\ &+ \binom{m}{2}2^2(m-2)!\,\mathscr{C}(z^{m-2}) + \cdots\Big\}\exp(e^z - 1)\end{aligned}$$

where $\mathscr{C}(z^r)$ means 'the coefficient of z^r in'. Therefore

$$\begin{aligned}\gamma_m + 3^m &= m!\left\{\mathscr{C}(z^m) + \frac{2}{1!}\mathscr{C}(z^{m-1}) + \frac{2^2}{2!}\mathscr{C}(z^{m-2}) + \cdots\right\}\exp(e^z - 1)\\ &= m!\,\mathscr{C}(z^m)\left\{\left[1 + \frac{2z}{1!} + \frac{2^2z^2}{2!} + \cdots\right]\exp(e^z - 1)\right\}\\ &= m!\,\mathscr{C}(z^m)\exp(e^z + 2z - 1).\end{aligned}$$

In other words

$$\sum_{m=0}^{\infty}\frac{\gamma_m}{m!}z^m = \exp(e^z + 2z - 1) - e^{3z} \tag{11}$$

which may be compared with

$$\sum_{m=0}^{\infty} \frac{\beta_m}{m!} z^m = \exp(e^z - 1) - e^z. \tag{12}$$

The values of γ_m for $m = 1(1)10$ are given in Table 2.

Asymptotic theory. Asymptotic values of considerable accuracy can be obtained for β_m and γ_m by using the saddlepoint method. The results are of some mathematical interest although perhaps not for the application to contingency tables. The saddlepoint method is applied in [4] to the problem of estimating the mean of n independent identically distributed random variables and the approximation is carried there to the second term. The calculation of the third term was given in [6]. The results of the heavy algebra in that work can be used for our purposes although our problem is different.

[15] and [3, pp. 104–108] consider the asymptotic expansion of b_m as an example of the saddlepoint method. Our method is close to that of the former authors but it seems worthwhile to record it because (i) it illustrates the general formulae of [6] and this makes the details more straightforward, (ii) the present results are taken to greater accuracy, and (iii) the method applies readily to γ_m.

By (2) and (11) we have

$$b_m = \frac{m!}{2\pi i e} \oint \frac{\exp(e^z)\,dz}{z^{m+1}}, \quad \gamma_m = \frac{m!}{2\pi i e} \oint \frac{\exp(e^z + 2z)}{z^{m+1}}\,dz \tag{13}$$

where the contours encircle the origin once in the positive direction. The integrands are of the form $\exp f(z)$ and $\exp g(z)$ where

$$f(z) = e^z - (m+1)\log z, \quad g(z) = e^z + 2z - (m+1)\log z. \tag{14}$$

Saddlepoints occur where the derivatives $f'(z)$ and $g'(z)$ vanish. Although there are an infinite number of saddlepoints, the important ones occur on the real axis, at say $z = \xi(m) = \xi$ and $z = \eta(m) = \eta$ respectively, if m is not too small. By the 'important' saddlepoints we mean the ones where the absolute values of the integrands (which are equal to the exponentials of the real parts of f and g) are maximal along suitable contours. [3] gives this for b_m and the comment is equally valid for γ_m. Moreover, again as pointed out in [3] for b_m, we can in both cases replace the

Table 2. Values of γ_m and of the first three terms of the asymptotic expansion

The last column gives the ratio of the sum of the three terms divided by the exact value. The 'first term' is $C_m^{(0)} - 3^m$, the second and third terms are $C_m^{(0)} g_1$ and $C_m^{(0)} g_2$ respectively.

m	γ_m	First term	2nd term	3rd term	Sum	Ratio
2	1	0.359	0.535	0.085	0.979	0.979
3	10	7.886	1.805	0.244	9.935	0.9935
4	70	62.563	6.539	0.739	69.841	0.9977
5	431	402.476	25.742	2.414	430.632	0.99915
6	2534	2414.532	110.055	8.541	2533.128	0.99966
7	14820	14276.477	508.820	32.647	14817.944	0.99986
8	88267	85598.026	2530.816	134.239	88263.081	0.999956
9	5 42912	5 28850.177	13472.523	590.002	5 42912.702	1.0000013
10	34 75978	33 96912.02	76393.79	2754.97	34 76060.78	1.000024
20	-	3.9923532(15)	3.95997(13)	-1.42(10)	4.0319529(15)	-
30	-	1.1710192(26)	6.7833(23)	-7.75(21)	1.1777249(26)	-
50	-	5.4792788(49)	1.51576(47)	-2.860(45)	5.4944364(49)	-

contour by a line through the main saddlepoint and parallel to the imaginary axis. The real saddlepoints are given by the equations

$$\xi = \log_e\left(\frac{m+1}{\xi}\right), \quad \eta = \log_e\left(\frac{m+1}{\eta} - 2\right) \tag{15}$$

each of which has a unique positive root. These transcendental equations can each be solved iteratively without meditation by substituting the 'old' value into the right side, computing an intermediate value from the equation, and averaging the old and intermediate values to obtain a new value. The new value of course becomes the 'old' value for the next step in the iterative process. An example that the reader might like to use as a check is $\eta(8) = 1{\cdot}44342422122$ correct to twelve significant figures (a curious-looking number). [15] uses the value of ξ with m decreased by 1.

Now we have

$$b_m \approx \frac{m!\exp f(\xi)}{2\pi e}\int_{-\xi'}^{\xi'} \exp\left\{-\frac{y^2}{2}f''(\xi) - \frac{iy^3}{6}f'''(\xi) + \cdots\right\} dy, \tag{16}$$

where ξ' is slightly smaller than ξ. (The radius of convergence of the expanded function is equal to ξ.) Thus

$$b_m \approx \frac{m!\exp f(\xi)}{2\pi e}\int_{-\xi'}^{\xi'} \exp\left\{(m+1)\left[-\frac{1}{2}\kappa_2 y^2 - \frac{i\kappa_3}{6}y^3 + \cdots\right]\right\} dy, \tag{17}$$

where

$$\kappa_r = \frac{\xi^{r-1} + (-1)^r(r-1)!}{\xi^r}. \tag{18}$$

The integrand now happens to be exactly the same as [6, p. 869] with $m+1$ written for t. It therefore follows from that paper that

$$b_m \sim B_m^{(0)}(1 + f_1 + f_2 + \cdots) \tag{19}$$

where

$$B_m^{(0)} = \frac{m!\exp(e^\xi)}{\xi^m e\sqrt{[2\pi(m+1)(\xi+1)]}} \tag{20}$$

$$f_1 = \frac{1}{24(m+1)}(3\lambda_4 - 5\lambda_3^2) \tag{21}$$

$$f_2 = \frac{1}{1152(m+1)^2}(168\lambda_3\lambda_5 + 385\lambda_3^4 - 630\lambda_3^2\lambda_4 - 24\lambda_6 + 105\lambda_4^2) \tag{22}$$

where

$$\lambda_r = \kappa_r\kappa_2^{-r/2} = \frac{\xi^{r-1} + (-1)^r(r-1)!}{(\xi+1)^{r/2}}. \tag{23}$$

The numerical approximations for $m = 1(1)15$ are shown in Table 1 from which it will be seen that the third approximation is remarkably accurate. The formula for $B_m^{(0)}$ is equivalent to a formula given in [3], but [3] does not give later terms explicitly. Also [3] is concerned with finding an asymptotic formula for ξ in terms of m, whereas our approach is to regard ξ as readily calculable and to regard it as a known constant once m is assigned, a method previously adopted in [15], with the gloss mentioned before. [15], formula (3.8) gives fractional errors of 0·00057, 0·00017, and 0·000033 for $m = 10$, 20, and 50 respectively. The manipulations required to obtain improved approximations are implicit in their work but would be very heavy. The fractional errors of our third-order approximations can be read off from the last column of Table 1.

The procedure for γ_m is similar, but when the factor $m+1$ is taken out in the exponential, as in (17), there remain some explicit occurrences of m within the brackets, in fact we now have

$$\lambda_r = \frac{\eta^{r-1} + (-1)^r (r-1)! - 2\eta^r (m+1)^{-1}}{[\eta + 1 - 2\eta^2 (m+1)^{-1}]^{r/2}} \tag{24}$$

$$\gamma_m + 3^m \sim C_m^{(0)}(1 + g_1 + g_2 + \cdots) \tag{25}$$

where

$$C_m^{(0)} = \frac{m! \exp\left(\dfrac{m+1}{\eta} + 2\eta - 3\right)}{\eta^m \sqrt{\{2\pi[(m+1)(\eta+1) - 2\eta^2]\}}} \tag{26}$$

and g_1 and g_2 are given by the right sides of (21) and (22) except that the λ's are now given by (24). Note that we have not rearranged the expansion in negative powers of m nor of $m+1$ because this would make the formulae more complicated. The results, given in Table 2, are again strikingly good. That the three-term approximation to γ_9 is better proportionately than that for γ_{10} must be because the error does not have a fixed sign. Where the error changes sign it has a 'chance' of being extra small.

REFERENCES

1. E. T. Bell, Exponential numbers, *Am. Math. Monthly* **41**, 411–419 (1934).
2. V. Broggi, Su di qualche applicazione dei numeri di Stirling, 1*st Lombardo Rend.* ser. II **66**, 196–202 (1933).
3. N. G. de Bruijn, *Asymptotic Methods in Analysis.* North Holland, Amsterdam (1958).
4. H. E. Daniels, Saddlepoint approximations in statistics, *Ann. Math. Statist.* **25**, 631–650 (1954).
5. R. A. Fisher and F. Yates, *Statistical Tables.* Oliver & Boyd, Edinburgh (1953).
6. I. J. Good, Saddle-point methods for the multinomial distribution, *Ann. Math. Statist.* **28**, 861–881 (1957).
7. I. J. Good, Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, *Ann. Math. Statist.* **34**, 911–934 (1963).
8. I. J. Good, On the application of symmetric Dirichlet distributions to contingency tables. Submitted for publication, 1974.
9. L. A. Goodman, Partitioning of χ^2, analysis of marginal contingency tables, and estimation of expected frequencies in multi-dimensional contingency tables. *J. Am. Statist. Assoc.* **66**, 339–344 (1971).
10. H. W. Gould, Research bibliography of two special number sequences, *Mathematica Monongaliae* No. 12, pp. 25 & iv (1971).
11. Charles Jordan, *Calculus of Finite Differences.* Chelsea, New York (1965).
12. M. A. Kastenbaum, Analysis of categorical data: some well-known analogues and some new concepts. *Communications in Statistics.* **3**, 401–417 (1974).
13. H. O. Lancaster, Contingency tables of higher dimensions, *Proc. 37th Session Bull. I.S.I.* **43**, 143–151 (1969).
14. Jack Levine and R. E. Dalton, Maximum periods, modulo p, of first-order Bell exponential integers, *Math. Computation* **16**, 416–423 (1962).
15. Leo Moser and Max Wyman, An asymptotic formula for the Bell numbers, *Trans. Roy. Soc. Canada* **49**, 49–54 (1955).
16. M. d'Ocagne, Sur une classe de nombre remarquables, *Am. J. Math.* **9**, 353–380 (1887).
17. R. L. Plackett, Multidimensional contingency tables: a survey of models and methods, *Proc. 37th Session Bull. I.S.I.* **43**, 133–142 (1969).
18. J. Riordan, *An Introduction to Combinatorial Analysis.* Wiley, New York (1958).
19. W. A. Whitworth, *Choice and Chance with One Thousand Exercises.* Hafner, New York (1901/1951).