BIOMETRICS 51, 1557–1561
December 1995

The Poisson Distribution and "Multiple Events" Cell Occupancy Problems

James A. Hanley[1] **and Brenda MacGibbon**[2]

[1]Department of Epidemiology and Biostatistics, McGill University,
1020 Pine Ave West, Montréal, Québec H3A 1A2, Canada
[2]Departement de mathématiques et d'informatique, Université du Quebec à Montréal,
C.P. 8888, Succursale "centre-ville," Montréal, Québec H3C 3P8, Canada

SUMMARY

This article studies a variant of the birthday problem called the "multiple events problem." It asks the following general cell occupancy question: if one distributes N items or units at random over c equally likely categories or cells, what is the number N of items required in order to ensure that the probability of at least one cell having k or more items is greater than or equal to P? The question is of particular interest to epidemiologists who investigate "clusters" of disease. Although the exact probability, with fairly sophisticated and numerically complicated solutions to this problem, has appeared in the literature, we prefer here to present an intuitive solution, readily explainable to non-mathematicians and easy to calculate using standard available statistical or spreadsheet software. The key to the result is to use two successive applications of the Poisson approximation and to take advantage of the underused relationship between the Poisson and χ^2 distributions. We illustrate the methods in two typical epidemiologic applications.

1. Introduction

Diaconis and Mosteller (1989) studied the number N required to have a probability P of k or more units in at least one category if one distributes N units at random over c equally likely categories. This "multiple events" variation on the birthday problem is not only of mathematical interest; for example, the answers can help epidemiologists who investigate "clusters" of disease and wonder if they could arise by chance alone. Thus, it would be useful if the solution to this problem were intuitive, readily explainable to non-mathematicians and easy to calculate using standard available statistical software.

For the case of $c = 365$ and $P = .5$, Diaconis and Mosteller give the required values of N for $k = 2$ to 13; the results were calculated for them by Levin, using an algorithm he developed by reversing the representation of the multinomial distribution as the conditional distribution of independent Poisson random variables given a fixed sum (Levin, 1981). The procedure allows one to calculate P for a given N. It may require an Edgeworth expansion, and in any case, one cannot directly solve for N as a function of c and P. Diaconis and Mosteller supply an approximate equation from their unpublished work that links N, c, k, and P when c is large; this equation does not yield a closed-form solution for N, but it can be solved by trial and error.

In the original "birthday problem," that is, in the calculation that for a given N and c there are no cells with $k = 2$ or more items, a single Poisson approximation works quite well; this remains true even when the cell occupancy probabilities are not necessarily equal (see Gail et al., 1979). The purpose of this note is to demonstrate how two successive applications of the Poisson approximation, together with the (not widely known) link between the Poisson and χ^2 distributions, can directly yield practically the same answers to the multiple events problem as those obtained by Levin and by Diaconis and Mosteller using much more complicated numerical methods. Moreover, the basis for the calculation can be explained easily and the calculations can be implemented using widely available software. We illustrate the methods of calculation using two examples from epidemiology.

Key words: Birthday problem; Clusters; Coincidences; Epidemiology.

2. Formulation and Solution

Suppose one distributes each of N items at random over c equally likely cells. Then the number, x, of items in a cell will have a binomial distribution with N trials and success probability $1/c$. The expected value of x, or equivalently, the average number of items per cell, is

$$\mu_x = N/c. \tag{1}$$

The expected proportion, π, of cells that have $\geqslant k$ units is the same as the probability that a cell has $\geqslant k$ items, and it can be calculated as the binomial tail area:

$$\pi = \sum_{x \geqslant k} \text{BinomialProb}(x|N, \text{ prob} = 1/c)$$

If c is large, π can be approximated using the Poisson distribution with mean μ_x, so that we can write it as

$$\pi = \sum_{x \geqslant k} \text{PoissonProb}(x|\mu_x) = Q[k|\mu_x], \text{ say.} \tag{2}$$

Denote by y the number of cells, out of c, that have k or more items. Then the expected value, μ_y, of y is

$$\mu_y = c\pi. \tag{3}$$

The probability P of k or more items in at least one cell is equivalent to the probability that y is greater than 0. This probability can be approximated, using the Poisson distribution with mean μ_y, as

$$P = 1 - \text{Prob}(y = 0|\mu_y) = 1 - \exp(-\mu_y). \tag{4}$$

These approximations depend on c being large enough that (a) the binomial distribution describing the number of items x in a cell is well approximated by the Poisson and (b) the 0:1 variable indicating whether, for any cell, $x < k$ or $x \geqslant k$, has negligible covariance with the corresponding indicator for any other cell.

To solve for N in terms of c, k, and P, we solve the four equations in reverse to get

$$N = cQ^{-1}[-\ln(1 - P)/c|k]. \tag{5}$$

Equation (2) can be solved for μ_x using the relationship (known to Fisher in 1935) between the tail area of the Poisson distribution and that of the χ^2 distribution with $2k$ degrees of freedom, that is:

$$\sum_{x \geqslant k} \text{Prob}(x|\mu_x) = \text{Prob}(\chi^2_{2k} \leqslant 2\mu_x). \tag{2'}$$

To do so, we invert equation (2′) to calculate μ_x directly from k and π as

$$\mu_x = 0.5 \text{ InverseChiSquare}(\pi, 2k), \tag{2''}$$

where InverseChiSquare(cumulative probability, d.f.) is the inverse cumulative function of the χ^2 distribution with d.f. degrees of freedom. It is available in many commonly used spreadsheets and statistical packages. For $k > 10$, it can also be approximated quite well by the Wilson-Hilferty equation

$$\mu_x = k[1 - (9k)^{-1} - z(9k)^{-1/2}]^3,$$

where z is the normal deviate corresponding to an upper tail area of π (Liddell, 1984).

Two versions of this problem can occur in epidemiology when a cluster of k cases of a disease occurs in a defined community or collection of individuals. In the first type, where one knows the number, c, of communities of this same size in the population, but the average number μ_x per community (i.e., the rate) is not known, the question often asked is: how high would μ_x have to be in order for there to be a reasonable probability P of getting at least k cases in at least one of the communities? The answer can be obtained by solving equation (5) for N (the total number of cases in the c communities) and then dividing N by c to get μ_x.

A different version of the question, when the "universe" of communities from which this "cluster" emerged is not known, can be stated as follows: given an average of μ_x cases per community, how many communities c of this same size would there have to be in order for there to be a reasonable probability P of getting at least k cases in at least one community? The probability can be written as

$$P = 1 - \exp(-cQ[k|\,\mu_x]).$$

the solution of which is given by

$$c = -\ln[1 - P]/Q[k|\,\mu_x] = -\ln[1 - P]/\text{Prob}(\chi^2_{2k} \leq 2\mu_x). \tag{6}$$

3. Applications

In order to validate the method, we compared our results with the published solutions to the questions in table 3 of Diaconis and Mosteller (1989). One of them dealt with the least number of people needed to ensure that the probability exceeds $P = 1/2$ that $k = 6$ or more of them will have the same birthday ($c = 365$). The four calculations are as follows:

$$\mu_y = -\ln(1 - 0.5) = .69315;$$

$$\pi = .69315/365 = .001899;$$

$$\mu_x = .5 \text{ InverseChiSquare}(.001899,\ 12) = 1.25826;$$

$$N = 365(1.25826) = 459.266.$$

Levin's method, based on an Edgeworth expansion, gives the required integer as 460. Solving Diaconis and Mosteller's approximate equation by trial and error gives $N = 458.8$. Our solutions for other values of k considered by Diaconis and Mosteller are equally close.

We now consider two variants on this problem in epidemiology. In the first, consider a case of $k = 4$ serious birth defects in children born to 220 women in a community. (We model this hypothetical case after real examples of clusters of adverse pregnancy outcomes in births to women who worked with video display terminals in a newspaper office in 1979 (Abenhaim and Lert, 1991). Most of the clusters considered by these authors involved mixtures of spontaneous abortions, premature births, stillbirths, etc., some of which are common enough that they involve the binomial rather than the Poisson distribution. They estimated the number c of workplaces in North America in 1979 with the same number of pregnant women, as in each reported cluster, working with display terminals).

In our example, suppose that it is estimated that the number of communities with 220 pregnant women was $c = 1000$. Rather than using known rates of serious birth defects, one can ask the question: how high would the rate of serious birth defects have to be in order for there to be a 50% probability of getting at least four cases in at least one community? An answer can be obtained by solving equation (5) for N (the total number of cases in pregnancies in the $c = 1000$ workplaces) and then dividing N by c and then by 220 to get the rate per pregnancy, i.e.,

$$Q^{-1}[-\ln(1 - P)/c]/220.$$

The steps are as follows:

$$\mu_y = -\ln(1 - .5) = .69315;$$

$$\pi = .69315/1000 = .00069315;$$

$$\mu_x = .5 \text{ InverseChiSquare}(.00069315,\ 8) = .387878;$$

$$\text{Rate per pregnancy} = .387878/220 = .00176.$$

This rate of 1.76 per 1000 pregnancies can be compared with the rate that might be considered typical of non-exposed women.

Our second example involves more benign outcomes. As described in Hanley (1992), a newspaper carried the following caption to a photograph showing six sets of twins

> Double trouble in Moose Jaw school: Every morning, teachers at Prince Arthur school in Moose Jaw, Saskatchewan see double—and it's not because of what they were up to the night before. Six pairs of identical twins attend the school, which has an enrollment of 375. Identical twins occur once in 270 births.

Here the situation is just a curiosity; it would have been much more serious if, instead of six twin pairs, it had been six cases of some disease. We are given that $\mu_x = 375/270 = 1.4$, but wonder how many schools of this size would there have to be to have, say, a 50:50 chance that one or more of them had $\geq k = 6$ sets of identical twins. Equation (6), with $P = .5$ and $\mu_x = 1.4$, yields

$$c = -\ln[0.5]/\text{Prob}(\chi^2_{12} \leq 2.8) = .693/.0032 = 217 \text{ schools.}$$

Considering the very large number of schools of this size in North America, the story is a case of what Diaconis and Mosteller call the "law of very large numbers."

Although the newspaper story did not say so explicitly, in our calculations we assumed that the six sets of twins are from six different families. Also, our use of 1.4, obtained by dividing a numerator of 375 children, rather than 369 births, by 270, is deliberately approximate. Exact calculations would involve recalculating the number of births for each k greater than 6 and still maintaining an enrollment at 375.

4. Discussion

The main contribution of this note is a simpler solution to the problem where the probability that an item would fall into each of the c categories is equal to $1/c$. In our epidemiologic illustrations, the categories were institutions or places of employment or communities. Thus, in reality, because of variations in the number of individuals in these categories, there will be considerable variation in the probabilities across the universe of c institutions or workplaces or communities that should be considered. Likewise, epidemiologists have to deal with multiple clusters, of possibly different sizes. It is also plausible that even if equal numbers of individuals were "at risk" in each category, the event rate per individual might itself vary across the categories. A few authors have addressed some of these variants on the birthday problem. For example, Gail et al. (1979) and Nunnikhoven (1992) have considered the problem where the probabilities that an item falls into the c different categories or cells are unequal, while Diaconis and Mosteller (1989) have considered other variants. However, more development is needed to extend the solutions to the real, but less structured, problems of this type encountered by epidemiologists.

Finally, we make a parenthetical remark about the distribution of the variable N, the number of units required for at least one instance of at least one cell having $k = 2$ or more units (Diaconis and Mosteller call this the "standard birthday problem"). They provided the expression $\sqrt{-2c \ln [1 - P]}$ to calculate approximate $100P\%$ percentiles for N. Thus, if $P = .5$, the median N is approximately $(6/5)\sqrt{c}$. It is also possible to calculate the two other measures of central tendency for N. The mode is approximately $\sqrt{c}$ (Hanley, 1984). The mean, obtained by integrating the curve Prob$[N \geq n]$ over n, can be shown to be $\sqrt{c\pi/2}$, which is quite close to $(5/4)\sqrt{c}$.

Acknowledgements

This study was supported by operating grants from the Natural Sciences and Engineering Research Council of Canada.

Résumé

Cet article traite d'une variante du problème des anniversaires de naissance que nous appelons "problème des événements multiples". Il s'agit du problème d'occupation suivant: si on distribue N objets au hasard dans c cellules, quel est le nombre N d'objets nécessaires pour que la probabilité d'avoir au moins une cellule contenant k objets ou plus soit plus grande ou égale à P? Cette question intéresse particulièrement les épidémiologistes qui étudient les maladies survenant "en grappes". La valeur exacte de la probabilité a été publiée mais les solutions proposées sont passablement sophistiquées et numériquement compliquées. Nous présentons ici une solution intuitive à ce problème, accessible à ceux qui ne sont pas mathématiciens, et où les calculs sont facilement réalisables à l'aide d'un progiciel statistique ou d'un chiffrier électronique. La clé du résultat consiste à appliquer deux fois l'approximation de Poisson et à tirer avantage de la relation (peu utilisée) qui existe entre la loi de Poisson et la loi du khi-deux. Nous illustrons les méthodes sur deux applications épidémiologiques typiques.

References

Abenhaim, L. and Lert, F. (1991). Methodological issues for the assessment of clusters of adverse pregnancy outcomes in the workplace: The case of video display terminal users. *Journal of Occupational Medicine* **33,** 1091–1096.

Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association* **84,** 853–861.

Gail, M. H., Weiss, G. H., Mantel, N., and O'Brien, S. J. (1979). A solution to the generalized birthday problem with application to allozyme screening for cell culture contamination. *Journal of Applied Probability* **16,** 242–251.

Fisher, R. A. (1935). The mathematical distributions used in common tests of significance. *Econometrica* **3,** 353–365.

Hanley, J. A. (1984). Lotteries and probabilities: Three case reports. *Teaching Statistics* **6,** 88–92.

Hanley, J. A. (1992). Jumping to coincidences: Defying odds in the realm of the preposterous. *American Statistician* **46,** 197–202.

Levin, B. (1981). A representation for multinomial cumulative distribution functions. *The Annals of Statistics* **9,** 1123–1126.

Liddell, F. D. K. (1984). Simple exact analysis of the standardized mortality ratio. *Journal of Epidemiology and Community Health* **38,** 85–88.

Nunnikhoven, T. S. (1992), A birthday problem solution for nonuniform birth frequencies. *American Statistician* **46,** 270–274.

Received November 1993; *revised May* 1995; *accepted May* 1995.