A NOTE ON BAHADUR'S EXPANSION IN BAYESIAN DIAGNOSTIC ALGORITHMS*

MARIJA J. NORUŠIS

Departments of Biostatistics and Internal Medicine, University of Michigan, Ann Arbor, Michigan, 48104 (*USA*)

(Received: 8 December, 1972)

SUMMARY

Scheinok's (1972) *empirical results, obtained from using Bahadur's expansion in Bayes's theorem, are explained by noting that the expansion is an exact representation of observed probabilities and thus no information was gained by its use. The calculated and observed joint probability distributions will always be equal. It is also demonstrated that posterior probabilities equal to the ratio of observed patients with a given profile in a disease category to the total number of patients with the symptom profile are always obtained when actuarial probability estimates are used in Bayes's theorem.*

SOMMAIRE

Les résultats empiriques de Scheinok obtenus grâce à la variante de Bahadur du théorème de Bayes s'expliquent par le fait que cette variante donne une représentation exacte des probabilités observées et que son usage ne permet pas de gagner de l'information. Les distributions de probabilités qu'elles soient observées ou calculées restant toujours à peu près identiques. On démontre également que des probabilités a postériori, égales au rapport du nombre de patients ayant un profil de symptomes et une maladie donnés au nombre total de patients ayant ce profil, sont toujours obtenues quand on utilise dans la formule de Bayes des estimations calculées de probabilités.

Bahadur (1961) has derived an exact finite series representation for the general multinomial model. Scheinok (1972) has considered its use in a diagnostic context. Since both authors present detailed discussion of the expansion, its derivation will not be considered here, except to point out that $P^*(X)$, the distribution under dependence, is *not*, in the general case, an approximation based on underlying

*Also see this Journal, 3 (1972) pp. 313–315—Ed.

assumptions. Any n-variate binary distribution can be *exactly* represented by Bahadur's expansion. When $P(X)$, the distribution under independence, holds:

$$S = \{1; z_1, \ldots, z_n; z_1 z_2, \ldots, z_{n-1} z_n; \ldots; z_1 z_2 \ldots z_n\}$$

is an orthonormal basis in the space of real valued functions on X. It follows (Bahadur (1961)) that each function, f, on X admits one, and only one, representation as a linear combination of functions in S. That is, any probability distribution is uniquely and completely specified by the set of α_i's and correlation coefficients of all orders.

Scheinok, failing to realise that Bahadur's expansion is exact, arrived at the observed probability distributions by calculating 12,282 parameters and then using:

$$P^*_i(X) = P_i(X)\,[1 + \sum_{i<j}\sum r_{ij}z_i z_j + \sum_{i<j<k}\sum\sum r_{ijk}z_i z_j z_k + \ldots + r_{12\ldots n}z_1 z_2 \ldots z_n] \quad (1)$$

Since Bahadur's representation is nothing more than an expansion of the observed probability distribution in an exact finite series, the estimates $P^*_i(X)$ are identical with the simple actuarial estimates:

$$P_i^a(X) = \frac{\text{number of persons with profile } X \text{ and disease } i}{\text{total number of persons with disease } i} \quad (2)$$

The determination of all higher order correlations was a circular, futile exercise which only demonstrated that, in spite of a million calculations, once an identity always an identity, to within a little rounding error, of course. An algebraic illustration of the equivalence of eqns. (1) and (2) is unwittingly provided by Scheinok in his appendix. For the simple, two-disease, two-symptom case, P^*_i (0·1) is derived as the ratio of persons with profile (0·1) in disease i to the total number of persons with disease i.

It can now be readily shown that, for any combination of m diseases and n symptoms, when actuarial estimates are used, the posterior probability depends only on the frequencies of occurrence of particular diagnoses for any set of profiles. Let n_{ik} be the number of persons in disease category k with symptom configuration S_i. Then, by Bayes's theorem:

$$P(D_k \mid S_i) = \frac{P(S_i \mid D_k)P(D_k)}{\sum_{j=1}^{m} P(S_i \mid D_j)P(D_j)}$$

$$= \frac{(n_{ik}/n_{.k})(n_{.k}/n_{..})}{\sum_{j=1}^{m} (n_{ij}/n_{.j})(n_{.j}/n_{..})}$$

$$= \frac{n_{ik}}{n_{i.}}, \qquad i = 1, \ldots, 2^n, k = 1, \ldots, m$$

where dot indicates summation over the subscript. Hence, the results obtained by Scheinok are hardly mysterious or surprising and the desired generalisation is trivial!

In order to understand the possible usefulness of Bahadur's representation in diagnosis, a brief overview of the general situation is necessary. Although clinicians are forced to make tacit assumptions concerning the correlation structure of symptoms, a satisfactory mathematical model describing correlation patterns, as well as their implications, has yet to be developed. When symptoms cannot be assumed independent, the likelihood of each possible configuration of observations must be estimated separately. When there are n discrete binary measurements a patient may have any of 2^n different combinations of signs and symptoms. If $n = 11$, as in Scheinok's data, 2048 estimates must be made for each disease. The size of available data bases usually precludes such a procedure. To permit the use of Bayes's theorem, the questionable assumption of symptom independence is often invoked, that is, the joint probability of symptoms is estimated as the product of their marginals. However, models intermediate to those of complete independence or total dependence can be postulated. Cox (1972) presents a brief review of some of the main models for the representation of multivariate binary data. Bahadur's expansion, when certain correlation parameters are assumed zero *a priori*, is one possible simplification of the general multinomial distribution for it allows a parameterisation using fewer than $2^n - 1$ parameters. This would properly be identified as Bahadur's *model*, for underlying assumptions have now been made.

In the common situation where the number of symptoms is fairly large while the available data base is sparse, the actuarial method provides many zero probability estimates. If the assumption of symptom independence is not tenable, as is often the case, the use of Bahadur's model could be considered. The joint probability distribution, determined by using only certain order correlation coefficients calculated from a data base, might be preferable to one derived from actuarial estimates alone, especially when the underlying assumptions are met and the sample is small. However, until investigations into the correlation structure of symptoms are conducted, the choice of reasonable assumptions remains unresolved.

REFERENCES

BAHADUR, R. R., A representation of the joint distribution of responses to n dichotomous items, in: *Studies in Item Analysis and Prediction*, ed.: H. Solomon, Stanford University Press, Stanford, California, 1961, pp. 158–68.

COX, D. R., The analysis of multivariate binary data, *Applied Statistics*, **21** (1972) p. 113.

SCHEINOK, P., Symptom diagnosis: Bayes's Theorem and Bahadur's Distribution, *Int. J. Bio-Med. Computing*, **3** (1972) p. 17.