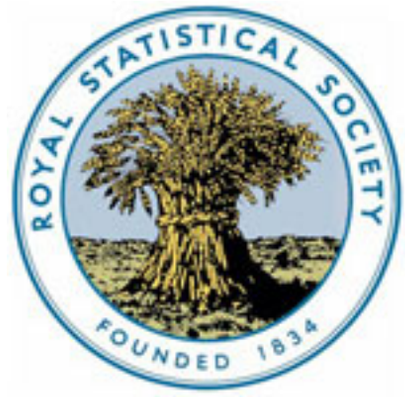

Appl. Statist. (1979),
28, *No.* 1, *pp.* 20–28

Maximum Likelihood Estimation of Observer Error-rates using the *EM* Algorithm

By A. P. DAWID† and A. M. SKENE‡

University College London

[Received October 1977. Revised May 1978]

SUMMARY

In compiling a patient record many facets are subject to errors of measurement. A model is presented which allows individual error-rates to be estimated for polytomous facets even when the patient's "true" response is not available. The *EM* algorithm is shown to provide a slow but sure way of obtaining maximum likelihood estimates of the parameters of interest. Some preliminary experience is reported and the limitations of the method are described.

Keywords: EM ALGORITHM; OBSERVER VARIATION; LATENT CLASS MODEL; MEDICAL EXAMPLE

1. INTRODUCTION

WHEN a patient's history is taken by different clinicians, different replies may be obtained to the same question. This may occur for a number of reasons; perhaps slightly different wording is used in each case, or perhaps the question is one the patient finds difficult to answer satisfactorily and so changes his reply from time to time. Similarly, in classifying a facet (sign or symptom) for type, severity, extent or duration, the patient and the clinicians may have different interpretations of the underlying scale of measurement. Such facets are said to be subject to observer error in that the response recorded may not be the "true" response as defined by some standard description of the facet or as implied by a consensus of medical opinion. The consequences of observer error are investigated by Good and Card (1971), where it is shown that a fairly low rate of error can lead to a considerable loss of diagnostic information.

In this paper we consider the problem of measuring observer error when the facet being recorded can take one of a finite number of values coded $1, 2, \ldots, J$. Two objectives can be distinguished. In some situations it is of interest to measure the degree of observer agreement. Landis and Koch (1977) give a general statistical methodology for the analysis of multivariate categorical data involving agreement among more than two observers. In fact the measurement of inter- and intra-observer agreement has received wide attention. Landis and Koch (1975a, b) give a review.

There are situations, however, where the performance of individual observers is more relevant. Let $\pi_{jl}^{(k)}$ be the probability that an observer, k, will record value l given j is the true response. The probabilities $\pi_{jl}^{(k)}, j = 1, \ldots, J, l = 1, \ldots, J$ are called the *individual error-rates* for the kth observer, although this set inclues $\pi_{jj}^{(k)}, j = 1, \ldots, J$, which are the probabilities that the observer records the true response in each of the J possible cases. Note that the error-rates are conditional probabilities where

$$\sum_{l=1}^{J} \pi_{jl}^{(k)} = 1 \quad \text{for each } j \text{ and } k,$$

† Now at Maths Dept, The City University, Northampton Square, London EC1V 0HB.
‡ Now at Maths Dept, University Park, Nottingham NG7 2RD.

and these quantities are mathematically independent of the marginal distribution of the true response.

Individual error-rates can be used to advantage in several situations:

(i) Through a knowledge of his own performance, an observer is able to recognize those facets where he makes the most frequent misclassifications and so reduce the number of errors in the patient's record. For example, he may reword a question or devote more time to the measurement of a particular sign. In those circumstances where a facet is recorded on an ordinal scale, error-rates serve to indicate how one observer's interpretation of the scale may differ from the general consensus, and thus show the way the observer should modify future assessments.

(ii) In the development of a data-base for purposes of diagnosis it is desirable to minimize errors of measurement. Knowledge of individual error-rates allows a contributor to that data-base to be monitored.

(iii) In some situations it might be recognized that a particular facet is subject to considerable observer error even when this is elicited by the most experienced clinicians. Nevertheless, it is highly desirable that this facet be accurately recorded and so, if several observers participate, a consensus judgement can be obtained. This judgement may be a simple majority opinion or a weighted consensus where the weights are functions of the individual error rates. In the latter case, each observer's contribution to the consensus is determined by his previous performance in eliciting that facet.

(iv) There has been recent interest in delegating certain tasks such as history-taking, which have traditionally been carried out by doctors, to ancillary personnel, or even to a computer terminal. It is important to know whether any loss (or gain) of accuracy is likely to ensue.

Henceforth all discussion is with reference to a single facet. When the true response can be obtained by some independent means, e.g. passage of time, X-ray or eventual operation, then sensible estimates of individual error-rates can be calculated very easily. For example, one possible estimator is

$$\hat{\pi}_{jl}^{(k)} = \frac{\text{number of times observer } k \text{ records } l \text{ when } j \text{ is correct}}{\text{number of patients seen by observer } k \text{ where } j \text{ is correct}}. \tag{1.1}$$

However, in many cases the true response can never be ascertained. Estimation of individual error-rates when the true response is not available has been discussed by Dawid (1971) for a facet having just two responses—yes or no, or 0 or 1. In the next section we extend the analysis proposed by Dawid to a facet having several possible responses. Section 3 mentions some of the computational problems and the practical limitations of this method of analysis. Section 4 gives an example.

2. Maximum Likelihood Estimation

Consider an experiment where K clinicians, indexed $k = 1, \ldots, K$, ask a single question of I patients indexed $i = 1, \ldots, I$. Not all clinicians need see every patient and a clinician may question the same patient more than once. The answer received is one of J possible replies. Let $n_{il}^{(k)}$ be the number of times clinician k gets responses l from patient i. The object of the experiment is to measure the individual error-rates $\pi_{jl}^{(k)}$ $(j = 1, \ldots, J;\ l = 1, \ldots, J;\ k = 1, \ldots, K)$. It is assumed that

(i) The responses given by a single patient to successive clinicians or to repeated questioning by a single clinician are independent, given the true response. Further, all patients respond independently.

(ii) There is no patient-by-clinician interaction. For example, one clinician does not obtain a more helpful response from a patient than is given to other clinicians.

The restrictive nature of these assumptions should be carefully noted in any application of the following analysis. In particular, in many studies there may be variables which intervene between the true response and the elicited responses and which, for each patient, are common to all clinicians. In the context of history-taking, such a variable might be the patient's vague memory of the true response. The assumption (i) above might more reasonably require independence of the responses given the true response and the intervening variables. However, assumptions (i) and (ii) are at present necessary if estimates of error rates are to be obtained.

Case 1. *True responses are available*

Let $\{T_{ij}: j = 1,...,J\}$ be a set of indicator variables for patient i. If response q is the true response for this patient then $T_{iq} = 1$ and $T_{ij} = 0$ $(j \neq q)$. Further, the patients in the experiment are considered to be a random sample from some population, where the probability that a patient drawn at random has true response j is p_j $(j = 1,...,J)$. Typically these probabilities are unknown.

Consider a single patient i and clinician k. Then, if q were the true response, the numbers of responses of each type actually obtained would be distributed according to a multinomial distribution and the likelihood would be proportional to

$$\prod_{l=1}^{J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}}.$$

As all clinicians elicit responses independently the likelihood for the responses of patient i when $T_{iq} = 1$ is

$$\prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}}$$

and unconditionally

$$\prod_{j=1}^{J} \left\{ p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} \right\}^{T_{ij}}. \tag{2.1}$$

That is, (2.1) is proportional to the probability of all the data (true class and responses) on patient i. It consists of a product of J terms, $J-1$ of which equal 1 (as $T_{ij} = 0$, $j \neq q$) and one term of the form

$$p(\text{responses obtained} \,|\, T_{iq} = 1)\, p(T_{iq} = 1).$$

As the data from all patients are assumed to be independent, the likelihood for the full data is

$$\prod_{i=1}^{I} \prod_{j=1}^{J} \left\{ p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} \right\}^{T_{ij}}. \tag{2.2}$$

In (2.2) the quantities $n_{il}^{(k)}$, T_{ij} and possibly p_j are all known. The maximum likelihood estimates can be calculated analytically, and we obtain estimators

$$\hat{\pi}_{jl}^{(k)} = \sum_i T_{ij} n_{il}^{(k)} / \sum_l \sum_i T_{ij} n_{il}^{(k)}. \tag{2.3}$$

The interpretation of (2.3) is simply equation (1.1). When the probabilities p_j $(j = 1,...,J)$ are unknown these can also be estimated:

$$\hat{p}_j = \sum_i T_{ij} / I. \tag{2.4}$$

We note, at this point, that when the individual error-rates $\{\pi_{jl}^{(k)}\}$ and the marginal probabilities $\{p_j\}$ are *known* but the true class of a given patient, i, is *unknown*, Bayes' Theorem may be used

to obtain estimates of the indicator variables T_{ij} $(j = 1, \ldots, J)$. *A priori*, $p(T_{ij} = 1) = p_j$. If data are then collected from the patients and the counts $n_{il}^{(k)}$ are obtained, by Bayes' Theorem

$$p(T_{ij} = 1 \mid \text{data}) \propto p(\text{data} \mid T_{ij} = 1) p(T_{ij} = 1)$$
$$\propto \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} p_j,$$

where all terms not involving j are absorbed into the proportionality sign. Thus

$$p(T_{ij} = 1 \mid \text{data}) = \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} p_j \Big/ \sum_{q=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}} p_q. \tag{2.5}$$

Case 2. True responses are not available

In this case the probability for the data observed on a single patient remains unchanged, given q is the true response. But as we do not know which response is the true response, unconditionally

$$p(\text{data on patient } i) \propto \sum_{j=1}^{J} p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}}. \tag{2.6}$$

Compare equations (2.6) and (2.1). In (2.1), as the T_{ij} $(j = 1, \ldots, J)$ were known, the distribution for the data on patient i was essentially multinomial. Equation (2.6) is a mixture of such multinomial distributions, the weights being the marginal probabilities p_j $(j = 1, \ldots, J)$. The likelihood for full data is therefore

$$\prod_{i=1}^{I} \left(\sum_{j=1}^{J} p_j \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} \right). \tag{2.7}$$

Calculation of the maximum likelihood estimates for the p's and π's in (2.7) is a much more difficult task. Dawid (1971) discusses the problems in detail but explicit expressions comparable to (2.3) and (2.4) cannot generally be obtained and numerical methods for the solution of simultaneous equations or function-maximization have to be employed.

In the past, the numerical methods available have not been satisfactory because of the large number of unknown parameters involved. However, Dempster *et al.* (1977) describe a numerical method of maximum likelihood estimation which is ideally suited to this particular problem. In certain circumstances missing data preclude the straightforward maximum likelihood estimation of the parameters of interest. However, if these parameters are known the missing data can be estimated. An iterative procedure is therefore proposed.

(i) Obtain some initial estimates of the missing data.

(ii) Calculate the maximum likelihood estimates for the quantities of interest as if the missing data had been found.

(iii) Now calculate new estimates of the missing data.

(iv) Repeat steps (ii) and (iii) until both the maximum likelihood estimates and the missing data estimates converge.

This procedure is known as the *EM* algorithm as each iteration consists of an Expectation (of missing data) step and a Maximization (maximum likelihood estimation) step. Dempster *et al.* give conditions under which the estimates obtained are those which should have been obtained if the quantities of interest had been estimated directly by maximum likelihood from the more complicated model which takes account of the missing data.

If, in the problem at hand, the indicator variables $\{T_{ij};\ i = 1, \ldots, I, j = 1, \ldots, J\}$ are treated as missing data then the conditions of the *EM* algorithm are satisfied. We therefore proceed as follows:

(i) Take initial estimates of the T's.

(ii) Use equations (2.3) and (2.4) to obtain estimates of the p's and π's.

(iii) Use equation (2.5) and the estimates of the p's and π's to calculate new estimates of the T's.

(iv) Repeat steps (ii) and (iii) until the results converge.

In step (iii) the estimate of T_{ij} is $E(T_{ij} \mid \text{data}) = p(T_{ij} = 1 \mid \text{data})$. That is, the estimate is expressed as a probability that the true response for patient i is j. Such probabilities may also be used as initial input in step (i).

The final estimates of the p's and π's are those values which maximize (2.7). The final estimates of the T's are the consensus probabilities for each patient from which the true response for each patient can be assessed.

3. Discussion

A number of comments relating to the execution of the algorithm and the interpretation of the results should be made at this point. Clearly, in any application, the concept of a true response must be meaningful, either as a hypothetically observable quantity or as the consensus agreement of qualified medical opinion given complete information on the patient. Even so, when the true responses are not known, there is no unique way in which observer error-rates can be estimated as there is no way in which the observations recorded by clinicians can be judged to be sensible or even relevant to the true state of the patient. This is reflected in the model, which has the structure of a latent class model (Lazarsfeld and Henry, 1968), and thus suffers from the lack of identifiability characteristic of factor-analytic models. Specifically, the likelihood (2.7) remains unchanged for any relabelling of the j index, and so it follows from the symmetry of the equation that there are $J!$ sets of estimates which correspond to the global maximum. Some will be more sensible than others, and it should be sufficient to assume that the correct estimates are those where $\pi_{jj}^{(k)} > \pi_{jl}^{(k)}$ $(j \neq l)$ for most k and j.

Closely allied to the issue of identifiability is the question of initial estimates. One possibility is to assign $T_{ij} = 1/J$ $(i = 1, \ldots, I, j = 1, \ldots, J)$. However, this action should be avoided as the initial estimates of the p's and π's correspond exactly to the centre of symmetry in (2.7) —a saddle point—and, as noted by Dempster *et al.*, the *EM* algorithm cannot converge from such a point. A second possibility is to use the data to calculate initial estimates. For example, one could use

$$\hat{T}_{ij} = \sum_k n_{il}^{(k)} / \sum_k \sum_l n_{il}^{(k)} \tag{3.1}$$

as starting values.

In practice, it is advisable to repeat the algorithm for several different sets of starting values. The *EM* algorithm is only guaranteed to converge to a local maximum and in situations where relatively few data are used to estimate a large number of parameters one must usually be content with chosing from a set of estimates corresponding to different local maxima on the basis of the magnitude of the likelihood. In our experience the starting values defined by (3.1) have been particularly useful in locating the best local maximum, if not the global maximum of interest. A maximum likelihood algorithm closely related to that above is given by Goodman (1974) who similarly advises on the use of several different starting values.

Little is known at present about the accuracy of the estimates. It is suspected that the standard errors of each of the π's and the p's are large and that these estimates are highly correlated. However, the large number of parameters makes the dispersion matrix difficult to obtain and, given that the maximum likelihood estimates will almost invariably lie on the boundary of the parameter-space, of limited usefulness. One approach has been to observe the stability of the estimates to the removal of individual patients or individual observers from the data. Our conclusion is that the Example below probably represents the smallest

experiment which will yield point estimates of any value unless the raw data indicate a very high level of agreement.

One consequence of probability estimates which are either zero or one is that unrealistically accurate estimates of the T's are obtained. However, inspection of the final estimates of the T's during repeated computation of the error-rates can lead to a clear appreciation of the "true" class of each patient in the experiment. While the vast majority of the final estimates of the T's remain virtually unaltered as different local maxima are discovered, some T estimates change markedly indicating the "reclassification" of one or two patients. A similar feature is noticed when the algorithm is repeated following removal of cases. Thus it is very easy to establish which patients are well classified and for which patients the consensus is more uncertain.

Dempster *et al.* prove that the algorithm exhibits first-order convergence. The significance of this result in our context depends largely on the proportion of cases in the data set where a consensus is not immediately obvious. If this proportion is substantial—say 30 per cent or more—a large number of iterations is usually required and repeated use of the algorithm becomes expensive. Nelder (1977) suggests that a substantial improvement is possible by exploring ahead along the apparent direction of convergence. This procedure has been used with good effect.

Finally, it must be emphasized that the method of maximum likelihood estimation has few advantages other than tractability. For a model as highly parameterized as this, having $(J-1)(JK+1)$ parameters, it would be naive to expect any of the theoretical large sample optimality properties to hold. It would be desirable to develop other estimation methods free of the general criticisms and caveats given above and to allow for prior opinions about the relative accuracies of the various questions. One possibility is to express the π's as functions of a much smaller number of parameters and estimate these by maximum likelihood. We hope to present this method in a future paper.

4. An Example

In the pre-operative assessment of a patient an anaesthetist must decide whether a patient is fit enough to undergo a general anaesthetic. In the trial reported here a standard form was completed on 45 patients by an independent party and contained information reflecting the

TABLE 1

Assessments of fitness for anaesthesia

	Observer						*Observer*						*Observer*				
Patient	1	2	3	4	5	*Patient*	1	2	3	4	5	*Patient*	1	2	3	4	5
1	111	1	1	1	1	16	111	2	1	1	1	31	111	1	1	1	1
2	333	4	3	3	4	17	111	1	1	1	1	32	333	3	2	3	3
3	112	2	1	2	2	18	111	1	1	1	1	33	111	1	1	1	1
4	222	3	1	2	1	19	222	2	2	2	1	34	222	2	2	2	2
5	222	3	2	2	2	20	222	1	3	2	2	35	222	3	2	3	2
6	222	3	3	2	2	21	222	2	2	2	2	36	433	4	3	4	3
7	122	2	1	1	1	22	222	2	2	2	1	37	221	2	2	3	2
8	333	3	4	3	3	23	222	3	2	2	2	38	232	3	2	3	3
9	222	2	2	2	3	24	221	2	2	2	2	39	333	3	4	3	2
10	232	2	2	2	3	25	111	1	1	1	1	40	111	1	1	1	1
11	444	4	4	4	4	26	111	1	1	1	1	41	111	1	1	1	1
12	222	3	3	4	3	27	232	2	2	2	2	42	121	2	1	1	1
13	111	1	1	1	1	28	111	1	1	1	1	43	232	2	2	2	2
14	222	3	2	1	2	29	111	1	1	1	1	44	121	1	1	1	1
15	121	1	1	1	1	30	112	1	1	2	1	45	222	2	2	2	2

patient's state of health. These forms were then independently assessed by five anaesthetists who classified each patient on a 1 to 4 scale, these categories having previously been defined. Anaesthetist 1 assessed the forms a total of three times, each time separated by some weeks. The data are shown in Table 1. Table 2 gives the estimates of the marginal probabilities and the individual error-rates. A range of starting values were explored although those described by (3.1) yielded the best maximum in this instance. Table 3 gives the matrices of probabilities $(p_j \pi_{ji}^{(k)})$ for each observer. In many circumstances these incidence-of-error matrices are of more immediate interest than the error-rates. The sum of the diagonal elements of such a matrix yields an estimate of the probability of a correct allocation by an observer and, when the feature under consideration is measured on an ordinal scale as in this example, the sum

TABLE 2

Maximum likelihood estimates

Marginal probabilities

Category:	1	2	3	4
	·40	·42	·11	·07

Error-rates

OBSERVER 1

Observed response:	1	2	3	4
True response 1	·89	·11	·0	·0
2	·07	·88	·05	·0
3	·0	·34	·66	·0
4	·0	·0	·56	·44

OBSERVER 2

Observed response:	1	2	3	4
True response 1	·78	·22	·0	·0
2	·06	·84	·10	·0
3	·0	·0	1·0	·0
4	·0	·0	·0	1·0

OBSERVER 3

Observed response:	1	2	3	4
True response 1	1·0	·0	·0	·0
2	·12	·79	·09	·0
3	·0	·40	·20	·4
4	·0	·0	·67	·33

OBSERVER 4

Observed response:	1	2	3	4
True response 1	·94	·06	·0	·0
2	·05	·84	·11	·0
3	·0	·0	·80	·20
4	·0	·0	·33	·67

OBSERVER 5

Observed response:	1	2	3	4
True response 1	1·0	·0	·0	·0
2	·16	·74	·10	·0
3	·0	·21	·79	·0
4	·0	·0	·33	·67

TABLE 3

Incidence-of-error probabilities

OBSERVER 1

Observed response:	1	2	3	4
True response 1	·36	·04	·0	·0
2	·03	·37	·02	·0
3	·0	·04	·07	·0
4	·0	·0	·04	·03

OBSERVER 2

Observed response:	1	2	3	4
True response 1	·31	·09	·0	·0
2	·02	·27	·13	·0
3	·0	·0	·11	·0
4	·0	·0	·0	·07

OBSERVER 3

Observed response:	1	2	3	4
True response 1	·40	·0	·0	·0
2	·05	·33	·04	·0
3	·0	·045	·02	·045
4	·0	·0	·05	·02

OBSERVER 4

Observed response:	1	2	3	4
True response 1	·38	·02	·0	·0
2	·02	·36	·04	·0
3	·0	·0	·09	·02
4	·0	·0	·02	·05

OBSERVER 5

Observed response:	1	2	3	4
True response 1	·40	·0	·0	·0
2	·07	·31	·04	·0
3	·0	·02	·09	·0
4	·0	·0	·02	·05

of the elements below or above the diagonal reflect the observer's optimism or pessimism relative to the other observers in the trial.

Table 4 gives the estimated probabilities for the T's for each patient. For most patients the posterior probability is 1·0 for a single response. In these cases the consensus appears obvious, but as mentioned above these results could be viewed with some suspicion. In fact,

TABLE 4

Final estimates of indicator variables for each patient†

	Category					*Category*			
Patient	1	2	3	4	*Patient*	1	2	3	4
1	1·0	0·0	0·0	0·0	24	0·0	1·0	0·0	0·0
2	0·0	0·0	0·0	1·0	25	1·0	0·0	0·0	0·0
3	0·0	1·0	0·0	0·0	26	1·0	0·0	0·0	0·0
4	0·0	1·0	0·0	0·0	27	0·0	1·0	0·0	0·0
5	0·0	1·0	0·0	0·0	28	1·0	0·0	0·0	0·0
6	0·0	1·0	0·0	0·0	29	1·0	0·0	0·0	0·0
7	0·986	0·014	0·0	0·0	30	0·999	0·001	0·0	0·0
8	0·0	0·0	1·0	0·0	31	1·0	0·0	0·0	0·0
9	0·0	1·0	0·0	0·0	32	0·0	0·0	1·0	0·0
10	0·0	1·0	0·0	0·0	33	1·0	0·0	0·0	0·0
11	0·0	0·0	0·0	1·0	34	0·0	1·0	0·0	0·0
12	0·0	0·0	1·0	0·0	35	0·0	0·948	0·052	0·0
13	1·0	0·0	0·0	0·0	36	0·0	0·0	0·0	1·0
14	0·0	1·0	0·0	0·0	37	0·0	1·0	0·0	0·0
15	1·0	0·0	0·0	0·0	38	0·0	0·021	0·979	0·0
16	1·0	0·0	0·0	0·0	39	0·0	0·0	1·0	0·0
17	1·0	0·0	0·0	0·0	40	1·0	0·0	0·0	0·0
18	1·0	0·0	0·0	0·0	41	1·0	0·0	0·0	0.0
19	0·0	1·0	0·0	0·0	42	1·0	0·0	0·0	0·0
20	0·0	1·0	0·0	0·0	43	0·0	1·0	0·0	0·0
21	0·0	1·0	0·0	0·0	44	1·0	0·0	0·0	0·0
22	0·0	1·0	0·0	0·0	45	0·0	1·0	0·0	0·0
23	0·0	1·0	0·0	0·0					

† Probabilities of 1·0 and 0·0 are correct to three decimal places.

repeated trials suggest that patients numbered 2, 12, 14 and 38 are the only cases which are sensitive to small changes in error-rate estimates, and thus a definite statement of true class is possible in many cases where consensus is not apparent in the raw data.

ACKNOWLEDGEMENT

We are grateful to Dr M. E. Wilson of the Bristol Royal Infirmary for permission to use the data in Section 4.

REFERENCES

DAWID, A. P. (1971). Estimation of error-rates in history taking. Preliminary analysis. Paper presented to the Royal College of Physicians Computer Workshop, November 1971.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm (with Discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

GOOD, I. J. and CARD, W. I. (1971). The diagnostic process with special reference to errors. *Meth. Inform. Med.*, **10**, 176–188.

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

LANDIS, J. R. and KOCH, G. G. (1975a). A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statist. Neerland.*, **29**, 101–123.
—— (1975b). A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). *Statist. Neerland.*, **29**, 151–161.
—— (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.
LAZARSFELD, P. F., and HENRY, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.
NELDER, J. A. (1977). In discussion of Dempster *et al.* (1977).