DOCUMENT RESUME

ED 114 412 TM 004 903

AUTHOR Shields, W. S.
TITLE Prediction from Contingency Tables Using Joint Likelihoods. MLM Research Report 74-1.
INSTITUTION Royal Military Coll. of Canada, Kingston (Ontario). Dept. of Military Leadership and Management.
SPONS AGENCY Defence Research Board, Ottawa (Ontario).
REPORT NO MLM-RR-74-1
PUB DATE [Oct 74]
NOTE 13p.; Paper presented at the Annual Meeting of the Military Testing Association (16th, Oklahoma City, Oklahoma, October 21-25, 1974)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Bayesian Statistics; Behavioral Science Research; *Classification; Higher Education; Hypothesis Testing; Matrices; Measurement Techniques; Prediction; Predictive Validity; *Predictor Variables; *Probability; *Questionnaires; Response Style (Tests); Statistical Bias
IDENTIFIERS Canada; *Contingency Tables; Moonan (W J)

ABSTRACT

A procedure for predicting categorical outcomes using categorical predictor variables was described by Moonan. This paper describes a related technique which uses prior probabilities, updated by joint likelihoods, as classification criteria. The procedure differs from Moonan's in that the outcome having the greatest posterior probability is selected as the prediction regardless of misclassification cost. It also differs in method of screening and weighting the predictor variables, and treats the problem of small-sample bias. Applications, to date, are in the analysis and use of questionnaire responses to predict categorical outcomes, namely, voluntary, academic, and military attrition from a Service College. Classification efficiency appears to be comparable to that of the Moonan technique. (Author)

MLM RESEARCH REPORT 74-1

PREDICTION FROM CONTINGENCY TABLES USING JOINT LIKELIHOODS

Lieutenant-Commander W.S. Shields

The research for this paper was supported in part by the Defence Research Board of Canada, Grant Number 9465-12. The views expressed are those of the author, and do not necessarily reflect official views or policy of the Royal Military College or the Department of National Defence.

Department of Military Leadership and Management
Royal Military College of Canada, Kingston, Ontario
Department Head: Lieutenant-Colonel G.J. Carpenter

PREDICTION FROM CONTINGENCY TABLES USING JOINT LIKELIHOODS

Lieutenant-Commander W.S. Shields
Department of Military Leadership and Management
Royal Military College of Canada

A procedure for predicting categorical outcomes using categorical predictor variables was described by Moonan (1972). This paper describes a related technique which uses prior probabilities, updated by joint likelihoods, as classification criteria. The procedure differs from Moonan's in that the outcome having the greatest posterior probability is selected as the prediction regardless of misclassification cost. It also differs in method of screening and weighting the predictor variables, and treats the problem of small-sample bias. Applications, to date, are in the analysis and use of questionnaire responses to predict categorical outcomes, — namely voluntary, academic, and military attrition from a Service College. Classification efficiency appears to be comparable to that of the Moonan technique.

Problem

The most common source of categorical data in behavioral research is the questionnaire. Whether questions are multiple-choice, or responses are grouped after-the-fact, it is usually difficult and frequently impossible to order responses along a metric scale. Even when an array of choices has a metric design, nonlinearity of relationships may make it preferable to treat responses as qualitative. The usual purpose of the questionnaire is to try to predict some criterion variable. Like the predictor variables (questionnaire responses) the criterion variable may be either quantitative (metric) or qualitative (categorical).

If either the criterion variable or its predictors are metric, discriminant analysis or a related technique may be used, unless relationships are nonlinear. If neither is metric some form of categorical analysis must be used. One option is to treat each category of each variable as a separate zero-one variable. The difficulty with this is proliferation in the number of variables. If each question has 5 possible responses and one asks 100 questions, 500 variables result. These have so many possible interactions that ones resulting from errors of measurement commonly dominate the analysis.

Moonan (1972) pioneered a strategy of analysis which treats the responses of a candidate to a selected subset of questions as a single unit and calculates the Bayesian conditional probability of an outcome to be predicted, given the candidate's particular response profile, under an assumption of independence of the

questionnaire items. A classification by outcome is then made which minimizes total misclassification cost.

Method

The strategy reported here is related to the first part of Moonan's procedure. It computes the joint likelihood of the candidate's responses under an assumption that a given outcome will occur, and multiplies this by the prior probability of the outcome, thus obtaining a quantity proportional to its posterior probability[1]. The outcome having the greatest posterior probability is then selected as the prediction. Because the joint likelihood of a given outcome, under an assumption of independence, is the product of the individual likelihoods derived from the predictor questions for a given set of responses, one may use the sum of the log-likelihoods as a sufficient statistic. The logarithm of the prior probability is then added to this sum and the result compared with that of other outcomes.

Because the concept of "likelihood" is used more often with continuous than with categorical variables, it should be defined carefully here. Suppose that an entire population could be entered into a contingency table such as that in Figure 1.

Figure 1

	a_1	a_2	a_3	
B_1	f_{11}	f_{12}	f_{13}	b_1
B_2	f_{21}	f_{22}	f_{23}	b_2
B_3	f_{31}	f_{32}	f_{33}	b_3
B_4	f_{41}	f_{42}	f_{43}	b_4
	A_1	A_2	A_3	

[1] This procedure was prescribed by Jeffreys (1939). He wrote: (p. 29) "The posterior probabilities of the hypotheses are proportional to the products of the prior probabilities and the likelihoods." Later (p. 133) he added "... when several estimates are combined the part from the prior probability enters only once, while that from the likelihood enters every time."

The B_i are responses to a question and the A_j are outcomes to be predicted. The frequency in each cell is f_{ij}; the row totals are b_i, the column totals a_j. The likelihood of outcome A_j given that a candidate has made response B_i is f_{ij}/a_j. This definition is in accord with that of Jeffreys, and differs from that of R.A. Fisher only in that the likelihood is required to *equal* the probability of the event which has been observed, given that the hypothesis under consideration is true, and not merely be proportional to it. This distinction makes it possible to perform any operation on the elements of a column vector, when considering likelihoods, that one would perform on a row vector when considering probabilities.

Independence

Before proceeding, some justification should be given for the assumption of independence of the categorical predictor variables. Moonan (1973) justifies it partly on the basis of computational feasibility. Certainly much less computer storage is required under this assumption, because only contingency tables relating to the outcome to be predicted need be stored, rather than a set which includes every predictor versus every other predictor. More important than this, if joint likelihoods were to be inferred directly from the data, the criterion sample would have to be extremely large, unless the number of questions were extremely small, for a sufficient cell content in the multidimensional contingency table that one could trust the results. For example, if 20 questions were to be used as predictors, a cell content greater than unity would exist only if two persons in the criterion sample had answered all 20 questions in exactly the same way, — a rare event indeed.

Another argument for the assumption of independence is that the consequences of failure of this condition are rarely serious. First of all, it is unusual for categorical variables to be intercorrelated as highly as metric ones. Commonly, one or two categories of a variable will correlate highly with one or two categories of another variable, ("Province of Birth in Canada" versus "Mother Tongue", for example) but overall correlations are usually low, unless one has accidentally (or deliberately) asked the same question twice.[2] Secondly, the effect of such correlation is merely to give some additional weight to highly correlated questions. Questions asked twice, for example, receive double weight. If this prejudices the prediction, it may do so in a favourable way because if a researcher has asked many questions centered in a particular area, or the same question in a number of different ways, it is presumably because he believes this area or question to be important.

---

[2] Moonan (1972) also used these arguments. He wrote "... many qualitative characters in practical problems are likely to be nearly independent and their dependencies poorly estimated."

Selection of Predictors

x^2 is a convenient criterion of variable selection because it is easily tested for significance. The author prefers the likelihood measure

$$x^2 \simeq -2 \ln \lambda$$

over the Pearson x^2 measure because of its stability when as many as 50% of the contingency table cells — not counting entirely blank rows and columns — are vacant. Wilks (1928) demonstrated that $-2 \ln \lambda$ is just as trustworthy as the Pearson approximation. The measure is identical to that derived from Shannon information theory:

$$x^2 \simeq 2N\hat{T}$$

where $\hat{T} = \hat{H}(A) + \hat{H}(B) - \hat{H}(A, B)$

$$= \hat{H}(A) - \hat{H}(A|B)$$

$$= \hat{H}(B) - \hat{H}(B|A)$$

and N is the sample size. A and B are two categorical variables, and the uncertainties (entropies) $\hat{H}$ are calculated in "nits" using:

$$\hat{H}(A) = -\sum_j \frac{a_j}{N} \ln \frac{a_j}{N}, \quad \hat{H}(B) = -\sum_i \frac{b_i}{N} \ln \frac{b_i}{N}, \text{ and}$$

$$\hat{H}(A, B) = -\sum_{ij} \frac{f_{ij}}{N} \ln \frac{f_{ij}}{N}$$

Having found a subset of predictors possessing significant relationships with the outcome to be predicted, it is suggested that the contribution of each predictor to the total log-likelihood be given a weight proportional to Newman and Gerstman's (1952) coefficient of constraint:

$$\hat{D}(A|B) = \frac{\hat{T}}{\hat{H}(A)} = \frac{\hat{H}(A) - \hat{H}(A|B)}{\hat{H}(A)}$$

Giving inferior weight to questions of low relevance can be justified on the basis of enhancing the "signal-to-noise" ratio. $\hat{D}(A|B)$ is the relative reduction in the uncertainty of A when B is known, and is asymmetrical. If desired, $\hat{D}$ can be corrected for bias using formulas developed by Miller (1954) for correction

of $\hat{H}$ and $\hat{T}$:

$$H \simeq \hat{H} + (K - 1)/2N$$

where K is the number of occupied categories in the variable, and

$$T \simeq \hat{T} - (R - 1)(C - 1)/2N$$

where R is the number of occupied rows, and C the number of occupied columns in the contingency table. (K-1) and (R-1) (C-1) will be recognized as the number of degrees of freedom appropriate to the χ^2 test. Negative values of T are set to zero.

When predictors are weighted proportional to D or $\hat{D}$, it is the practice of the writer to weight the prior distribution equally with the strongest predictor. This amounts to giving each a weight of unity, because a constant factor will not affect the comparison of a number of posterior probabilities. The decision to weight the prior in this manner is supported by the fact that it is based on a sample of the same size as is each predictor. The prior is simply $\ln (a_j/N)$.

Bias

Suppose k cell members occur out of a possible (column total) n. The "maximum-likelihood" estimate of the likelihood is k/n. This estimate, although bias-free for addition, has negative bias for multiplication (addition of log-likelihoods). It was decided to seek a formula for multiplicative use which would be as free as possible of bias over a broad range of population likelihoods.

If the sampling method resembles a Bernoulli process, and p represents the true population likelihood, two consecutive cell frequencies r and r+1 will have the same expectation when

$$\frac{n!}{r!(n-r)!} p^{r}(1-p)^{n-r} = \frac{n!}{(r+1)!\,(n-r-1)!} p^{r+1}(1-p)^{n-r-1}$$

or when $p = \frac{r+1}{n+1}$. These will also be the two most commonly observed cell frequencies, and will contain bias of opposite sign. If these biases are to be mutually cancelling then the likelihood estimates $\hat{p}$ should be such that

$$\hat{p}_r \hat{p}_{r+1} = p^2$$

Letting $\hat{p}$ have the form $\frac{k+a}{n+b}$, the above becomes:

$$\left(\frac{r+a}{n+b}\right)\left(\frac{r+1+a}{n+b}\right) = \left(\frac{r+1}{n+1}\right)^2$$

Because r and n will vary somewhat independently, we may set b=1 and solve for a, obtaining:

$$a = \sqrt{r^2+2r+1.25} - r - .5$$

Unfortunately, r is unknown. However a is very nearly .5 over a broad range of values of r and approaches .5 as r increases. A test of the formula $p = (k+.5)/(n+1)$ reveals that its performance for large values of k is improved if it is modified slightly to:

$$p = \frac{k+.5}{n+.5} \qquad \text{----------1}$$

This has very little effect on its performance for small values of k. Table 1 compares the expected value of joint likelihoods, calculated from random samples using formula 1, with population values. The estimated likelihoods are seriously biased only for small values of both p and n. In these instances sample likelihoods are bound to be poorly estimated. The bias is in the direction of avoiding the rejection of a hypothesis purely on the basis of a very small sample.

Formula 1 was adopted for estimating *both* likelihoods and prior probabilities, and was found to perform slightly better than the traditional k/n.

Missing Data

The method of this paper lends itself to a convenient and profitable treatment of missing data. One category of each variable is reserved for missing responses (or outcomes, as appropriate). The missing data category sometimes conveys important predictive information. In a study relating responses on a dental questionnaire to clinical dental examinations, the response most indicative of an unsatisfactory oral environment was refusal (or inability) to answer some of the questions.

Confidence

Because a posterior log-probability is calculated for each outcome, the difference in this quantity for the two most probable outcomes, less the difference in their priors, equals the logarithm of their likelihood ratio. This likelihood ratio provides an excellent indication of the confidence with which a given prediction can be made.

Test of the Method

The above procedure was developed for predicting categorical outcomes of Cadets at a Military College, such as academic failure, voluntary resignation, military failure, achievement of distinction, etc., based on a questionnaire written on their first day at the College. A total of 596 Cadets from four College years are currently under study. Their graduation years range from 1974 to 1978.

Rather than report results prematurely, the method will be demonstrated instead using R.A. Fisher's (1950, p 32.180) Iris data. It is chosen partly because it has already been used by several researchers, including Moonan (1972), and is widely available to future users for purposes of comparison.

Categories identical to those defined by Moonan were used in grouping the data:

Category:	1	2	3
Sepal Length:	SL $<$ 4.9	4.9 $\leq$ SL $\leq$ 5.7	5.7 $<$ SL
Sepal Width:	SW $<$ 2.3	2.3 $\leq$ SW $\leq$ 3.0	3.0 $<$ SW
Petal Length:	PL $<$ 3.3	3.3 $\leq$ PL $\leq$ 4.9	4.9 $<$ PL
Petal Width:	PW $<$ 1.0	1.0 $\leq$ PW $\leq$ 1.5	1.5 $<$ PW

Moonan used all 150 Iris plants as both criterion and prediction samples. As he pointed out, this stacks the odds heavily in favour of the classification algorithm. His program produced 9 misclassifications of the 150 Iris plants compared with 7 misclassifications made by using the same procedure with the method of this paper. The difference is nonsignificant; however, the result supports an opinion that the algorithms are of comparable quality. The method of this paper was tested also by shuffling the data cards and using the first 75 as a criterion sample from which to calculate prior probabilities and joint likelihoods for the classification of the remaining 75 plants. This resulted in 69 correct classifications and 6 incorrect ones. Misclassifications occurred only between Iris versicolor and Iris virginica.

Summary

A procedure has been described which, like Moonan's, represents a radical departure from conventional questionnaire analysis. The formula developed for making unbiased estimates of joint likelihood is equally applicable to the calculation of sample-derived joint probabilities. There are several areas in which the strategy is open to further refinement, particularly in regard to the "independence" assumption.

The comparison of hypotheses through the calculation of joint likelihoods is somewhat analogous to putting them through a long filter. A single very low likelihood, anywhere along the length of the filter, can cause a hypothesis to become "clogged" and fall hopelessly behind its competitors. Consider, for example, the following flow of information and its effect on the likelihoods of three competing hypotheses:

Hypothesis:	Mammal	Bird	Fish
Information Flow	Approximate Likelihood		
a. It has a head.	1.0	1.0	1.0
b. It has eyes.	1.0	1.0	1.0
c. It can fly.	.01	.99	.001
d. It has no feathers.	.01	0	.001
e. It has no fur.	0	0	.001

One can now classify the subject as a "flying fish" with reasonable confidence. That the choice is "unlikely" is not nearly as important as the fact that it is many times more likely than any of the available alternatives. The procedure is well summarized by a statement made by "Inspector Maigret" on a radio mystery program by that name some two decades ago. Asked the secret of his uncanny success, the famous detective replied: "Having eliminated all of the possibles, whatever remains - however improbable - must be the truth".

Table 1

EXPECTED VALUE OF A SAMPLE-DERIVED JOINT LIKELIHOOD ESTIMATE
USING THE FORMULA (K+.5)/(N+.5) FOR ITS FACTORS

SAMPLE SIZE	POPULATION LIKELIHOOD								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	0.14906	0.22228	0.30733	0.40041	0.49811	0.59797	0.69853	0.79914	0.89963
10	0.11452	0.20221	0.29914	0.39891	0.49921	0.59948	0.69967	0.79980	0.89991
15	0.10515	0.19964	0.29928	0.39951	0.49968	0.59978	0.69986	0.79991	0.89996
20	0.10180	0.19945	0.29958	0.39974	0.49983	0.59988	0.69992	0.79995	0.89998
25	0.10049	0.19957	0.29975	0.39984	0.49989	0.59993	0.69995	0.79997	0.89999
30	0.09998	0.19970	0.29983	0.39989	0.49993	0.59995	0.69997	0.79998	0.89999
35	0.09980	0.19978	0.29988	0.39992	0.49995	0.59996	0.69997	0.79998	0.89999
40	0.09975	0.19984	0.29991	0.39994	0.49996	0.59997	0.69998	0.79999	0.89999
45	0.09976	0.19988	0.29993	0.39995	0.49997	0.59998	0.69998	0.79999	0.90000
50	0.09978	0.19990	0.29994	0.39996	0.49997	0.59998	0.69999	0.79999	0.90000
55	0.09981	0.19992	0.29995	0.39997	0.49998	0.59998	0.69999	0.79999	0.90000
60	0.09984	0.19994	0.29996	0.39997	0.49998	0.59999	0.69999	0.79999	0.90000
65	0.09986	0.19995	0.29997	0.39998	0.49998	0.59999	0.69999	0.80000	0.90000
70	0.09988	0.19995	0.29997	0.39998	0.49999	0.59999	0.69999	0.80000	0.90000
75	0.09990	0.19996	0.29998	0.39998	0.49999	0.59999	0.69999	0.80000	0.90000
80	0.09991	0.19997	0.29998	0.39999	0.49999	0.59999	0.70000	0.80000	0.90000
85	0.09993	0.19997	0.29998	0.39999	0.49999	0.59999	0.70000	0.80000	0.90000
90	0.09993	0.19997	0.29998	0.39999	0.49999	0.59999	0.70000	0.80000	0.90000
95	0.09994	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
100	0.09995	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
105	0.09995	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
110	0.09996	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
115	0.09996	0.19998	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
120	0.09997	0.19999	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
125	0.09997	0.19999	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
130	0.09997	0.19999	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
135	0.09997	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000
140	0.09998	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000
145	0.09998	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000
150	0.09998	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000

REFERENCES

Bayes, Thomas. "An Essay Towards Solving a Problem in the Doctrine of Chances" (1773). In: Deming, W.E. *Facsimiles of Two Papers by Bayes*. New York: Hafner Publishing Co., 1963.

Bowser, S.E. "Applications of Predictor Ordering and Selection by a Bayesian-Decision Technique", *Proceedings of the 1973 Military Testing Association Conference*, San Antonio, Texas, 28 October - 2 November, 197[illegible]

Edwards, A.W.F. *Likelihood*. Cambridge: University Press, 1972.

Fienberg, S.E. and Holland, P.W. "Simultaneous Estimation of Multi-nominal Cell Probabilities", *Journal of the American Statistical Association*, 68 No. 343 (September, 1973), 683-91.

Fisher, R.A. *Contributions to Mathematical Statistics*. New York: John Wiley and Sons, Inc., 1950.

_____. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd, 1956.

Good, I.J. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, Mass: The M.I.T. Press (Research Monograph No. 30), 1965.

Harville, D.A. "Assigning Probabilities to the Outcomes of Multi-Entry Competitions", *Journal of the American Statistical Association*, 68 No. 342 (June, 1973), 312-16.

Jeffreys, H. *Theory of Probability*. Oxford: Clarendon Press, 1939.

Laplace, Marquis de. *Theorie anatytique des probabilites*. Paris: Courcier, 1820.

Miller, George A. "Note on the Bias of Information Estimates", in Quastler, H. (Ed.), *Information Theory in Psychology*, Glencoe, Ill.: The Free Press, 1955.

Moonan, W.J. "ABCD: A Bayesian Technique for Making Discriminations with Qualitative Variables", *Proceedings of the 1972 Military Testing Association Conference*, Lake Geneva, Wisconsin, 18-22 September, 1972.

_____. "Charosel: A Computer Program which Selects Qualitative Predictors for Qualitative Criterion Prediction Problems", *Proceedings of the 1973 Military Testing Association Conference*, San Antonio, Texas, 28 October - 2 November, 1973.

Newman, E.B. and Gerstman, L.J. "A New Method for Analyzing Printed English", *Journal of Experimental Psychology*, 44 (1952), 114-25.

Shields, W.S. "The Use of Qualitative Information in the Prediction of College Attrition", *Proceedings of the First Annual Conference; Canadian Association of Administrative Sciences*, Queen's University, Kingston, Ont., 31 May - 1 June, 1973.

_____. "The Use of Information Tehory in the Selection and Weighting of Cluster Variables for Prediction", *Proceedings of the Southeastern Regional Meeting, The Institute of Management Sciences*, Atlanta, Georgia, 18-19 October, 1973.

Wilks, S.S. "The Likelihood Test of Independence in Contingency Tables", *Biometrika*, 20A (1928), 263-94.