Estimation and Interaction in a Censored 2 x 2 x 2 Contingency Table

Joel E. Cohen

Biometrics, Vol. 27, No. 2. (Jun., 1971), pp. 379-386.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197106%2927%3A2%3C379%3AEAIIAC%3E2.0.CO%3B2-T

Biometrics is currently published by International Biometric Society.

BIOMETRICS 27, 379–86
June 1971

ESTIMATION AND INTERACTION IN A CENSORED $2 \times 2 \times 2$ CONTINGENCY TABLE

JOEL E. COHEN

Society of Fellows, Harvard University, Cambridge, Mass. 02138, U. S. A.

SUMMARY

An iterative procedure is presented for obtaining the maximum likelihood estimates of the probabilities of three noninteracting attributes when the available observations are the number of individuals having none of the attributes, the numbers of individuals having each one and only that one of the attributes, and the number of individuals having two or more of the attributes. The procedure is applied to observations of the prevalence of single and mixed infections of human malaria. The results are interpreted with caution.

1. INTRODUCTION

If each of N individuals may or may not have each of A attributes, then the numbers of individuals having each possible combination of attributes form the entries in a 2^A contingency table. In the analysis of such a table, the two questions of estimation and interaction frequently arise. Under the assumption that the N individuals are a random sample from a universe of individuals in which the probability of an individual's having one attribute is independent of his having any others, what is the best estimate of that probability? Are the entries in the contingency table consistent with the assumption that an individual's possession of one attribute is independent of his possession of any other?

Present answers to these questions, including cases when some entries in the contingency table may be missing or unobservable (truncated), are provided by Bhapkar and Koch [1968], Goodman [1968], Mantel [1970], Mosteller [1968], and the authors they cite.

A censored contingency table is one in which some cells have been pooled or some frequencies summed so as not to reveal individual entries in finest detail. Of interest here is the form of censoring of a 2^A table which leaves $A + 2$ numbers or cells: the number N_0 of individuals who have none of the A characteristics, the numbers N_i , $i = 1, 2, \cdots A$, of individuals who have the ith attribute but who do not have any of the other attributes, and the number N_{A+1} of individuals who have two or more attributes. Under the assumption of no interaction among attributes, the expectation of each N_i is NP_i , where α_i is the probability of the ith attribute (that is, the probability of success on a Bernoulli trial involving the ith attribute), $0 < \alpha_i < 1$, and

$$P_0 = \prod_{j=1}^{A} (1 - \alpha_j)$$

$$P_i = \alpha_i \prod_{\substack{j=1 \\ j \neq i}}^{A} (1 - \alpha_j) = \frac{\alpha_i}{1 - \alpha_i} \prod_{j=1}^{A} (1 - \alpha_j), \qquad i = 1, 2, \cdots, A, \tag{1}$$

$$P_{A+1} = 1 - \sum_{j=0}^{A} P_j = 1 - \left(1 + \sum_{i=1}^{A} \frac{\alpha_i}{1 - \alpha_i}\right) \prod_{j=1}^{A} (1 - \alpha_j).$$

Since $P_i/P_0 = \alpha_i/(1 - \alpha_i)$, it follows that $\alpha_i = P_i/(P_0 + P_i)$, $i = 1, 2, \cdots A$. If the observed N_i were exactly equal to their expectations NP_i, then the α_i could be calculated as $N_i/(N_0 + N_i)$.

For $A = 3$ this paper gives an easy iterative solution for the maximum likelihood (ML) estimates of the α_i and discusses the problem of measuring interaction. The same ML methods apply for arbitrary A but are not required by the application in view.

2. MAXIMUM LIKELIHOOD ESTIMATION

The ML estimates of the α_i, which in this case are also the modified minimum χ^2 estimates, are the solution of the system of A equations (Cramér [1946] p. 426):

$$\sum_{i=0}^{A+1} \frac{N_i}{P_i} \frac{\partial P_i}{\partial \alpha_i} = 0, \qquad i = 1, 2, \cdots, A. \tag{2}$$

For $A = 3$, substituting (1) into (2) gives, for $i = 1, 2, 3$, and $i \neq j \neq k$,

$$0 = -N_0 - N_i - N_k + \frac{N_i(1 - \hat{\alpha}_i)}{\hat{\alpha}_i} + \frac{N_4(1 - \hat{\alpha}_i)}{\hat{\alpha}_i + Q_{ik}}, \tag{3}$$

where

$$Q_{ik} = \hat{\alpha}_i\hat{\alpha}_k/[\hat{\alpha}_i + \hat{\alpha}_k - 2\hat{\alpha}_i\hat{\alpha}_k]. \tag{4}$$

Equation (3) is a quadratic in $\hat{\alpha}_i$ with coefficients involving Q_{ik}. The roots of (3) are given by

$$-2N\hat{\alpha}_i = Q_{ik}N - N_i - N_4(1 + Q_{ik}) \pm \{[-Q_{ik}N + N_i + N_4(1 + Q_{ik})]^2 + 4Q_{ik}NN_i\}^{\frac{1}{2}} \tag{5}$$

and a positive value of $\hat{\alpha}_i$ is obtained only by taking the root with the negative sign in front of the radical. Since Q_{ik} contains the product of probabilities in its numerator and their sum in its denominator, and hence should be small compared to 1, one initial estimate $\hat{\alpha}_i^{(0)}$ of α_i is obtained by setting $Q_{ik} = 0$ in (5) or (3), giving

$$\hat{\alpha}_i^{(0)} = (N_i + N_4)/N. \tag{6}$$

An alternative initial estimate is suggested by the observation after (1) that $\alpha_i = P_i/(P_0 + P_i)$:

$$\hat{\alpha}_i^{(0)} = N_i/(N_0 + N_i). \tag{7}$$

Given some initial estimates, values of Q_{ik} then follow from (4) and improved $\hat{\alpha}_i^{(1)}$ follow from (5). Further iteration proceeds back and forth between (4) and (5).

3. INTERACTION

In modeling the censored 2^A contingency table introduced above, a parameter s may be introduced to measure the deviation of the number N_{A+1} of individuals with multiple attributes from the expected number of such individuals assuming no interaction among attributes. Under this new model, the N_i are multinomially distributed in cells with probabilities π_i, where π_i is proportional to P_i for $i = 0, 1, \cdots, A$, and π_{A+1} is proportional to sP_{A+1}. Thus if s is greater (or less) than 1, an individual has a greater (or less) chance of having two or more attributes under this model than he would under the assumption of no interaction. Normalizing so that the π_i sum to 1, we have

$$\begin{aligned} \pi_i &= \frac{P_i}{1 + (s - 1)P_{A+1}}, \qquad i = 0, 1, \cdots, A, \\ \pi_{A+1} &= \frac{sP_{A+1}}{1 + (s - 1)P_{A+1}}, \end{aligned} \tag{8}$$

where the P_i are the functions of α_i given by (1). Since this model has $A + 1$ independent probabilities and $A + 1$ parameters (s and the α_i), it has no degrees of freedom; s may be expressed exclusively as a function of the π_i.

The parameter s is a weighted average of the first-order up to $(A - 1)$-order interactions in the 2^A table in a way which may be illustrated by the case $A = 3$. For $i \neq j \neq k$, if first-order interactions make the probability of the two attributes i and j in the absence of k proportional to $s_k(1 - \alpha_k)\alpha_i\alpha_j$ instead of simply to $(1 - \alpha_k)\alpha_i\alpha_j$ and if a second-order interaction makes the probability of all three attributes proportional to $s_0\alpha_1\alpha_2\alpha_3$ instead of simply to $\alpha_1\alpha_2\alpha_3$, then s is the weighted average of these interaction coefficients s_i which satisfies $sP_4 = s_0\alpha_1\alpha_2\alpha_3 + s_1(1-\alpha_1)\alpha_2\alpha_3 + s_2\alpha_1(1-\alpha_2)\alpha_3 + s_3\alpha_1\alpha_2(1-\alpha_3)$. Clearly when all the s_i are equal, $s = s_i$. The extension to $A > 3$ is obvious.

For $A = 2$, the number N_3 of individuals who have two or more attributes equals the number of individuals who have exactly two attributes. No censoring of the usual 2×2 table remains to be done. From (8) it follows that

$$s = \pi_0\pi_3/\pi_1\pi_2 . \tag{9}$$

The estimate $\hat{s}$ of s obtained by replacing each π_i with N_i is just the relative odds in a 2×2 table. Approximate confidence limits for the estimate $\hat{s}$ have been obtained by Goodman [1964] and the authors he cites. Hence the value of $\hat{s}$ may be used to measure approximately the goodness of fit of the no-interaction model (1) instead of calculating all the expected frequencies and using Pearson's X^2.

For $A = 3$, it may be shown from (8) that

$$s = \frac{\pi_4\pi_0^2}{\pi_1\pi_2\pi_3 + \pi_0\pi_1\pi_2 + \pi_0\pi_2\pi_3 + \pi_0\pi_1\pi_3}. \tag{10}$$

An estimate $\hat{s}$ of s is obtained by replacing each π_i in (10) with N_i . Asymptotically for large N, $\hat{s}$ is normally distributed (Cramér [1946] pp. 354, 366) with $E(\hat{s}) = s$ and with variance (11) which depends on the multinomial variances and covariances, var $(N_i) = N\pi_i(1 - \pi_i)$ and for $i \neq j$, cov $(N_i, N_j) = -N\pi_i\pi_j$:

$$\widehat{\text{var}}\,\hat{s} = N \sum_{i=0}^{4} \left(\frac{\partial \hat{s}}{\partial N_i}\right)^2 \pi_i(1 - \pi_i) - 2N \sum_{i>j}^{4} \left(\frac{\partial \hat{s}}{\partial N_i}\right)\left(\frac{\partial \hat{s}}{\partial N_j}\right)\pi_i\pi_j \ , \qquad (11)$$

where

$$\begin{aligned}
\partial\hat{s}/\partial N_0 &= N_0N_4/D + N_0N_1N_2N_3N_4/D^2, \\
\partial\hat{s}/\partial N_1 &= -(\hat{s}/D)(N_0N_2 + N_0N_3 + N_2N_3), \\
\partial\hat{s}/\partial N_2 &= -(\hat{s}/D)(N_0N_1 + N_0N_3 + N_1N_3), \\
\partial\hat{s}/\partial N_3 &= -(\hat{s}/D)(N_0N_1 + N_0N_2 + N_1N_2), \\
\partial\hat{s}/\partial N_4 &= (N_0^2/D), \quad \text{and} \\
D &= N_1N_2N_3 + N_0N_1N_2 + N_0N_2N_3 + N_0N_1N_3 \ .
\end{aligned} \qquad (12)$$

Though unattractive analytically, the approximate variance (11) presents no computational difficulties. The 100 α% confidence interval around $\hat{s}$ may be approximated by $\hat{s} \pm z_x(\widehat{\text{var}}\,\hat{s})^{\frac{1}{2}}$, where z_x is the xth percentile of the standardized normal distribution, $x = 50(1 + \alpha)$. This approximation to the confidence interval assumes that $\hat{s}$ is symmetrically distributed in the region of estimation. Because $\hat{s}$ ranges over $(0, +\infty)$ this assumption is not likely to be valid for values of $\hat{s}$ near zero unless N is very large.

The transformed variable $s' = \log \hat{s}$ is distributed over $(-\infty, +\infty)$ and may be preferable. The asymptotic variance of s' is given by replacing $\hat{s}$ with s' everywhere in (11). Then since $\partial s'/\partial N_i = \partial \log \hat{s}/\partial N_i = (1/\hat{s})(\partial\hat{s}/\partial N_i)$, the asymptotic variance of the log-transform is $1/\hat{s}^2$ times the asymptotic variance of $\hat{s}$. Confidence intervals around s' on the logarithmic scale are given by $s' \pm z(\widehat{\text{var}}\,s')^{\frac{1}{2}}$ for the appropriate z. A confidence interval on the logarithmic scale may be exponentially transformed back to the arithmetic scale to yield a second approximation to a confidence interval around $\hat{s}$.

Another way to set confidence limits on $\hat{s}$, given data which do fit (1) according to the χ^2 test at some probability level, is suggested by a reviewer (N. Mantel): find the greatest (and least) parameter values s^* such that when the α_i are fitted conditional on s^* from (8), the fit worsens just significantly. See Mantel and Patwary [1961].

The following numerical example will show that neither the arithmetic nor the transformed (approximate) confidence intervals around $\hat{s}$ are especially satisfactory for moderate sample sizes N.

4. NUMERICAL APPLICATION: HUMAN MALARIA

A computer program written to carry out the estimation procedure of section 2 calculates the current estimates of α_i initially from (6) and subse-

quently from (4) and (5). It then derives the expected values of N_i from these estimates using (1) and compares them with the observed values of N_i by calculating Pearson's X^2. Iteration terminates when the changes in X^2 and in $\hat{\alpha}_i$ between two successive iterations are each less than 0.01. Then a second set of initial estimates of α_i is computed from (7) and the entire procedure is repeated.

This double calculation permits a comparison of the two initial estimators (6) and (7). In the examples of real data analyzed below, none of the corresponding final values of $\hat{\alpha}_i$ and X^2 in the two analyses differed by as much as 10^{-5}, and hence only a single set of calculations will be presented for each set of data. In only one case did the two initial estimators require different numbers of iterations to quasi-converge: in the third set of data (for the whole sample), estimators (6) required three iterations, while (7) required two. The other data required two or three iterations. In general, when the data agreed poorly with the model (1) of no interaction, the initial estimates (6) were closer to the final estimates of α_i than were (7), while when the data agreed well with (1), the initial estimates (7) were closer to the final.

Examples not reported here show that the procedure recovers accurately and quickly the parameter values used to generate artificial data.

Since there are 5 data cells and 3 fitted parameters, the value of X^2 is compared with the distribution function of χ^2 with $5 - 1 - 3 = 1$ D.F.

An example of the biological data (from Downs *et al.* [1943] p. 22) which gave rise to this estimation problem is given in Table 1. School children

TABLE 1

PREVALENCE OF SINGLE AND MIXED MALARIAL INFECTIONS OF SCHOOL CHILDREN IN TRINIDAD AND TOBAGO COMPARED TO EXPECTATIONS FROM A MODEL OF NO INTERACTION

	Normal spleens			Enlarged spleens			All children		
	Observed[b]	Fitted	MLE[a] $\hat{\alpha}$	Observed	Fitted	MLE $\hat{\alpha}$	Observed	Fitted	MLE $\hat{\alpha}$
No infection	5856	5856.4		1053	1172.0		6909	6920.8	
P. falciparum only	298	297.6	.0484	592	495.5	.2971	890	880.1	.1128
P. vivax only	127	126.7	.0212	290	214.2	.1545	417	409.0	.0558
P. malariae only	41	40.9	.0069	205	146.1	.1109	246	240.5	.0336
Mixed infection	9	9.4		78	190.3		87	98.6	
X^2	0.023			147.742			1.782		
	$0.8 < P < 0.9$			$P \ll 0.01$			$0.1 < P < 0.2$		
$\hat{s}$	0.949			0.213			0.852		
99% conf. int.[c]	Lower	Upper		Lower	Upper		Lower	Upper	
Arithmetic	.9383	.9597		.2113	.2144		.8487	.8544	
Transformed	.9384	.9599		.2113	.2146		.8487	.8579	

[a] Calculation described in section 2.

[b] Data from Downs et al. ([1943] p. 22).

[c] Calculation described in sections 3 and 4.

(aged 5 to 15) on Trinidad and Tobago were examined and classified as having an enlarged spleen (spleen positive) or a spleen of normal size (spleen negative). Blood smears from children in both groups were examined for the presence or absence of each of three species of malarial parasite (genus *Plasmodium*). However, only the numbers of individuals whose blood smears had no parasites (N_0), *Plasmodium falciparum* only (N_1), *P. vivax* only (N_2), *P. malariae* only (N_3), or a mixed infection (N_4) were reported. This form of censorship—the pooling of all mixed infections—is found in most reports of malaria surveys in areas where more than two species of *Plasmodium* occurred.

The analysis is presented in Table 1 for spleen-negative children, spleen-positive children, and all children combined. For the spleen-negative children and the total sample, the values of X^2 (0.02 and 1.78, respectively) are too small to permit rejection at the 1% level of the assumption that there is no interaction between species of infection. For the spleen-positive children, the value of $X^2 = 147.74$ makes the hypothesis of no interaction improbable. Some of the difficulties of interpreting these findings will be discussed in the final section.

In the presentation (section 3) of a model for measuring interaction via the parameter s, two confidence intervals around a sample estimate $\hat{s}$ were suggested. Using $z_{99.5} = 2.58$ to obtain a 99% confidence interval, these two confidence intervals are the usual one calculated on the arithmetic scale, $\hat{s} \pm 2.58\ (\widehat{\text{var}}\ \hat{s})^{\frac{1}{2}}$, and the exponentially transformed interval based on the variance of the log-transform $s' = \log \hat{s}$, namely $\exp\ (s' \pm 2.58\ (\widehat{\text{var}}\ s')^{\frac{1}{2}})$. These two intervals, which may be called the 'arithmetic' and the 'transformed,' respectively, are given along with the sample estimate $\hat{s}$ for each set of data.

Assuming that the χ^2 test is a true measure of whether or not the data reveal interaction, the confidence intervals around $\hat{s}$ should include 1 (the value of s when there is no interaction) for the spleen-negative children and for the total sample, while they should exclude 1 for the spleen-positive children. In fact, as Table 1 shows, both the arithmetic and the transformed confidence intervals exclude 1 in all three cases. Thus, if the χ^2 test is a good standard, the approximate variance in (11) is not a satisfactory source of information about the distribution of $\hat{s}$.

5. INTERPRETING THE RESULTS

When statistical analysis of data on the joint prevalence of parasitic species fails to reveal any association (as in spleen-negative children or in the whole sample), or reveals a negative association (as in spleen-positive children) or a positive association, circumspection is required in interpreting the results.

The absence of apparent interaction in the contingency table does not necessarily imply that the parasitic species do not interact within individual hosts. Though infections may occur independently, prior infection with one species may inhibit the level of infestation with a second. Such an interaction can be revealed only by counts of parasites in the blood, and is known to occur among related malaria strains in rodents (Cox and Voller [1966]).

A negative association among parasites—fewer mixed infections than expected from no interaction—could arise from at least five different causes. First, the parasite species could be localized in different, largely disjoint regions, giving no opportunities for mixed infections, but the sample in the contingency table could be pooled over these regions. Second, children with mixed infections may suffer from increased rates of death or sickness, and hence be less likely to survive for or appear in a malaria survey. Third, malarial species may appear alternatively in the peripheral blood which is sampled in a survey even though both are present in the body at once; hence single-species infections may be observed when mixed infections actually occur. Fourth, in poor laboratory technique, diagnosis of a blood smear may stop or tend to stop after the discovery of a single parasite or single parasite species, especially if infection with one species is much more massive than infection with others. Fifth, resistance to infection mobilized by infection with one species of malaria may apply cross-specifically to diminish the rate or duration of infection with another species.

That the negative association between malarial species was found in children with enlarged spleens favors the fifth explanation because enlargement of the spleen is one common indicator of a well developed immune response. Splenomegaly is not an unfailing indicator of immune response, however, both because the spleen may grow smaller again after development of an immune response and because it may enlarge for reasons other than the development of malaria-specific immunity. But if the preceding four alternative explanations can be ruled out by stratification of the data by (1) locality and (2) age, and by appropriate (3) sampling and (4) diagnostic technique then, in combination with experimental demonstration of cross-specific immunity to malaria in rodents (Cox and Voller [1966]), the fifth explanation becomes plausible.

A positive association between parasitic species, though not observed in the example given, could appear in the contingency table even though different species were infecting hosts independently for three reasons, at least. First, differences in overall levels of exposure of different fractions of the population sampled may be due to: differences in duration of exposure because of age differences, differences in altitude, differences in occupation and daily habits, and differences in natural immunity or susceptibility to all species of malaria. Second, if diagnostic procedures are variable in quality, poor examinations of blood slides may miss all parasites while good examinations may be more likely to find parasites of all species. Third, infection with one species may weaken the host's resistance and pave the way for other species. This last possibility can be taken seriously only when the contributions of the first two explanations have been controlled by appropriate stratification and diagnostic technique.

Finally, the probability of *having* one infection given another says nothing by itself about the probability (per unit time) of *getting* one infection given another. Inferences from prevalences to incidences require supplementary information and arguments (Cohen [1970]).

The sum of these caveats is that the statistical analysis is only the begin-

ning of understanding. Yet it can be a beginning. The many malaria surveys carried out in this century contain a wealth of epidemiological information about the interactions of malarial infections that remains to be exploited.

An example of a censored 2^{16} contingency table of the form considered in this paper appears in Yao *et al.* [1935].

6. ACKNOWLEDGMENTS

I thank M. Drolette, J. J. Feldman, N. Mantel, T. H. Weller, and an anonymous referee for helpful criticisms and suggestions; and the School of Public Health and the Milton Fund of Harvard University as well as The RAND Corporation, for support.

ESTIMATION ET INTERACTION DANS UNE TABLE DE CONTINGENCE TRONQUEE $2 \times 2 \times 2$

RESUME

On présente une procédure itérative pour obtenir les estimateurs par le maximum de vraisemblance des probabilités de trois attributs sans interaction quand les observations disponibles sont le nombre des individus n'ayant aucun des attributs, le nombre des individus ayant chacun un et seulement un de ces attributs et le nombre des individus ayant deux attributs ou plus. La procédure est appliquée à des observations faites sur l'étendue des infections simples et mélangées de la Malaria chez l'homme. Les résultats sont interprétés avec précaution.

REFERENCES

Bhapkar, V. P. and Koch, G. G. [1968]. On the hypothesis of 'no interaction' in contingency tables. *Biometrics 24*, 567–94.

Cohen, J. E. [1970]. A Markov contingency-table model for replicated Lotka-Volterra systems near equilibrium. *American Naturalist 104*, 547–60.

Cox, F. E. G. and Voller, A. [1966]. Cross-immunity between the malaria parasites of rodents. *Ann. Tropical Med. Parasitology 60*, 297–303.

Cramér, H. [1946]. *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Downs, W. G., Gillette, H. P. S., and Shannon, R. C. [1943]. A malaria survey of Trinidad and Tobago, British West Indies. *J. Nat. Malaria Soc. 2* (1) Suppl., August.

Goodman, L. A. [1964]. Simultaneous confidence limits for cross-product ratios in contingency tables. *J. Roy. Statist. Soc. B 26*, 86–102.

Goodman, L. A. [1968]. The analysis of cross-classified data: independence, quasi-independence and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Ass. 63*, 1091–131.

Mantel, N. [1970]. Incomplete contingency tables. *Biometrics 26*, 291–304.

Mantel, N. and Patwary, K. M. [1961]. Interval estimation of single parametric functions. *Bull. Int. Statist. Inst. 38*, 227–40.

Mosteller, F. [1968]. Association and estimation in contingency tables. *J. Amer. Statist. Ass. 63*, 1–28.

Yao, Y. T., Hsu, S. C., and Ling, L. C. [1935]. On the occurrence of intestinal parasites in man in different combinations. A statistical study of the results of 9853 fecal examinations. *Far East Ass. Tropical Med., Trans. 9th Congr. Nanking 1934. 2*, 31–8.

Received April 1970, *Revised November* 1970