Synthetic Retrospective Studies and Related Topics

Nathan Mantel

Biometrics, Vol. 29, No. 3 (Sep., 1973), 479-486.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28197309%2929%3A3%3C479%3ASRSART%3E2.0.CO%3B2-2

Biometrics is currently published by International Biometric Society.

BIOMETRICS 29, 479-486
September 1973

SYNTHETIC RETROSPECTIVE STUDIES AND RELATED TOPICS

NATHAN MANTEL

Biometry Branch, National Cancer Institute, Bethesda, Maryland 20014, U.S.A.

SUMMARY

Prospective and retrospective approaches for estimating the influence of several variables on the occurrence of disease are discussed. The assumptions under which these approaches would tend to yield the same estimates as would be given by an ideal but unattainable experimental design approach are stated. It is then brought out that in a large prospective study in which comparatively few cases of disease have occurred, computational problems can be so burdensome as to preclude a comprehensive and imaginative analysis of the data. The prospective study can be converted into a synthetic retrospective study by selecting a random sample of the cases and a random sample of the noncases, the sampling proportion being small for noncases, but essentially unity for cases. It is demonstrated that such sampling will tend to leave the dependence of the log odds on the variables unaffected except for an additive constant.

The use of a discrimination function noniterative method of analysis is noted and is indicated to be not generally appropriate. The reverse suggestion is made that normal data can be analyzed by a log-odds approach, this yielding alternative tests to those ordinarily used for comparing two or several means or mean vectors, or two or several variances or variance-covariance matrices.

INTRODUCTION

The usefulness of retrospective studies for identifying potential causal factors in human disease has been well established; at the same time it is recognized that the results forthcoming from any such study can be interpreted only with caution and with recognition of the limitations of the retrospective study approach (see Mantel and Haenszel [1959]).

An ideal, but for all purposes unattainable, study would be one paralleling laboratory experimentation. At some appropriate time after birth, children (or, reasonably in some cases, adults) would be randomly assigned to one of several different treatment groups and, in addition, any relevant characteristics about the individual are measured; the ith individual will thus be characterized by the vector X_i, which includes levels of both treatment variables and individual characteristic variables. Assuming no problems of discontinued treatment or early death, we will eventually be able to identify which individuals, develop the disease by a particular age, ($Y = 1$), or remain disease free, ($Y = 0$). If the proper parametric model is employed the resulting data could be used for estimating the parameters relating the probability, $P(Y_i = 1 \mid X_i) = F(\theta, X_i)$. (A more sophisticated model than is ordinarily used would be

required if one wished to take into account the exact age at which the disease occurred or as of which individuals remained disease free.)

A practical alternative to the laboratory-style study is the prospective study, one in which we measure the self-selected variables in X_i, follow the individuals after they have made their selections, and analyze the resulting data as though they had arisen in a properly-randomized experiment. With this approach no individual is subject to enforced treatment and, where there is a long delay from time of treatment initiation to time of response, the time-duration of the study can be reduced by confining the study to individuals who have attained the age at which responses begin to occur. The basic assumption for the validity of this approach is that, whether or not the self-selected variables influence the eventual outcome, the individual's choice of levels for the self-selected variables is not influenced by factors about the individual which are determining in whether or not he develops disease. Even if this assumption is met, the prospective study can be subject to limitations avoidable in a designed laboratory-style study. In the designed study we can assign levels for treatment variables in a way so as to minimize, or even eliminate by orthogonalization, confounding of effects. In the prospective study self-selected variables can be highly intercorrelated with each other as well as with the person's individual characteristics; estimates of the contributory effects of each variable studied will consequently have reduced precision.

A retrospective study approach has two particular advantages in comparison with the prospective study approach. In the prospective study of a rare or uncommon disease, one must follow a large number of individuals to get a limited number of cases of disease, and the period for which individuals need be followed may be a prolonged one. Retrospectively, one merely identifies recent cases of disease and obtains historical information about them, doing the same for a group of negative controls which can be considered representative of the many individuals who failed to develop the disease. The propriety of the retrospective approach is premised on the same basic assumption about self-selected variables and is subject to the same limitations as obtain for the prospective approach. The added further assumptions are that the cases selected are a random sample of all cases (which likely obtains if we are taking a 100% sample of cases) and that the controls selected are a random sample of the disease-free individuals from the population in which the cases arose. This latter assumption can sometimes be difficult to validate (see Mantel and Haenszel [1959], for a discussion) and, perforce, retrospective study analyses are made *as though* the controls were appropriate even when questions about appropriateness could be raised. However, some remedial devices exist which can allow correcting for some kinds of inappropriateness. Thus if controls differ from cases in their distribution by some variables, e.g. age, race, sex, one can analyze the data adjusting for any such differences using procedures like those suggested by Mantel and Haenszel. (It is, of course, further assumed that correct information is obtained in a retrospective study. Because of faulty recall or for

other reasons this may not always be true and the above devices may not be remedial for such faults. Where faulty information is suspected it is important to consider whether it leads only to weakened associations or whether it might also lead in some instances to biased or false associations.)

THE SYNTHETIC RETROSPECTIVE STUDY—SAMPLING FROM A PROSPECTIVE STUDY

My present purpose is to propose, discuss, and validate use of retrospective approach procedures in a prospective study situation. A particular prospective-study situation which I encountered gave rise to only 165 cases of a particular condition in a cohort of about 4,000 individuals.[1] Preliminary analyses were undertaken using a limited number of the variables on which data had been collected. But even with simple maximum likelihood analyses of the form used involving only one or two of the study variables, the computer time required was somewhat prolonged. This could then preclude making analyses as comprehensive and as extensive as we should have liked.

As it turned out, the key cause for prolonged computer time was the large number of observations involved. Computation was simple and rapid once the necessary totals were obtained for all 4,000 individuals. But time was consumed for entering all the information and for computing at each iterative stage certain quantities appropriate for each individual. The number of iterative cycles for convergence could be reduced by a device for obtaining suitable entering approximations (see below), but even this would not resolve our problem.

A possible remedy envisaged was to convert the study, in principle, to a retrospective one. Suppose we included in the analysis a random proportion, π_1, of our cases and another random proportion, π_2, of the negatives. If we chose π_1 as 1 and π_2 as 0.15, we would have all the cases and 3.5 negatives per case. By the reasoning that $n_1 n_2/(n_1 + n_2)$ measures the relative information in a comparison of two averages based on sample sizes of n_1 and n_2 respectively, we might expect by analogy, which would of course not be exact in the present case, that this approach would result in only a moderate loss of information. (The practicing statistician is generally aware of this kind of thing. There is little to be gained by letting the size of the control group, n_2, become arbitrarily large if the size of the experimental group, n_1, must remain fixed.) But the reduction in computer time would permit much more effective analyses. Ostensibly we would be meeting the additional conditions assumed for validity of the retrospective study approach; that is the retained individuals would be a random sample of the cases and disease-free individuals arising in the prospective study.

Suppose the randomness conditions are met. Still it seems that we are selecting our retained individuals on the basis of their response variable,

[1] The actual number of individuals was substantially less than 4,000. An initial cohort of about 1,350 men was studied to evaluate the short-term prognostic value of various factors in coronary heart disease. Men remaining free of disease for two years could be reentered into the analysis for the next two years using their new X_i values.

diseased or disease-free, a selection procedure which could invalidate the results of normal regression analysis. The question of validation was resolved by the following reasoning. The possible outcomes for individual i with vector X_i are: 1) he can develop disease and be in the sample, with probability $\pi_1 P(Y_i = 1 \mid X_i)$; 2) he can develop disease and not be in the sample, with probability $(1 - \pi_1)P(Y_i = 1 \mid X_i)$; 3) he can remain disease free and be in the sample, with probability $\pi_2 P(Y_i = 0 \mid X_i)$; 4) he can remain disease free and not be in the sample, with probability $(1 - \pi_2)P(Y_i = 0 \mid X_i)$.

We now make use of the fact that for any truncated multinomial, that is a multinomial in which the frequencies for certain outcomes cannot be observed, the probability, P', for a particular observable outcome is its unconditional probability divided by the total of probabilities for observable outcomes. Thus we may write

$$P'(Y_i = 1 \mid X_i) = \pi_1 P(Y_i = 1 \mid X_i)/[\pi_1 P(Y_i = 1 \mid X_i) + \pi_2 P(Y_i = 0 \mid X_i)] \quad (1)$$

in consequence of which

$$\frac{P'(Y_i = 1 \mid X_i)}{P'(Y_i = 0 \mid X_i)} = \frac{\pi_1 P(Y_i = 1 \mid X_i)}{\pi_2 P(Y_i = 0 \mid X_i)} \quad (2)$$

or the log odds

$$\log \frac{P'(Y_i = 1 \mid X_i)}{P'(Y_i = 0 \mid X_i)} = \log \frac{\pi_1}{\pi_2} + \log \frac{P(Y_i = 1 \mid X_i)}{P(Y_i = 0 \mid X_i)}. \quad (3)$$

What this implies is that the conditional log odds for being a case has the same dependence on X_i as the unconditional log odds; only the intercept is changed. Whatever parametric model we may assume for this dependence, we can obtain equally valid estimates of the parameters involved from retrospective or synthetic retrospective studies as from prospective studies. Also, in the synthetic case since we know π_1 and π_2, we can adjust our estimate of the intercept to take them into account. In particular, if we have a linear parametric model of the form

$$\log \frac{P(Y_i = 1 \mid X_i)}{P(Y_i = 0 \mid X_i)} = \sum_{j=0}^{k} \beta_j z_j(X_i) \quad (4)$$

in which the $Z_j(X_i)$, $j = 0, \cdots, k$, are various known functions of the vector X_i, we can proceed with obtaining estimates of the β_j from the results of our contrived retrospective study. The same rationale underlies the analysis of actual retrospective studies.

In (4) the $Z_j(X_i)$ are general and include functions of either single or several of the variables in X_i, whether those variables are continuous, discrete, or categorical.

(The development above under which conditional log odds have the same dependence, except for the intercept, on X_i as unconditional log odds can still hold for certain forms of nonrandom sampling. Suppose that instead

of random sampling of cases and controls, we sampled in a way which depended on the value of X, $\pi_1(X)$, and $\pi_2(X)$ replacing π_1 and π_2 respectively. If $\pi_1(X)$ and $\pi_2(X)$ factor respectively into $\pi_1 \times f(X)$ and $\pi_2 \times f(X)$, the common factor $f(X)$ would cancel out so that the demonstrated result would continue to hold. Essentially then, the retrospective method does not require random sampling, but only unbiased sampling—the nonrandomness with respect to X should be the same for both cases and controls. However, if a 100% sample of cases is taken, then the sample of controls should be a random one.

An interesting extension is to the case of partial dependencies—suppose we wish to get at the influence of a particular x in X, say x_1, holding constant x_2 the remaining variables in X. If we use the Mantel–Haenszel device of keeping x_2 constant by stratification, then even if sampling is biased relative to x_2, we can still get properly at the influence of x_1 if the following factorizations of the sampling rule hold—$\pi_1(x_1, x_2) = \pi_1(x_2) \times f(x_1 \mid x_2)$ and $\pi_2(x_1, x_2) = \pi_2(x_2) \times f(x_1 \mid x_2)$. That is we can have a biased sample with respect to x_2 provided we have a conditionally unbiased sample with respect to x_1.)

PARAMETER ESTIMATION USING A DISCRIMINATION APPROACH—QUESTIONS ABOUT PROPRIETY

An interesting approach to estimating the β_i was proposed by Cornfield *et al.* [1961] for continuous $Z_i(X_i)$. (They were considering the case where X_i was a vector of continuous variables and that the log odds could be expressed as a linear function of those variables.) Their position was that variables were of two kinds—that while some variables measured factors which would influence the development of disease, other variables might only be descriptive or discriminatory of individuals who were or were not going to develop disease. In a mixed situation of causative and descriptive variables, they proposed that it would be as reasonable to make an analysis assuming a purely descriptive situation as one assuming a purely causative situation. Their descriptive approach led them to the problem of having to discriminate between two multivariate distributions which were assumed to be normal in the several variables measured. A postulation of equality for the variance-covariance matrices in the two populations would lead to equality of variances for any linear function of the variables, including specifically the readily-obtained best linear discriminator between the two populations. But where two normal distributions differ only in their averages, it is readily shown that the log ratio of their two densities at any value is linear in that value; linearity in all the variables then obtained for normal linear combinations of several variables. The Cornfield–Gordon–Smith device was then to obtain the best linear discriminator and to use the resulting coefficients in the log-odds expression as estimates of the β_i. These results could be used alternatively as an entering estimate in a causative-type analysis.

This device of Cornfield, Gordon, and Smith has been successfully applied in practice, having the great advantage of permitting direct rather than

iterative solution. In some cases, application has been mechanical even when the variables measured clearly departed from normal, being in some instances discrete, categorical, or even dichotomous. In one instance, Kahn *et al.* [1971] reported rather substantial agreement between the discriminatory and the causative approaches.

Unquestioned acceptance of the Cornfield–Gordon–Smith discriminatory approach would of course avoid the problems of excessive computer time referred to above. But reservations as to the general suitability of the approach do exist, and some have been voiced by Halperin *et al.* [1971]. The practical success to date of that approach may have been largely happenstance. In part, the success may be due to the large degree of variation occurring in the self-selection of X values by individuals. But uniformly good success should not be expected. To see this, consider that we have conducted a designed laboratory experiment with a single variable employed at only two levels. The resulting data can be analyzed validly by the causative approach, but the normal discriminatory approach will fail since the X values will have dichotomous distributions among both those developing and those remaining free from the disease.

LOGISTIC REGRESSION AS AN ALTERNATIVE TO NORMAL VARIABLE ANALYSIS

Some interesting possibilities arise, however, if we consider using the Cornfield–Gordon–Smith device, but in reverse; that is we employ causative-type analyses even when we are in a discriminatory or descriptive-type situation. In fact, this reverse use is valid even if not so efficient or powerful as the discrimination approach would be if the normal parametric model held. A causative-type analysis can be used in place of normal-assumption tests for a variety of situations, as follows:

1. As a replacement for the t test in comparing two normal averages. This is done by regressing the log odds of being from one or the other of the populations against the variable measured and testing the significance of the fitted regression coefficient;
2. As a replacement for the F test in comparing several normal averages. This is an extension of the preceding using Mantel's [1966] generalization of the logistic from dichotomous to polychotomous situations;
3. Each of the foregoing tests is immediately expandable to the comparison of more than two vectors of normal averages;
4. A test can be made of the equality of 2 or more variances or variance-covariance matrices. This follows from a readily-made demonstration that the log-relative densities for two normal distributions is linear in X if the variances are equal, but is quadratic if they are unequal. If a quadratic function fitted to the data leads to significant improvement over a linear function, this would then be indicative of inequality in the variances.

Fuller implementation of such an approach would have to depend on the development of correct significance tests, but the use of likelihood chi-squares may reasonably be suitable.

What justification might there be for using this reverse causative approach, since suitable tests do exist? An answer occurs when we consider that there are other situations for which this approach can be suitable, but for which again suitable tests already exist. Thus it is also true that the log relative densities or log relative probabilities of two exponential distributions, or two binomial distributions, or two Poisson distributions are also linear in the observation with the coefficient dependent on some measure of difference in the parameter values. This can, in fact, be shown to be the case for any exponential-type distribution for which the sum of the observations is sufficient.

What seems then to be the case is that the reverse causative approach permits us to make comparisons when distributions are of such an exponential type, but without having to postulate just which distribution of the type. In general, we do not, for instance, know that we are dealing with normal distributions; we just make the normal assumption for convenience. The causative analysis would have greater validity by virtue of not requiring so restrictive an assumption, but it would result in some loss in power if normality did in fact hold. If investigations into this revealed only minor or moderate loss in power, the increased validity could justify more general use of this approach.

In fact, the use of logistic-regression analysis as an alternative to normal variable analysis has already been suggested by Cox ([1970] exercises 49 and 50, p. 121). Cox suggests the validity of such an approach for comparing two multivariate normal means, and indicates the more general appropriateness for exponential-type distributions.

(We may note a parallel to the problem of nonrandom but unbiased sampling relative to retrospective studies. Suppose that our samples come not from the original normal distributions, but these distributions modified by a censoring function $\pi(X)$, so that the modified density of X is given by $\pi(X)f(X)/\int_{-\infty}^{\infty} \pi(X)f(X)\,dx$. If $\pi(X)$ is the same for all the samples, the logistic-regression approach would be sensitive to differences in the underlying $f(X)$. Where 2 normal distributions have equal variances but unequal means, the censored distributions may have unequal variances. Alternative logistic-regression approaches can be taken so as either to assume no censoring or to allow for any censoring in making tests on variances.)

ETUDES RETROSPECTIVES SYNTHETIQUES ET PROBLEMES ASSOCIES

RESUME

On discute des approches prospectives et rétrospectives pour l'estimation de l'influence de plusieurs variables sur la présence de maladies. On établit les hypothèses sous lesquelles ces approches tendent à donner les mêmes estimateurs que ceux d'une approche expéri-

mentale idéale mais inaccessible. On prouve que dans une grande enquête prospective où l'on a observé relativement peu de cas de la maladie, les problèmes de calcul peuvent être si accablants qu'ils empêchent une analyse imaginative et complète des données. L'enquête prospective peut être reconvertie en une étude rétrospective synthétique en sélectionnant un échantillon aléatoire des cas de maladie et un échantillon aléatoire des non-cas, l'intensité d'échantillonnage étant petite pour les non-cas mais très forte pour la population des cas. On démontre qu'un tel échantillonnage tendra à laisser la dépendance des "log odd" avec les variables inchangée à une constante additive près.

On note l'utilisation d'une méthode d'analyse non itérative pour la fonction de discrimination et on indique qu'elle n'est pas en général appropriée. On fait la suggestion inverse que les données normales peuvent être analysées par une approche en "log odd" et que cela fournit d'autres tests que ceux utilisés dans la comparaison de deux (ou plusieurs) vecteurs de moyennes, ou deux (ou plusieurs) matrices de variance-covariance.

REFERENCES

Cornfield, J., Gordon, T., and Smith, W. W. [1961]. Quantal response curves for experimentally uncontrolled variables. *Bull. Int. Statist. Inst. 38*, part 3, 97–115.

Cox, D. R. [1970]. *Analysis of Binary Data.* Methuen, London.

Halperin, M., Blackwelder, W. C., and Verter, J. I. [1971]. Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *J. Chron. Dis. 24*, 125–58.

Kahn, H. A., Herman, J. B., Medalie, J. H., Neufeld, H. N., Riss, E., and Goldbourt, U. [1971]. Factors related to diabetes incidence: a multivariate analysis of two years observation on 10,000 men. The Israel ischemic heart disease study. *J. Chron. Dis. 23*, 617–29.

Mantel, N. [1966]. Models for complex contingency tables and polychotomous dosage response curves. *Biometrics 22*, 83–95.

Mantel, N. and Haenszel, W. [1959]. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst. 22*, 719–48.

Received May 1972, *Revised February* 1973

Key Words: Retrospective studies; Prospective studies; Synthetic retrospective studies; Logistic regression; Log-linear models; Discrimination analysis; Exponential-type distributions; Linear discriminators; Polychotomous responses.