PREFACE

Twenty years have elapsed since Mantel and Haenszel published their seminal article on statistical aspects of the analysis of data from case-control studies. Their methodology has been used by thousands of epidemiologists and statisticians investigating the causes and cures of cancer and other diseases. Their article is one of the most frequently cited in the epidemiological literature, and there is no indication that its influence is on the wane; on the contrary, with the increasing recognition of the value of the case-control approach in etiological research, the related statistical concepts seem certain to gain even wider acceptance and use.

The last two decades have also witnessed important developments in biostatistical theory. Especially notable are the log-linear and logistic models created to analyse categorical data, and the related proportional hazards model for survival time studies. These developments complement the work done in the 1920s and 1930s which provided a unified approach to continuous data *via* the analysis of variance and multiple regression. Much of this progress in methodology has been stimulated by advances in computer technology and availability. Since it is now possible to perform multivariate analyses of large data files with relative ease, the investigator is encouraged to conduct a range of exploratory analyses which would have been unthinkable a few years ago.

The purpose of this monograph is to place these new tools in the hands of the practising statistician or epidemiologist, illustrating them by application to *bona fide* sets of epidemiological data. Although our examples are drawn almost exclusively from the field of cancer epidemiology, in fact the discussion applies to all types of case-control studies, as well as to other investigations involving matched, stratified or unstructured sets of data with binary responses. The theme is, above all, one of unity. While much of the recent literature has focused on the contrast between the cohort and case-control approaches to epidemiological research, we emphasize that they in fact share a common conceptual foundation, so that, in consequence, the statistical methodology appropriate to one can be carried over to the other with little or no change. To be sure, the case-control differs from the cohort study as regards size, duration and, most importantly, the problems of bias arising from case selection and from the ascertainment of exposure histories, whether by interview or other retrospective means. Nevertheless, the statistical models used to characterize incidence rates and their association with exposure to various environmental or genetic risk factors are identical for the two approaches, and this common feature largely extends to methods of analysis.

Another feature of our pursuit of unity is to bring together various methods for analysis of case-control data which have appeared in widely scattered locations in the epidemiological and statistical literature. Since publication of the Mantel-Haenszel procedures, numerous specializations and extensions have been worked out for particular types of data collected from various study designs, including: 1-1 matching with binary and polytomous risk factors; 1:M matching with binary risk factors; regression models for series of 2×2 tables; and multivariate analyses based on the logistic function. All these proposed methods of analysis, including the original approach based on stratification of the data, are described here in a common conceptual framework.

A second major theme of this monograph is flexibility. Many investigators, once they have collected their data according to some specified design, have felt trapped by the

intransigences of the analytical methods apparently available to them. This has been a particular problem for matched studies. Previously published methods for analysis of 1:M matched data, for example, make little mention of what to do if fewer controls are found for some cases, or how to account for confounding variables not incorporated in the design. The tendency has therefore been to ignore the matching in some forms of analysis, which may result in considerable bias, or to restrict the analysis to a subset of the matched pairs or sets, thus throwing away valuable data. Such practices are no longer necessary nor defensible now that flexible analytical tools are available, in particular those based on the conditional logistic regression model for matched data.

These same investigators may have felt compelled to use a matched design in the first place in order simultaneously to control the effects of several potential confounding variables. We show here that such effects can often be handled adequately by incorporation of a few confounding variables in an appropriate regression analysis. Thus, there is now a greater range of possibilities for the control of confounding variables, either by design or analysis.

From our experiences of working with cancer epidemiologists in many different countries, on projects wholly or partly supported by the International Agency for Research on Cancer, we recognize that not all researchers will have access to the latest computer technology. Even if such equipment is available at his home institution, an investigator may well find himself out in the field wanting to conduct preliminary analyses of his data using just a pocket calculator; hence we have attempted to distinguish between analyses which require a computer and those which can be performed by hand. Indeed, discussion of the methods which require computer support is found mostly in the last two chapters.

One important aspect of the case-control study, which receives only minimal attention here, is its design. While we emphasize repeatedly the necessity of accounting for the particular design in the analysis, little formal discussion is provided on how to choose between various designs. There are at least two reasons for this restriction in scope. First, the statistical methodology for estimation of the relative risk now seems to have reached a fairly stable period in its development. Further significant advances in this field are likely to take place from a perspective which is quite different from that taken so far, for instance using cluster analysis techniques. Secondly, there are major issues in the design of such studies which have yet to be resolved completely; these include the choice of appropriate cases and controls, the extent to which individual matching should be used, and the selection of variables to be measured. While an understanding of the relevant statistical concepts is necessary for such design planning, it is not sufficient. Good knowledge of the particular subject matter is also required in order to answer such design questions as: What factors are liable to be confounders? How important are differences in recall likely to be between cases and controls? Will the exposure influence the probability of diagnosis of disease? Are other diseases liable to be related to the same exposure?

We are indebted to Professor Cole for providing an introductory chapter which reviews the role of the case-control study in cancer epidemiology and briefly discusses some of these issues.

N.E. Breslow and N.E. Day