MODERN EPIDEMIOLOGY

KENNETH J. ROTHMAN

Professor, Department of Family and Community Medicine,
University of Massachusetts Medical School,
Worcester, Massachusetts

LITTLE, BROWN AND COMPANY
Boston/Toronto

1986

CONTENTS

PREFACE

The tenets of epidemiology, like those of every other science, have become established piecemeal. Some are more useful than others, and some exist in mutual conflict. In this book my aim has been to weave the diverse threads of epidemiologic concepts and research methods into a single fabric. I have tried to reconcile conflicting ideas and unify the conceptual foundation, omitting needless partitions. In particular, I have labored to tie the statistical topics of epidemiologic analysis—which have a way of generating their own special goals, momentum, and lingo—to the basic goals of epidemiologic research. I have also ventured to reconcile epidemiologic principles with the broader goals and methods of scientific inquiry, as I understand them. In sewing the final cloth, I have been mindful that I cannot succeed fully, but rather must fail in my attempts to varying degrees. Intent readers will surely find holes in the fabric and an incorrect stitch here and there. Some of these irregularities undoubtedly reflect inadequate understanding or communication on my part. Some mark conceptual areas, such as confounding and interaction between causes, where development is progressing rapidly. I hope that such problems are few, and small enough not to impair the overall usefulness of the work.

Throughout this book I have strived to make the material accessible to a novice to the field. Whenever possible the descriptions are verbal rather than mathematical, despite the quantitative objectives of research. The first eight chapters deal with fundamental issues of epidemiologic conceptualization, measurement, and study design, and should be comprehensible even to those who lack previous training in epidemiology or statistics; the second eight chapters address the somewhat more technical issues of epidemiologic data analysis, but even these topics are presented with step by step explanations and simplicity as a central objective.

Chapters 1 through 5 form an introductory unit on basic epidemiologic concepts and tools. Chapter 1 places epidemiology in its historical perspective. Chapter 2 ventures into the philosophic foundation for epidemiology, providing a model for causal action that serves as a platform for understanding etiology and its quantitative description. Chapters 3 through 5 continue with the fundamental measures of epidemiology (incidence, prevalence, and risk) and the measures derived from them to quantify causal actions.

Chapters 6 through 8 form a second unit that deals with epidemiologic studies. The basic types of studies are presented in Chapter 6, where I have pursued steadfastly the objective of a unified approach, stressing the theoretical connections among study types. Chapters 7 and 8 explore the issues of study design without resorting to mathematical notation. They emphasize the sources of error in effect estimates as well as the quantitative nature of most aspects of study design.

Chapters 9 through 16 deal with data analysis. In this section some reliance on mathematical formulations has been unavoidable, and I have assumed a basic knowledge of the relevant statistical distributions. Nevertheless, the fundamental statistical principles are introduced and explained in Chapters 9 and 10 using as little notation as possible. Chapter 11 introduces the basic analytic formulations for crude data, which are extended in Chapters 12 and 13 for stratified and matched data. Chapters 11 through 13 cover the routine analytic tasks that an epidemiologist faces; consequently, these are the most technical chapters in the book. Various approaches are described in detail, so that these chapters can be used as a reference for researchers, as well as an instructional guide to the fundamental analytic methods.

The final three chapters turn to more advanced analytic topics, but the emphasis is not so much on formulas as on analytic strategies. Thus, Chapter 14 on multivariate analysis is probably the least technical description of multivariate analysis in any textbook; it provides practical guidance on choosing, constructing, and interpreting multivariate models. Chapters 15 and 16 deal with the advanced topics of interaction and "dose-response" evaluation, but the emphasis once again is on the principles and pitfalls of such analyses, rather than on the technical aspects of the requisite calculations. I could not avoid formulas entirely and still provide an adequate discussion of these topics, but the formulas presented illustrate approaches of conceptual simplicity amenable to a pencil-and-paper solution.

In my efforts to tie together epidemiologic concepts for all these topics, I have encountered some fossilized divisions that I consider no longer useful. For example, a rift has separated the traditional area of infectious disease epidemiology from the more recent and growing area of "chronic" disease epidemiology. I have never been persuaded of any rationale for this distinction. The terms "infectious" and "chronic" are neither mutually exclusive nor collectively exhaustive alternatives. Many diseases are both infectious and chronic; some, such as fatal traumatic injury, are neither. "Chronic" has sometimes been taken to mean a long induction period, rather than a long period of manifestation, but this redefinition still fails to make a meaningful distinction between two conceptually different types of epidemiology. Although some specialized methods have been developed solely to study the spread of infectious illness, whatever distinctions exist between traditional and modern areas of epidemiology are certainly less important than the broad base of concepts that are shared. This book does not deal with models for epidemic spread, but focuses on the general epidemiologic concepts that apply to all diseases, infectious or not, chronic or not, and to causes that have short or long induction periods.

Another distinction that has been used to categorize epidemiologic work is its classification into descriptive and analytic epidemiology. My view is that this demarcation is also best forgotten. It has been used in reference both to specific study variables (so-called "descriptive" variables being distinguished from putative causes) and to entire studies, but in neither context does it hold as a sensible classification scheme. No qualitative distinction, other than a completely arbitrary one, distinguishes "descriptive" variables from more fundamental risk factors. Any disease determinant can be specified in terms of more proximal determinants or previously unsuspected confounding factors. The division of epidemiologic research into descriptive and analytic compartments has given rise to the illusion that there are different sets of research principles that apply to descriptive and analytic studies. This notion devolves from a mechanical view of scientific research, and diverges from prevailing doctrines of scientific philosophy. For example, the view that "descriptive data" from "exploratory studies" generate hypotheses, whereas the data from "analytic studies" are used to test hypotheses, does not cohere with a broader understanding of science. Hypotheses are not generated by data; they are proposed by scientists. The process by which scientists use their imagination to create hypotheses has no formal methodology and is certainly not prescriptive. Any study, whether considered exploratory or not, can serve to refute a hypothesis. It is not useful to regard some studies merely as "hypothesis generating" and others as "hypothesis testing," because the inexorable advance of scientific knowlege cannot be constrained by such rigidities.

I believe that epidemiology is much more coherent than these traditional divisions would suggest. Even the stark contrast between follow-up studies and case-control studies has been softened as understanding of the basic principles of epidemiology has progressed. In writing this book, my greatest hope is to convey to the reader the conviction that epidemiologic principles can be understood as an integrated substrate of logical ideas, rather than as a jumble of isolated and sometimes conflicting postulates.

K. J. R.

10. FUNDAMENTALS OF EPIDEMIOLOGIC DATA ANALYSIS

In a well-planned study, the raw observations that constitute the data contain the information that satisfies the objectives of the study. In Chapter 7 it was emphasized that a study is a measurement exercise and that the overall goal for a study is accuracy in measurement. Accordingly, the goal in data analysis is to extract the pertinent measurement information from the raw observations.

Typically, there are several distinct stages in the analysis of data. In the preliminary stage, the investigator should review the recorded data for accuracy, consistency, and completeness; this process is often referred to as *data editing*. Next, the investigator should summarize or transform the data into a concise form for subsequent analysis, usually into contingency tables that tabulate the distribution of the observations according to key factors; this stage of the analysis is referred to as *data reduction*. Finally, the edited and reduced data are used to generate the epidemiologic measures of interest, typically one or more measures of effect (such as relative risk estimates), with appropriate confidence intervals. This last stage of analysis is sometimes considered the analysis proper, but it is more convenient to refer to it as *effect estimation* (or perhaps just *estimation*, if the goal of the analysis is to estimate disease frequency rather than to measure an effect). For some investigators, the last stage of analysis inevitably includes statistical hypothesis testing. The previous chapter explained why hypothesis testing is an undesirable feature of data analysis in most epidemiologic situations. Since the statistical theory behind interval estimation is closely related to statistical hypothesis testing, however, it is useful to consider the issues described in statistical hypothesis testing as a foundation for understanding epidemiologic data analysis.

DATA EDITING

There is no excuse for failing to scrutinize the raw data intensely for errors and to correct such errors whenever possible. Errors are routinely introduced into data in a variety of ways; some errors are detectable in editing and some are not.

The data in an epidemiologic study usually derive from a self-administered or an interviewer-administered questionnaire or from existing records that are transcribed for research. The data from the questionnaire or record-abstraction form may be transcribed from this primary form to a code form for machine entry, usually by keypunching. Coding of responses is often necessary. For example, occupational data obtained from interviews need to be classified into a manageable code, as does drug information, medical history, and many other types of data. Data such as age or year of birth (year of birth is usually preferable to age, since it tends to be reported more accurately and does not change with time), although often grouped into broad categories for reporting purposes, should be recorded in a precise form rather than grouped because the actual values will allow greater flexibility later in the analysis. For example, different groupings may be necessary for comparisons with several other studies. Some nominal scale variables that have only a few possible values can be precoded on the primary forms by checking a designated box corresponding to the appropriate category. For nominal scale variables with many possible categories, however, such as country of birth or occupation, precoded questions are not practical. If all data items can be precoded, it may be feasible to collect the data in a primary form that can be read directly by a machine, by optical scanning, or by some comparable method. Otherwise, it will usually be necessary to translate the information on the primary data form before it is stored in a machine or in machine-readable form.

It is possible and usually desirable to avoid rewriting the data onto a secondary data form during the coding process. Rather than generating additional transcription errors, it is preferable to code the data while simultaneously keying them into a computer storage system. A computer program can be devised to prompt data entry item by item, displaying category codes on a terminal screen to assist in coding. If the data are coded and rewritten by hand, they will often require keypunching anyway, unless they are coded onto optical scanning sheets; consequently, direct data entry during coding reduces both costs and errors. The fewer the number of rewriting operations between the primary record and the machine-stored version, the fewer the errors that are likely to occur. If rewriting is unavoidable, it is useful to assess the extent of coding errors in the rewritten form by coding a proportion of the data forms twice, independently. The information thus obtained can be used to judge the magnitude of bias introduced by misclassification from coding errors.

Basic editing of the data involves checking each variable for illegal or unusual values. For example, gender may be coded 1 for male and 2 for female. Usually a separate value, perhaps 3, is used to designate an unknown value. It is preferable not to assign a code of zero if it can be avoided because missing information or non-numeric codes may be interpreted by some machines or programs as a zero. By not assigning zero as a specific code, not even for unknown information, it may be possible to detect keypunching errors or missing information. The distribution of each variable should be examined in the editing process. Any inadmissible values should be checked against the primary data forms. Unusual values such as unknown gender or unusual age or birth year should also be checked.

In addition to checking for incorrect or unusual values, the distribution of each variable should be examined to see if it appears reasonable. Would you expect about half of the subjects to be males, about 80 percent (a reasonable figure if the subjects have, say, upper respiratory cancer), or about 2 percent (if the subjects are nurses)? Such an evaluation may reveal important problems that might not otherwise come to light. For example, a programming error could shift all the data in each electronic record by one or more characters, thereby producing gibberish that nevertheless might not be detectable in, say, a multivariate analysis (an important drawback of the multivariate approach). The potential for such a disaster heightens the need to check carefully the distribution of each variable during the editing of the data.

The editing checks described so far relate to each variable in the data taken singly. In addition to such basic editing, it is usually desirable to check the consistency of codes for related variables. It is not impossible, but it is improbable that a person 18 years of age will have three children. Males should not have been hospitalized for hysterectomy. People over 2 meters tall are unlikely to weigh less than 50 kilograms. Thorough editing will involve many such consistency checks and is best accomplished by computer programs designed to flag such errors [MacLaughlin, 1980]. Occasionally an apparently inconsistent result may appear on checking to be correct, but many errors will turn up through such editing. It is important, also, to check the consistency of various distributions. If exactly 84 women in a study are coded as premenopausal for a variable, "type of menopause," then it is reassuring that exactly 84 are likewise coded as premenopausal for the variable "age at menopause" (for such a variable, the code "premenopausal" should take a different code number from that assigned to unknown—e.g., 98 for premenopausal and 99 for unknown).

An important advantage of coding and entering data through a computer program is the ability to edit the data automatically during the entry process. Inadmissible or unusual values can be screened as they are entered. Inadmissible values can be rejected and corrected on the spot by programming the machine to print an error message on the screen and give an audible message as well to alert the operator about the error. Unlikely but legal values can be brought to the operator's attention in the same way. A sophisticated data-entry program can also check for consistency between variables and can eliminate some potential inconsistencies by automatically supplying appropriate codes. For example, if a subject is premenopausal, the program can automatically supply the correct code for "age at menopause" and skip the question. (On the other hand, some investigators may prefer the redundancy of the second question to guard against an error in the first.)

Even with sophisticated editing during data entry, it is still important to edit the stored data before analysis, to check on the completeness of the data and the reasonableness of the distribution of each variable. Neither of these features can be evaluated by a data-entry program.

Every experienced investigator knows that even the most meticulous data collection efforts suffer from errors that are detectable during careful editing. If editing is planned as a routine part of handling the data, the existence of such errors is usually not a serious problem. If editing is ignored, momentous problems can result.

DATA REDUCTION

The notion fundamental to data reduction is that certain observations in a set of data are equivalent, and it is easier to deal with equivalent observations after they have been summarized. The summary form usually is a contingency table in which the frequency of subjects (or units of observation) with every specific combination of variable values is tabulated for variables of interest. Such a table is presumed to contain, in summary form, essentially all the relevant information in the data. From the contingency table, the investigator can proceed with effect estimation. In addition, the table displays the distribution of subjects according to key variables and thus conveys directly to the investigator an intimacy with the data that is not easily obtained in any other way.

Data reduction into a contingency table is predicated on an analysis in which there is no concern for confounding or effect modification or there are at most only a small number of variables that might be confounders or effect modifiers. If the analysis must take account of a large number of variables, a multivariate analysis using mathematic modeling will be necessary. For such multivariate analyses, it is not necessary to reduce the data into a contingency table. Nevertheless, to ensure that the investigator acquires some familiarity with the data, it is advisable, even when planning a multivariate analysis, to reduce the data into contingency table format for the variables of central interest. Indeed, proceeding with an abridged analysis based on the contingency table data is a good idea even if the need for the multivariate analysis is certain.

Collapsing the edited data into categories for the contingency table may necessitate some decision making. The process is straightforward for nominal scale variables such as religion or race, which are already categorized. For continuous variables, however, the investigator must decide how many categories to make and where the category boundaries should be. The number of categories will usually depend on the amount of data available. If the data are abundant, it is always preferable to divide a variable into many categories. On the other hand, the purpose of data reduction is to summarize the data concisely and conveniently; creating too many categories would defeat this purpose. For control of confounding, it is rarely necessary to have more than about five categories [Cochran, 1968]. If an exposure variable is categorized to examine effect estimates for various levels of exposure, again it would be unusual to require more than about five categories. Frequently, however, the data are so sparse that it is undesirable to create as many as five categories for a given variable. When the observations are stretched over too many categories, the numbers within categories become statistically unstable and produce large random errors in the effect estimates.

Since most of the confounding from a given factor can be removed by a stratified analysis based on only two categories of a continuous variable [Cochran, 1968], it is desirable with sparse data to keep the number of categories small, perhaps two or three. Even a large body of data can be spread too thin if the contingency table involves too many dimensions, that is, if too many variables are used to classify the subjects. With three variables, apart from exposure and disease, and three categories for each variable, there will be 27 2 × 2 tables (assuming that both exposure and disease are dichotomous). With an additional two variables of three categories each, there will be a total of 243 2 × 2 tables, enough to stretch even a considerable body of data too thin, since a study of 10,000 people would average only about 10 subjects per cell of the multidimensional table. If a stratified analysis is planned and it is necessary to stratify by several variables, it is probable that only a few, perhaps as few as two, categories can be used for each variable. With only two categories per variable, stratification by five variables requires 32 rather than 243 2 × 2 tables, and a study of 10,000 subjects would average 78 subjects per cell rather than 10, thereby gaining precision at the cost of some potential residual confounding within categories.

The investigator must also decide where to draw the boundary between categories. There is no accepted method for doing this. A frequently expressed concern is that boundaries might be "gerrymandered," that is, shifted after a preliminary examination of the effect estimates in such a way that the estimates are altered in a desired direction. This concern imputes a level of dishonesty to the investigator that is presumably uncommon. Furthermore, the shift of a boundary in categorization rarely has a substantial effect on the magnitude of an estimate and then only because of a large random error component. On the other hand, it is frequently useful to inspect the distribution of a variable before deciding at which points to carve categories. There may be "natural" categories if the distribution has more than one mode. The distribution may be sufficiently skewed that preconceived category boundaries would lead to an inefficient separation of subjects, with too few in some categories and too many in others. For these reasons, it is often preferable to define the final categories after reviewing the data, notwithstanding the common advice that it is somehow more "objective" to do so in ignorance of the distribution of observations in hand. Nevertheless, if meaningful category boundaries are inherent in the variable, these can and should be specified a priori. For example, in categorizing subjects according to analgesic consumption, it is desirable to create categories that contrast the various therapeutic indications for analgesic use, the recommended doses for which can be specified in advance. It is often desirable, especially for an exposure variable, to retain extreme categories in the analysis without merging these with neighboring categories, since the extreme categories are often those that permit the most biologically informative contrasts.

A common problem in creating categories is the question of how to deal with the ends of the scale. Open-ended categories can provide an opportunity for considerable residual confounding, especially if there are no theoretical bounds for the variable. For example, age categories such as 65+, with no upper limit, allow a considerable range of variability within which the desired homogeneity of exposure or outcome may not be achieved. Another example is the separation of the effects of alcohol consumption and tobacco smoking on the risk of oral cancer; within categories of heavy smoking, it is a reasonable possibility that the heaviest smokers drink more alcohol than those who smoke less within that category [Rothman and Keller, 1972]. When residual confounding from open-ended categories is considered likely, strict boundaries should be placed on every category, including those at the extremes of the scale.

A convenient method of assembling the final categories is to categorize the data initially much more finely than is necessary. A fine categorization will facilitate review of the distribution for each variable; more usable categories can then be created by coalescing adjacent categories. The coalescing of adjacent strata for a rank-ordered confounding variable can be justified by the lack of confounding that is introduced by merging the categories; this merging will not introduce confounding if the exposure distribution is the same among the controls or person-time denominators between the strata, or if the proportion of cases or the disease rate is the same among nonexposed subjects between the strata [Miettinen, 1976b]. The advantage of starting with more categories than is ultimately necessary is that the merging of categories can be conveniently accomplished with pencil and paper in seconds or minutes, whereas separating categories into subcategories cannot be done without reading through the entire data file, thus adding another computer run.

EFFECT ESTIMATION (AND HYPOTHESIS TESTING)

Hypothesis Testing

In data analysis, as opposed to the broader area of scientific inference, hypothesis testing generally refers to the evaluation of a null hypothesis. The introduction of the concepts of statistical evaluation early in the twentieth century led to an appreciation of the importance of assessing the role of random error in observations. Hypothesis testing is directed at the question of whether random error might account entirely for an observed association. The statistic used to evaluate this question is the *P-value*.

The *P*-value is usually interpreted as the probability that an association at least as strong as that actually seen in the data might have arisen if the null hypothesis were true, that is, by chance alone. Because a low *P*-value indicates a low probability, under the null hypothesis, of results as extreme or more extreme than those observed, low *P*-values are taken as an indication that the data are more compatible with the alternative hypothesis of a nonzero effect than with the null hypothesis. A *P*-value should not be confused with the probability that the null hypothesis is correct; it is calculated *on the assumption* that the null hypothesis is correct. Extremely low *P*-values can occur even when the null hypothesis is true; in fact, they are guaranteed to occur a small proportion of the time. The informativeness of the *P*-value derives solely from the interpretation that small *P*-values indicate relatively less consistency between the data and the null hypothesis and relatively more consistency with the alternative hypothesis of a nonzero effect.

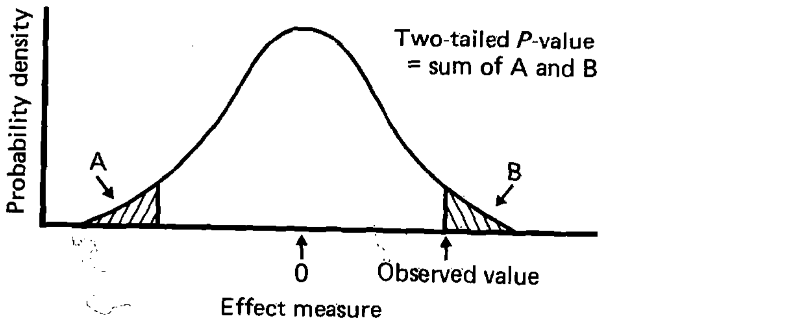

Fig. 10-1. Distribution of effect estimates under the null hypothesis in large studies (continuous distribution).

Imagine that an estimate had a continuous sampling distribution on its scale of measurement, with a value of zero corresponding to the null hypothesis of no effect. Figure 10-1 illustrates the hypothetical probability density of the estimated effect; the bell shape of the curve is ensured for large studies by the central limit theorem in statistics. Values of the estimate equal to or more extreme than that observed correspond in the likelihood of their outcome to the shaded area in the diagram. The definition of *more extreme* can be unidirectional, in which case the *P*-value is said to be "one-tailed" or "one-sided" and is represented only by the shaded area under one end of the curve, or it can be bidirectional, in which case the "two-tailed" *P*-value corresponds to the sum of the shaded areas under both ends of the curve.

To calculate the *P*-value, it is necessary to postulate a statistical model that describes the probability distribution of the data on the assumption of the null hypothesis. If the distribution of effect estimates that are calculable from the data were actually continuous, it would be inconsequential whether the tail area of the curve is defined as the area corresponding

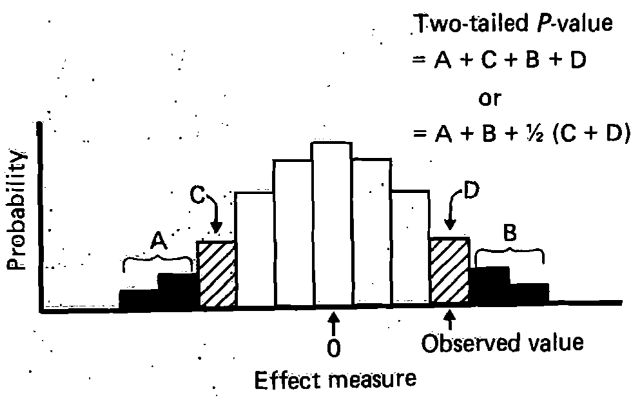

Fig. 10-2. Distribution of effect estimates under the null hypothesis (discrete distribution).

to effect estimates equal to or more extreme than those actually observed, or simply the area corresponding to estimates more extreme than those observed. Typically, however, in epidemiology, the data from which the effect estimates are calculated are discrete frequencies, and the distribution of effect estimates is discrete rather than continuous. The area representing a *P*-value for a discrete distribution is illustrated in Figure 10-2. Traditionally, the *P*-value has been defined as the sum of both the lightly shaded areas and the heavily shaded areas in Figure 10-2. The lightly shaded areas correspond to the probability of the actual observations (and the corresponding value in the opposite direction), whereas the darkly shaded areas correspond to the probability of more extreme departures from the null value than those actually observed. Obviously, for discrete distributions it does matter whether the *P*-value is defined as including the probability of the observed outcome or just the more extreme values.

The problem with the traditional definition of the *P*-value is that it leads to inconsistencies. For example, what if the observed value of the effect estimate were in the center of the distribution, right on the null value? In the traditional definition, each tail would then include more than half the distribution, and the two-tailed *P*-value would be greater than 100 percent, which is inconsistent with the view that the *P*-value represents a probability. An alternative definition of the *P*-value that overcomes this problem is one in which the probability of the observed value of the effect is partitioned, generally by splitting it into equal parts [Lancaster, 1949; Lancaster, 1961]. Thus, the one-tailed *P*-value would correspond to the probability of the more extreme values plus one-half the probability of the observed value. This definition of the *P*-value has been referred to as the "mid-*P*" [Lancaster, 1961]. The two-tailed *P*-value is generally obtained by doubling the one-tailed *P*-value, however the *P*-value is defined.

With discrete data, the probability distributions used to calculate the *P*-value can give rise to intricate calculations; *P*-values calculated directly in this way are referred to as *exact P-values.* Usually it is simpler to use an approximation to the discrete distribution, relying on the fact that a normal curve will approximate the shape of the distribution reasonably well; the larger the frequencies involved in the discrete data, the greater the number of values that can be assumed by the effect estimate and the better the normal approximation to the discrete distribution. The advantage of using the normal distribution is that the calculations necessary to obtain the *P*-values are considerably simpler than those needed to get the exact *P*-value.

In an attempt to make the normal approximation better when frequencies are small, Yates [1934] suggested a "correction" procedure that amounts to shifting the observed value of the effect estimate toward the null value by a distance that corresponds to half of the probability of the actual data under the null hypothesis. This adjustment is intended to compensate for the fact that the observed value of the effect actually represents the central value of a range that corresponds to the region on the scale of the effect measure representing each discrete value. Since the probability of the entire range for the observed value is included in the definition of the traditional *P*-value, the Yates "correction" usually improves the approximation to the traditionally defined exact *P*-value. If, however, the mid-*P* definition were used, then the Yates "correction" would actually make the approximation worse, since the observed value already represents the central value of its discrete range. In this text, the Yates "correction" is ignored.

The general form for statistical testing based on a normal distribution around the null value is given by equation 10-1:

$$\chi = \frac{\mathrm{A} - \mathrm{E}}{\sqrt{\mathrm{V}}} \qquad [10\text{-}1]$$

A is the observed value of the effect estimate, E is the expected value for A under the null hypothesis, and V is the variance of A under the null hypothesis. Provided that under the null hypothesis A is normally distributed, then under the null hypothesis χ will also be normally distributed but with a mean of zero and a standard deviation of unity. A normally distributed random variate with a mean of zero and a standard deviation of unity is referred to as a *standard normal deviate;* synonyms are *critical ratio* and *Z-value.* In this text, χ is used as the notation in the formula to emphasize that the square of the standard normal deviate has a chi-square distribution with "one degree of freedom"—indeed, that is how the one degree of freedom chi-square statistic is defined. (Chi-square with n degrees of freedom is simply the sum of n independent chi-squares with one degree of freedom.) The *P*-value is obtained from the χ value from tables (or computational formulas) of the standard normal distribution. In essence, equation 10-1 converts a normally distributed statistic with a calculated expectation and variance into a standard normal deviate (expectation of zero and standard deviation of unity) for which detailed tables are conveniently available to obtain *P*-values. It would be possible to square the χ and obtain the *P*-value from tables of chi-square, but since these usually have considerably less detail than tables of the standard normal distribution, there is no reason to do so.

To this point this discussion has presumed that the observation of interest is the estimate of effect derived from the data. Although this is generally so, in calculating the χ it is usually more convenient to postulate for the random variable A a measure that contains all the essential statistical information about the effect but for which the variance is more easily and accurately calculated. It is convenient to designate A as the number of exposed subjects with disease in the study; with this substitution, the expected number for A under the null hypothesis will not be zero but must be calculated from the data based on the relevant probability model. The models relevant to epidemiologic studies will be described in Chapters 11 and 12.

Estimation of Effects

The single best numerical estimate of an effect from a set of data is referred to as a *point estimate*. Because a point estimate is only one point on a continuous scale with an infinite number of possible values, there is essentially zero probability that it is correct, even if there is no source of bias. Therefore, although point estimates serve as useful indicators of the magnitude of an effect, it is important to supplement the information that they provide with a measure of the random error in the data. Hypothesis testing can accomplish this goal, but the *P*-value is an undesirable statistic for evaluating random error because it provides no information about magnitude of effect and only indirectly allows assessment of the extent of random error in an estimate. As was emphasized in Chapter 9, the greatest drawback of *P*-values is that they tend to be used for "significance" testing as an analytic goal, diverting the focus away from the proper goal of estimation of effects. A better approach is the use of confidence intervals, which have none of the drawbacks of *P*-values.

A confidence interval denotes a range of values surrounding the point estimate that amounts to a "sampling range" for the estimate. The level of confidence, which is arbitrarily selected by the investigator, is the frame of reference by which the sampling range can be interpreted. Most investigators repeatedly use the same level of confidence to ease comparison; 90 and 95 percent are commonly used values.

The connection between confidence intervals and *P*-values, described in Chapter 9, should be expressed in more formal terms. Like a confidence interval, the *P*-value also measures a sampling range, but it specifically measures the sampling range of the data under the null hypothesis. The null point on the effect scale is the reference point for hypothesis testing, and the *P*-value is a measure of the discrepancy of the data with the reference point in probability terms. A confidence interval, in contrast, fixes the probability to an arbitrarily chosen value, which is dependent on the desired level of confidence, and varies the reference point, which becomes the limit to the confidence interval. Thus, in determining the lower boundary of a 90 percent confidence interval, the reference point is adjusted until the upper tail area is exactly 5 percent (Fig. 10-3). For 90 percent confidence limits, the direction of the adjustment of the reference point will be from the null value toward the point estimate if the one-tail *P*-value is less than 5 percent, leading to a lower confidence bound above the null value (for positive effects). If the one-tail *P*-value is greater than 5 percent, the reference point must be adjusted away from the null value in the direction opposite the point estimate to bring the tail area down to 5 percent, resulting in a confidence interval that will bracket the null value. If the one-tail *P*-value is exactly 5 percent, then one boundary of the 90 percent confidence interval will be equal to the null value.

The most accurate way to determine a confidence limit is to use exact calculations analogous to the exact calculations used to calculate *P*-values. The calculations for confidence limits are considerably more difficult, however, for two reasons. First, the adjustment of the reference point in calculating the tail area amounts to the testing of a non-null hypothesis.

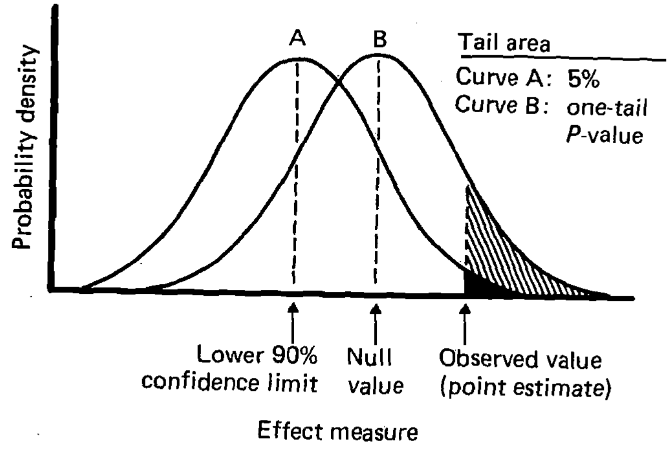

Fig. 10-3. Sampling range of the data in reference to the null value and the lower 90 percent confidence limit.

The statistical models that describe the non-null situation are highly complicated in comparison with the null-hypothesis models and demand much more involved calculations. Second, these intricate calculations have to be repeated in an iterative process for trial values of the reference point until the tail area conforms with the desired level of confidence. Therefore, calculation of exact confidence limits is practically infeasible without programmable electronic computing equipment.

Fortunately, many simple techniques exist, analogous with formula 10-1, to obtain approximate confidence limits. As with hypothesis testing, the accuracy of all the approximate techniques depends on the number of observations because all the methods depend on the normal distribution of effect estimates guaranteed by the central limit theorem for observations that are sufficiently numerous.

A simplifying assumption that is often made is that the sampling variability of an effect estimate is constant along its scale of measurement, that is, the variance of the effect estimate is a constant, independent of the value of the estimate. This assumption is not necessary for hypothesis testing, since the *P*-value is calculated on the assumption that the null hypothesis holds, and therefore the concern in hypothesis testing is to estimate the variance only at the null value. With a large set of observations, the sampling range for the effect estimate is narrow enough to make this assumption appropriate; even if the variance changes substantially along the scale of measurement of the effect measure, in a narrow enough range it will be nearly constant. Therefore, the simplifying assumption that the variance is constant is asymptotically correct; that is, the assumption becomes more appropriate as the number of observations used in the estimation process increases.

The usual and simplest approach to calculating approximate confidence limits is to estimate the standard deviation of the normal curve that represents the approximate sampling distribution of the effect estimate. The area under a symmetric segment of a normal curve is a specific function of the standard deviation; in fact, this relation provides the only interpretability for the standard deviation as a measure of variability: If the distribution is not normal, there is no meaningful interpretation of standard deviation, though confidence intervals might nevertheless be obtained by exact calculation. For any normal curve, 68 percent of the area under the curve lies in the region within one standard deviation (SD) of the central point. Thus, measurement values reported with ± SD as a measure of variability amount to a point estimate with an accompanying 68 percent confidence interval, provided that the sampling distribution is indeed normal. When a level of confidence is chosen, usually the value is not 68 percent but commonly 80, 90, or 95 percent. These levels of confidence correspond to regions that are bounded by points 1.282, 1.645, and 1.960 standard deviation units, respectively, from the central value in either direction (Fig. 10-4).

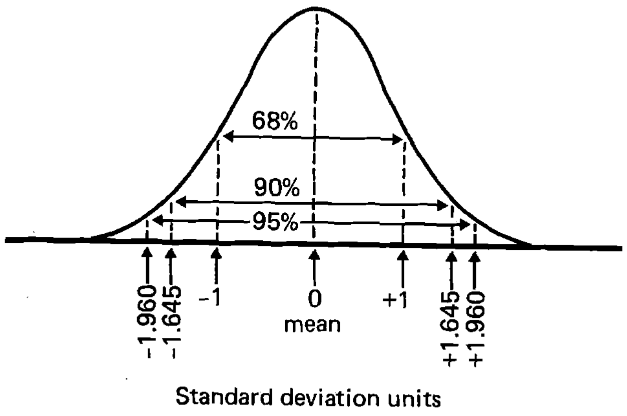

Fig. 10-4. Area under a normal curve.

To obtain a confidence interval based on the assumption of a normal sampling distribution it is necessary to estimate both the expected (mean) value of the effect and the standard deviation. The expected value is estimated by the point estimate, and the standard deviation is usually also estimated simply from the observed data. To construct confidence intervals, with rate difference as the effect measure, the resulting formula would be

$$\widehat{RD} \pm Z\cdot\widehat{SD}(\widehat{RD}) \qquad [10\text{-}2]$$

in which $\widehat{RD}$ indicates the point estimate of rate difference (the caret signifies an estimate), Z is the multiplier for the standard deviation corresponding to the desired level of confidence, and $\widehat{SD}(\widehat{RD})$ indicates the estimated standard deviation of the point estimate; the minus sign gives the lower limit for the interval, and the plus sign gives the upper limit. The point estimate and the standard deviation are derived from the data, and a value for Z is arbitrarily selected to give the desired confidence level, for example, 1.645 for 90 percent confidence, and so on. Frequently, in a formulation such as that given in equation 10-2, the standard deviation is referred to as the *standard error* (abbreviated SE). In some circumstances there is an important distinction to be made between standard deviation and standard error: The standard error is the standard deviation of the sampling distribution of mean values; if the original observations come from a normal distribution, it is important to distinguish the standard deviation of the latter from the standard deviation of mean values, thus giving rise to the need for a separate term, standard error. In the context of this book, however, we shall generally be interested in the sampling distribution of point estimates, which corresponds to the standard error, although it is also perfectly acceptable to use the term standard deviation, since a standard error is a specific type of standard deviation.

If the effect measure of interest were rate ratio rather than rate difference, it might be reasonable to use formula 10-2 and simply substitute $\widehat{RR}$ for $\widehat{RD}$. It is preferable, however, to use a different equation because the sampling distribution for $\widehat{RR}$ is asymmetric, and consequently the sampling distribution of rate ratio estimates is not normally distributed unless a relatively large number of observations is available. Why is the sampling distribution for $\widehat{RR}$ asymmetric? The minimum value for RR is zero, whereas the maximum value is infinity. Random errors can lead to larger discrepancies on the high side of the mean than corresponding discrepancies on the low side of the mean. Notice that for $\widehat{RD}$ the sampling distribution is symmetric. Although the sampling distribution for $\widehat{RR}$ approaches a normal curve for a sufficiently large number of observations, it is customary to use a scale transformation to introduce symmetry and to set confidence limits on a scale of measurement that gives a better approximation to the normal distribution when the observations are relatively sparse. This is conveniently accomplished by using a logarithmic transformation. For setting confidence limits after logarithmic transformation of rate ratio, the formula is

$$\ln(\widehat{RR}) \pm Z\cdot\widehat{SD}(\ln(\widehat{RR})) \qquad [10\text{-}3]$$

This is analogous to formula 10-2, differing only in that $\ln(\widehat{RR})$ has been substituted for $\widehat{RD}$. Having set confidence limits on the logarithmically transformed scale, it is necessary to reverse the transformation so that the limits can be interpreted on the original scale. To do so requires taking the antilogarithm of the limits resulting from formula 10-3. The whole process can be summarized by the formula

$$\exp[\ln(\widehat{RR}) \pm Z\cdot\widehat{SD}(\ln(\widehat{RR}))] \qquad [10\text{-}4]$$

Whereas formula 10-2 gives confidence limits that are equidistant from the point estimate, formula 10-4, because of the scale transformation, gives confidence limits that are asymmetric about the point estimate. The limits are symmetric on the logarithmic scale, but on the original scale the point estimate is the geometric mean between the lower and upper limits; that is, the ratio of the upper bound to the point estimate equals the ratio of the point estimate to the lower bound.

Formulas 10-2 and 10-4 are the simplest general formulas for deriving approximate confidence limits for the rate difference and rate ratio measure of effect, respectively. Many specific techniques have been proposed, each striking a different balance between computational ease and accuracy. Some formulas discard the assumption that the standard deviation is uniform along its scale of measurement and use iterative techniques to estimate the value of the standard deviation at the boundary of the interval; the method of Cornfield [1956] for calculating confidence limits for the odds ratio is an example of this approach; Miettinen and Nurminen [1985] have extended Cornfield's approach to the risk ratio and risk difference measures. Iterative calculations usually require programmed computing assistance, so that the theoretical advantages are accompanied by practical disadvantages.

The simplest specific technique for performing interval estimation is the "test-based" method [Miettinen, 1976a], which assumes that the estimate of the standard deviation of the sampling distribution of the effect estimate obtained at the null value is a reasonable estimate of the standard deviation of the distribution elsewhere along the scale. This assumption differs slightly from the usual assumption that the estimated value of the standard deviation at the point estimate will be appropriate at the bounds of the interval; although both approaches assume that the value of the standard deviation estimated at one point along the scale will apply for both lower and upper bounds, the value estimated at the point estimate is more or less centrally placed between the limits of the interval, whereas the null value is not and might even be outside the interval. If the standard deviation changes along the effect-measure scale, the degree of error in the approximate limits is probably less severe if the standard deviation is estimated at a point central to the confidence interval rather than at the null point, which has no connection to the location of the limits. On the other hand, by choosing the null point as the point at which the standard deviation is estimated, the resulting confidence limits will tend to be more accurate when they fall in the vicinity of the null point, and it may be argued that it is worth obtaining greater accuracy in the vicinity of the null value even if it means sacrificing some accuracy for limits calculated to be far from the null value.

Applying the assumption of test-based limits leads to a concise formulation for obtaining confidence limits, based on the test statistics from equation 10-1. Consider the reformulation of equation 10-1 for rate difference:

$$\chi = \frac{\widehat{RD} - E}{\widehat{SD}(\widehat{RD})}$$

where E, the expectation of $\widehat{RD}$ under the null hypothesis, is zero, and $\widehat{SD}(\widehat{RD})$ is calculated on the assumption that the null hypothesis is true. This gives

$$\chi = \frac{\widehat{RD}}{\widehat{SD}_0(\widehat{RD})} \qquad [10\text{-}5]$$

where $\widehat{SD}_0(\widehat{RD})$ indicates that the SD is estimated at the null value. Equation 10-5 can be rewritten as

$$\widehat{SD}_0(\widehat{RD}) = \frac{\widehat{RD}}{\chi}$$

and substituted into formula 10-2, giving, for the lower and upper limits

$$\widehat{RD} \pm Z\frac{\widehat{RD}}{\chi}$$

or

$$\widehat{RD}(1 \pm Z/\chi) \qquad [10\text{-}6]$$

The χ in formula 10-6 was assumed to be a test statistic evaluating RD per se. Miettinen recommended inserting into formula 10-6 any χ statistic that represents an equally efficient test of the null hypothesis based on the same data. For example, the usual χ based on the distribution of the number of exposed cases could be substituted (see Chap. 11 for the specific application).

The counterpart of equation 10-5 using the rate ratio measure of effect, after logarithmic transformation, is

$$\chi = \frac{\ln(\widehat{RR})}{\widehat{SD}_0(\ln(\widehat{RR}))} \qquad [10\text{-}7]$$

which can be rewritten as

$$\widehat{SD}_0(\ln(\widehat{RR})) = \frac{\ln(\widehat{RR})}{\chi}$$

and substituted into formula 10-4 to give

$$\exp\left[\ln(\widehat{RR}) \pm Z\frac{\ln(\widehat{RR})}{\chi}\right]$$

which simplifies to

$$\widehat{RR}^{(1 \pm Z/\chi)} \qquad [10\text{-}8]$$

As with formula 10-6, the attraction of formula 10-8 rests with the substitution for the χ statistic based on $\widehat{RR}$ an alternative and more convenient χ testing the null hypothesis. Indeed, the same χ statistic can be used in formulas 10-6 and 10-8 to generate confidence limits for rate difference and rate ratio. Note that when the χ value equals the Z multiplier, the lower bound should and does correspond exactly to the null value, which is zero for rate difference and unity for rate ratio.

The test-based formulas for approximate confidence limits given in formulas 10-6 and 10-8 are exceedingly easy to apply and produce usable confidence intervals in a wide variety of situations. The only numbers required from the data are an appropriate point estimate of the effect estimate and the χ statistic from hypothesis testing. Indeed, the use of the χ statistic in these test-based formulas is the main justification for any detailed discussion of statistical hypothesis testing in modern epidemiology, since the estimation of a confidence interval is preferable to the use of *P*-values to evaluate random error, and the *P*-value adds very little information if a confidence interval is given.

Unfortunately, the principle of test-based limits is invalid as a general method of interval estimation [Halperin, 1977; Gart, 1979]. Simulations have borne out the predictably poor performance of the method for large departures of the odds ratio from the null value [Brown, 1981; Gart, 1982], and Greenland [1984] has provided a counterexample with the SMR that refutes the general validity of the approach. Greenland [1984] states

> [T]he problem with test-based limits is not (as has been suggested) lack of variance stabilization in specific applications, but rather that the principle requires us to equate two different large-sample test statistics. Since these statistics are equivalent only in the neighborhood of the null hypothesis, the principle itself is fallacious. . . . Unfortunately, the size of the neighborhood for which the principle holds will vary from parameter to parameter.

Despite the theoretical drawbacks, test-based limits can be useful as a "quick-and-dirty" method of interval estimation. The method is known to perform well for odds ratio limits when the odds ratio is between 0.2 and 5.0, and it can also be an acceptable tool in other situations. A comparison of the various methods of confidence interval estimation is illustrated for some simple data in the next chapter.

Adjustment for Multiple Comparisons

Many statisticians have voiced concern about the interpretation of *P*-values or "significance" tests when multiple comparisons are made. The basis for concern rests on the following argument: Suppose a complex set of completely random numbers were evaluated for 1,000 associations. The premise is that there are no real associations in the data but that 1,000 different measures of association are examined. If "significance" testing is performed, at the 5 percent level of "significance" there would be about 50 "significant" associations in the data, all representing type I or alpha-errors, that is, "statistically significant" associations that occur only by chance.

The point is that chance guarantees a certain proportion of such associations, and when many associations are studied, many false positive associations are possible.

The traditional statistical approach to this "problem" has been to make the "significance" test more stringent, either by changing the criterion to a more stringent value, such as 1 percent instead of 5 percent, or by actually inflating the calculated *P*-values by some factor that depends on the number of comparisons made. Since epidemiologists, in their usually thorough evaluation of expensively obtained data, typically make multiple comparisons, they have frequently been admonished to be wary of the problem.

It is not clear, however, that the recommended solution is an improvement. In the first place, the above argument, like all hypothesis testing, starts from the premise that the explanation for all the so-called "significant" results is chance, a sort of grand null hypothesis. But why should we assume that chance is a likely explanation for the associations that are observed? Indeed, one might argue that it seldom is (some would say never is) the explanation for findings. If chance is not the explanation for a "significantly" positive association, then the finding does not represent a type I or alpha-error. By making the screening criterion for statistical "significance" more stringent, a penalty is paid: Real non-null associations may go undetected (a type II error) because they fail to meet the more stringent criterion. An elementary consideration of screening principles, which apply here, makes it clear that from a *single criterion* (the "significance" level) the number of false positives can be reduced only at the expense of an increased frequency of false negatives. Is it worthwhile to reduce false positives at the expense of false negatives? The question cannot be answered generally; it requires a deeper understanding of the consequences of false positive and false negative results in the context of the research setting. One thing, however, is extremely clear: Whatever the arguments might be for reducing the chance of a false positive in favor of a false negative, they have nothing to do with multiple comparisons; they would apply equally well to a single comparison.

The crux of the multiple comparison problem seems to be that in performing many comparisons and reporting only those that are "statistically significant," it is difficult to impute the intended interpretation to the *P*-value; in the null hypothesis, a well-defined proportion of tests would be "significant," but if the denominator, the number of comparisons, is large and unknown, a reasonable interpretation of the *P*-values reported is hindered.

If many comparisons were made and each one were reported individually, let us say in a separate publication, it would be absurd to make adjustments to the reported *P*-values in each report based on the total number of such reports. If such adjustments were indicated, it would also follow that an investigator should keep a cumulative total of comparisons made during a career, and adjust all "significance" tests according to the current total of comparisons made to date. The more senior the investigator, the more the *P*-value would have to be inflated. For that matter, wouldn't such adjustments have to take into account the anticipated number of future comparisons as well as those already made? It should be obvious that these concerns are irrelevant to the research problem; they convert the *P*-value from a statistic conveying information about a specific association in the data to one that depends on the unrelated experiences or psychologic state of the investigator. No one has yet suggested making adjustments for multiple comparisons if the results are reported individually in separate publications. But is it not inconsistent then to consider making such adjustments if the same results are aggregated into one or several publications? Would a review paper of individually reported associations have to adjust the *P*-values? If no adjustments should be made to *P*-values when they are reported individually in separate publications, it follows that the process of lumping the results together in one place should not affect the results themselves, regardless of when and how the lumping is done. Therefore, no adjustments for multiple comparisons should be made even if a large number of comparisons are reported at one time, provided that it is clear how many comparisons have been made and that all "negative" (that is, "nonsignificant") results have been reported along with the "positive" or "significant" results.

A problem does exist when the negative results are not reported; it is then more difficult to interpret properly the *P*-values for the positive findings that are reported. It is still a mistake, however, to believe that interpretation can be improved by adjusting the *P*-value or changing the criterion for "significance." The adjusted values are also impossible to interpret, since they divulge even less about the actual association; changing the criterion for "significance" does not actually solve the problem; as discussed earlier, it merely produces a smaller type I error at the expense of a greater type II error.

As usual, some clarity is gained by considering the use of confidence intervals rather than "significance" tests. The equivalent of multiplying the *P*-value by some adjustment factor to compensate for multiple comparisons would be broadening the confidence interval. But the broader interval has no relation to the amount of information in the data about the effect in question; it depends instead on the number of comparisons that the investigator might have made. The problem with this approach is that it seems to defy the logical presumption that the reported results about an effect should reflect the amount of information about the effect in the data, nothing more and nothing less. If broader confidence intervals were reported to compensate for multiple comparisons, a reader with an interest focused solely on the one item would pay an unnecessary penalty in terms of the information imparted by the reported findings simply because the original investigator did not also focus solely on that problem.

Since no problem calling for any adjustments seems to exist unless the positive results from a large number of comparisons are reported without any information about the total number of comparisons, and since even then it appears that adjustments in the results only make them more difficult to interpret, the best course for the epidemiologist to take when making multiple comparisons is to ignore advice to make such adjustments in reported results. Each finding should be reported as if it alone were the sole focus of a study. If a large number of comparisons makes it infeasible to report all findings, it is important to make it clear how many associations were evaluated. If it cannot be determined how many comparisons were made, then associations not previously reported should be considered merely suggestive. It is worth emphasizing, however, that any new findings should always be considered only suggestive, even if only one comparison is made. Findings that address a previously reported association or lack of association should not become a weaker confirmation or refutation simply because they are accompanied by many other unrelated comparisons, since the previously reported findings on the question amount to a prior hypothesis.

REFERENCES

Brown, C. C. The validity of approximate methods for interval estimation of the odds ratio. *Am. J. Epidemiol.* 1981;113:474–480.

Cochran, W. G. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:295–313.

Cornfield, J. A statistical problem arising from retrospective studies. In J. Neyman (ed.) *Proceedings Third Berkeley Symposium,* Vol. 4. Berkeley: University of California Press, 1956, pp. 135–148.

Gart, J. J. Statistical analyses of the relative risk. *Environ. Health Perspect.* 1979;32:157–167.

Gart, J. J., and Thomas, D. G. The performance of three approximate confidence limit methods for the odds ratio. *Am. J. Epidemiol.* 1982;115:453–470.

Greenland, S. A counterexample to the test-based principle of setting confidence limits. *Am. J. Epidemiol.* 1984;120:4–7.

Halperin, M. Re: "Estimability and estimation in case-control studies." Letter to the Editor. *Am. J. Epidemiol.* 1977;105:496–498.

Lancaster, H. O. The combination of probabilities arising from data in discrete distributions. *Biometrika* 1949;36:370–382.

Lancaster, H. O. Significance tests in discrete distributions. *J. Am. Stat. Assoc.* 1961;56:223–234.

MacLaughlin, D. S. A data validation program nucleus. *Comput. Prog. Biomed.* 1980;11:43–47.

Miettinen, O. S. Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* 1976a;103:226–235.

Miettinen, O. S. Stratification by a multivariate confounder score. *Am. J. Epidemiol.* 1976b;104:609–620.

Miettinen, O. S., and Nurminen, M. Comparative analysis of two rates. *Statistics Med.* 1985;4:213–226.

Rothman, K. J., and Keller, A. Z. The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. *J. Chron. Dis.* 1972;25:711–716.

Yates, F. Contigency tables involving small numbers and the chi-square test. *J. R. Statist. Soc.* Suppl. 1934;1:217–235.