A STATISTICAL PROBLEM ARISING FROM RETROSPECTIVE STUDIES

JEROME CORNFIELD
OFFICE OF BIOMETRY, NATIONAL INSTITUTES OF HEALTH

1. Introduction

A recurring problem in medical statistics is the determination of the relative risks of developing a disease, say lung cancer, among two or more subclasses of a population, say, smokers and nonsmokers. Ordinarily, the risk for any subclass is estimated as the ratio of the number of cases of the disease developing in that subclass to the total number of persons in it, while an estimate of the risk for one subclass relative to another is provided by the ratio of the estimated absolute risks. Studies which start in this fashion with populations classified into subgroups, for each of which one counts the number of new cases of a disease which develop during some subsequent period of time are ordinarily referred to as "forward-looking" or "prospective" studies.

One may also be concerned with other types of relative risk, for example, the relative risk of dying from a disease or of having a disease. These different relative risks need not be the same for any one disease, and in cases where they are not it is customary to attempt to estimate all three. The relative risk of developing a disease is usually referred to as the relative incidence, the relative risk of dying from it as the relative mortality, and the relative risk of having it during some specified interval of time as the relative prevalence. In diseases, such as lung cancer, where the outcome is usually fatal and the interval between detection and death is relatively constant, the difference between these three different measures of relative risk will be small. In such cases it is common to choose that relative risk which can be estimated most easily. Thus, in prospective studies of lung cancer an estimate of relative mortality is usually preferred to one of relative incidence or prevalence because (a) the death registration system provides a complete enumeration which is lacking for newly developed or for existing cases and (b) diagnosis of cause of death is usually more accurate.

The risk of developing, having, or dying from any one disease in any one year is small. For this reason prospective studies designed to supply estimates of any one of the three relative risks must cover large numbers of persons, usually kept under observation for several years. An alternative method of gathering data, which avoids the necessity of observing large numbers of persons without the disease, but which, as usually done, supplies only estimates of relative prevalence is now commonly referred to as a retrospective study. In such a study one starts with a population (or a sample of it) classified into groups having and not having the disease, and determines for each group the proportion belonging to some subclass. Thus, one might classify a population into those having and not having lung cancer and