

PROBLEMS IN THE SELECTION OF A UNIVERSE FOR THE STUDY OF CHRONIC ILLNESS

FELIX E. MOORE

MANY of the statisticians present will at some time have been confronted with the following situation. The Chief calls you in and says, "Dr. Doe was telling me the other day that he has records on 1,000 cases of Roe's disease that he has seen over the past ten years. There's a gold mine of information there and I told him that I thought he ought to get some help to work it up statistically. I want you to drop around to see him one of these days." This not uncommon approach to medical research is what I like to call "A Universe in Search of a Problem."

There is an alternative approach—to start with a problem and search for a universe. I am sure that this group would agree that the specification of a problem for study must logically be antecedent to the specification of a universe. Indeed, if the problem is properly specified, the ideal universe will be implied. As a corollary, it might also be pointed out that there is no single universe for the study of chronic illness; there are as many universes (in the statistical sense) as there are problems.

Suppose we take, as an illustration of these points, the problem of prognosis after a coronary attack. Certainly it would be of considerable value to the physician at the present time to be able to state with a fair degree of assurance what is the proportion of persons who will, having had a first coronary attack, be alive after any specified period of time has elapsed. The first thing that is immediately obvious is that, if we want to study this problem, we must sample from a special universe, namely the universe of all persons who have had a coronary attack. Secondly, it is clear that this universe is unmanageable from the research standpoint. We must agree to narrow it with respect to both time and area. We might, therefore, re-state the prob-

lem as applying to white males in the United States at the present time.

At this point we see the necessity of making other compromises which are not, strictly speaking, statistical. If it is not feasible to study experience in the United States as a whole—and it clearly would not be in this problem—we should probably narrow our range of investigation to a more restricted geographical area—a city, or a small medical service area. But this narrowing of the universe from which we sample will of necessity be accompanied by restrictions on the generality of application of our conclusions. And so we have a process of give and take as our aspirations are adjusted to our capabilities.

Step by step, then, we approach a problem something like this:

1. Define the problem, i.e., state the question for which answer is required.
2. Describe the ideal universe from which a sample should be drawn to provide the answer.
3. Find that universe most closely approximating the ideal from which it is in fact practicable to sample.
4. State the restrictions which must be made on the generality of the answer which the practicable sample will provide.
5. Decide whether, in the face of these restrictions, the study is worth doing.

It is clear that these steps involve not only statistical problems—the whole research team (*vide* Gordon's discussion) will have to participate in their solution, since each member will have something to contribute. In passing, it may be worth noting that the art of statistics as well as the science will be involved, notably at steps 3 and 4, and this is perhaps the reason that there are statisticians specially labelled as "health" or "medical" statisticians.

I would like now to complete this discussion with a few illustrations, all well known to all of you, of the ways in which the problem determines our universe of study and the ways in which we cut our universe down to manageable size.

Expectation of Life. One of the elementary questions that a man can ask is how long he can expect to live. We don't, as statisticians, try to answer that for individuals, but we do undertake to answer it for classes of individuals. We accept, then, for research the question, "How long, on the average, will a class of persons of given age, sex, and race live?" The ideal universe in this case would be the completed lives of all persons in the given class now alive, but it is obvious that we can't wait until the last person dies to give an answer. In practice we go to the life table for our answer, which is to say that we accept as our universe a single year of experience. This considerably restricts our generalization, and we must now qualify our statement as to expectation by saying that it holds if mortality rates do not change. We have only an approximate answer to the original question, but we often find it useful.

Prevalence of a Disease. The question here is, "What proportion of the population is ill of a given disease in a specified interval of time?" The ideal universe is the total population of an area observed over the time interval. The practicable universe for study in this case will depend in large measure on the type of disease and the amount of information we require about it. If we are interested primarily in disease known to the individual, we may be willing to depend on the household survey using lay interviewers. In theory, we may not have to restrict our universe except to the extent that a certain group of non-cooperators will not be interviewed. In practice, we have apparently found it economically unfeasible since 1937 to survey anything like a sample of the United States, and so we must of necessity depend upon studies of more restricted areas such as the Baltimore Eastern Health District, or Hunterdon County, or San Jose. If we are interested in the prevalence of all stages of a disease, whether known to the afflicted individual or not, there must be some form of medical screening, and this will necessarily restrict our practicable universe geographically, and make the problem of non-response more important.

Etiological Factors in Disease. Because such chronic diseases

as arthritis, cancer, and arteriosclerotic heart disease have an insidious onset and long duration, the study of their etiology requires long-term longitudinal studies. Certain characteristics of these longitudinal studies make it almost mandatory to depart in at least two respects from the ideal universe which would be the entire adult population. First, the diagnostic procedures are difficult and require highly specialized medical and technical personnel. This makes it necessary that a team be assembled to work in a fairly limited geographical area. Second, the longitudinal study requires continued cooperation by a group of respondents who are willing to be re-examined over a period of years if we are not to depend solely on cause of death as the criterion for our study. Since it is not possible to get a completely representative sample of the population of any area to make themselves available for repeated examinations, we may find that our practicable universe has to be restricted to the point where there is serious question as to whether the data we can secure from it will answer the questions we have asked. This is a problem which has faced the National Heart Institute in its decision to carry out a twenty-year study of etiological factors in heart disease on a group of approximately 5,200 respondents in the town of Framingham, Massachusetts. While we can never reach generalizations which will be applicable with known error to the entire population of the United States, or Massachusetts, or even Framingham (because of non-cooperators), we feel that much that is valuable can be learned from an intensive study of this restricted group.

Prognosis. The study of prognosis after a coronary attack was mentioned in an earlier section where the desirability of geographical localization was discussed in general terms. In point of fact, prognosis in most diseases is dependent on treatment to such an extent that meaningful statements can be made only on the basis of known treatment. It is often desirable, on that account, to restrict the universe in studies of prognosis to persons with a comparable treatment regime, thus restricting the generality of statements that can be made, but causing

those which can be made to be more meaningful in the situations where they apply.

We have thus seen that the universe, which we actually select to sample from for any given study, will in practice be determined by the problem at hand and the practical considerations which lead us to depart from the ideal universe. The decision as to the extent to which departure from the ideal is allowable will be based on the state of our knowledge at any given moment and the amount that new knowledge is worth. In a state of ignorance, we are willing to accept data with many qualifications. As our knowledge increases, we require more precision and thus increasingly closer approximations to the ideal universe implicit in the problem. In our own field we have so little exact data about most problems that relatively crude approximations are still in order.