The Planning of Observational Studies of Human Populations

By W. G. Cochran

Harvard University

[Read before the Royal Statistical Society on February 17th, 1965,
the President, Mr S. Paul Chambers, C.B., C.I.E., in the Chair]

1. Introduction

Since this introduction was written during a period of exposure to the hypnotic effects of election campaign oratory in two countries, I shall not apologize unduly if my title seems to promise more than the paper will attempt to deliver. A more accurate title would be "Comments on some aspects of the planning of certain types of observational studies of human populations". I have in mind studies with two common characteristics:

(i) The objective is to elucidate cause-and-effect relationships, or at least to investigate the relationships between one set of specified variables x_i and a second set y_i in a way that suggests or appraises hypotheses about causation.

(ii) It is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures. Some randomization may be employed, however, e.g. in selecting for measurement a random sample from a population that seems suitable for the enquiry at hand.

In recent years such studies have become increasingly common in medicine, public health, education, sociology and psychology. Examples are the studies of the relationship between smoking and health, studies of factors that affect the probability of injuries in motor accidents, studies of the differences in behaviour of school children under permissive and authoritarian régimes, and studies of the effects of new social programmes such as replacing slum housing by public housing.

My experience in this area comes from service on advisory groups that assist in the planning of individual studies and on refereeing teams whose function is to recommend whether the money requested for a proposed study shall be granted. Typical of the latter are the Study Sections of the National Institutes of Health in the United States. In handling applications for statistical studies, a Study Section usually consists of about eight medical specialists and two statisticians. To put it simply, the examination of a proposal boils down to a judgement on two issues. (1) If the investigator succeeds in answering the questions that he proposes to answer, will this be a worthwhile contribution to our knowledge of health and illness? (2) If the investigator does what he proposes to do, is he likely to answer the questions that he proposes to answer? Although the statisticians are presumably placed on the Study Sections in order to help primarily with the second question, the division of labour is informal. In time the statisticians pick up at least some superficial medical expertise by osmosis, and the medical specialists often bring out statistical points that the statisticians have overlooked.

Appraisal of a large number of proposals leaves two general impressions. There is a regular procession of what might be called elementary mistakes. The proposal provides no clue as to what the investigator is trying to find out, if anything. The proposed sample size is too small to offer a reasonable hope of success even under the most optimistic assumptions. The population chosen for the study, although convenient and accessible, is one in which the variables whose effects are to be disentangled and measured just do not vary much. The study is to be based on routine records, found on inspection to be incomplete, frequently illegible and to contain numerous measurements that are either gross errors or wild guesses. The study is liable to a large amount of initial non-response plus loss of subjects as the study proceeds, but the proposal reveals no awareness of this problem. No provision is made for a control or comparison group of subjects, although one is obviously required if any inferences are to be drawn. The two groups of subjects to be compared will obviously differ in some variables that are not under investigation, and the disturbing effects of these variables are likely to be greater than those of the variables that are under investigation. The objective is to discover whether differences in the behaviour of two groups can be explained by certain variables, yet the investigator proposes to match the groups with regard to these same variables. And so on.

Secondly, at the other extreme, there is a different class of proposal. The objective is important, the study will be difficult and costly, and the plan has been carefully thought out. The results will, however, be subject to several sources of bias, and neither the investigator nor the appraising committee can suggest a method of reducing these biases (except that in some cases a completely different type of study might be less vulnerable to bias). The appraisers' judgements about the seriousness of these biases vary widely. This type of proposal leaves the statistician frustrated, because he seems unable to help an obviously capable investigator and because he faces a troublesome decision on whether to recommend support in view of a distinct possibility of misleading results (though he may resolve this by adopting a more general principle that men who are good deserve support, if his terms of reference allow him this latitude).

Drawing on this experience I would like to discuss some of the problems in planning observational studies and some of the current strategies for overcoming them. This type of research, dealing with the acquisition of knowledge that may enable us to enjoy healthier and more harmonious lives, is potentially important, yet I have the impression that it has been somewhat neglected by the statistical profession. (In this remark I am echoing Dr Wold (1956), who stressed the same point when addressing this Society some years ago.) Books on this subject have been written mainly by groups of subject-matter specialists, with the statistician perhaps contributing a chapter on sampling or tests of significance. Good examples from sociology are the books by Jahoda, Deutsch and Cook (1951) and by Festinger and Katz (1953). Similarly, at least in the United States, the statistics departments in universities present courses on the design of experiments and of sample surveys, but instruction on the planning of observational studies, if given, is usually found in the subject-matter departments, often under the title "Research Methods".

It is natural and commendable that subject-matter experts are developing and teaching their own research strategies. One reason is that effective planning of observational studies calls for considerable mastery of the state of research in the subject-matter field, for instance in forming a judgement as to which potential sources of bias are major and must somehow be controlled and which are minor and can be ignored, in knowing whether a proposed method of measurement has been tested for reliability and validity; in appraising which theories of behaviour are consistent with the results of a completed study, or more generally in deciding which type of study it is most fruitful to attempt next. Nevertheless, statisticians have much to contribute —particularly their training and experience in the conditions needed to provide sound inferences and their ability to develop efficient techniques for the analysis of the untidy data that are hard to avoid in observational studies.

Section 2 presents a brief account of the major difficulties in the planning of observational studies. Since detailed discussion of all aspects of planning would be lengthy, later sections are confined to three of these problems—the setting up of comparisons that are to throw light on the causal hypothesis, the handling of disturbing variables, and the step from association to causation—that seem to me to differentiate observational research most clearly from controlled experimentation. In discussing these topics it is relevant to indicate how the problem is tackled in controlled experimentation, because to a large extent, workers in observational research have tried to copy devices that have proved effective in controlled experiments. For instance, Dorn (1953) recommended that the planner of an observational study always ask himself the question, "How would the study be conducted if it were possible to do it by controlled experimentation?".

2. Major Difficulties

In the discussion of Dr Wold's (1956) paper on observational data, Dr Barnard was quoted as having said that a paper of this kind is useful in showing the younger statisticians what difficulties they may be up against. In writing this paper I have been conscious of the danger that a random member of the audience, if asked later for a concise summary of the paper, may quite properly report: "He said that it's all very difficult". A listing of common difficulties is, however, helpful in giving an overall view of the problems that must be overcome if this type of research is to be informative. In multi-variable experimentation he can apply combinations of levels, experience would doubtless construct a different list or at least change the emphasis given to the items on my list.

2.1. *Setting Up the Comparisons*

In controlled experimentation the investigator decides on the procedures or treatments whose effects he wishes to compare, and takes steps to apply them. The ability to do this gives him great flexibility and power. With a quantitative variable he can choose the number of levels and the intervals between them that will be most informative. In multi-variable experimentation he can apply combinations of levels, as in factorial design, selected so as to disentangle the effects of the different variables or to map a response surface effectively.

In observational studies the investigator, having decided on the types of comparison that he would like to make, often has to search for some environment in which it may be possible to collect data that provide such comparisons. As examples, this search would probably be the first step if it is desired to study the effects of air pollution on the health of urban dwellers, the type of protection afforded by seat belts under actual accident conditions, differences between the social outlooks of girls who attended co-educational schools and those who attended girls' schools, or the relative effectiveness of surgery and radiation for the treatment of malignant conditions in which ethical considerations forbid randomized experimentation. Often the investigator makes do with comparisons that are far from ideal for his purpose, and sometimes he postpones the study, hoping that later a more suitable environment will be found.

2.2. *The Handling of Disturbing Variables*

There is the familiar problem that the response or dependent measurements are usually influenced by many variables other than those under investigation. In controlled experimentation the investigator has three types of weapon at his disposal for handling such disturbing variables: (1) the experiments may be carried out under specialized conditions, e.g. on small plots in agriculture, or in laboratories with temperature and humidity control and highly precise instrumentation, in which some of the principal disturbing variables are absent or have greatly reduced effects; (2) blocking or adjustments made in the analysis can remove the disturbing effects of known major variables; (3) randomization and replication can diminish to a tolerable level the average effects of the remaining disturbing variables, including some whose presence is unknown to the investigator.

In an observational study the research worker can attempt to use the first device by looking for an environment in which some of the most important disturbing variables happen to be absent. This freedom of choice, however, is likely to be limited by the requirement that the environment shall also provide the types of comparison that he wants. If the subjects have been carefully selected so that they are similar as regards the principal disturbing variables, this process may have made them also similar on the variables whose effects we wish to study. Blocking (or matching) and adjustments in the analysis are frequently used. There is, however, an important difference between the demands made on blocking or adjustments in controlled experiments and in observational studies. In controlled experiments, the skilful use of randomization protects against most types of bias arising from disturbing variables. Consequently, the function of blocking or adjustment is to increase precision rather than to guard against bias. In observational studies, in which no random assignment of subjects to comparison groups is possible, blocking and adjustment take on the additional role of protecting against bias. Indeed, this is often their primary role. The extent to which they are capable of doing this will be discussed in Section 3.

2.3. *The Step from Association to Causation*

As mentioned, this paper deals with studies whose aim is to elucidate cause and effect relationships. I hope that I shall not be asked to explain exactly what is meant by cause and effect, since writers on the philosophy of science seem unanimously to discard this concept sooner or later as more confusing than helpful in complex situations. But to illustrate situations in which the concept is clear enough, the ultimate goal in applied studies may be to be able to predict the consequences of a new social programme, or of an experience that individual subjects may undergo, or of changes in the subject's living habits. Even in theoretical studies designed mainly to increase our understanding of people's behaviour, the idea of cause and effect is useful in the simpler situations.

In controlled experimentation the investigator who wishes to learn the effects of some procedure can usually go ahead and apply it, if necessary under a variety of other conditions, obtaining a direct answer to the question. A similar approach is sometimes possible in observational research. If a governmental agency has a programme of building new low-cost public housing or a new type of living accommodation for old people in several towns, this may provide the opportunity for studying the effects of the new living conditions on the people who enter them. Indeed, sociologists and economists are sometimes scolded for not being more enterprising in making plans in advance for the direct study of the effects of new public programmes, the co-operative efforts of astronomers and geophysicists in this respect being held out as an example.

For the most part, however, an observational study is a study of the associations between two sets of variables. Attempts to interpret these associations as causal or non-causal must rely heavily on information not supplied by the study, though some information may come from previous studies of a different type. To cite a simple example, suppose that an economist is interested in the question: if families of a certain size and with a certain income received an increase in income, how much of this increase would be spent on food? The data likely to be available or readily collectible relative to this question are a cross-sectional study of the amounts currently spent on food by families with different incomes. The increase in food expenditure per unit increase in income, as computed from this study, may or may not predict the increase that the economist wishes to estimate. In speculating on whether to trust this estimate or how to revise it, he would doubtless use any previous studies or reasonable theories about family spending habits that appeared relevant.

2.4. *Inferences from Sample to Population*

In most studies of human beings, the population to which we would like inferences to apply is real, not hypothetical. It is often extensive; the investigator hopes that his conclusions are valid for all males in the country with certain specified characteristics (e.g. of age or marital status). But the population actually sampled is frequently different. It is often narrow in scope, either for financial reasons or because the environment is particularly opportune for providing the appropriate comparisons. Random sampling may not be feasible. Sizeable amounts of non-response may occur. The subjects may be essentially volunteers if the measuring process is troublesome.

Standard statistical methods supply inferences from the sample data to a population of which the data can be regarded as a random sample. This sampled population is often hypothetical and sometimes hard to describe. Judgement as to how far the inferences apply to the target population involves trying to describe the relevant ways in which the sampled and target populations differ, and using any information that gives a clue as to the manner in which these differences will change the inferences. In research in which all sampled populations have to be specialized, a useful safeguard is a series of studies on sampled populations that have different peculiarities. For instance, in the studies comparing the death rates of smokers and non-smokers, the sampled populations were chosen in part because of social forces that facilitated getting good co-operation and accurate data. Fortunately, there are seven large studies, from three countries, all having broad sampled populations. The degree of agreement between studies in the relative death rates of cigarette, cigar, pipe and non-smokers and in the causes of death that show the greatest elevations in the death rates of smokers is impressive. On the other hand, the interpretation of these results is impeded by the fact that five of the studies had sizeable amounts of non-response, while in the remaining two studies no meaningful non-response rates can be calculated. Further, little was done, possibly because the studies were already complex enough in execution, to try to measure the influence of some of the disturbing variables that come to mind.

The deliberate use of sampled populations differing from the target populations is likely to remain a standard practice. This practice is followed also in controlled experimentation, both in agriculture and industry, in which much of the early screening or developmental research is conducted on a small scale that gives precise comparisons and saves money. The most promising results are checked by experiments that more closely approximate the conditions of application. The role of this approach in observation studies has been discussed, but opinion is divided as to the extent to which investigators should attempt to work in the target population itself, despite the extra expense and complexities of execution. Of course, in astronomy and meteorology, and to some extent in clinical medicine, conclusions that seem universal in scope are obtained from highly restricted observational studies. Perhaps in time an increasing body of simple fundamental laws of human behaviour will be uncovered, but in this area there is much to learn about the extent to which, as claimed in the old cockney song "it's the sime the 'ole world over".

2.5. *Measurement*

Much of the research on human behaviour and adjustment to life faces formidable problems of measurement. The investigator may want to study concepts like "feelings of ability to cope", "degree of frustration" or "strength of maternal affection". Different investigators develop different measuring instruments (often a series of questions) but it may not be known to what extent they are measuring the same thing, so that the combination or comparison of findings from different studies is rendered uncertain. The problem of varying definitions and measuring instruments is also familiar in clinical medicine. Psychologists and sociologists rightly devote much attention to clarifying the idea of measurement, to sophisticated analyses of the types of bias that may enter with human observers and human reporters and to the study of errors of measurement. In some areas of research, little progress seems in prospect until a substantial improvement in measuring technique is discovered.

The problem of measurement errors also affects the handling of disturbing variables. In many studies it is considered essential that the groups being compared shall be equated or adjusted for differences in socio-economic status. Numerous measures of socio-economic status have been developed, but it is hard to be sure that we have adjusted for the really relevant variable; further, errors of measurement decrease the effectiveness of the adjustment.

2.6. *Multiple Variables*

Multiplicity of variables is common, either in the response variables, the potentially causal variables or the important disturbing variables. An example that is far from extreme is the study by Neel and Schull (1956) of the effects of the parents' exposure to atomic radiation in Hiroshima and Nagasaki on the subsequent children. The measure of amount of exposure was a single 4-class variable, except that it was necessary to rate fathers and mothers separately. The indicators of radiation effects on children were the frequencies of stillbirths, neo-natal deaths and gross malformations, the sex ratio, and four measurements of the bodily development of the children. Disturbing variables that were judged important were maternal age, parity and degree of consanguinity in the marriage. Other disturbing variables were carefully considered, although for various reasons no adjustments for them were made. Much larger lists of variables may be present. In studies designed to measure the effects of something on family patterns of life, the response variables may include a health questionnaire, measures of the social activities of the members and of the relations between parents and children, and attempts to assess the degree of satisfaction that the family members derive from their way of living.

The presence of multiple variables raises a number of issues. Despite the advances in multi-variate analysis and in computing aids, summarization of a complex set of tables is still largely an art. Some investigators are appalled to realize how many tables they have to digest. Underestimation of the time and resources required for analysis of results is one of the most frequent features of proposals for research studies. The old maxim that the outlines of the analysis should be carefully sketched as a part of the research plan has lost none of its force. There is also, I believe, useful work to be done by statisticians in learning what some of the newer multivariate techniques really accomplish when applied to data and in explaining this to investigators, many of whom have no clear understanding of the techniques that they are trying to use. We need good expository papers of this type.

Participation in multi-variable studies leaves the impression that a series of lengthy questionnaires weakens the quality of the measurements. One is reminded of Bradford Hill's dictum (1953) that for every question asked of the respondent the investigator should ask himself three questions, one of which should always be: "is this question really necessary?" But when dealing with an imaginative investigator I do not find it easy to determine at what point one should adamantly oppose all further questions, however ingenious and interesting.

2.7. *Long-term Studies*

Some studies, e.g. of child growth and development, of chronic disease, or of social programmes whose effects are slow to appear, occupy the full-time energies of research teams for periods of 5, 10 or 15 years. Keeping track of the subjects and persuading them to be measured repeatedly requires much organizational skill, which may be only partially successful. Maintaining the interest of the research team, especially if there are no opportunities for publication for long periods, is another task. In short, this type of study produces a series of administrative and financial problems that are new to most research workers. It also naturally raises the question: is there any quicker way of obtaining useful results? For instance, Kodlin and Thompson (1958) give a useful analysis of the circumstances in which cross-sectional studies conducted at a single time will provide some of the results of long-term studies in growth.

As a final comment, many of these problems arise because investigators are beginning to study a series of new and probably complex phenomena, not because the investigator is restricted to observational methods.

3. THE HANDLING OF DISTURBING VARIABLES

Although it may seem to be putting the cart before the horse, some repetition is saved if the discussion of disturbing variables precedes that of the setting up of comparisons. The first step is to construct a list of known disturbing variables. Usually, these are arranged in three classes. (1) Major variables for which some kind of matching or adjustment is considered essential. Their number is kept small in view of the complexities involved in matching or adjusting for many variables simultaneously. (2) Variables for which, ideally, we would like to match or adjust, but content ourselves with some verification that their effects produce little or no bias. (3) Variables whose effects, thought to be minor, are disregarded. I shall consider the comparison of two groups, as occurs in the simplest type of observational study.

In the handling of disturbing variables there are two objectives. We want to protect against bias entering into the estimate of the difference between the two group means for the dependent variable y. Secondly, even if there seems no danger of bias, the presence of the disturbing variable may inflate the variance of $\bar{y}-\bar{y}'$ to an extent that makes the comparison imprecise.

3.1. *Variables for which no Matching or Adjustment is made*

For a disturbing variable in class (2), it is good practice to measure the variable and check that $\bar{y}-\bar{y}'$ is unlikely to be biased. If the disturbing variable is categorical (e.g. religious affiliation) a condition for absence of bias is that each class has the same frequency in the two populations of which the groups are random samples. The χ^2 test for a $2 \times k$ contingency table is the standard check.

If the disturbing variable x is continuous, let the relations between y and x in the two populations be

$$y = \mu_y + \xi + e; \quad y' = \mu_{y'} + \xi' + e'$$

where the residuals e and e' have zero population means, while $\xi = \phi(x)$ and $\xi' = \phi(x')$ is the regression of y on x, assumed the same in both populations. Hence,

$$\bar{y}-\bar{y}' = \mu_y - \mu_{y'} + \bar{\xi} - \bar{\xi}' + \bar{e} - \bar{e}'.$$

A necessary condition for the validity of the usual method of testing the significance of $\bar{y}-\bar{y}'$ or constructing confidence limits for $\mu_y - \mu_{y'}$ is that the population mean of $\bar{\xi}-\bar{\xi}'$ be zero, because the computed variance of $\bar{y}-\bar{y}'$ assumes that ξ acts like a random variable with zero mean.

Assurance that the distribution of x is the same in both populations guarantees that $E(\bar{\xi}-\bar{\xi}') = 0$ for any shape of regression function. A comparison of the frequency distributions of x in the two groups, usually made by the χ^2 test for a $2 \times k$ contingency table, is therefore relevant.

If $\phi(x)$ can be approximated by the polynomial

$$\xi = \phi(x) = \beta_0 + \beta_1 x + \beta_2(x^2) + \beta_3(x^3) + \ldots$$

the population mean of $\bar{\xi}-\bar{\xi}'$, i.e. the bias in $\bar{y}-\bar{y}'$ due to x, is

$$\beta_1(\mu_1 - \mu_1') + \beta_2(\mu_2 - \mu_2') + \beta_3(\mu_3 - \mu_3') + \ldots$$

where $\mu_i = E(x^i)$, $\mu_i' = E(x'^i)$ are the ith moments of x in the two populations about zero. Thus, verification that the sample means $\bar{x}$, $\bar{x}'$ do not differ by more than sampling error gives assurance only that bias arising from a *linear* regression of y on x is absent or small. A check that the two samples have the same means and variances in x (apart from sampling errors) is assurance against a quadratic regression, and so on.

The polynomial approximation is useful because sometimes the general shape of the regression of y on x is known from previous studies. Many relations are nearly linear or quadratic. Consequently, comparison of the means and variances of x may be more to the point than the χ^2 comparison of the whole frequency distributions. An extreme example is that in which the death rates of two groups of men are being compared and x is age. If y is a (0, 1) variable that denotes survival or death of a man during a year, the regression of y on x in the range 35–80 years is far from linear and is not a polynomial. However, a cubic in which the linear and quadratic terms dominate is often a fair approximation. Thus, unless at least the means and variances of x agree well in the two samples, there is a danger of substantial bias from differences in age. Of course, the relation between death rate and age is so marked at the upper ages that adjustment or matching for age is advisable.

With several x-variables, the common practice is to compare the marginal distributions in the two groups for each x-variable separately. The above argument makes it clear, however, that if the form of the regression of y on the x's is unknown, identity of the whole multi-variate distribution is required for freedom from bias. Similarly, the polynomial approach indicates that cross-product moments may be involved as well as univariate moments. In view of these extra complexities, it would be useful to know whether a check confined to marginal distributions is in practice likely to give a misleading impression.

Although these checks on the x distribution are usually made by tests of significance, it is not clear what kind of assurance is given by the finding of a non-significant result, nor that a test is the appropriate criterion. An alternative approach will be illustrated for the case in which the regression of y on x is linear, with a residual denoted by e. In repeated samples in which $(\bar{x}-\bar{x}')$ is fixed, $\bar{y}-\bar{y}'$ is normally distributed with mean

$$\mu-\mu'+\beta(\bar{x}-\bar{x}')$$

and variance $2\sigma_e^2/n$, this holding whether the mean value of $(\bar{x}-\bar{x}')$ is zero or not. If we assume that there is no bias due to x, we regard $\bar{y}-\bar{y}'$ as normally distributed with mean $\mu-\mu'$ and variance $2(\sigma_e^2+\beta^2\sigma_x^2)/n$. In large samples, 95 per cent confidence limits for $\mu-\mu'$ are therefore calculated by the formula

$$\bar{y}-\bar{y}' \pm 1{\cdot}96\sqrt{\left\{\frac{2}{n}(\sigma_e^2+\beta^2\sigma_x^2)\right\}}.$$

From the conditional distribution of $\bar{y}-\bar{y}'$ as given above, the probability that these limits include $\mu-\mu'$ is easily seen to be the probability that a normal deviate lies between the limits

$$-\sqrt{\left(\frac{n}{2}\right)}\frac{\beta(\bar{x}-\bar{x}')}{\sigma_e} \pm 1{\cdot}96\sqrt{\left\{1+\frac{\beta^2\sigma_x^2}{\sigma_e^2}\right\}}.$$

In the preliminary test of significance of $\bar{x}-\bar{x}'$, the test criterion is

$$t=\sqrt{(n/2)}\,(\bar{x}-\bar{x}')/\sigma_x.$$

Hence, the above limits may be written

$$-tv \pm 1{\cdot}96\sqrt{(1+v^2)} \qquad (3.1)$$

where $v=\beta\sigma_x/\sigma_e$.

Now if variations in x could somehow be removed, the unconditional variance of y would be σ_e^2. Thus the quantity v^2 represents the relative increase in the variance of y due to variations in x. This result is a reminder that even if there is no danger of bias from x, there is a loss of precision. Placing an x variate in class (2) instead of class (1) implies a judgement that this loss is small, say that $v^2<0{\cdot}2$, or v does not exceed 0·45.

Table 3.1 shows the conditional probabilities that the 95 per cent confidence limits actually include $\mu-\mu'$ for $v=0{\cdot}3$, 0·4, 0·5, 0·6 and 1·0 and $t=0{\cdot}5$, 1·0, 1·5 and 2·0, computed from (3.1) above.

Note that the value of t is known to the investigator from the preliminary check. If $t < 1$, the conditional probabilities of coverage are greater than the stipulated 95 per cent. If $t = 1{\cdot}5$, the probabilities are not too far below 95 per cent, provided that the initial guess that v is small was correct. For $t = 2$, its 5 per cent significance level, the coverage is unsatisfactory even if v is small.

TABLE 3.1

Probability that the 95 *per cent limits include* $\mu - \mu'$

t \ v	0·3	0·4	0·5	0·6	1·0
0·5	0·957	0·961	0·966	0·972	0·988
1·0	0·950	0·950	0·951	0·953	0·963
1·5	0·948	0·931	0·924	0·917	0·898
2·0	0·922	0·903	0·882	0·862	0·785

If a single x variate shows a value of t above 1·5, these results suggest that we have another look at this variate when the values of y become known. At that time we can estimate v and also compare the unadjusted $\bar{y} - \bar{y}'$ with an estimate adjusted for the linear regression, to see whether there is a material difference. If several x-variables show t values substantially above 1·5, this raises a question whether the groups are suitable for comparison.

Finally, even equality in the frequency distributions of x does not guarantee absence of bias if the regression function differs in the two populations. This point can be checked when the values of y become known.

3.2. *Matching and Adjustment*

For the major disturbing variables, four methods used in practice will be considered.

1. *Matching.* Pairs are drawn, one from each population, such that x_i and x_i' are identical within some small tolerance. This equates the frequency distributions of x in the two samples. The practical difficulties of matching vary with the situation. If the available populations are much larger than the desired size of sample and the distribution of x differs little in the two populations, matching gives little trouble. If the population reservoirs are limited and show markedly different x-distributions, or if several groups or several variables are to be matched, the process can be extremely tedious and may necessitate reducing the desired sample size. A much-quoted example is that of Chapin (1947), who compared the later economic adjustments of boys who completed high school with boys who dropped out. Starting with reservoirs of 671 and 523 boys, he ended with samples of size 23 after matching on six major disturbing variables.

2. *Equal sample sizes within sub-classes.* This procedure attempts to gain most of the advantages of matching with less expenditure of time. Each population is stratified into subclasses by the values of x. Within a given sub-class, samples of the same size are drawn from each population, but are not individually matched. To illustrate, suppose that only 100 subjects in group 1 are available, but group 2 has a larger

reservoir, and that there are five sub-classes. To see how things look, 100 subjects are also drawn from population 2. The numbers in each sub-class are found to be as follows.

	Sub-class					*Total*
	1	2	3	4	5	
Group 1	8	24	30	25	13	100
Group 2	20	34	28	15	3	100

Group 2 is to have the same sample size as group 1 in each sub-class. In sub-classes 3, 4 and 5, additional group 2 members will have to be drawn from the population to reach the goals of 30, 25 and 13. There may be difficulty in subclass 5 unless the group 2 reservoir numbers over 400.

In this approach, $\bar{y}-\bar{y}'$ is freed from bias due to differences between means of x in different sub-classes. Some bias may remain from variation of x within sub-classes.

3. *Adjustment for sub-class differences.* In this method the sample sizes are not equalized within sub-classes. However, the estimate of the mean difference is of the form $\Sigma W_j(\bar{y}_j-\bar{y}_j')$, where j refers to a sub-class. If the variance of y appears constant within sub-classes, W_j may be taken as $n_j n_j'/(n_j+n_j')$ by the usual least-squares principle. This method and the previous one have approximately the same properties as regards removal of bias, but differ in precision.

4. *Adjustment by regression.* This familiar method, used mostly when x is continuous, consists of computing the within-group regression of y on x, and adjusting $(\bar{y}-\bar{y}')$ to remove the effects of this regression. In practice, linear regressions are much the most common.

3.2.1. *x categorical or discrete*

The simplest case is that in which x is categorical or discrete and is known without error. Methods 1 and 2 become identical. They are free from bias, provided that the relation between y and x is the same in both populations. With samples of size n in each group, the variance of $\bar{y}-\bar{y}'$ is $2\sigma^2_{y.x}/n$, where $\sigma^2_{y.x}$ is the variance of y in arrays in which x is fixed. Method 3 (adjustment by sub-classification) gives the same protection against bias, but the estimate $\Sigma W_j(\bar{y}_j-\bar{y}_j')$ has a variance

$$\sigma^2_{y.x}/\Sigma W_j = \sigma^2_{y.x}/n\Sigma p_j p_j'/(p_j+p_j')$$

where p_j is the proportion of the group 1 sample falling in sub-class j. The relative precision of method 3 to methods 1 and 2 is therefore $2\Sigma p_j p_j'/(p_j+p_j')$. Calculation of this quantity helps in a choice between methods 2 and 3. In the numerical example given with the description of method 2, the initial sub-class sample sizes were 8, 24, 30, 25 and 13 in group 1, and 20, 34, 28, 15 and 3 in group 2. If method 3 is used instead of equating sample sizes within sub-classes, the relative precision works out at 0·92. It might not be considered worth while to go to the extra trouble of equating sample sizes.

If the values of $(\bar{y}_j-\bar{y}_j')$ differ from sub-class to sub-class or if the residual variance of y is not constant, a different estimate of the mean difference may be adopted, and the result for the relative precision of method 3 will be modified. Adjustment by regression is not possible when x is categorical. Comments on the properties of this technique when x is continuous (next section) apply also when x is discrete.

If x represents an ordered classification (e.g. none, mild, moderate, severe), the remarks in this section do not apply, because x cannot usually be assumed known without error. In such cases it is often more realistic to regard the classification as formed by subdividing the frequency distribution of an underlying continuous variable u into four distinct parts. The situation may indeed not be quite as clear-cut as this, since mistakes in classification may have occurred. As a first approximation I suggest that the results given in section 3.2.2. for x continuous be regarded as applying to ordered classifications also, method 2 corresponding to matching or use of equal sub-class numbers and method 3 to adjustment. This issue has also been discussed by Kihlberg and Narragon (1964).

3.2.2. *x continuous*

Individual matching removes any bias arising from x, provided that the relation between y and x is the same in both populations and that x is measured without appreciable error. Methods 2 and 3 equate the distributions of x in the two groups only partially, since these distributions may differ within sub-classes. Some residual bias therefore remains in $\bar{y}-\bar{y}'$ and $\Sigma W_j(\bar{y}_j-\bar{y}_j')$, and the variances of these estimates are increased relative to method 1, since the corresponding functions of x are not zero. Thorough examination of these methods requires an investigation covering different frequency distributions and types of initial bias. The following illustration indicates the general properties of the methods.

Suppose that x is normally distributed in both groups with $\sigma = 1$. In group 1, $\mu = 0$. In group 2, $\mu = -\frac{1}{2}$ and $\mu = -1$ are considered. With $\mu = -\frac{1}{2}$, the initial bias in x in group 2 might be called moderate: a test of $(\bar{x}-\bar{x}')$ in samples of size 100 gives about a 95 per cent chance of finding the difference significant. With $\mu = -1$ the bias is striking, being detectable in samples of size 25.

TABLE 3.2

Properties of methods 2 *and* 3

No. of sub-classes	$\mu = -\frac{1}{2}$			$\mu = -1$			*Remaining variance*
	Remaining bias		*R.P. of Method* 3	*Remaining bias*		*R.P. of Method* 3	
	Method 2	*Method* 3		*Method* 2	*Method* 3		
2	0·184	0·190	0·96	0·382	0·430	0·87	0·39
3	0·105	0·109	0·95	0·218	0·259	0·84	0·23
4	0·071	0·076	0·95	0·147	0·183	0·82	0·15
5	0·052	0·055	0·95	0·109	0·135	0·81	0·10

The sub-class boundaries were constructed so that the sub-classes have equal frequencies in population 1. For 2, 3, 4 and 5 subclasses, Table 3.2 gives the relevant results.

The remaining bias in x is $E(\bar{x}-\bar{x}')$ for method 2 and $E\Sigma W_j(\bar{x}_j-\bar{x}_j')$ for method 3. The original biases were $\frac{1}{2}$ and 1 for the two cases. Method 2 leaves about 36–38 per cent of this bias remaining if there are only two sub-classes and 10–11 per cent remaining with five sub-classes. With moderate bias ($\mu = -\frac{1}{2}$), method 3 has about the same effectiveness in removing bias, and gives only a slightly higher variance, as indicated by the relative precision figures (R.P.) computed as in section 3.2.1. Method 3 is noticeably less effective, when $\mu = -1$, both as regards bias and variance.

The extreme right column of Table 3.2 shows for method 2 the quantity

$$nV(\bar{x}-\bar{x}')/2.$$

Since this quantity would be unity if random samples had been drawn without any sub-classification, it might be described as the proportion λ of the original variance of x that remains. The corresponding values for method 3 are only slightly higher and are not shown. How much this variance increases the variance of $\bar{y}-\bar{y}'$ relative to that given by individual matching depends on the correlation between y and x. The comparable quantities for $V(\bar{y}-\bar{y}')$ are $(1-\rho^2+\lambda\rho^2)$ for methods 2 and 3 and $(1-\rho^2)$ for method 1. With five sub-classes ($\lambda = 0{\cdot}10$) the increase in variance with methods 2 and 3 does not reach 10 per cent until ρ exceeds 0·7. With two sub-classes the loss of precision may be substantial.

Evidently, sub-division into two or three classes has limited effectiveness both in controlling bias and reducing variance. Naturally, investigators prefer to use only a few sub-classes, especially when there are several x-variables. Methods 2 and 3 also reduce differences in the higher moments of the distribution of x, but illustrations will not be given.

With $\mu = -1$ and five sub-classes, the highest sub-class contains only 3·3 per cent of population 2 as against 20 per cent of population 1. Method 2 therefore requires a reservoir from population 2 that is around six times the sample from population 1. For individual matching a much larger reservoir would be needed.

Assuming the same linear regression in each population, adjustment by linear regression removes the bias. As regards precision, there is the well-known result

$$V\{\bar{y}-\bar{y}'-b(\bar{x}-\bar{x}')\} = \frac{2\sigma_{y.x}^2}{n}\left\{1+\frac{n(\bar{x}-\bar{x}')^2}{2\Sigma}\right\}$$

where Σ, with $2(n-1)$ d.f. is the pooled sum of squares for x. In repeated sampling, when x is $N(0,1)$ and x' is $N(\mu,1)$,

$$E(\bar{x}-\bar{x}')^2 = \mu^2+2/n; \quad E(1/\Sigma) = 1/2(n-2)$$

and the two are independent. Hence the average variance of the adjusted mean is

$$\bar{V} = \frac{2\sigma_{y.x}^2}{n}\left\{1+\frac{n\mu^2}{4(n-2)}+\frac{1}{2(n-2)}\right\} = \frac{2\sigma_{y.x}^2}{n}\left\{1+\frac{\mu^2}{4}\right\}$$

when n is large.

With a linear regression, the relative precision of covariance to individual matching is therefore $16/17 = 0{\cdot}94$ when $\mu = \frac{1}{2}$ and $4/5 = 0{\cdot}8$ when $\mu = 1$. Covariance is superior to method 3 as regards both bias and variance. It is superior to method 2 in the control of bias. Method 2, with at least five sub-classes, is likely to have a smaller variance unless μ is small.

If the true regression is quadratic but adjustment is made by a linear regression, the remaining bias in large samples works out as

$$\beta_2\left[(m_2-m_2')-\frac{(m_1-m_1')(m_3+m_3')}{m_2+m_2'}-\frac{(m_1-m_1')^2(m_2-m_2')}{m_2+m_2'}\right]$$

where $m_i = E(x-\mu_x)^i$ and $m_1 - m_1' = -\mu$ in the above notation. We see again how helpful is equality of the first two moments of the distributions of x and x', since in this event covariance is not needed to remove bias and can concentrate on increasing precision. If the low moments are unequal, methods 2 and 3 would be expected to be more potent than a linear regression in removing non-linear bias, but further investigation is required to appraise whether this superiority is material.

The regression approach has the advantage that separate regressions can be computed in the two populations and used in the adjustment, this being a situation in which bias remains even with individual matching. If the two regressions differed markedly, however, one would be inclined to reconsider whether the groups are suitable for comparison.

If x is subject to appreciable errors of measurement, none of the methods succeeds in complete removal of bias. Suppose that $X = x+d$ is the recorded value of x, where d is the error of measurement. If x and d are normally and independently distributed, the expected reduction in x due to a reduction of amount μ in X is $\mu\sigma_x^2/(\sigma_x^2+\sigma_d^2)$. Thus the fraction of the original bias that remains in X after individual matching is $\sigma_d^2/(\sigma_x^2+\sigma_d^2)$. Similarly, if β_1 is the linear regression of y on x, the regression of y on X is $\beta_1\sigma_x^2/(\sigma_x^2+\sigma_d^2)$, so that in large samples the fraction of bias remaining in y after regression adjustment is again $\sigma_d^2/(\sigma_x^2+\sigma_d^2)$. It is worth remembering that the basic quantity is this ratio. A value of σ_d^2 that looks large to someone accustomed to highly precise measurements might be small relative to the total variance of x.

With several x-variables the relative properties of the four methods appear to remain as in the above example. The difficulties of matching and of equating sample sizes mount steadily. Under the adjustment methods, the variance of the adjusted mean difference tends to increase and the analysis becomes more complex. With method 3, a point may be reached at which the investigator wonders whether it is worth adjusting for one or more extra x-variables, since the reduction in bias may not compensate for the loss of precision and extra complexity. A criterion for forming a judgement on this question from the sample data has been sketched by Cochran, Mosteller and Tukey (1954), though details need to be worked out.

To summarize, the similarities among the methods are greater than their differences. When feasible, matching is relatively effective. Overall, covariance seems superior to adjustment by sub-classification, though the superiority will seldom be substantial. If the original x-distributions diverge widely, none of the methods can be trusted to remove all, or nearly all, the bias. This discussion brings out the importance of finding comparison groups in which the initial differences among the distributions of the disturbing variables are small.

4. Setting Up the Comparisons

4.1. *The Choice between Different Types of Study*

As mentioned previously, the investigator often has to search for some environment in which a comparison relevant to the causal hypothesis can be made. Sometimes he faces a choice between different types of study. For instance, in the work on the relation between smoking and lung cancer, the crudest approaches were a comparison of the time trends in the lung cancer death rate and in the consumption of tobacco per head within a country, or an examination of the relation between these two figures in different countries at the same time. Then there were numerous studies in which the percentage of smokers among lung-cancer patients was compared with that among patients with other diseases or among the general public. In another approach the lung-cancer death rates of groups of smokers and non-smokers were recorded over a period of years. These comparison groups were obtained either by classifying the members of a large population into different smoking classes by an initial questionnaire or by finding a smaller homogeneous group whose members did not smoke, and constructing a comparison group of smokers. An attractive possibility would be to compare twins, preferably identical, of whom one smoked and the other did not, although it seems highly unlikely that enough pairs could be located.

In making a choice between different studies that he might undertake, the investigator should consider the resources and time needed and the status of each study with regard to the handling of disturbing variables and to the quality of the measurements. Other relevant factors are: (1) The quantities that can be estimated from the study. Sometimes one study yields only a correlation coefficient while another gives an estimate of the response curve of y to variations in x. (2) The range of variation of the suspected causal variable. In general, a study that furnishes a wider range of variation may be expected to give more precise estimates of the effect on y. (3) The relation to previous work. One study may be a new approach to the problem, another a repetition of studies done elsewhere. Both have their uses, but preference would normally be given to a new approach, especially if it seems free of some of the biases thought to be present in previous studies. Naturally, all these questions involve judgement. As research by observational methods becomes more widespread and familiar, we should be able to make better appraisals of the productivity of different approaches.

4.2. *Some Common Types of Comparison*

In this and the following sections some common types of comparison are presented from the viewpoint of their statistical structure. The simplest plan is a direct comparison of a few groups (often two) that differ in the hypothetically causal variable. Frequently, only one of the groups is clearly demarcated in advance, and the investigator must construct one or more control or comparison groups. For instance, radiologists, particularly the older ones, were formerly exposed to repeated small doses of radiation in their practices. Studies have been made to try to see whether this exposure produced an increase in their probability of dying. For this purpose they must be compared with some other non-exposed group. Ideally, a control, while lacking the suspected causal factor, should have the same distribution as the chosen study group with regard to all major disturbing variables. Sometimes the investigator is not sure whether a specific disturbing variable affects his study group. This situation may require more than one control. In discussing controls for hospitalized lung-cancer patients for a comparison of the proportion of smokers, Mantel and Haenszel (1959) point out that hospital patients in general are known to yield a higher proportion of smokers than members of the general public. One possible reason is that smoking histories collected in hospital are more accurate, those obtained from the general public being underestimates. If so, a hospital control is indicated. But if smokers have higher rates of hospitalization, the smoking data being equally accurate, the control should come from the general population. With uncertainties like this, use of both controls is advisable.

4.2.1. *The "before–after" study*

This plan is much used in investigating the effects of new social, economic or medical programmes. If the programme applies to everyone, there is no possibility of finding a control group that does not experience it. At a minimum, the y-variables and the principal disturbing x-variables are measured before and at appropriate times after the initiation of the programme. This enables us to investigate whether changes in the y-variables have occurred over and above those expected from any changes in the x-variables. This estimate of the effect of the programme is liable to two types of bias: people's behaviour immediately prior to the start of the programme may be affected by knowledge that the programme is about to begin, and some disturbing variables that affect time changes may be unknown. An estimate that would be free from the first bias is a comparison of the residual y changes during a period after the start of the programme and a period prior to the announcement of the programme, but it is not often feasible to obtain the necessary data.

When the programme (e.g. of new public housing) is available only to certain people, it becomes possible to measure the y- and x-variables, before and after, both for the group that undertakes the programme and for a control group that does not. The effect of the programme is estimated by the difference between (After–Before) for the programme participants and (After–Before) for the controls, the variable being the residual of y after adjustment for the disturbing variables. One advantage of this design is that we can verify, at the start of the study, whether the two groups are similar in their y-variables, instead of having to guess about this from measurements on the x-variables alone, as is the usual situation. Despite this, a fully satisfactory control may not come easily. If much initiative is required to get one's family on the eligible list for new housing, families that have shown no such initiative are a dubious control even though they state in a questionnaire that they would like to be in new housing. For certain programmes, e.g. an educational one to improve family health practices, a suggestion is sometimes made of a second control in which y is measured only afterwards, to guard against the possibility that the initial questionnaire alerts the control families to deficiencies in their health practices which they proceed to remedy. My own view is that an educational programme that cannot improve health practices more than can a single questionnaire is not wrongly considered a failure, and that this enlargement of the study is seldom justified.

4.2.2. *The* ex post facto *or retrospective study*

Sometimes an unexpected event has occurred, and the question is "what caused it?" An outbreak of nausea and vomiting follows a picnic meal. If all left-over food has been destroyed, eliminating laboratory analysis, a list of the foods eaten by those who became ill and those who did not, and a description of the symptoms and age and sex distributions of affected and unaffected persons and of the preparation of the dishes served are the main clues to the responsible organism and food. The same approach can be used to investigate why riots have occurred in certain communities while others, at first sight similar, have had no disturbances. The strategy is to set up groups that differ in the y-variable, and examine whether they differ in the suspected causal variables.

Because of its relative cheapness and high efficiency under favourable circumstances, this approach is often used in problems in which a direct approach is also feasible. In the smoking–lung cancer relation, numerous studies have been done by both approaches. Another frequent application is a comparison of the successes and failures in some occupation or task, in the hope of discovering the causes, or at least useful predictors, of success and failure.

In the simplest case in which both the cause and the event in question are dichotomous, the two approaches are different ways of sampling the following 2×2 table, where the N's are the numbers in the population.

Postulated cause	*Event*		
	Present	*Absent*	*Total*
Present	N_{11}	N_{12}	$N_{1.}$
Absent	N_{21}	. N_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	$N_{..}$

Under the direct approach, a sample is selected from each row and the proportions $N_{11}/N_{1.}$ and $N_{21}/N_{2.}$ are compared. This is expensive if the event is rare, since large samples are needed, or if the event takes years to manifest itself. In the retrospective approach a sample is drawn from each column. If rare, the "event present" column can be sampled at a much higher rate—frequently, all such cases that can be found in the population are taken. Both methods furnish a χ^2 test of the null hypothesis that there is no association. In large samples, if the two methods are to have equal power in detecting a small departure from the null hypothesis, it may be shown that $n_R/n_C = N_{1.}N_{2.}/N_{.1}N_{.2}$, where n_R, n_C are the sample sizes used in the direct and retrospective approaches, respectively. For example, if the event is present in only 1 per cent of the population but the postulated cause occurs in half the population, $n_R/n_C \simeq 25$. Moreover, in the retrospective method the data on the causal variables lie in the past and can be collected without waiting for the effect to develop.

In appraising the size of the effect of the postulated cause, we need to compare the quantities $N_{11}/N_{1.}$ and $N_{21}/N_{2.}$, i.e. the frequencies with which the event occurs when the cause is present and absent, respectively. These are the quantities that are estimated in the direct approach. They can also be estimated by the retrospective approach, provided that

(i) the samples from the columns are random samples and,

(ii) the relative sampling fractions in the two columns are known.

As the retrospective method has been applied in practice, neither condition is usually fulfilled. Cornfield (1951) pointed out that if the event is rare, the ratio of the frequencies, $N_{11}N_{2.}/N_{1.}N_{21}$, will be close to $N_{11}N_{22}/N_{12}N_{21}$. From a retrospective study a consistent estimate of this quantity, called the *relative risk* of the event, is given by the sample cross-product ratio $n_{11}n_{22}/n_{12}n_{21}$, provided that no bias has arisen from non-randomness in sampling. Confidence limits for the relative risk and methods for comparing and combining results from different studies have been given by Cornfield (1956), while Mantel and Haenszel (1959) present methods for estimating an overall relative risk in retrospective studies in which the data have been subclassified by another variable (e.g. age or location).

The retrospective approach has various weaknesses. The obtainable data about postulated causal factors may be of poor quality, especially if they lie in the distant past or have to be taken from routine records, and the sampling of the columns may be far from random. As more studies in different fields are done by the two methods, the ability of the retrospective approach to estimate relative risk can be more soundly judged. In the smoking–lung cancer relation, the two approaches agreed well on the whole.

4.2.3. *Multiple causal variables*

Typical examples are the studies to investigate the roles of measures of blood pressure, obesity and cholesterol levels in the individual as predictors of later heart disease (although these variables would not necessarily be viewed as causal, but perhaps as indicators of the presence of some deeper cause). For the most part, often because there seems no choice, investigators have taken the postulated causal variables as they come in the selected sample, with no deliberate attempt to borrow the idea of factorial design. The disadvantage in this approach is that if the variables are highly correlated it becomes difficult to disentangle their effects. Further, if one of two correlated variables has a high error of measurement while the other does not, the regression coefficient on the first variable is an underestimate and on the second an overestimate.

Exceptions can be cited. In his studies of sexual behaviour, Kinsey's (1948) independent variables (again not necessarily considered causal) were arranged in a multiple classification with something over 300 cells. His announced plan was to obtain a sample of size 300 in each cell, the planned sample size being 100,000. Since, however, he regarded his major problem as that of getting people to tell the truth, his standards for the selection and training of interviewers were exacting, so that his field force was very small. This, plus a haphazard method of sampling, made the realization fall far short of the goal.

The opportunity of using a factorial approach presents itself if there is a population reservoir of size N in which the causal variables have already been measured, and a much smaller sample of size n is to be selected for the actual study. This reservoir might come from some other investigation. Alternatively, if y is expensive to measure, it may be worth while to expend some of the resources on measuring the causal variables in a sample of size N, of which n will later be selected for measurement of y. Further work on this approach is required. Given the results of the large sample it is not obvious how best to select the sub-sample, with say three or four variables, nor how great a gain in precision over a random sample of size n can be expected. An example of this method was given by Keyfitz (1952) in studying the relation between fertility and five dichotomous demographic variables by sampling from records, although owing to the nature of the variables, he was not able to obtain complete orthogonality in his 2^5 factorial. Use of an initially larger sample in this way also makes it possible to obtain a better estimate of the response curve of y as the level of a causal variable change.

4.2.4. *Population laboratories*

One device that has been tried in a number of large research centres, primarily in public health and sociology, is to select some area, perhaps of 50,000–100,000 persons, that seems appropriate for the type of field research carried on. The background characteristics of the people are measured in an initial census. This is repeated at intervals. Supplementary questions may be added to these censuses to provide reservoirs of data, as in the previous section, for individual studies. The studies, which are mostly carried out on sub-samples, may have widely different foci of interest.

An early example is the Eastern Health District of Baltimore (Fales, 1951; Cochran, 1952), though in this case financial considerations limited the scope of the background data.

A population laboratory of this kind is expensive to maintain. Its advantages are that the background information facilitates the drawing of efficient subsamples and that these can be probability samples from a known population with a broad coverage of urban and rural conditions and of the different social classes.

5. The Step from Association to Causation

This issue is naturally of great concern to workers in observational research and has received much discussion in individual subject-matter fields. I shall confine myself to a few comments on statistical aspects of the problem.

First, as regards planning. About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate". The reply puzzled me at first, since by Occam's razor the advice usually given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many *different* consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold. If a hypothesis predicts that y will increase steadily as the causal variable z increases, a study with at least three levels of z gives a more comprehensive check than one with two levels. A secondary consequence of a hypothesis may be that the relation between y and z changes in a known direction as we move from low to high educational levels. If the study can be made large enough, a verification on this point can be included as well as a determination of the overall relation between y and z. The comparisons of the death rates of cigarette smokers and non-smokers are rich in opportunities for this kind of verification. In the largest studies, we can compare the death rate (i) of men who smoked different amounts for the same time, (ii) among smokers of the same amount, of men who had been smoking for different lengths of time, (iii) of ex-smokers and current smokers of the same amount, (iv) among ex-smokers, of those who had previously smoked different amounts, (v) among ex-smokers of the same amount, of those who had stopped recently and those who had stopped for longer periods. The causal hypothesis predicts the direction in which the results should lie for each of these comparisons.

Of course, the number and variety of the consequences depends on the nature of the causal hypothesis, but imaginative thinking will sometimes reveal consequences that were not at first realized, and this multi-phasic attack is one of the most potent weapons in observational studies. In particular, the task of deciding between alternative hypotheses is made easier, since they may agree in predicting some consequences but will differ in others.

Since the initial work on a problem is often done in a restricted population, repetition of the study plan in different environments by different workers has its value, especially in forming a judgement whether results from the sampled population can be extended to a broader target population. Since studies that follow essentially the same plan may be subject to the same biases, an approach with a different plan that escapes some of these biases is highly useful.

When summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him. This advice may sound trite, but in practice is often neglected. A model is the section "Validity of the results" by Doll and Hill (1952), in which they present and discuss six alternative explanations of their results in a study. Since this discussion may require use of data extraneous to the study and may be assisted by supplementary observations that can be taken in the study if the investigator realizes the need for them, it is well to anticipate these alternative explanations when the study is being planned.

Similarly, if the study gives an estimate of the size of the effect, possible biases in the estimated size should be discussed. For instance, the responses of crop yields to fertilizers can be estimated by a survey in which farmers report their yields and the amounts of fertilizers used. Clearly, the response will be over-estimated if the better farmers use more fertilizers, but under-estimated if fertilizers are applied primarily to poor soils. If controlled experimentation were impossible in agriculture, so that such surveys were the only means of information about the effects of agricultural practices, we could at least ask supplementary questions to permit adjustment for some known sources of bias and to aid our judgement in discussing remaining biases. As Yates and Boyd (1951) have pointed out, there is no assurance that this process will bring us close to an unbiased estimate. On date of planting of potatoes, surveys over three years gave an average reduction of 0·45 tons per acre for each week's delay, as against 0·5 tons per acre from a limited number of experiments—a good agreement, Boyd (1957). But in these surveys, the response of potatoes to farmyard manure (in the presence of complete fertilizers) averaged 0·1 tons per acre. Statistical adjustments for region, variety, class of seed, date of planting, and acreage grown raised this average to 0·4 tons per acre. Large numbers of controlled experiments give an average of 1·4 tons per acre—more than three times as great. The specialist potato grower has limited available manure but obtains high yields, while from the livestock farmer, who has ample manure, potatoes evidently receive less skilled attention. Although this example is a dampening reminder of the hazards of observational studies, thoughtful judgement about the direction and size of biases is better than no judgement.

In interpreting the results of regression studies on several variables, some postulated as causal and some as disturbing, I believe that more attempts should be made to bring in the investigator's ideas on the directions of causal paths. This type of analysis has been developed to considerable lengths in economics, with much clarification as well as some differences of opinion, and in genetics, under the leadership of Sewall Wright, but has been little tried elsewhere. For those unfamiliar with the technique, illustrations of its potentialities are the discussion by Tukey (1954) of the relations between birth weight, gestation period and litter size in guinea pigs, by Yates (1960) of the relation between health, total income, rent and housing quality, and by Wold (1956) in economics.

The combined evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis. If some results appear to support the hypothesis, some to contradict it and some are neutral, reaching a verdict demands much skill. Obviously, the investigator should consider whether some revision of his hypothesis will remove the contradictions. In default of this, he cannot avoid an attempt to weigh the evidence for and against, since some results are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance. He should state such judgements forthrightly, remembering his duty to maintain even standards and, if possible, an air of calm detachment. An example that can be recommended is the competent discussion by Cornfield *et al.* (1959) of some of the controversial results in the smoking–lung cancer problem. Even here, some readers may detect a slight shifting of standards that hints at a departure from complete objectivity, though I mention this with trepidation, since the paper has six authors and I am alone. The situation is, of course, more comfortable for the pure scientist, who can always reserve final judgement while stating that there is a strong *prima facie* case, than for the applied scientist who must decide at what point a call for action should be made.

I have written as if the observational statistical studies are the only evidence. Fortunately, there will often be a corpus of laboratory-type research, perhaps using controlled experiments, that endeavours to probe more deeply into the nature of the causal mechanism. In many areas, such research is the primary hope of reaching a full understanding.

In conclusion, much of this paper was written while I enjoyed the hospitality of the Statistics Department, Rothamsted Experimental Station.

REFERENCES

BOYD, D. (1957). "A scrutiny of the British potato crop", *Oper. Res. Quart.*, **8**, 6–21.

CHAPIN, F. S. (1947), *Experimental Designs in Sociological Research.* New York: Harper.

COCHRAN, W. G. (1952), "An appraisal of the repeated population censuses in the Eastern Health District, Baltimore". In *Research in Public Health*, pp. 255–265. New York: Milbank Memorial Fund.

COCHRAN, W. G., MOSTELLER, F. and TUKEY, J. W. (1954), *Statistical Problems of the Kinsey Report*, pp. 246–253. Washington, D.C.: American Statistical Association.

CORNFIELD, J. (1951), "A method of estimating comparative rates from clinical data, applications to cancer of the lung, breast and cervix", *J. Nat. Cancer Inst.*, **11**, 1269–1275.

—— (1956), "A statistical problem arising from retrospective studies", *Proc. Third Berkeley Symp.*, **4**, 135–148.

—— HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A., SHIMKIN, M. B., and WYNDER, E. L. (1959), "Smoking and lung cancer: recent evidence and a discussion of some questions", *J. Nat. Cancer Inst.*, **22**, 173–203.

DOLL, R. and HILL, A. BRADFORD (1952), "A study of the aetiology of carcinoma of the lung", *Brit. Med. J.*, **2**, 1271–1286.

DORN, H. F. (1953), "Philosophy of inferences from retrospective studies", *Amer. J. Public Health*, **43**, 677–683.

FALES, W. T. (1951), "Matched population records in the Eastern Health District, Baltimore, Md.", *Amer. J. Public Health*, **41**, 91.

FESTINGER, L. and KATZ, D. (Eds.) (1953), *Research Methods in the Behavioral Sciences.* New York: Holt, Rinehart & Winston.

HILL, A. BRADFORD (1953), "Observation and experiment", *New England J. Med.*, **248**, 995–1001.

JAHODA, M., DEUTSCH, M. and COOK, S. W. (1951), *Research Methods in Social Relations.* New York: Dryden.

KEYFITZ, N. (1952), "Differential fertility in Ontario. An application of factorial design to a demographic problem", *Population Studies*, **6**, 123–134.

KIHLBERG, J. K. and NARRAGON, E. A. (1964). "A failure of the accident severity classification", *Cornell Aeronautical Lab. Report.* VJ–1823–R8, 62–70.

KINSEY, A. C., POMEROY, W. B. and MARTIN, C. E. (1948), *Sexual Behavior in the Human Male.* Philadelphia: Saunders.

KODLIN, D. and THOMPSON, D. J. (1958), "An appraisal of the longitudinal approach to studies in growth and development". *Monographs of the Society for Research in Child Development*, No. 67. Lafayette, Indiana.

MANTEL, N. and HAENSZEL, W. (1959), "Statistical aspects of the analysis of data from retrospective studies of disease", *J. Nat. Cancer Inst.*, **22**, 719–748.

NEEL, J. V. and SCHULL, W. J. (1956), *The effect of exposure to the atomic bombs on pregnancy termination in Hiroshima and Nagasaki.* Washington, D.C: Atomic Bomb Casualty Commission.

TUKEY, J. W. (1954), "Causation, regression and path analysis", Ch. 3 in *Statistics and Mathematics in Biology*. Ames: Iowa State College Press.

WOLD, H. (1956), "Causal inference from observational data", *J. R. statist. Soc.* A, **119**, 28–61.

YATES, F. (1960), *Sampling Methods for Censuses and Surveys*, 3rd ed. London: Griffin.

—— and BOYD, D. (1951), "The survey of fertilizer practice: an example of operational research in agriculture", *Brit. Agric. Bull.*, **4**, 206–209.

DISCUSSION ON PROFESSOR COCHRAN'S PAPER

Sir AUSTIN BRADFORD HILL: I suspect that in our approach to observational studies of the human population, there is only one material difference between Professor Cochran and myself. He, as he points out in an early paragraph of his paper, has (in this situation) largely served as a referee or, at the very least, as a linesman. Over the last 40 years I have had to rush feverishly around the field of play, and in this particular field, unfortunately, most of the missiles are aimed at the players; indeed it is not unknown for the referee to join in. The only comfort I can get is that I suspect that some of those applying for research grants and ruled "offside" by Professor Cochran have probably not quoted the first line of that cockney ditty that he quotes, "it's the sime the 'ole world over". They are more likely to have used the last line!

I agree with him and the random member of the audience whom he quotes that it is all very difficult. I would, however, emphasize—and this arises from Professor Cochran's discussion of the experimental approach—that in this field the difficulties of experiment are no less. That tends sometimes to be overlooked. Take, for instance, the work quoted by Professor Cochran of Neel and Schull on atomic radiation at Hiroshima and Nagasaki and its effect on the human population. If one could make an experiment, surely all the disturbing variables would still have to be recognized? Would they not still have to be considered in setting up the experimental design? Mere randomization of the experimental units would, I believe, be unlikely to give a valuable answer or one that was practically useful.

Later in his paper Professor Cochran rightly quotes with approval the use of a population laboratory. But we are still faced with the problem of how far we can pass from the particular to the general. And in some instances with population laboratories, the particular is *very* particular. The same is true of the clinical trial of a new treatment or a field trial of a vaccine. It is only in quite limited circumstances and with deliberately courted statistical dangers that we can make experiments at all.

How far we can extrapolate from those experiments must always be a matter of concern. And so, as with observations, it is all very difficult. Heaven forbid that in saying so I should appear to denigrate the experimental approach—I have been passionately devoted to it throughout my life; but to statisticians who do not work in this field, and to some who do, I emphasize that it also needs very great care and thought.

With observational studies such as Professor Cochran has described we are, in the last analysis, having to take decisions on circumstantial evidence. Usually, no one sees the murderer slip the arsenic into the teacup and no one sees the *Bacillus typhosus* slink into the tin of corned beef. And so, just as in everyday life, and nearly every day, we take decisions on circumstantial evidence, so must we do in preventive medicine. We cannot escape it.

More often than not the retrospective enquiry, with all the weaknesses that Professor Cochran emphasized, is inevitable. As he himself says, it is only after the victims have appeared that we can start to explore the origin of the epidemic. It is in this way that classical epidemiology has been built up and advanced, but, as Professor Cochran rightly stresses—and this is the important point—along with the results of cognate enquiries of quite other types.

What is essential here is a profound knowledge of the field of work under discussion. Merely to go into the origin of a modern epidemic of typhoid fever from a knowledge of statistical records, their strength and weaknesses, will not take one far. One must have a close familiarity with the habits of the bacillus and of the strength and weaknesses of the laboratory evidence.

Similarly, although this is less well recognized, we will not get far in discussing cancer of the lung and smoking unless we know the whole field, a field with which Professor Cochran became familiar through his membership of the magnificent Advisory Committee to the Surgeon General of the United States Public Health Service. There is indeed a very wide range of data to consider. We have the statistical association in man with all its ramifications in both retrospective and prospective enquiries. We have the histo-pathological evidence of cellular changes in the respiratory mucosa of smokers and non-smokers. We have animal experimentation by the "back-room boys" in the laboratories, using smoke or smoke products, and there is semi-experimental evidence in man when smoking is given up. All this has to be taken into account in reaching a conclusion. It is ignorance of the field, or perhaps neglect of it, that has led some scientists up such curious garden paths that one might almost think that they were random walks!

There comes a time, as Professor Cochran finally observes, when the decision must be made, when the applied scientist must believe that the time for action has arrived—not that I believe that *he personally* should necessarily do anything whatever about it. That time, I believe, may quite rightly and justly be made to vary with the circumstances.

On that subject, I was recently delivered of a presidential address to the Section of Occupational Medicine of the Royal Society of Medicine. In it, I considered nine different ways in which one should study an observed association before passing to causation. In conclusion I said:

> "In passing from association to causation, I believe in 'real life' we shall have to consider what flows from that decision. On scientific grounds we should do no such thing. The evidence is there to be judged on its merits and the judgement (in that sense) should be utterly independent of what hangs upon it—or who hangs because of it. But in another and more practical sense we may surely ask what is involved in our decision. That almost inevitably leads us to introduce differential standards before we convict.
>
> "Thus on relatively slight evidence we might decide to restrict the use of a drug for early-morning sickness in pregnant women. If we are wrong in deducing causation from association no great harm will be done. The good lady and the pharmaceutical industry will doubtless survive.
>
> "On fair evidence we might take action on what appears to be an occupational hazard, e.g. we might change from a probably carcinogenic oil to a non-carcinogenic oil in a limited environment and without too much injustice if we are wrong.
>
> "But we should need very strong evidence before we made everyone burn a fuel in their homes that they do not like, stop smoking cigarettes that they do like or eating the fats and sugar that they enjoy. In asking for very strong evidence I would, however, repeat emphatically that this does not imply crossing every 't', and swords with every critic, before we act.
>
> "All scientific work is incomplete, whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have nor ever to postpone the action that it appears to demand at a given time."

In the end the decision must turn, I suspect, upon our personalities. Only very limited aid can be sought in those subtle partitions of χ^2 to which the reader of the paper has so eruditely contributed.

We are delighted to have Professor Cochran in this Society again and it gives me a very real and personal pleasure to move this vote of thanks to him.

Mr C. B. WINSTEN: Professor Cochran's paper is a timely one. A large amount of work is being done (and always has been done) on non-experimental data, and statisticians should, of course, make a special contribution. The situation as he sees it in the United States, with "the statistician perhaps contributing a chapter on sampling or tests of significance" (Professor Cochran does not say whether the chapters in different books are all the same) is, on the face of it, depressing. But the picture may look different if we think more about the nature of this type of research.

As R. A. Fisher's enigmatic remark emphasizes, the principal characteristic of non-experimental data is the wealth of possible hypotheses available to explain the patterns of variability which we observe. Even the most innocent-looking scatter diagram, showing an encouraging tendency to correlation, may have a host of different causal schemes which can explain it. Cause may go either way, as we all know. Or there may be some underlying cause influencing both variables. Or an underlying variable may influence one of the variables, and just happen to be correlated with another. Or one variable may have some influence in one part of the range of variation, and another elsewhere.

The patterns of variation may arise from self-selection of the population, as in Professor Cochran's example of the housing list. Or, more subtly, and very important in cross-section studies in economics, the units may adapt, or be forced to adapt, to their own peculiarities. For example, we might expect firms with low productivity to capture less of the market than those with high productivity. If this happens, observed data will over-emphasize the advantages of large units. Again, people presumably adapt to their own peculiarities in choosing or doing a job. We would have to guard against any simple causal scheme in studying the effects of type of work conditions or on those doing the work.

Now such complexity should not be discouraging: it is in fact a challenge. But it does mean, as Professor Cochran says, that research in non-experimental studies should be done by people who have a wide grasp of the field of study and an understanding of the plausibility of different theories. In other words, if statisticians take to this work they must stop being "methods" men and become research workers in the field being studied. To some extent this should be true in experimental studies too, but the need there is not so great. Perhaps, therefore, both teaching and observational studies are best carried out in "subject matter" departments. Perhaps the statisticians interested in such subjects have adapted themselves and migrated to these departments, and this explains Professor Cochran's observations.

Because of the richness of possible explanation, any particular set of data needs far more careful study than experimental data. In experimental data each reading is a sample representing many other readings and therefore often remains anonymous. But in non-experimental data each reading may be a clue to some neglected cause at work. Thus it should be anything but anonymous. At the very least this means that one should identify the individual points in scatter diagrams and get to know why they are there. For example, in studies of variations between regions, or towns, in which I am interested, if one looks at the points which are particularly high or particularly low on the diagram one may be able to pick out particular patterns which suggest a new cause: this may be a useful guide to useful multiple regressions. But such methods mean that one must have, or acquire, detailed knowledge of the towns or regions being considered. Incidentally, it is rather rare to find a linear regression in this type of study.

It is also important to see how much one's measurement of causes might depend on one or two exceptional units. How often, for example, is a study saved from the perils of multi-collinearity by a single point "off the line"? But if this reading is exceptional in this way it may be exceptional in many other ways too.

And one of these other ways may be associated with an unsuspected causal factor. This danger leads to another rule of action for the statistician: always study how far your regressions depend on particular readings: do them several times with and without points where the causal factors are in some way extreme or unusual.

There are dangers, in the light of this, in giving an air of false precision and scientific exactness to this type of work, a precision appropriate to a carefully randomized experiment but inappropriate for data collecting from that untidy laboratory, the real world. It is easy enough to take our data along to a computer, where a regression will be fitted, standard errors will be calculated for the coefficients and, for good measure, a usually inappropriate *t* test will be performed. This will look as though it gives a full measure of the uncertainty of the theory, but it does not, of course, Not only are there nearly always (*pace* Occam, a bad guide, as the quotation from Fisher shows) the possibilities of non-linearities of many different sorts, but also of all sorts of different causal schemes. Standard errors do not measure these sorts of uncertainties. Here I would like to stress how useful the likelihood function is as a mode of passing on to other investigators the objective knowledge and the objective uncertainty found from a particular experiment. The likelihood function can always be extended to include newly conceived hypotheses: it is always clear to what range of hypotheses it applies. What is more, in regression analysis, since the likelihood function is of such a simple form, the sum of squared residuals away from the hypothetical regression, it can even be estimated by eye. Indeed, where a two- or three-dimensional scatter diagram can be used, the eye, when trained, is probably the best computer of likelihood functions. It also has the great advantage of being able to see almost immediately what the computer can only do elaborately, the sensitivity of the likelihood to different shapes of regressions, different points, etc. Unfortunately we cannot use scatter diagrams in four or more dimensions so easily. Here is a field where we must make the computer help to scan the data in the way that is possible without its help in simpler cases.

Professor Cochran has given us a paper on a most important topic. I hope that it will lead to more discussion by statisticians and comparison of their experience in the very wide range of subjects touched upon to-day. I have much pleasure in seconding the vote of thanks.

The vote of thanks was put to the meeting and carried unanimously.

Mr F. D. K. LIDDELL: I should like to add my thanks to Professor Cochran for his paper, which I have found most stimulating, particularly because, in Sir Austin Bradford Hill's terminology, I have just been brought into a particular field of play as a substitute half-way through a game and I have a feeling that the players who have already been taking part will throw bottles, as well as the spectators and the referee.

I should like to say something about the pneumoconiosis field research project carried out by the National Coal Board. I want to do this because there is a belief that if an investigation is on a large enough scale and well enough intentioned, it must provide adequate data.

The population that we are looking at in this research is the complete labour force of a sample of 25 collieries spread throughout Great Britain; at the start of the research 10 years ago, it consisted of 35,000 coal miners. This study falls within Professor Cochran's definitions: the principal objective is to elucidate a cause and effect relationship (it is to discover what level of respirable airborne dust can be tolerated in mining without leading to an unacceptable prevalence of the disease which is called pneumoconiosis—and that means simply dust in the lungs) and it is quite clear that we cannot carry out controlled experimentation.

The response to dust exposure is measured on medical surveys carried out five years apart at each of the 25 collieries. In the interval between the surveys, dust exposure itself is being measured.

The main response variable is change in X-ray appearances but, unfortunately, assessment of such change is subjective and no truly reliable method has yet been found anywhere in the world. We like to think that we are furthest advanced in this country, but even we have a long way to go.

We obtain a second response variable from a questionnaire on respiratory symptoms. This has all the usual difficulties, confused here perhaps even more by many miners' thoughts about compensation when they are answering the questions.

Perhaps the most reliable response we have is that measured in lung function tests. But, unfortunately, lung function deteriorates markedly with the age of the subject—much more than with the degree of pneumoconiosis; further, a more than usually rapid deterioration of lung function may be due not to pneumoconiosis alone but to bronchitis, to smoking, or to some other cause completely unrelated to dust exposure.

Levels of airborne dust in the mining environment vary greatly between jobs and depend on such matters as the mining machines and conditions, so we have no worries that at least we shall find differences in the cause being examined. However, dust levels are highly variable from time to time in the same working place and measurement is difficult. Despite a massive effort in absolute terms, it leads in relative terms to only about one complete shift sampled per man in about 10 years.

The study has to be long-term because the disease takes many years to develop, and the field units are to be congratulated on having solved many of the problems that Professor Cochran mentions on this particular subject. The response rate of the subjects is of the order of 95 per cent of those employed at the time of each medical survey.

Our main concern, however, is with the cohort—that is, the survivors from one survey to the next—and these become only 60 per cent of the initial population. That sounds fine—20,000 subjects still appear an adequate population for study. The real blow is that on the first serial examination of these 20,000 subjects we find only about 800 showing signs of radiological progression and, therefore, of direct concern to us. It is all very fine that the management of the 25 collieries may have been particularly zealous in dust suppression but that does not help us very much from the research point of view, particularly when we bear in mind the many disturbing variables, e.g. smoking habits, obesity, atmospheric pollution.

This material, I am told by physicists and doctors, is better than anything available anywhere else. Nevertheless, the correlations that my colleagues have been able to establish are disappointingly poor.

I have no doubt that in speaking so briefly of the pneumoconiosis field research so soon after having been thrown into its deep end I have distorted considerably, possibly through oversimplification as much as anything else. Nevertheless, I cannot but feel a degree of the frustration that Professor Cochran mentions in his introduction as arising in admittedly rather different circumstances. Here, the basic planning was carried out a decade ago and now we have a mass of data. Quite clearly, we must obtain the maximum benefit from them. I should therefore welcome greatly any suggestions on strategies that Professor Cochran, or anyone else in this distinguished audience, may care to make. Meanwhile, his paper will be most useful in placing the difficulties of our own research in a wider context.

Miss E. M. BROOKE: The Ministry of Health is, naturally, concerned with observations on human populations. In particular, it is concerned with the subject of mental illness, which accounts for nearly half the occupied hospital beds in this country.

Compared with previous speakers, I come almost from the Stone Age because, while it is comparatively easy to talk about miners, about people moving to new towns, or about people with lung cancer, we must at the outset ask who is mentally ill.

Therefore, the population which we want to observe is not at all clearly defined. Except in conditions like general paralysis of the insane, or phenylketonuria, we have no tests. For diagnosis and for detecting the presence of the illness we rely on verbal communication and observation of behaviour. These differ in different cultures and what is normal in one is abnormal in the other.

We are told that mental illness is a deviation from the normal, but no one defines what the normal is. We are also told that we must have tensions if we want to live satisfactory, healthy lives and, therefore, some symptoms of mental illness, such as anxiety, should naturally occur. And so we have to face the problem of when a symptom ceases to be healthy and becomes morbid.

As soon as we have tried to define the population, diagnostic labels are attached to people. Here again we are at a great disadvantage because it is obvious that, even in this country, the diagnostic label must vary very much, as can be seen from the map on p. 149 in *The Geography of Life and Death* by L. Dudley Stamp.

This variation is even greater between countries. In America there are three times as many schizophrenics as depressives. In this country the depressives greatly outweight the schizophrenics. The question is: are the Americans more schizophrenic and are we more depressed or is this a diagnostic variation? A great deal of time has already been devoted to attempts to find the beginnings of an answer to this question.

For an actual investigation of a group of mentally ill people, cases have to be assembled and, therefore, populations should be screened. This is too much for one person and, in order to get a sufficient number of cases, we need an instrument which will enable any number of workers infallibly to find the same case of illness. It is this kind of instrument—a questionnaire which will enable different investigators to arrive at the same kind of case—that has to be developed. So that we are, in fact, at the beginning of shaping our tools.

In the past we have, unfortunately, had to fall back upon hospital populations, but these, of course, are determined by the numbers of beds and by the existence of other facilities. The study of these can lead to considerable fallacies. For example, we notice that the admission rates for single people are much higher than for married and, therefore, the hypothesis was set up that there was a selection against people who were liable to mental illness so that they were not selected as marriage partners. But once one extends one's observations to the people receiving other services, we find that a great many of these missing married people are, in fact, attending the out-patient clinics.

When it comes to cause and effect, we are very much struck in mental illness by the effects which are startling, especially to people who see patients. We are extremely ignorant, I think, of the causes, and there is a great tendency to take coincidences for causes.

This happened, for instance, at the time when a decrease in mental hospital populations coincided with the advent of the tranquillizing drugs. The conclusion straight away appeared to be that one was the effect and the other the cause. In this country, however, the run-down in the mental hospital population started before tranquillizers had gained ground and, at the same time, a great many other factors were coming in, such as the opening of closed wards, bringing the patients into touch with the community and introducing industrial work which would fit them to take their place in society again. Thus we are faced with the problem of discovering which is the causative factor, and in this we have not made very much progress so far.

Retrospective studies are, of course, very popular in this exercise: one sees the problem and one hopes to look back and find the cause. One of the particular joys has been the broken home. This is a quite popular subject and easy to establish, judging by the number of studies in which it features.

People look at the life histories of delinquents or mentally ill people and find the broken home. No one appears to explain why a home that is as broken for Tom as it is for Jim, causes Tom to go into a mental hospital but not Jim. Nor has anyone apparently so far started a longitudinal study to find out how many normal people come from broken homes, how many people from broken homes do not become abnormal, etc.

We are, of course, concerned in the Ministry with the effects of administrative policy and in particular we have adopted the policy that people should, as far as possible, be treated in the community and in general hospitals instead of in large mental hospitals away from the community.

The charge that has been made is that we have not demonstrated the advantage of this. Of course, we cannot demonstrate the advantage until we have worked the system. But it means that we have to organize studies which will enable us to compare the effects of treating people in the general hospital and in the large mental hospital. Here, it is desirable to match groups.

One is then faced with the variables on which to match, and those usually selected are the first suggested: age, sex, marital status, occupation, social class and so on. But it is by no means certain that these are all that are necessary. There is the variation in conditions within the same diagnosis. For example, Leonhard has suggested that schizophrenics can be classified into 20-odd groups and, obviously, even if you compare cases with a slow, insidious onset with those having a sharp onset, considerable variation in their reaction can be observed.

Altogether we are trying to match on some 60 items and we hope to find out by cluster analysis which of these are really significant. At present we are in the position of preparing the tools and hoping that at some future time we may be able to do some useful work with them.

Mr A. S. C. Ehrenberg: Professor Cochran is concerned with "problems . . . that seem . . . to differentiate observational research most clearly from controlled experimentation" and in Section 2 he lists what he calls the major difficulties. Running through these briefly, we have:

In Section 2.1 the specific difficulty mentioned is that the "observational investigator" has to know where to find the phenomena he wants to study; it can, however, be equally helpful if the "controlled experimenter" knows something about his subject-matter.

In Section 2.3 Professor Cochran describes his concept of cause and effect as essentially predictive, but so it would have to be in controlled experimentation.

Under Sampling, in Section 2.4, it is thought that the deliberate use of populations to be sampled which differ from the target populations will remain standard "observational" practice, but this practice is said also to be followed in controlled experimentation.

The "formidable measurement problems" of Section 2.5 must, of course, apply equally well in controlled experimentation, and are to be overcome only by hard work and very little talk.

If in Section 2.6 there is difficulty in telling "imaginative investigators" to stop adding more and more "ingenious" variables to be measured, such investigators can be relegated to their science fiction department irrespective of whether they are "observational" or "controlled".

Again, if in Section 2.7 observational long-term studies are said to lack opportunities for quick publication, we ought to thank God and remember Fisher, who left Rothamsted years before his famously *controlled* long-term experiments were even completed.

Only in Sections 2.2 and 3 dealing with "disturbing variables" is there a suggestion—at first sight—that "observational" and "controlled" studies do differ in the way stated, i.e. that in the latter "the function of blocking or adjustment is to increase precision rather than to guard against bias". However, the "adjustment" procedures given in Section 3 generally do not work. They largely turn on the "dependent" variable y being regressed on the "disturbing" variable x, where (i) the distribution of x must differ in the populations to be compared (otherwise there would not be any effective disturbance), and (ii) the regressions are assumed to be the same. This assumption is known to be universally false under condition (i)—a point which was made in Section 5.11(B) in Lindley (1947, p. 234)† and which has been belaboured since in *Applied Statistics* (Vol. XII, 1963 No. 3) to be published this year.

My conclusion is that the problems which Professor Cochran has discussed have nothing to do with any distinction between observational and controlled studies. I am not alone here since Professor Cochran, having led us up the garden path, says exactly the same in his final "comment" at the end of Section 2.

† Lindley, D. V. (1947), "Regression lines and the functional relationship", *J. R. Statist Soc.* B, **9**, 218–244.

Perhaps I have now prepared the way for the outright rejection of the good statisticians' pseudo-traditional myth which is always implicit and often explicit in Professor Cochran's own paper, namely, that observational studies' are at best a second best to so-called controlled experimentation. However, if we survey the work of thousands of scientists over hundreds of years, a conservative estimate might be that at the very least 99·9 per cent of the work has been done by methods which statisticians would never deem fit for their canons of "controlled experimentation". And, of course, the scientists were right, as shown by history and as may be very briefly illustrated by the smoking–lung-cancer problem to which Professor Cochran himself frequently refers in his paper, where I now echo one or two of the remarks made by Sir Austin Bradford Hill.

Suppose that we adopt Professor Cochran's "attractive possibility" (Section 4.1) of comparing twins, and add full-blooded randomization of which twin is to smoke. Then the differing incidence of lung cancer in this "controlled experiment" may be caused by differential rates of contact with cigarette paper or with the noxious metal foil in the packets, by differential exposure to fumes from matches, etc., used for lighting up, by different frequencies of visiting tobacconists, by smokers being either more or less relaxed because they can puff and suck, by non-smokers eating more sweets, etc., and therefore dying early of heart disease, by non-smokers having nothing to do with their hands and getting up to neurotic substitute malpractices, by the well-known fact that smokers "note" cigarette advertisements more and may therefore be impelled to emulate their virile suggestions such as mountain climbing and sailing with ghastly effects on their health such as getting lung cancer (it's the "ads" that did it!), and so on.

In other words, the number of potential "disturbing variables" left in this "perfectly controlled experiment" are enough to give a statistician heart failure, especially if he also smokes. Many of the factors can no doubt be, or have already been, eliminated, but this could never be done within the narcissistic confines of the *one* controlled experiment. Controlled experimentation is no more than a potentially convenient and just occasionally applicable tactical method for eliminating one or two of the many disturbing variables present in any situation—here people's initial differences, if any, in their predisposition to lung cancer.

Anyway, having found a five times higher incidence of lung cancer amongst non-smokers in this experiment, and for many other good reasons which are well known to *scientists*, the experiment is "repeated" in some sense. Unfortunately then, two or more independent studies of any kind can no longer make up a controlled situation. It follows immediately that controlled experimentation can play no central role in scientific methodology, which concerns itself with the building up of generalized and integrated knowledge. A controlled experiment is necessarily self-contained, and there is room for a *first* study or an *isolated* experiment at most once in any one field of study.

Apart from wasting their own time by attempting to transmute controlled experimentation into a philosopher's stone instead of treating it as an occasional convenience, statisticians are—far more importantly—making practical workers feel guilty about "not doing experiments". The stultifying influence of statistical teaching in this particular respect is exceeded in its ill-effects by at least two other aspects of modern statistics, and could therefore be even worse than it is.

Clearly, there is more that one could say, but I conclude still within the general temper of our discussions in this Society by admitting that when I now quote Professor Cochran's own transferable quotation of Professor Barnard's reported comment on Dr Wold's 1956 paper on observational data—namely, that "a paper of this kind is useful in showing the younger statisticians what difficulties they may be up against"—I do feel younger in spirit. Such, we humbly learn, are the workings of cause and effect.

Mr V. SELWYN: As the instigator in the market research world of a large-scale series of observation studies may I in contrast to the previous speakers quote from the hard world of commerce. I should like to contrast observational studies in the market research world, not with controlled experiments, but with verbal techniques.

So much market research work turns on asking people questions and the very act of asking questions produces a series of unknowns. We are dependent upon investigators and upon the words that people use. Therefore, there is a virtue in any technique which avoids questioning.

The exercises with which I have been connected have been concerned with brand consciousness in the motor-oil world. We have taken one measure of brand consciousness, and that is the degree to which the motorist asks for a brand when he buys his motor oil. We will not argue about how far that measures his brand consciousness. We merely take it as a measure.

When we first looked at this particular problem, we realized the cheapest way of finding out was to ask people what they do in their buying situation. But this is not the most desirable way. When we ask we are dependent on respondents' memories, their beliefs, and so on, and memories of occasional purchases can be hazy. Instead, we decided to base a study on observing people at the point of purchase. What does the motorist actually do when he pulls up to a garage forecourt and buys his motor oil?

This study has been repeated over a series of years. Each year 25,000 motorists are observed, of whom about 4,000 buy motor oil.

Now the technique raises a formidable sampling problem. It is the main objection to this technique in market research. We try to base our sample of garages, where we observe, on unknown data—mainly, the degree to which various brands of petrol are sold through various garages. In this garage world, however, there are political factors: i.e. garages are controlled as to what they sell, they are tied to various petrol companies. Some co-operate and some will not. In the final analysis, one is dependent upon a sample of co-operators.

However, we do not depend on the absolute figures of each exercise. By repeating the exercise we get a trend from year to year. On the basis of 4,000 oil purchasers we may find, for example, 71 per cent in one year ask for a brand of oil by name. But instead of depending on that as absolute, we measure what happens the year later and the year following that so as to eliminate the effects of certain variables in the sampling. We may have too many garages of a certain type, small or large, too many on a main road or too many where only pleasure or only business motorists go. These are factors which may affect the numbers asking. Over a period of years we try to minimize as many of these variables as possible.

On the other hand, this operation introduces new variables. Although we try to hold the test on the same week of each year, in June, when we hope to have the same weather, we still have the effects on traffic flows of a variable Whitsun in this country. We may also find that the garage which we visited last year will not co-operate this year, or, alternatively, that the garage changes its policy or the attendants have changed. Instead of finding a lackadaisical attendant who does not mind what brand of oil a motorist has, we might find an aggressive one who tries to force sale.

These points illustrate my contention that one can apply various statistical methods—but everything turns on the raw material. It proves impossible to apply tests to all the variables—for some we do not even know. We try to learn all the possible factors which might affect our answers—but we can only hope to allow for all of them.

Any research technique must have its limitations. In an observation test we can only analyse our motorists by observable factors. We can see whether a motorist is in a saloon or a sports car, or whether he is driving a lorry or a van. But we cannot classify him by age. Investigators only record and do not ask questions. We cannot ask them, for instance, to classify motorists into young, old or middle-aged. For this would prove a subjective assessment on the part of an investigator, on which we do not want to depend.

Yet within all these limitations we have produced very useful results, which, obviously, I cannot go into now. It is a technique which, I feel, is neglected in the market research world, perhaps because of the limitations I have described or perhaps because people have not thought about it sufficiently.

Professor COCHRAN subsequently replied in writing as follows:

I am grateful to the speakers for gently correcting an imbalance that my method of presentation may have created and for rounding out the paper with descriptions of the nature of observational research in pneumoconiosis, mental illness and marketing. Sir Austin Bradford Hill and Mr Ehrenberg have reminded us that controlled experiments are faced with much the same set of problems as observational studies and demand an equally high degree of skill in their execution. My reason for bringing controlled experiments into the paper was that the formal study of techniques for planning and analysis has proceeded much further in experimentation than in observational work, and I believe that when grappling with an observational study a knowledge of the lessons learned over the years in experimentation is likely to be useful.

Mr Winsten and Sir Austin have emphasized that the statistician involved in observational studies must become a "subject-matter" expert, because subject-matter knowledge is essential to effective planning and because the information to be summarized on a complex problem is usually widely miscellaneous, ranging from statistical studies to casual but suggestive hints. The degree of attention to detail and imaginative insight that must be used if one is not to be misled in scrutinizing the results of an observational study have been well brought out by Mr Winsten. I am glad that these points were made, because in a statistical paper it is easy to leave an exaggerated impression of the importance of the statistician's standard bunch of tricks.

In the studies described by Mr Liddell, Miss Brooke and Mr Selwyn, I was interested to note that problems of measurement are a major concern. With new tools of measurement, a significant part of the research must be to learn their strengths and weaknesses. This is particularly so when measurement is subjective, for despite Mr Selwyn's reluctance to rely on subjective appraisals, they are often the only ones available. Fortunately, the importance of errors of measurement in the sample surveys taken in many countries is now much better appreciated, with a consequent increase in the research effort devoted to this problem. Although lack of good instruments of measurement is sometimes the chief stumbling-block to progress, there is always room for hope that a major advance may come from an unexpected quarter. The awards of Nobel prizes for developments in tissue culture and paper chromatography are a sign of the importance rightly attached to advances in the science of measurement.

The approach described by Mr Selwyn is an ingenious way of circumventing an erroneous assumption that still persists in some quarters—namely that if one asks people how they would react to a hypothetical situation with which they are not now faced, their replies will be trustworthy for predictive purposes. It is necessary to devise some means of observing how they actually react when faced with the situation.

The areas of research in which Mr Liddell and Miss Brooke are engaged are especially difficult. The list of problems given by Miss Brooke is very familiar to me from my limited acquaintance with the mental illness field, although, alas, I have no good suggestions to offer. I have the impression that the degree of success attained in the long-term follow-up of subjects is greater in this country than in the U.S., perhaps because of the higher mobility in the latter. I was a little puzzled by the phrase "disappointingly poor" applied by Mr Liddell to the correlations obtained. If low correlations are a sure indication that the measurements are inadequate, the word "disappointing" is appropriate. But if they merely imply that the results do not agree with preconceived notions, the adjective hints at an emotional involvement that may obstruct clear thinking.

Since I may already have written too much about controlled experiments, I am reluctant to deal fully with Mr Ehrenberg's comments. His account of their limitations is, I think, much overdrawn. In many countries, large programmes of controlled fertilizer and variety trials are in progress, because past information from such experiments has led to substantial increases in the world's supply of food. The flood of useful new products coming from the research and development sections of industry relies heavily on the skilful use of multi-variable experimentation. A proposal that scientists in these fields should in future confine themselves to observational studies because experimentation is "no more than a potentially convenient and occasionally applicable tactical method" would be considered ridiculous.

With regard to his amusing report on the hypothetical twin-smoking experiment, I know of no one engaged in this field who would regard an identical twin study as more than a small additional link in the evidence, helpful because, as Mr Ehrenberg points out, it removes a genetic difference that is hard to eliminate from most of the other available comparisons. In my own experience I cannot recall instances of statisticians bullying practical workers for not attempting experiments that would have been a waste of time, though they sometimes scold workers for doing bad experiments when they could have done good ones.

Mr Winsten's account of the usefulness of examination of the likelihood function interested me. Now that machines are becoming available for graphical presentation of the results from electronic computers, we may, with some practice, be able to grasp results in more than three dimensions.

On a technical point, Mr Ehrenberg claims that if the distribution of x differs in two populations, the assumption that the regression of y on x is the same in the two populations is "universally false". On the contrary, models in which the assumption holds are easily constructed. We need, for example, merely make $y = \alpha + \beta x + e$ in both populations, where e is a random variable with zero mean for given x. His supporting citation from Lindley's well-known paper deals with quite a different problem. Lindley's Section 5.11(B) on p. 234 reads as follows: "The regression line is required for prediction of either true or observed values of one variate from observations of the other *whether or not this latter is in error*: it being understood that the x_1 from which x_2 is to be predicted comes from the same population as those x_1's used in the estimation of the regression line." Lindley deals with a *single* population in which (i) the regression of x_2 on a variate ξ_1 is linear, (ii) we cannot measure ξ_1 but only $x_1 = \xi_1 + d_1$, where d_1 is an error of measurement independent of ξ_1. He gives necessary and sufficient conditions for the linearity of the regression of x_2 on x_1, and shows that if this regression is linear, it should be used for predicting x_2 for a new specimen with measured value x_1. Lindley's final phrase is a warning that if x_1 is measured much more or much less accurately in the new specimen than in the specimens from which the regression was constructed, his results do not apply.

The fact that regressions can be mathematically the same in two populations does not mean that they are the same in applications. Numerous cases can be found in which they are not the same. The appropriate method of adjustment in this situation presents problems, which for limitations of space I did not discuss.

In conclusion, my objectives in writing a mainly descriptive paper on this extensive subject were twofold. I believe, as stated, that statisticians have much to contribute to this challenging field. Secondly, it is a commonplace to find the same problem in study after study. If means can be found for bringing them together for the discussion of tough methodological issues, practical workers in diverse areas of investigation can profit greatly from one another's experience and wisdom, both as regards what to do and what to avoid.

As a result of the ballot taken during the meeting the candidates named below were elected Fellows of the Society:

Allen, Alfred John, B.Sc. (Econ.)
Ayers, Kenneth Edwin
Brew, John Morland, B.A.
Bristow, Alan John, B.A. (Ind. Econ.)
Bunn, Richard Herbert
Davies, Allan, B.A.
Gaitonde, Shrikrishna Yashwant, B.A.
Gamble, Elaine Lilian
Guyton, David Firth, B.A.
Hall, Alan Vivian, B.Sc.
Hope, Adery Catherine Alison, B.Sc.
Jones, Patricia Ann
Knott, Martin, B.Sc. (Econ.)
Nicol, David Peter, B.A.
Nicol, Helen Isabella, B.Sc.
Ord, John Keith, B.Sc. (Econ.)
Shrubbs, David John, B.Sc.
Stockham, John Evans, B.Sc.
Stracey, Brian Alexander, B.A.
Taylor, Humphrey John Fausitt, B.A.

Corporate Representatives

Balint, Magda, *representing* International Co-operative Alliance.
Insull, Anthony David, *representing* Associated Fisheries Ltd.