The Statistician, Vol. 18, No. 2

Longitudinal Studies and the Measurement of Change

H. GOLDSTEIN

Introduction

IN the study of change in the characteristics of a population, two basic sampling designs can be distinguished. The one, known as the "cross-sectional" design, gathers information using a different sample of individuals at each point on a time scale, and the other the "longitudinal" design, gathers information using the same sample of individuals at each point. Thus, in a cross sectional study, if the change with age is of interest, then the study would consist of different groups of individuals at selected points on the age scale. If it is the changes over the calendar time scale (secular trends), which are being measured, a cross sectional study would consist of individuals of one particular age, examined at different points in time. We may wish to measure both age and secular trend effects, to make allowance for one in analysing the other, or to study the "interactions" between secular and age trends; that is, whether the change in reading ability between 1955 and 1965 say, of seven year olds is different from the change in ability of eight year olds during the same period. An appropriate design would involve combinations of ages at different points in time.

This article is mainly concerned with longitudinal studies, where the same individuals are measured at each point on the time scale (referred to as an "occasion"). Such studies are almost always concerned with individual development, although the interactions of secular and age trends may also be studied by choosing samples from populations originating at different points in time, for example by following one sample of children born in 1946 and one sample born in 1958.

In between these two basic types of study there are what are often referred to as "mixed longitudinal" studies. For example, in the study of change in adult characteristics different groups of individuals may be followed over different age periods determined so that the period for each group overlaps the period for the group before and after, and thus cover the whole age range without having to wait to complete the study until a single age group of individuals had passed

from the beginning to the end of the whole period. In fact, most longitudinal studies are not "pure" longitudinal, because some individuals are invariably lost and new ones acquired during the course of the study, so that not every individual is measured on every occasion.

Some general problems of definition are now discussed, and this is followed by consideration of a general design structure for the various types of study. Problems of sampling and data processing are then discussed in relation to one particular longitudinal study, and some of the more interesting results from this and other studies are presented. The final sections deal with aspects of the statistical analysis of longitudinal studies.

The Definition of a Longitudinal Study

Longitudinal studies may be classified according to the method of collecting information and to the type of hypotheses which are to be tested. A discussion of the different kinds of study is contained in the report of the N.I.C.H.D. Colloquium on Longitudinal studies.[1] Three main kinds of study are distinguished and these are referred to as the "retrospective", the "prospective" and the "longitudinal". In the retrospective study, information for occasions prior to the one on which the individuals are measured, is obtained by questioning and by making use of any records which may have been kept. The limitations of this method of studying change are imposed by the reliability of the records and the individuals' memories. The method is probably of more use in individual case studies where more can often be done to overcome these limitations, than in large scale surveys.

The distinction between "prospective" and "longitudinal" studies seems to be a little uncertain. According to Yerushalmy[1], the prospective study is usually concerned with the "outcome" (the value of a selected characteristic measured on the last occasion) in a "deviant" group of individuals, subject to some form of treatment. For example, a study of the adult intelligence (outcome) in a group of illegitimate children (a "deviant" group) who have been adopted (treatment). A control group is usually present. The true "longitudinal" study is defined as one in which the whole "pattern of development" is of interest, and it is usually concerned with a sample of "normal" individuals. Aside from the difficulty of defining precisely what is meant by "pattern of development", it seems that, in practice, the distinction between these two types of study becomes blurred

since "normal" children are sometimes the subject of studies where the outcomes are related to an initial point in time (e.g. one could select a sample of "normal" children who happen to be illegitimate, regarding the "treatment" as the fact of their illegitimacy) and, furthermore, such a study may change so that the total "pattern of development" comes under scrutiny. This uncertainty of definition seems to be largely a semantic problem, and for simplicity I will use the term "longitudinal" to describe any study where information is repeatedly collected, over time, on the same sample of individuals. The term will also be taken to include mixed longitudinal studies which are designed around a "pure longitudinal" core of individuals. Thus a longitudinal study may include some retrospective information (e.g. daily consumption of cigarettes during pregnancy, asked of a mother at the birth of her child), since it may be too difficult to obtain the information in any other way, or where, for example, an individual has missed a measuring occasion and some information, obtained by retrospective questioning, may be better than no information at all.

Cohort Studies

Most longitudinal studies, especially those concerned with physical growth, have involved rather small numbers of individuals, up to about 500. The samples have usually been selected according to convenience afforded for making a study, and carefully controlled measurements made on each individual. (E.g. the Harpenden Growth Study[8] uses children in a Children's Home where a well equipped measuring laboratory is set up and measurements made by trained measurers, and is one of the largest physical growth studies, having "mixed longitudinal" information on about 650 children from a few years old to about 20 years of age.)

So called "Cohort" Studies are longitudinal studies involving large numbers of individuals, selected according to an easily defined characteristic, usually the time of birth. The two major child development studies in Britain, the National Survey of Health and Development[2,3,4] and the National Child Development Study[5,6] both initially took all the babies born in the week 3rd–9th March in 1946 and 1958 respectively. One of these studies, the National Child Development Study, will be described in more detail later on.

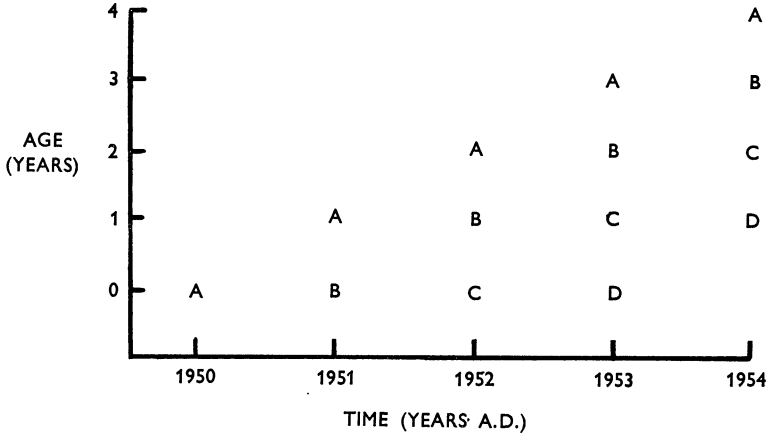
Generalized Design Considerations

The two cohort studies mentioned above were selected from the same week's births partly in order to provide reliable comparisons

between two generations 12 years apart. Such comparisons may be viewed as part of a general population model which includes cross sectional comparisons as well.

Diagram 1 shows successive "cohorts" (defined by year of birth) A,B,C,D, moving through the age-time plane.

Diagram 1



This diagram illustrates the fact that any one of the three variables (age, time, and cohort) is uniquely determined by the other two (e.g. cohort A, born 1950, defines the age of the cohort at any time). If there is a real difference between cohorts A and B at one year of age on a particular measurement, then this may have arisen from the fact that the cohorts, having been born at different times, have (possibly) been subject to different environmental influences during the first year of life. One can either choose to refer to this difference as a difference between cohorts or as a difference between the times of measurement (1950 and 1951), but it means the same thing; and the difference is often referred to as a secular trend (at a given age) between the times of measurement.

Schaie[7] proposes a general model, using essentially diagram 1, in which he postulates three "effects" on an individual measurement; namely the cohort effect, the age effect and the time of measurement effect. However, he seems to fail to recognize the fact, pointed out above, that since any one of these effects can be uniquely defined in terms of the other two, in reality only two of these "effects" can be considered to act on an individual, and his resulting description and

classification of research designs not only results in a somewhat circular argument, but also serves to complicate a relatively straightforward idea.

Returning to Diagram 1, if any diagonal is followed through, this describes the conventional cohort study which selects a group of individuals born in a specified time period. If a vertical line is followed, i.e. at a fixed time, the conventional cross sectional age study is described, and if a horizontal line is followed, this describes a study concerned with secular trends at a given age. If all the entries in the table for the years 1952, 1953 and 1954 are taken this gives interpenetrating age groups over the age range 0 to 4 and both age and secular trend differences can be studied. If however, a study includes say, cohort A at age 1 and cohort B at age 2, the age and secular trends would be confounded by the design and could not be separately estimated. Thus research designs can be classified simply by specifying the region of space in the diagram that the measurements occupy. For example, if a study measures, over the period between the points 1950 and 1951 on the Time axis (x), all children at age 1, this gives the horizontal line segment

$$y = 1,\ 1950 \leqslant x \leqslant 1951.$$

The National Child Development Study, for example, would be described by segments of the space between two parallel straight lines a week apart at an angle of 45° to the x axis.

In addition to its being a convenient method of describing a study design, this also enables one to plot, along a third axis, estimates for a variable or a function of variables obtained at different ages and times, and thus to obtain an age-time surface. (The multivariate case is a natural extension.)

In view of the importance of secular trends in educational variables, the above approach would seem to be a useful one to use when presenting the results of a study. If, in addition, a third "space" dimension is added to the diagram, a design could be classified by the population which is being sampled (e.g. England).

A further consideration in designing a study to measure change is the efficiency of the design. If, for example (ignoring secular trend), one is interested in the change in the stature of London Schoolboys between seven and eight years of age, then for a given level of precision, and with a knowledge of the distributions of stature at the two ages and the relation between stature at 7 years and stature at 8 years, one can estimate the number of children needed at each

age, (*a*) if different individuals are used at each age and (*b*) if the same individuals are used at each age. According to Tanner[8], for most body measurements it takes twenty times the number of children measured cross sectionally (design *a*) as longitudinally (design *b*) to achieve the same precision for the estimates of change. The successive occasion to occasion correlations for psychological and educational measurements are usually less than those for body measurements, so that this ratio would be somewhat smaller in educational studies. The quantitative aspects of the most efficient sampling design to be adopted for educational studies do not seem to have been examined, and could well be profitably investigated. In most longitudinal studies, of course, the mean change is only one aspect of the study, and primary interest usually centres on the relationships between variables on different occasions, which can be examined only by using a longitudinal design.

It should be noted that the main advantage of longitudinal over cross sectional studies lies in the efficiency of the estimation of change. For cross sectional parameters (e.g. the mean stature at age 7), longitudinal studies which have a high coefficient of longitudinality (i.e. the proportion of "pure" longitudinal elements; see section on "Sampling"), provide no advantage over cross sectional studies with the same number of individuals, and indeed, because of the serial correlation from occasion to occasion, a longitudinal study will provide less information on the population means for the set of occasions, than independent cross sectional samples with the same number of individuals at each occasion. Hence, for the purpose of measuring secular trends at particular ages, cross sectional studies will generally be more efficient than longitudinal ones.

Narrowly defined cohort studies, such as those quoted above, which use one week's births, are sometimes criticized on the grounds that, in view of the importance of secular trends, sampling should be spread out, and smaller cohorts should be taken, for example, three months apart. Such cohorts, it is argued, would enable secular trends to be measured and allowed for. If, however, this argument refers to trends in cross sectional parameters, then as pointed out above, longitudinal studies do not necessarily provide the best means for doing this, and in overall terms may be less efficient than repeated cross sectional studies. If, on the other hand, interest centres on trends in the changes or in the relationships between ages for successive cohorts, then longitudinal methods must be used. Such trends, however, are usually slower than trends in cross sectional

parameters, and are likely to be of less import for policy decisions. Thus the measurements of these trends could be made using more widely spaced cohorts, and unless seasonal differences are expected (see section on "Sampling") there seems to be little reason for not concentrating the major cohort studies into, for example, one week's births with all the administrative convenience that this brings, supplementing them with cross sectional studies for monitoring specific secular trends in cross sectional parameters.

The National Child Development Study (1958 Cohort)

For many different reasons, the importance of Cohort Studies in educational, social and psychological research is increasing, and it will be helpful to discuss some of the problems involved in carrying out longitudinal studies, by describing the experience of one large scale Cohort Study, the National Child Development Study (N.C.D.S.).

In 1958 the Perinatal Mortality Survey[9] was carried out on all babies born in the week 3rd–9th March, in England, Wales and Scotland. The survey included 17,205 births and it was estimated that this was about 98 per cent of all the births in that week.

In 1964 it was decided to set up the N.C.D.S. to re-examine this cohort and the field work for this was done in the first six months of 1965 (for details see (6)). Initially, the children were traced through the co-operation of local authority departments, who returned the names, addresses and birthdays of all children born in the cohort week. This information was then linked with the records of the children in the perinatal survey, as far as this was possible. From the original sample, 14,862 children were measured, an additional 639 children being included who were not in the 1958 survey. The remaining children had either died (809), emigrated (423), refused to participate (84) or had remained untraced (1,238). Thus the total loss from the original sample was a little over 8 per cent. Work started in October 1964 and an interim report, covering educational, social and medical aspects of 7 year old children, was completed in a little over 18 months. The interdisciplinary nature of the research was reflected in the composition of the research team consisting of an educational psychologist, a sociologist, a medical research officer, and a statistician (part time).

Large scale longitudinal studies have sometimes been criticized for accumulating large amounts of data which are then never fully analysed. It is indeed true, that the ease with which computers can

process large quantities of data has made the physical task of handling data much simpler, and this may sometimes produce a temptation to collect data for their own sake. This fault, however, is not implicit in this type of study, as the two British Cohorts are demonstrating. In defence, it should be pointed out that the extra cost involved in measuring a few more variables on each individual is relatively small, and since the relevance of much of the data in a longitudinal study may only become apparent at a later stage in development, it is worth collecting a certain amount of ancillary information which is not of immediate relevance, but which could become so. Furthermore, large scale Cohort Studies can be used in the investigation of special groups. Since, in a sample of 15,000, there will be sizeable groups of deviant children, useful detailed studies of these can be mounted, using the remainder of the children as a control group. This is illustrated by the several special sub studies described below, now taking place within the framework of the N.C.D.S.

There is a study of "word blind" children which involves a detailed analysis of the children in the cohort with severe reading disabilities. The quality of the birth data and the representative nature of the sample make this unique among studies of dyslexia. In similar ways, studies of illegitimate children, children in care, adopted children, physically handicapped and gifted children are also being undertaken. All these studies take the data collected on the main cohort as a basis for selecting the deviant children and then persuing intensive research on them, in a way which has not previously been possible on a representative sample.

I shall now discuss, with particular reference to the N.C.D.S., the four major aspects of a longitudinal study, namely sampling, editing and processing data, setting up of hypotheses, and statistical analysis.

Sampling

One may regard a particular cohort as a reference population such that any statements concerning it, based on observations on a sample drawn from it, apply only to this cohort. This is of course much too restrictive and in practice one may regard this cohort as a representative (but strictly non-random) sample from a larger population born during an (unspecified) period of time which includes the actual survey time. This rather imprecise formulation does not take account of secular trends, but the magnitude of these may be estimated from cross sectional studies, or possibly from successive cohorts, and allowance made. There may be a further complication if the time of

the year chosen has a particular effect. Aside from astrological considerations, there is some evidence for an association between educational attainment and date of birth (Pidgeon[10]). It seems however that any effect is likely to be very small, and it will therefore be understood in what follows that the N.C.D.S. target population is not the cohort, but "children born around 1958".

The survey of the 1946 Cohort began by attempting to include all the children in the cohort, but due to restrictions on resources, only about one third of the children were followed up[3]. The aim of the N.C.D.S. has been to retain as many children as possible of the cohort, and the retention of about 92 per cent of the original sample compares favourably with other longitudinal studies. (The International Children's Centre's five European Growth Studies[11]; for example, had retained about 70 per cent of their samples by the age of 7.) Whether resources should be allocated to further attempts to include missing individuals is a difficult problem, and will be returned to below.

In all surveys of human populations, a major problem is the bias introduced by "non-response", and certain special problems arise in longitudinal studies.

In such studies, it is desirable to measure all individuals at the same ages, and in a cohort study it is important to carry out measurements over as short a time interval as possible, and information gathered at a later time on non-respondents will not necessarily be comparable unless some adjustment for the change which may have taken place in this time interval is made. Such adjustment will only be possible if some knowledge of the development pattern is available, for example, if some individuals who are measured at the "correct" time are remeasured at the same time as the non-respondents. If such a procedure is not feasible (and carrying this out for a small subsample in a large cohort study would present difficult problems) one may still be able to estimate the change over the period between the "correct" time and the time at which the late information is gathered, by using the estimates of change between the longer periods from occasion to occasion. Since the "non-respondents" may also have a different developmental pattern from the remainder this should also be taken account of, and individuals who return to a study after missing occasions may provide relevant information. (In studies such as the N.C.D.S., the method of tracing children means that a number of children can be expected to return to the study on subsequent occasions.)

Since individuals may be lost for different reasons, these will give rise to different kinds of bias associated with different variables. For the variables where the "lost" individuals differ from the remainder on the initial measuring occasion, a function of these variables, obtained for example, from a discriminant analysis, could be used to allow for bias. As yet, there seems to be little information on the nature of biases in longitudinal studies due to non-response.

Many of the variables which may be expected to be associated with "non-response", Social Class for example, are relatively unaffected over short periods of time and may be utilized in the conventional way to eliminate bias due to non-response. Other methods, involving, for example, building special questions into questionnaires can also be used. (See Cochran[12].)

Where biases are non-developmental, these may be partly allowed for by utilizing previous measurements to fill in the missing observations. Techniques for utilizing all the measurements on every occasion in a mixed longitudinal study to give efficient estimates at each occasion, are given by Patterson[13] and Gurney and Daly[14]. Where the missing individuals have different mean values from the remainder, these techniques can be used with an adjustment for a known bias. Gurney and Daly discuss the effects of two different bias patterns on the estimated values given by their procedures.

Editing and Processing Longitudinal Data

No data, collected on many different individuals by different measurers, are ever entirely free from error. Errors may occur either at source, for example by incorrect filling in of an assessment or interviewing schedule, during the transfer of information to a suitable medium for analysis such as punched cards, or in the analysis itself. Since any analysis is only as good as the data on which it is based, the preparation of comparatively error free data has a very important role in the data processing.

Assuming that a computer is available, errors may be checked in two ways. Firstly, most of the obvious errors can be eliminated by the use of skilled clerks to check questionnaires etc. and skilled machine operators to punch and verify cards or tape. Secondly, the computer can be used to scan the data. In large scale studies such as the N.C.D.S., involving about 10^7 items of information, one has to rely heavily on automatic computer editing procedures. Such schemes will be designed essentially to indicate "suspicious" values. The computer cannot, in most cases, take a final decision on the

validity of a particular value, which will then have to be examined and a decision made either to alter or to retain it.

Three kinds of editing procedure are available. Some variables will have a well defined range (e.g. a test score which must lie in the interval 0 to 10), and checks for values outside this can readily be made. Secondly, logical cross checks can be made for variables where the value of one variable is dependent on the response to another variable (e.g. if a child is attending a grammer school, then his age should be at least 11 years). Thirdly, there remain those variables where useful well defined limits for the range do not exist (e.g. Stature). For the scanning of large quantities of such data, the most practical scheme involves scanning and evaluating each individual "record" in turn. Limits may be set for individual variables or functions of several variables, under specified conditions (for example age and sex), within which nearly all the values may be expected to fall. These limits may either have been determined in previous studies, or, for example, by examination of a subsample from the current study. In longitudinal studies, there is information on change for individuals, and by using limits based on patterns of development we have a more powerful editing procedure than by using "cross sectional" data only, since limits can be assigned for each individual based on his or her own status and not simply as a random member of a population at a given occasion.

The resources available in any study for investigating "suspicious" values detected by the computer, are necessarily limited and it is important that an efficient scheme is used; that is, one which maximizes the ratio (for the "suspicious" values) of true errors to correct (but extreme) values. Using a simple "cross sectional" editing procedure on some selected body measurements in a carefully controlled study with about 200 children a value for this ratio of about 1:3 has been found[15], and can probably be considerably increased with more efficient longitudinal procedures. Once the general efficiency of a method for a variable or variables has been established, the human resources and the computer time available will determine the limits to be used.

A more detailed account of a longitudinal editing scheme, and a computer program for it is given by Goldstein[15].

There seems to have been little done on practical editing schemes, and one promising approach towards the development of efficient schemes may be by computer simulation, where distributions can be

generated with different types of "errors" and the performance of different editing procedures investigated.

Turning now to the general processing of longitudinal data, it becomes clear that the addition of a time dimension to the data calls for a rather different approach to that used when analysing cross sectional data. The added difficulty arises from the need to relate measurements from occasion to occasion within individuals. Where longitudinal analysis is carried out using punched cards and conventional punched card machines, such as a sorter, collater and reproducer, no essentially new problems arise, since relationships between occasions are obtained by transferring information from several cards onto a single one, a standard procedure whenever more than one card per individual is used. If a computer is used, such a card could of course be used for the analysis, but it will generally be more efficient to program the computer to sort, collate, and update data files.

There are five basic requirements of a computer program to handle longitudinal data. Firstly, a data file (e.g. on magnetic tape) must be set up. This will consist initially of a set of measurements on each of several occasions for each individual. It should be noted that the number of occasions may not be the same for each individual. Secondly (and this may be done at the time of setting up the data file), there should be facilities for editing the data file and defining further variables, e.g. as functions of the original variables which are measured. The full longitudinality of the data needs to be exploited here, both for editing and for defining new variables in terms of functions of variables measured on past (and future) occasions. Thirdly, as fresh data arrive, either from new individuals, or on new occasions for individuals already in the study, these must be edited, and the current data file updated. Fourthly, data may become available at some stage on further variables and have to be inserted into the existing file (e.g. a particular test may be scored only some time after a file has been set up, or X-rays of children may only be measured several years after having been taken). This implies a program for collating files. Finally, a program is needed to read data files and produce suitable statistical output. In practice, the output program is probably best limited to producing summary information which can be used as input to programs designed for special analyses, and it should again have facilities for defining new variables making use of the longitudinality of the data. In addition to the above,

useful features of such a program are the ability to handle multi-punched cards, and produce punched card and graphical output.

A set of programs[16] which meets nearly all of the above requirements has been developed in the Department of Growth and Development at the Institute of Child Health, London, for an I.B.M. 7094, and has been used mainly for the analysis of physical growth studies.

Some Results of Cohort Studies

In addition to describing development, large representative cohort studies also supply valuable cross sectional information. The first report of the N.C.D.S.[5] on the Cohort at 7 years consisted largely of cross sectional results. Information was obtained, for example, on the age at which formal arithmetic and phonics were begun in school for each of several large geographical regions. One conclusion, that about a quarter of children in the final term of infant school were rated by their teachers as poor or non-readers, has important implications for teaching methods. At seven years, head-teachers estimated that 8 per cent of children were in need of special help in school who were not receiving it, in addition to the 5 per cent already receiving it. The fact that the children in the Cohort are all the same age makes possible comparisons of such things as differential lengths of schooling on attainment, and it was found that those children in school for longer periods scored higher on tests (although allowance was made for social class in this analysis, other factors such as the staffing position in schools may be associated with early starting and the cause-effect relationship may not be a direct one). Comparisons between sexes indicated clearly that girls do better at reading, are socially better adjusted, and have fewer illnesses than boys.

These rather isolated results illustrate ways in which cohort studies can be used to monitor the population and provide national and regional cross sectional information.

The main function of longitudinal studies, however, and the basis on which they must be evaluated is in the study of development and some results from British cohort studies will now be discussed. For a detailed account of the results from some of the major American longitudinal studies the reader is referred to Bloom[17].

The 1946 Cohort has contributed a great deal to our knowledge of the relative effects of home and school environment on attainment. In "The Home and School"[4] the effects of the condition of the home, parental encouragement, academic record of the school and

streaming are examined in relation to social class and the age of the child. The interactions among these effects are shown to be complex, and the analysis is in some ways incomplete, but it does seem, for example, that children from working class homes who start with disadvantages such as poor housing and cultural background, fall further behind as they grow older.

Both the 1946 and 1958 perinatal studies have given unrivalled information on the influence of maternity factors on immediate outcomes of pregnancy as measured by birthweight, gestation and mortality. The 1946 Cohort has investigated the effect of prematurity on subsequent physical and mental development and concluded that, whereas these children are physically more vulnerable only up to about 2 years, they have a mental handicap which persists. The first report of the 1958 Perinatal Mortality Survey[9] analysed the effects of maternal factors on mortality rates, and was largely descriptive with no attempt at statistical hypothesis testing and it concentrated mainly on the effects of factors taken one at a time. The data, however, have proved so valuable, that 10 years after the original survey, further analyses of the material have been carried out[18]. One of these will illustrate some of the problems involved. This analyses the effects of parity, social class, age, hypertension, smoking, and maternal height on birth weight and mortality. It is found that while the partial effects of these factors on perinatal mortality (in a "main effects" linear model) are all significant, this is not so when the dependent variable is birthweight. The effect of social class on birthweight becomes non-significant when allowance is made for maternal height. Since maternal height is strongly associated with social class, it would appear that the effect of social class is operating largely through its effect on the size of the mother. It is also known that children of low birthweight tend to be mentally and physically backward (see, e.g. Illsley[19]), and a preliminary (unpublished) analysis of N.C.D.S. data indicates that although the partial effects of birthweight, gestation, and social class on test scores at 7 years are all statistically significant, the social class effect is by far the most important. These results, then, point to the importance of different influences at different stages of development, and more detailed analysis of the material should make these clearer.

In addition to attempting a theoretical understanding of the influences on development, the N.C.D.S. data is being used to evaluate the functioning and practicability of the "at risk" register[20]. One of the aims in the care of physically and mentally handicapped

children, is the detection of conditions in early life which predispose children to develop these handicaps. Much use has recently been made of the "at risk" concept and "at risk" registers of children with abnormal early life experiences (largely perinatal) have been set up by local authorities, so that these children may be given special attention.

There has been, to date, little research on the efficiencies of the various methods adopted. The most common criterion for placing a child on the register is whether he or she exhibits any one of up to about 40 conditions ranging from low birthweight to maternal hypertension. Beacuse of the interrelationships among all the conditions the efficiency of this approach is low, involving the inclusion of a large number of "normal" children in addition to the children who will become handicapped. (Rather surprisingly, occupation of the father is often not taken into account, although this is one of the most important single factors, even when only classified into manual and non-manual occupations.)

The prediction may be improved if, having defined a handicap outcome, a function of (perinatal) variables is derived which gives the most efficient prediction of which individual will have this handicap. In this way a more precise evaluation of the usefulness of such registers becomes possible. The question of causal relationships is irrelevant in this context so long as the prediction is reasonably stable. One of the aims of the N.C.D.S. is to carry out a study of handicap prediction and this will be an important section of its next report (now in preparation).

It is relevant at this point to make some remarks concerning social class. The most often used occupational classification, that of the Registrar General, was devised on the basis of occupational mortality. Such a classification into (usually) five or six groups has repeatedly been shown to be associated with physical and mental status. The "causal" relationships which give rise to these differences are complex and not very well understood, and in analysing the interrelationships of several variables including social class, there are two possible ways of regarding social class.

On the one hand, one may attempt to isolate the variables which give rise to social class differences, as, for example, in the effect of maternal factors on birthweight, where it appears that social class has its effect through its association with maternal size. On the other hand one may regard social class as a "nuisance" variable and attempt

to partial it out or otherwise allow for it. This second approach is aimed at discovering relationships which are present not because of the all pervading influence of social class (or some other measure of socio-economic environment), but because of independently acting biological mechanisms, which would continue to operate in a constant social environment. One may question whether this approach is a useful one in the study of human populations, and maintain that what are really interesting are the interactions of variables with the social environment since, in practice, it is never really possible completely to eliminate the environment. Assuming, however, that one wishes to allow for social environment, one may take account of, in addition to the basic occupational group classification, such things as housing and neighbourhood conditions, size of family, etc. As Douglas has shown (see above), social environment is associated with intellectual development. His analysis was not concerned primarily with allowing for social environment, but with investigating its effect on change. If one wished to investigate the effects of further variables on intellectual development, which were also associated with social environment, one would probably wish to make allowance for the "nuisance" relationship of intellectual development with social class.

So far, in studies of mental development, attention has been given mainly to answering specific questions about individual stages in development, rather than to describing the general developmental patterns. Bayley[21], for example, in a very carefully controlled study is concerned with the age to age correlations between test results, and the prediction of adult intelligence.

In physical growth studies, on the other hand, when analysing patterns of development, the importance of studying individual growth curves, as distinct from average populations curves, has long been recognized. It is known that individuals enter the various stages of development at different ages and the population average will therefore tend to smooth out the interesting details (see, e.g. Tanner[8]). It is true, of course, that in studying mental development, difficulties arise from having, in many cases, to use different measuring instruments at different stages of development. A further complication is introduced by the relatively large "measuring error" involved in mental testing.

In the following sections, some of these problems and techniques for handling them will be discussed.

Statistical Analysis

Longitudinal Studies are usually concerned with answering two related kinds of question. The one kind deals with the nature of the change between occasions in measured variables, and with comparisons of these changes between groups of individuals. The other kind is concerned with the relationship between variables at one occasion and the same or different variables at other occasions.

The second kind of question is illustrated by Bayley's study[21] of Adult–Child correlations, and by the "at risk" prediction problem discussed in the previous section. Here, a single variable at one occasion is related to a function of one or more variables at another occasion, and the analysis involves univariate regression and correlation techniques. In the analysis of neonatal mortality rates using the 1958 survey data[18], the dependent variable was taken to be the proportion of perinatal deaths and a logit transformation of this was related to a linear function of selected maternal factors. An alternative approach to a similar problem is illustrated by Berendes *et al.*[23] who used a discriminant function of maternal variables to classify each baby into a perinatal death or a survivor. These approaches may be generalized to the case where a function of several variables at one occasion is related to a function of several variables at another occasion, and here multivariate regression or canonical analysis could be useful; but little use of these techniques seems to have been made in longitudinal studies.

In classical regression analysis, the independent variables are assumed to be free from measurement error. Many mental measurements are, however, subject to quite large errors of measurement and where such variables are used in regression or correlation analysis, some adjustment for this may be necessary. Observed correlation and regression coefficients will tend to be reduced when there are errors in the measurements of the independent variables (see, e.g. Lord[22] and Kendall[24]). In these cases, failure to take account of errors of measurement may not have very serious consequences. When analysing the change in a variable over occasions however, failure to take account of these errors may easily lead to false conclusions, and this situation will now be discussed in more detail.

Regression Effect

Many of the difficulties associated with analysing the change in a variable from one occasion to another, arise when the variable is "fallible", i.e. it is subject to relatively large errors of measurement.

These difficulties are illustrated by the so called "regression effect".

Consider a test, administered to a sample of individuals, such that each individual obtains a single score on the test. If a similar test designed to measure the same abilities, is administered later, and if there are no carry over effects from the first test, and the ability being measured has undergone no change in the time between the tests, then each individual can be expected to obtain very similar scores (properly scaled) on the two tests. If one accepts the concept of a true underlying ability of which the score is an unbiased estimate, then the two values obtained for each individual will give an estimate of the errors of measurement of the test, which is the difference between the observed score and the (hypothetical) true value. This may be written as,

$$x = X + \varepsilon \qquad \text{—(1)}$$

Where X is the true value, x the observed value and ε a random measurement error with expectation zero. If two measurements x_1, x_2 with errors ε_1, ε_2 are made on an individual, then the expected values of both ε_1 and ε_2 are zero, and ε_1 and ε_2 are independent.

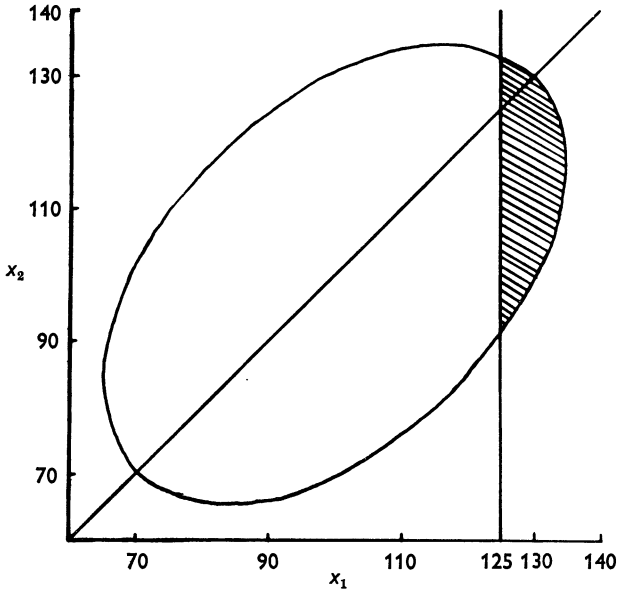
Consider now the distribution of scores for all individuals on the first test, and in particular consider those individuals with an observed score above a chosen fixed value Z. These individuals can be divided into two groups, those with true values less than Z and those with true values greater than Z. The second group of individuals will on average contain a higher proportion of positive errors of measurement than negative ones, since the individuals in the sample with true values greater than Z will, on average, have an equal number of positive and negative errors of measurement and some of those with negative errors will actually have observed scores less than Z, thus leaving an excess of positive errors in the group of individuals who have both observed and true values greater than Z. Since this group of individuals will on average have zero errors of measurement on retesting, their mean observed score will tend to be lower on the second test. Similarly, the first group of individuals, all of whom have positive errors of measurement on the first test, will also tend to have a lower mean score on the second test. The net effect therefore, is that those individuals with an observed score on the first test greater than a fixed value will tend to have a lower mean score on the second test than on the first test.

Although discussed above in terms of measurement errors, this regression effect may be seen acting in any population which is in

"dynamic equilibrium", where the distribution of the measurement in the population is the same on each occasion, but individuals within the population may change their values; individuals with large values on one occasion tending to have smaller values on a second occasion, and vice versa.

The following diagram summarizes the situation.

Diagram 2



The ellipse represents a sample scatterplot of scores obtained on the two tests discussed above. The line $x_1 = x_2$ represents the relation between the true values on the two tests. If one considers all individuals with a score greater than 125 (Z) on the first test (shaded area), it is clear that their mean score on the second test is less than 125. A similar effect takes place at the lower end of the scale. Although this diagram may not be entirely realistic (for example all the individuals scoring over 130 on the first test score under 130 on the second test) it can be thought of as representing, say, a bivariate normal distribution, and the above remarks can be expressed in terms of the conditional distribution of x_2, given $x_1 > 125$.

It should also be noted that the observed mean change in score of a chosen extreme group, will be greater the further away from the

population mean that the value of Z is chosen. This fact will be used to illustrate how a failure to take account of the regression effect can lead to wrong conclusions.

Suppose that a sample consists of two groups of children, those whose fathers have non-manual occupations, and those whose fathers have manual occupations. Assume further that the children of fathers in manual occupations have the same bivariate distribution (in the above two-test situation) as those children of fathers in non-manual occupations, except for a downward shift in location. The scatterplot is now represented by two congruent overlapping ellipses along the line $x_1 = x_2$. If for example, the mean of the manual group is 96 and that of the non-manual group is 104, and $Z = 125$ is chosen as before, then the "manual" children scoring above this value are further away from the mean of their own distribution than the "non-manual" children. Hence the mean drop in score from the first to the second occasion for these manual children will be greater than for these non-manual children. The opposite is true at the lower end of the scale. The analysis of extreme groups of children would therefore lead to the false conclusions that the deterioration in scores for high scoring manual children is greater than for high scoring non-manual children, and the improvement in score for low scoring manual children is less than for low scoring non-manual children. An analysis of the mean change in scores of all children would show no difference between the two groups.

In the above example, individuals were divided into two groups on the basis of a variable which was independent of the errors of measurement. Using equation (1) an unbiased estimate of the mean change in true score $\bar{X}_1 - \bar{X}_2$, is given by $\bar{x}_1 - \bar{x}_2$. Comparisons between groups may therefore be made on the basis of observed changes.

Referring to equation (1), one may ask whether it is possible to obtain a more efficient estimate than $x_1 - x_2$ for the change $X_1 - X_2$ in an individual score? Given a sample of individuals and assuming that the change in true score is a function of the change in observed score, a more efficient estimate of true change can be found. The simplest case is where a simple linear relationship between observed and true scores is assumed, and Lord[26] discusses this and alternative models. It should be noted however that the estimate depends on the function chosen, and if, for example, a quadratic relationship is assumed, different estimates would be obtained, and these would not necessarily maintain the same rank order among individuals.

Changes over Several Occasions

The above analysis of mean score change may be described in terms of the analysis of variance, and in this case is the usual two-way mixed model with one observation per cell. The fixed effect (A) is the occasion at two levels, and the random effect (B) is the individual. The "paired comparison" t-test of the mean change in score is equivalent to the comparison of A with the AB interaction.

If, retaining the mixed model design, the number of occasions is increased, then the sum of squares for occasions and for the interaction between occasions and subjects may be partitioned into quadratic and higher trend terms (see e.g. Scheffe[27]). Thus, if the usual analysis of variance assumptions are satisfied, this approach may be used to study linear, quadratic, etc. time trends. The aim of this approach is to summarize the change over several occasions by a small number of polynomial coefficients and methods which use this approach will be discussed.

Consider the following matrix which is the variance covariance matrix of responses for a design where each individual is measured on four occasions.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

The usual analysis of variance assumptions state that the variances at each occasion, which are the diagonal terms of this matrix, are equal and the covariances, which are the off-diagonal terms, are zero. The presence of correlation between occasions in a repeated measurement design violates the second of these assumptions, and the general case which leads to a multivariate analysis is discussed below. The mixed model case where the covariances are all equal, although not zero, may be handled by univariate procedures (see Scheffe[27]), but this assumption is unlikely to be true in growth studies, where the correlation between occasions far apart in time will usually be smaller than that between occasions closer together.

In the simplest case with just two occasions, the covariance matrix has only one off-diagonal term, so that the usual univariate analysis applies.

In the multivariate case where there are p occasions the model may be written,

$$y = A\beta + \varepsilon$$

where y is a $p \times 1$ column vector of mean responses at each occasion, A is a $p \times m$ design matrix, β is a $m \times 1$ column vector of unknown coefficients, and ε is a $p \times 1$ column vector of errors which are assumed to follow a multivariate normal distribution with covariance matrix Σ. The coefficients β are now fixed effects related to the occasions, and individuals are treated as replicates. For example, if the response is linearly related to time, then $m = 2$ and we have for the r^{th} occasion at time t_r

$$y_r = \beta_0 + \beta_1 t_r + \varepsilon_r$$

The coefficient β_1 can be estimated and the adequacy of the model tested (in the above case whether a straight line is sufficient to describe growth) using the methods of Rao[29]. The above example may be readily extended to the comparison of different groups, and a discussion of this with a worked example, in which different tests are used at each occasion, is given by Bock[22].

The use of univariate procedures in the general case has been studied and compared with multivariate procedures. Elston and Grizzle[28] make a comparison of two univariate models with a multivariate one in the analysis of a set of growth measurements. A straight line growth curve is fitted for each model, and a test made for the adequacy of a straight line to describe the growth. The first model due to Rao[29], assumes that the errors follow a multivariate normal distribution with no restriction on the matrix Σ. The second model assumes the complete independence of the responses for a single individual and is the usual fixed effects analysis of variance model, the off-diagonal terms of Σ being zero. The third model is the mixed model described above. The authors conclude that whereas the first and third models give similar confidence bands and similar results for tests of significance, method two gives rather different results. Estimates of the parameters of the straight line growth curve are similar for all three methods.

Danford *et al*[30] discuss the problems involved in repeated measurement designs and compare the univariate model with a multivariate one and come to conclusions which are similar to those of Elston and Grizzle. A fuller discussion of the use of univariate procedures is given by Gaito and Wiley[22].

More recently Rao[31] has considered a multivariate extension of a univariate mixed model for growth data, and provides a test of significance for the adequacy of a chosen degress polynomial to describe the growth curve. Rao's model includes models 1 and 3 of

Elston and Grizzle as special cases, and re-analysing their data he shows that a slight narrowing of the confidence intervals for the parameters can be made using a high order polynomial coefficient as a concomitant variable to the assumed linear trend.

In the analysis of growth curves it is convenient to describe the response over time by using orthogonal polynomials. If the p measurements on each individual are replaced by p orthogonal polynomial regression coefficients, then without loss of information, these may be used instead of the original mesurements. Wishart[32] first used this approach, taking each coefficient in turn, starting with the zero order coefficient, to investigate treatment differences for mean, linear, quadratic, etc. effects in separate univariate analyses. A systematic approach to the estimation and use of orthogonal polynomials in growth data is described by Rao[31].

Missing Data

In the above discussion, it has been assumed that all individuals are measured at the same time points. In a few longitudinal studies this may be the case but, in general, individuals will miss or will be late for particular occasions. This problem does not seem to have been explicitly discussed with reference to the above procedures, but it would seem reasonable to fit orthogonal polynomials to each individual's measurements and to utilize these coefficients as if the measurements had been taken at the same time points. This would not lead to the covariance matrix Σ because the covariances will depend in general on the time interval between occasions, and since the number of occasions determines the order of the curve that can be fitted, the analysis will also be limited by the individuals with the smallest number of occasions. Where the number of occasions varies little between individuals and where the variability of the time of measurement at each occasion is small, this approach is probably acceptable, but the effects of this procedure on the results of the analyses need to be properly examined.

This discussion of statistical problems is not intended to be exhaustive, but to illustrate some of the difficulties involved in, and some of the suggested ways of dealing with, the analysis of repeated measurement designs. The interested reader is referred to the proceedings of the Madison conference on "Problems in the Measurement of Change"[22] for a discussion of other topics, including the use of factor analysis for describing patterns of development.

In conclusion, I would like to stress that this article has concentrated mainly on those aspects of the subject with which the author is most familiar and finds most interesting. Space has not allowed the more extensive coverage which ought to have been given to many of the topics dealt with, and in particular to the problems of statistical analysis. It is hoped however that the list of references will provide a useful means for pursuing these in greater depth.

Acknowledgements

I would like to thank the following individuals for their helpful comments during the preparation of this article: Professor J. M. Tanner, Professor M. J. R. Healy, Dr. M. Kellmer Pringle, Mr. R. Davie, Dr. P. Levy and Miss B. Collinge. I would also like to thank the Co-directors of the National Child Development Study for permission to quote hitherto unpublished material.

REFERENCES

1. "Colloquium on Longitudinal Studies". (1965), National Institute of Child Health and Human Development, U.S.A.
2. "Maternity in Great Britain". (1948) O.U.P.
3. DOUGLAS, J. W. B. and BLOMFIELD, J. M. (1958). "Children Under Five". George Allen and Unwin, London.
4. DOUGLAS, J. W. B. (1964). "The Home and the School". MacGibbon and Kee, London.
5. KELLMER PRINGLE, M. L., BUTLER, N. R. and DAVIE, R. (1966). "11,000 Seven Year Olds". Longmans, London.
6. "Children and their Primary Schools". Report of the Central Advisory Council for Education (England). (1967) Vol. 2. Appendix 10. H.M.S.O. (This is identical with (5) but without the medical information.)
7. SCHAIE, K. W. (1965). "A general model for the study of Developmental problems". *Psychological Bulletin*, **64**, No. 2, 92.
8. TANNER, J. M. (1962). "Growth at Adolescence". Blackwell, Oxford.
9. BUTLER, N. R. and BONHAM, D. G. (1963). "Perinatal Mortality". Livingstone, Edinburgh and London.
10. PIDGEON, D. A. (1965). "Date of Birth and Scholastic Performance". *Educational Research*. Vol. VIII, No. 1. (National Foundation for Educational Research.)
11. FALKNER, F. (ed.) (1960). "Child Development. An International Method of Study". Vol. V. *Modern Problems in Pediatrics*, Karger, Basel.
12. COCHRAN, W. G. (1963). "Sampling Techniques". Wiley, New York.
13. PATTERSON, H. D. (1950). "Sampling on Successive Occasions with Partial Replacement of Units". *J. Roy. Stat. Soc.*, **B. 12**, 241.

14. GURNEY, M. and DALY, J. F. (1965). "A Multivariate Approach to Estimation in Periodic Sample Surveys". *Proc. Social Stats. Sect. of Amer. Stat. Assn.*, 242.
15. GOLDSTEIN, H. (1968). "The detection of errors in data from longitudinal studies". "*Proceedings of Annual Reunion of Child Growth Studies, Brussels*, Feb. 1968". Centre International de L'enfance, Paris.
16. GOLDSTEIN, H. and MANNING, M. (1968). "Longitudinal Survey Program". Dept. of Growth and Development, Institute of Child Health, University of London.
17. BLOOM, B. S. (1964). "Stability and Change in Human Characteristics". Wiley, New York.
18. BUTLER, N. R. and ALBERMAN, E. (1968). "Second Report of the Perinatal Mortality Survey". Livingstone, Edinburgh and London.
19. ILLSLEY, R. (1966). "Early Prediction of Perinatal Risk". *Proc. Roy. Soc. Med.*, **59,** 181.
20. SHERIDAN, M. D. (1962). "Infants at Risk of Handicapping Conditions". *Ministry of Health Monthly Bulletin*, **21,** 238.
21. BAYLEY, N. (1949). "Consistency and Variability in the Growth of Intelligence from Birth to 18 years". *J. Gen. Psychol.*, **75,** 165.
22. HARRIS, G. W. (ed.) (1963). "Problems in Measuring Change". University of Wisconsin Press, Madison.
23. BERENDES, H. W. *et al.* (1965). "Factors Associated with Breech Delivery". *Am. Jour. Pub. Hlth.*, **55,** 708.
24. KENDALL, M. G. and STUART, A. (1961). "The Advanced Theory of Statistics". Vol. 2, Chap. 29. Griffin, London.
25. FISHER, R. A. (1935). "The design of Experiments". Oliver and Boyd, Edinburgh and London.
26. LORD, F. M. (1959). "Statistical Inferences about True Scores". *Psychometrika*, **24,** 1.
27. SCHEFFE, H. (1959). "The Analysis of Variance". Wiley, New York.
28. ELSTON, R. C. and GRIZZLE, J. E. (1962). "Estimation of Time-Response Curves and Their Confidence Bands". *Biometrics*, **18,** 148.
29. RAO, C. R. (1959). "Some Problems Involving Linear Hypotheses in Multivariate Analysis". *Biometrika*, **46,** 49.
30. DANFORD, M. B., HUGHES, H. M. and MCNEE, R. C. (1960). "On the analysis of replicated measurements experiments". *Biometrics*, **16,** 547.
31. RAO, C. R. (1965). "The Theory of Least Squares when the Parameters are Stochastic and its Application to the Analysis of Growth Curves". *Biometrika*, **52,** 447.
32. WISHART, G. (1938). "Growth Rate Determination in Nutrition Studies with the Bacon Pit, and their Analysis". *Biometrika*, **30,** 16.