QUALITY OF OBSERVATIONAL DATA

BY A. R. KAMAT

Gokhale Institute of Politics and Economics, Poona

[Reproduced here are abstracts from the Address of Professor A.R. Kamat, delivered as President of the Section of Statistics at the 56th Session of the Indian Science Congress held in Bombay early this year. The subject of quality of data has been exercising the minds of statisticians increasingly and it is hoped that these abstracts will be found stimulating.—Editor.]

I

The last three decades have witnessed rapid advance in statistical theory and statistical techniques : multivariate analysis, construction and analysis of experimental designs, sampling theory and applications, sequential analysis, linear programming, information theory, queuing theory, spectral analysis of time series, to name only a few of the recent developments. Simultaneously with the growth of techniques the foundations of statistics are also being critically examined as is evidenced by the discussions on the general theory of inference and the theory of finite sampling in particular. All this churning of thought over the basic formulations is in a sense inevitable. During the last thirty years the use of statistical methods both at primitive as well as highly sophisticated level, has spread to diverse realms of knowledge, realms even like political science where the application of statistical methods was unheard of before (see *e.g.* Deutch, 1963). It is true that the statistician is not yet 'admired' nor 'understood' by the poet who therefore classifies him with

> .... the bat
> holding on upside down or in quest of something to eat,
> elephants pushing, a wild horse taking a roll,
>   a tireless wolf under
> a tree, the immovable critic twitching his skin
>   like a horse that feels a flea, the base-
> ball fan......   (Marianne Moore in *Poetry*)

But even when the poets would like to keep them at an arm's length the statisticians have not left them alone and are using their techniques to investigate the claims of disputed authorships !

The burst of knowledge is not a special feature of statistics alone but is a part of the accelerated progress in all sciences, sometimes described as the scientific explosion, and is one of its necessary consequences. The rapid progress in other fields of learning have in their turn confronted statisticians with altogether novel situations challenging the ingenuity of applied statistics and the analytical and integrating faculties of the theorisers in statistics. The advent of high-speed electronic computers on the scene has added still another dimension to the theory and practice of statistics. The computer era has just started but this new technology has already made its presence felt and it has compelled statisticians to modify considerably their techniques and to reshuffle the order of importance conventionally attached to different elements in the statistical methodology. All these advances have forced statisticians both pure and applied to rethink and readjust. It is well-known that this process of readjustment is taking place in all branches of science. And statistics is no exception. For the central problem of knowledge, as always, is the problem of the growth of knowledge, and when a scientist speaks about the growth of knowledge he means the growth aud advance of scientific knowledge (Popper 1959).

The growth and advance in statistical theory and statistical techniques are no doubt fantastic. Certainly it is most welcome. But all these advanced techniques have ultimately to operate on observational data. Observational data are at the very basis of all statistical analysis. It is the actual raw material input in every statistical process. No refinements in techniques and theoretical, methodological, technical or technological advances, however powerful and efficient, can hope to compensate in a fundamental manner for the shortcomings of this basic raw material. To give an illustration, the solution of a system of simultaneous equations to a desired degree of accuracy has become a very simple operation on an electronic computer. But small changes in the coefficients produce large deviations in the solution when the system is *ill-conditioned*, that is, when the determinant of the coefficient matrix is small in magnitude in comparison with certain of its cofactors (Hildebrand, 1956). This will be clear from the solutions of the following two systems :

$$x-y=1, \quad x-1.001\,y=0 \qquad \dots \quad (1)$$

$$x-y=1, \quad x-0.999\,y=0 \qquad \dots \quad (2)$$

which have solution sets $x = 1001$, $y = 1000$ and $x = -999$, $y = -1000$ respectively. (See : Milne, 1949.) In many of our problems where highly sophisticated mathematical models are used the coefficients are very often estimated from observational data. Although the example given above is purposefully extreme, it illustrates the great distortions which may arise from even minute inaccuracies of data. Although such critical situations occur only very infrequently they indicate that it is always necessary and desirable to have a closer look at the type of raw data which statisticians collect, handle or interpret, and to consider in general the problems of the accuracy of statistical data and their improvement. It will be my attempt to discuss these problems in this paper in which, apart from general formulations, the discussion will centre round the quality of Indian data in the socio-economic field.

II

The scientific progress is the result of our belief that the world exists and is knowable and that it is both interesting and useful to seek relationships and, if possible, causal relationships between different phenomena. Most philosophers and scientific thinkers agree that the development of scientific knowledge takes place in three broad stages : collection of facts or empirical data, formulation of a theory or hypothesis by which they can be understood or explained, and further development of the theory to deduce or predict consequences which can be empirically verified. Thus all our scientific endeavour, (perhaps with the exception of pure mathematics,) whether it is the study of physical and biological phenomena or of human activities, is based on induction, on empirical facts ; and we can be aware of facts only through observation, by the use of our five senses. Theorising and the use of deductive reasoning including the use of mathematical models come later. Even in the case of pure mathematics which is supposed to be governed by purely internal criteria such as consistency, economy of assumptions, attempts to generalise and aesthetic considerations, it should be remembered that Gödel (1931) has knocked out consistency as a purely internal criterion. Moreover one of the greatest mathematical minds of this century holds that "it is a relatively good approximation to truth...that mathematical ideas originate in empirics, although the genealogy is sometimes long and obscure." (Von Neumann, 1947.)

All theories or models used in science are constraints imposed on the empirical reality by the theoriser. In other words, we never *know* anything absolutely or finally, we can only guess and continue to do so. A regression relationship, for instance, does not provide an *explanation* ; it is simply a mathematical representation of the regularity of observations. Einstein and most other scientific thinkers agree that propositions about the physical world cannot be *proved* by mathematical models or reasoning. As regards social sciences, and economics in particular, Barbara Wooten (1950) underscores the same view when she endorses Ritchie's remark that the attempt to deduce laws from definitions "may possibly be pure mathematics but it is more likely to be pure nonsense", and Oscar Lange (1945-46) reiterates it when he simply states that "economic theory is empirical science."

What is said above applies to science in general, to both natural and social sciences, to the study of physical universe as well as to that of human activities. Thus the methodological problems faced by social scientists are not essentially different from those which canfront other scientists. This is not to say, however, that they are not more complicated, because in many respects they *are* more complicated, and this poses a number of questions, peculiar to the social sciences, in the collection, analysis and interpretation of empirical data.

××× ...... ××× ...... ×××

III

Let me now turn to the quality of observational data which one has to deal in the scientific study of human affairs. This is important not only because such data are the only method of knowing or determining the directions of social change but also because they are taken as the basis of policy-making in a modern state. Most sophisticated mathematical models nowadays used in social sciences have also ultimately to lean heavily on the aggregated empirical data. The observational data in social sciences are likely to suffer much more from errors than those in natural sciences and yet the realization of this fact is much more recent. Although statisticians and social scientists have become aware of errors in social surveys or of errors of nonresponse ever since the large-secale sample surveys began to be used (*e.g.* Deming, 1944 ; Mahalonobis, 1944) it is only during the last ten or fifteen years that a systematic and comprehensive treatment is being attempted, as for instance by Morgenstern (1950 and 1963) who has discussed the problem as it concerns economics and by Zarkovich (1966) who has done it from the standpoint of statistics.

Whether the empirical data are *satisfactory* depends not so much on their intrinsic accuracy as on use to which they are put, and on the decisions which are to be taken by their use. Some data may be of a quality high enough to allow the user to build more or less accurate summing up of the status. Some others may fall short of this standard but may provide a more or less accurate picture of the changes or differences in time or space. And some others, perhaps quite a large number, may be good for neither of the above, but may be useful for giving only a dimensional idea about the phenomena or indicate other broad structure and tendencies. Again some data may have a degree of accuracy high enough to allow its use in refined mathematical models but some others cannot be used for analysis which is sensitive to errors in data and consequently only robust mathematical models can be used with them.

How does one determine the accuracy or quality of empirical data ? An error is the deviation of the actually observed value from the true value. But like Pontius Pilate, but not cynically nor skeptically, one may ask : what is truth ? (St. John, 18, 38.) What is a true value ? The true value is that value which would be observed if the observer had followed strictly the prescribed concepts, definitions and procedures and evoked a faithful response from the subject about the phenomenon under observation. I am aware this definition suffers to a certain extent from solipsis and one may raise the question whether the true value can ever be observed in natural or social phenomena, especially the latter, where the observational procedure is affected inevitably because of the observer, the observed, their interaction as well as the very process of observation. But every scientist has to live with this problem and we shall not enter into its discussion here.

Now the observed value will often differ from the true value and in natural sciences where the experiment can be repeated in almost identical, that is, controlled circumtances the variability of the repeated observations indicates the margin of error. There, it is a fairly established practice to report this error while reporting the empirical data. Moreover the producer of data in natural sciences is very often also their user and even when he is not the same person the user is aware of the existence of such error and very often also of its magnitude.

The situation in the social sciences is entirely different. Here repetition of observations in identical or almost identical circumstances is difficult. Errors due to the subjective element in the observer, the observed and their interaction are likely to be quite substantial, and while errors from the first and the third sources can be reduced by following closely the prescribed procedure of observation, errors from the second source, the response of the observed, may not be amenable to any control. In the observation of social phenomena, unlike that of natural phenomena, the observer may have to face not only indifference on the part of the observed but also counterwork or positive hostility. Men can and often do lie ; nature does not, or very rarely does. Nature at the most misleads, but there is hardly any motivation in this misleading. It is not also true, although it is a common misconception, that errors at various stages, or from different sources at the same stage, always cancel out. They often accumulate ! This does not mean that satisfactory empirical data can never be obtained in social seiences. It only means that the problem of ascertaining the error is much more difficult in the study of human affairs. But the problem is essentially the same and the method of investigating into them are also the same as in the natural sciences : repetition of observations wherever possible in more or less similar circumstances, comparison with similar or related data and study of internal discrepancies and inconsistencies.

Whatever the difficulties it is absolutely essential while presenting socio-economic data to give some idea of the margin of error involved, if possible explicitly, and if that is not possible, at least indirectly by describing in detail the manner of its collection and by pointing out its shortcomings, imperfections and incompleteness. In fact it is desirable to give thought to the possible errors of observations and their relative importance before one undertakes large-scale data collection. Unreliable statistics should not see the light of the day. As everywhere else, in statistics also, it is better to keep quiet than say wrong things and mislead. If a physicist or chemist does not think it *in fradig* to say 'I do not know' when he does not know, why should the statistician not do the same ? But one rarely finds this wisdom heeded in the cartloads of figures which are uncritically unloaded on the innocent users day after day, month after month, year after year. On the contrary often there is an attempt at displaying spurious accuracy by quoting figures to the last digit or decimal that is collected or calculated, even when one knows that they are of doubtful reliability. Perhaps it has some justification in census reports where there has been actual counting of heads. But how does one justify the projection of India's population or industrial production or exports to the last head or the last ton, based on broad assumptions of uncertain validity ? If one could save the expenses of producing, processing, printing and computing unnecessary digits of basically doubtful statistics, it should be possible from such saving to finance a great deal of research in social sciences and statistics, and especially in the problems of improving the quality of data.

As a first step, it is necessary to cultivate 'error consciousness' among those who produce, publicise or utilize observational data. It should be insisted that it is necessary in any given project of collection of data to investigate into the likely sources of error, so as to avoid gross blunders and to control the inevitable error and attempts must be made as a matter of course to estimate the magnitude of the resulting errors. If such estimates of error are given wide publicity that will not only warn the users against the unwarranted and indiscriminate use of the data material but the ensuing discussion will also stimulate future researchers in similar projects to try and reduce the error further.

At the same time it is also necassary to cultivate error consciousness among the lay users, politicians and public men. They must be frankly told that no data are free from error. This is not the same thing as the glib mouthing of the common denigration of statistics and data collection that statistics are all lies of various degrees and that statistics can be collected or calculated, that is, fabricated, and statistical techniques can be used, that is, misused, to support any hypothesis or policy, although statisticians have sometimes, through their own doings, laid themselves bare to this charge (Nandi, 1968). But it means that all data suffer from inaccuracy and are bound to suffer. The question is that of ascertaining the degree of inaccuracy from which they suffer and the point is that of minimising inaccuracy and of making due allowance for it when inferences or policy formulations are based on them.

As mentioned earlier research workers have now started probing into the quality of the data, to detect and discuss the various sources of errors and to attempt to quantify their magnitudes. In-spite of its fundamental importance, however, the 'error consciousness' is still far from adequate. It should be clarified, for convenience, that the concept of error in observational data is different from the concept of error as used in statistics. While the latter is a technical concept and denotes the error due to sampling, the former is much wider and is more a common-sense concept, the inevitable deviation of the observed value from the true value, a deviation which may be due to many other factors mentioned above besides fluctuations due to random sampling. One of the earliest papers to systematise thinking in this direction is that by Deming (1944) wherein he lists thirteen different factors which affect the ultimate usefulness of the data collected in a survey. During the last twentyfive years and especially during the last fifteen, with the wide-spread use of large-scale surveys for initiating and evaluating state policies firmly established, the subject is assuming great importance in many countries. In advanced countries the study of errors in survey data has developed into a specialized branch in survey technology, sampling and non-response errors forming a part of such study. It has not been possible so far, nor does it seem likely in future, because of the very nature of the problem, to evolve a comprehensive statistical theory of errors in observational data to cover all types of human activities. But various aspects of the different sources of error that affect the data material are being studied and attempts are being made to estimate their direction as well as their likely magnitudes. For instance, to name only a few of its important aspects, valuable work has been done concerning interviewer variability, the most suitable period for observation, reference and recall, the memory factor, the errors due to instruments of survey such as the questionnaire, the schedule and the type of questions asked, and several other questions related to errors in observational data, as will be seen from the following few selected references. (Mauldin and Marks, 1950; Moser 1959; Gray, 1955 ; Gales and Kendall, 1957 ; Hanson and Marks, 1958; Jaeger and Pennock, 1961; Turner, 1961; Belson and Duncan, 1962 ; Coale and Stephan, 1962; Kish, 1962; Neter and Waksberg, 1964; Bailer, 1968; etc.).

In India even after the adoption of countrywide large-scale sample inquiries like the agricultural labour inquiries, rural credit surveys, crop-estimate surveys and, of course, the National Sample Surveys (NSS), the efforts in this field have been relatively rather limited. The most important amongst them are those by Mahalanobis and his colleagues in the Indian Statistical Institute (ISI), the researchers of the Institute of Agricultural Research Statistics (IARS) and very recently a few of us in the Gokhale Institute. They also include a few foreign scholars who have carried out investigations in this country. But it cannot be said that Indian statisticians and survey methodologists have earnestly applied their minds to this problem, commensurate to the extent that they are using the largescale sample surveys in this country.

IV

In the remaining paper, I shall briefly indicate and discuss some aspects of the accuracy of Indian data and their improvement. I shall first give a few illustrations of the difficulties with which one is confronted when one proposes to use Indian official data.

The spread of literacy is one of the directive principles of the Indian constitution and the one (and perhaps the only) instrument of measuring its progress is the decennial census. The definition of literacy and instructions in respect of this item have remained almost the same since 1911. According to the current census definition (Census of India 1951, 1961) literacy is defined to mean the ability to read and to write a simple letter. It is quite clear that in the hurried large-scale operation of the population census this cannot be tested in the many doubtful cases which are bound to arise with the result that all adults who assert that they are literate are probably recorded as such. While this is understandable it raises a difficulty in the case of school-going children in the first two or three standards, whether and whom to include. In some of our recent investigations (Kamat 1967; 1968a, 1968b) it was found that an arbitrary number from amongst them is included as literates in the census figures and for some villages in Maharashtra it led to the curious result that literacy decreased from 1951 to 1961 when in fact it is well-known that literacy and education have progressed considerably all over the State during this period. It is true that not all those who are in school can be strictly classified as literates since there is a sizable lápse into illiteracy of the early leavers from the primary school, and there is a fairly large number of such drop-outs On the other hand it is clear that with a definition such as the one above and its arbitrary application by the enumerators the final figures not only suffer from inaccuracy but it is also difficult to quantify the error that enters into them.

For planned industrial development, and even otherwise, it is of fundamental importance to collect factual information about industrial growth and industrial production in India quickly and efficiently and also to publish it without much time-lag. The two major sources in this respect are the Annual Survey of Industries (ASI) and the Monthly Statistics of Production of Selected Industries of India (MSP). The ASI has two sectors, the census sector which covers all big establishments divided into some 230 industries and the sample sector which covers, on a sample basis, the medium establishments grouped under about 90 industries. It gives annual production in physical quantities and also in value terms, and other information about industrial production. But the ASI covers only the registered factories registered under factory legislation, the smaller unregistered establishments being left out of its coverage. The summary figures see the light of the day two years after the reference year and the detailed survey is published four years after.

The MSP is based on the monthly figures obtained, mostly on voluntary basis, from the factories that are listed with the Directorate General of Technical Development (DGTD) and about a dozen other government agencies. This coverage is different from that of the ASI and does not claim to cover all those enterprises covered by the ASI. The MSP gives production figures in physical units, but not in value terms, for about 325 industrial products; and it gives the series of monthly index of production for certain groups of products with 1960 as the base year. It also provides figures for monthly capacity of production. The monthly index series is published within about three months of the reference month and the complete details regarding production are out within six to eight months.

That this is not a very happy state of affairs is admitted by all concerned. The ASI figures leave out all unregistered establishments, do not give all relevant details (such as capacity), and are inconvenient for constructing an annual series of index of production since the classification is by enterprises and not by products. Moreover they suffer from a long period of delay in publication. On the other hand the coverage of the MSP cannot be considered as satisfactory because it is confined only to those establishments which are listed with the DGTD or particular government agencies and the returns collected are at least partly on a voluntary basis, with the result that one cannot say how representative it is for a particular industry or a particular product. And yet one has to rely on the index series supplied by the MSP for considering the growth of industrial production.

It is high time that this country should have regular collection and publication of industrial production based on adequate coverage spread over all sectors of industry. Along with the census and sample sectors of the present ASI it is desirable to have a quinquennial sample census of unregistered enterprises. The information should give at least the number of units in the industry, broad details of different products produced in physical quantities as well as in value terms, estimates of the installed capacity for different products, the capital investment position and figures about employment and earnings. It has been suggested by many (see *e.g.* Tata Quarterly 1966) that there should be two series, one on the monthly basis or at least on the quarterly basis covering the items mentioned above and the other, more detailed, on the annual basis. Perhaps an element of compulsion may be necessary for obtaining regular returns and there is no reason why it cannot be introduced under the factory legislation. Many suggestions have been made from time to time for improvements as regards coverage, weighting patterns, more detailed groupings, of products and industries, selection of a more recent base year for index construction, and also for expeditious publication of this information. It is understood that a committee is deliberating over these and other proposals for the last many months. But the impression that one gathers from its deliberations is that nobody seems to be in any great hurry to alter the old estalished procedures even when they have been shown to be incomplete, inadequate and therefore not very useful, either to the private sector or to the planners.

The problem of food is perhaps the most vital problem in the context of Indian economy and yet we have not been able to devise procedures to obtain quick and reliable crop-estimates even after twenty years of Independence. At present statistics of area and production of food-crops are obtained from two sources. One set of figures called the "Official Estimates" is based on the figures supplied by the State governments which estimate area under crop by field to field enumeration and yield rates by means of crop-cutting experiments carried out on sampling basis according to the scheme initially developed by the Indian Council of Agricultural Research (ICAR). Since the starting of the multi-purpose National Sample Surveys (NSS) another set of figures is available based on the area estimates obtained from land-use surveys, and yield rates from crop-cutting surveys both carried out on sampling basis. Thus there are two series of estimates where the agencies, methods of estimating areas under crops and crop-cutting methods (choice of plots) are different. Moreover the NSS sample consists of two matched sub-samples, the State sample and the Central sample, where data collection and processing agencies are different. Finally the NSS also tried in some of their earlier rounds to estimate the actual food consumption by the population by means of sample surveys and thus to generate another set of figures, food consumption figures, to compare with the production figures.

One would have thought that this proliferation of estimates and estimating agencies would operate as a healthy check on one another and thus work towards better accuracy. But the net result seems to have been more confusion instead of precision. The check by means of consumption figures could never be taken very seriously in spite of one very elaborate exercise and involved piece of argument (Mahalonobis 1963), as the estimation of consumption by the interview method (even for the reduced period of 30 days), is bound to suffer very greatly from inaccuracies. It is well-known that accurate figures of consumption can never be built except by means of direct observation or short-period surveys, that is, elaborately planned dietary surveys or expenditure surveys. In all other surveys carried out by the interview method the estimates become very often national rather that actual. (See : Sukhatme 1962 ; Panse 1961.) In any case this kind of check was never very convincing ; and it became difficult to operate and had to be given up when the NSS figures for consumption began to be reported (after 1958) in terms of expenditure and not actual quantities.

The comparison of the Official Estimates of production with the NSS estimates was also not easy for many years. (That the NSS estimates had not covered summer crops may be ignored for this purpose.) There was considerable time-lag in the publication of the NSS estimates and even the published NSS report did not always bother to compare their figures with the Official Estimates except on a few occasions. (NSS Reports No. 38, 73 and 106.) Those who cared to compare the two sets of figures found that the NSS estimates of production were consistently and considerably higher than the Official Estimates. Since this state of affairs was far from satisfactory a Technical Committee was appointed in 1960 to go through the entire procedures in detail. The report of the Technical Committee submitted in the middle of 1967 is an unhappy commentary on the quality of data in this sphere and our efforts to improve them. Table 1 extracted from the above-mentioned report gives the total production

TABLE 1

Comparison of the NSS and Official Estimates of production of seven cereals for India from 1957-58 to 1965-66 (in thousands of tons)

Year	*NSS*	*Official*	$\frac{\text{NSS-Official}}{\text{Official}} \times 100$
1957-58	68,064	52,180	30.4
1958-59	82,283	60,830	35.3
1959-60	83,862	61,856	35.6
1960-61	90,472	66,340	36.4
1961-62	82,852	67,812	22 2
1962-63	72,571	74,120	13.2
1963-64	72,242	67,094	7.7
1964-65	78,712	73,427	7.2
1965-66	67,608	59,687	13.2

N.B. (*i*) NSS estimates for 1957-58 to 1961-62 do not include summer season.
(*ii*) Official estimates for 1962-63, 1963-64 and 1964-65 are partially revised estimates and those for 1965-66 and other years are final estimates.

estimates from the two sources for the seven cereals, rice, wheat, jowar, bajra, maize, ragi and barley, taken together, for the period 1957-58 to 1965-66. The gap between the two estimates in the first five years from 1957-58 to 1961-62 is simply enormous even when one ignores the fact that the NSS estimates do not cover the summer crops which may raise them by at the most five per cent or so The NSS estimates are higher than the Official Estimates for all these years and although the differences tend to narrow down during the next five years, in the last year for which the figures are given, 1965-66, the difference is still quite considerable *viz.* 13 per cent.

At the State level the differences are even more breath-taking. The next table gives the estimated production figures for jowar for Maharashtra State, where jowar accounts for 60 to 70 per cent of the State's total cereal production. The figures are for the period 1959-60 to 1965-66 and they give separate and pooled estimates of the NSS, and the Official Estimates. Comments are superfluous.

TABLE 2

Estimated production of Jowar (in thousand tons) in Maharashtra State from 1959-60 to 1965-66

Year	NSS Estimates			Official Estimates	NSS—Offi.	S—C
	State	*Central*	*Pooled*		*Offi.* *Per cent*	*Pooled* *Per cent*
1959-60	2,912	5,903	4,408	2,835	55.5	67.9
1960-61	4,182	5,739	5,110	4,157	22.9	24.6
1961-62	5,227	3,145	4,186	2,920	43.4	49.7
1962-63	4,987	4,251	4,619	3,294	40.2	15.9
1963-64	4,188	3,896	4,042	3,151	28.3	7.2
1964-65	5,219	4,961	5,090	3,249	56.7	5.1
1965-66	5,098	3,981	4,539	2,292	98.0	24.6

N.B. The last column gives the difference between estimates from State and Central samples divided by the pooled estimate and multiplied by 100.

It is clear that this situation suits nobody except perhaps the policy-makers who can justify any policy which suits their prejudices, predilections or political interests ; procurement or non-procurement, PL 480 imports or no imports. So far as statistics and statisticians are concerned this sort of situation can only create the meanest kind of dis-respect and cynicism. What makes one most sad however is that the Technical Committee mentioned above in spite of its seven-year labour has not been able to decide which of the two methods is better nor to arrive at a general consensus and devise agreed methods by which the existing procedures can be improved and these enormous differences narrowed down. In the meanwhile the controversy continues (Naqvi, Pillai and Saha, 1968 ; Counter-Statistician, 1968) and it seems the present state of affairs will continue in the foreseeable future !

So the current situation in Agricultural Statistics is, to put it mildly, not very happy (see *e.g.* Divatia 1963, Subramanian 1967). But this does not deter some econometricians and model-builders from assuming that official estimates of the area and production figures for 1960-61 as "reasonably accurate" and then proceeding to fit a neat growth-model to the data for the period 1951-54 to 1958-61 (Minhas and Vaidyanathan 1965). It is also interesting to note in this connection that the 'spokesmen' of the Agriculture Ministry publicise their forecasts (in millions of tons of food production) every fortnight or so, right from the first week of the rainy season Although their intention is perhaps laudable—to boost and raise the morale of the people (including their own)—it passes one's understanding how they do it ; perhaps they have also a crystal, in addition to the estimating procedures discussed above.

××× ...... ××× ...... ×××

V

The sources of Indian survey data can be broadly classified into two categories : institutional records, such as those of the government, local bodies, schools and co-operative societies, and personal interviews. Used with discretion the institutional sources are generally reliable although they suffer from the usual faults of incomplete and dated information, careless and slipshod manner of keeping the records and the general slow-moving character of the entire bureaucratic set-up. Recent probes have shown that they may also suffer sometimes from deliberate distortions. For instance records such as village land records and books of co-operative societies are likely to hide realities in order to circumvent the relevant legislation and thus be technically on the right side of the law. It is said that this is happening on considerable scale where the recent laws have affected the entrenched interests.

In India, as far as personal or household data are concerned, use of mail inquiries and questionnaires to be filled by the respondent has to be ruled out as a method of inquiry except for certain very small segments of the population. So the main and perhaps the only instrument of inquiry is interview and schedule Even here purely personal or attitudinal questions become difficult because the interview, although planned to be personal, becomes in practice a group interview with some members of the family, neighbours as well as sometimes the village *Patil* or chief joining the respondent. Moreover because of scarce resources the inquiries are very often multi-purpose with bulky schedules which take hours to complete.

What can we say about the quality of information collected in such interviews ? Information about the status, that is the actual position as it then obtains, is generally accurate, such as family members, their present occupations, relationships, acreage under crops, crop patterns, cattle, housing and major items of household, farm or craft equipment. But when it comes to recalling the events of the past, even the recent past, the area of uncertainty grows. On the basis of this data it is hazardous to estimate the small changes that may have taken place because the order of error involved may dominate all apparent differences. According to many surveyors it is difficult to obtain reliable detailed information about purchases and sales, about farm operations, or about production of milk or vegetables for more than a week. For major capital investments such as wells, farm-houses and pumps you can stretch back the memory of the respondent by several years but for small scale credit transactions and their utilization it cannot go back for more than a year without seriously affecting the accuracy. Such details can be obtained in the case of only those very few households which run their farms or enterprises on strictly business lines and maintain accounts provided they co-operate, which is not always. The quality of information on assets and liabilities and even crop production is often very poor ; and this is not due to the memory factor or lack of maintaining proper accounts. To quote Mukherjee and Gupta (1959), ".. information about financial assets is either supplied reluctantly, evasively or wrongly or is completely concealed from the investigator....Data about liabilities, though relatively complete are also of doubtful nature, ...some households have given false statements about their debts or have concealed them altogether."

Then there is the problem of matching the estimates obtained for difficult periods of reference. For instance there are weekly figures, monthly figures, quarterly or thirteen-week figures and annual figures and it is not at all surprising that they do not check because of the deficiencies in the data mentioned above. Neale (1958) tried to match the annual figures of crop-yields for a few households in certain better-conducted farm-accounts studies against the sum of four thirteen-week figures and found that they often differed by 25 to 50 per cent variation, and it was not possible to "explain" or "rationalize" them as due to specific factors. That the accuracy of data in this respect leaves much to be desired is also evident from the fact that the annual balance-sheets constructed for entire villages or groups of cultivators often showed inexplicably huge surpluses or deficits. (See *e.g.* Dhavale, 1962.)

In fact studies like farm-accounts studies which base themselves on information about fairly long periods obtained through interviews, and not by direct observation, suffer from inaccuracies which are more or less inevitable and inherent in the situation. How does one ascertain the annual egg-yield of a hen, total milk yield of a cow for an entire lactation period, the number of bullock-hours and man-hours which go into different farm operations ? In the absence of direct observation and day-to-day records such questions are bound to elicit replies involving rotional averages such as 200 or 150 eggs for every hen and so on. Superimposed on this is the problem about the vagueness of the measures used such as bundles or armfuls of hay and cart-loads or baskets of manure and similar measures for dung-cakes or other fuel. So there is the inevitable tendency on the part of the investigator to put in imputed figures based on notional norms. Thus one often gets a situation where the averages computed from the mass of the schedules cannot be far different from those on the basis of which the figures were put into the schedules. This is one of those production processes where the output closely matches the input !

Apart from the problems mentioned above there is the problem about the quality of work by the field investigator, which is sometimes very low. Consider, for instance, the following figures obtained from a duplicated complete enumeration where 332 fields were surveyed by investigator A for noting the crop in each field and the same fields were surveyed by another investigator B a fortnight later (Mahalonobis, 1946).

TABLE 3

Comparison of duplicated complete enumeration

A B	Jute	aman rice	Jute-*aman*	No crop	Total
Jute	4	15	4	3	26
aus rice	4	12	1	4	21
Jute-*aus*	17	66	2	9	94
Jute-*aus-aman*	—	2	—	—	2
aus-aman	1	—	—	—	1
No crop	37	45	4	102	188
Total	63	140	11	118	332

There is no agreement except in 4 fields under jute and 102 fields under no crop. It is clear that the work of A or B or both has been very unsatisfactory even in this simple enumerative exercise. Woe to the researcher who has to depend on such low-quality field work. Fortunately not all survey work is as bad as this which is an

TABLE 4

Summary income and expenditure figures for 98 owner-cultivator families of village Khandali

Expenditure Group (Rs.)	*No. of families*	*Reported*		*Difference (Rs.)*	*Under-reported cash receipts of cash-crops (Rs.)*	*Percentage gap explained*
		Receipts (Rs.)	*Expenditure (Rs.)*			
Below 1000	9	2,662	7,463	4,801	1,973	41·1
1000 to 2000	37	25,300	54,549	29,249	17,548	60·0
2000 to 3000	24	31.932	57,231	25,299	22,979	90·9
3000 to 4000	13	23,000	41,645	18,645	18,317	98·2
Above 4000	15	138,682	314,196	175,514	144,944	82·6
Total	98	221,576	475,084	253,508	205,761	81·1

revealed substantial under-reporting of cash-receipts thus explaining a large part of the gap (columns 5, 6, and 7 in the table). It is seen that in the case of medium farmers this under-reporting accounted for most of the difference. In the case of the big cultivators it was found, again from other checks and comparisons, that the expenditure on labour and manure was over-stated. In the case of the two lowest groups however the gap was still considerable which led to the conclusion that income from other sources than farm was under-reported.

In another methodological scrutiny of the survey data of farm incomes the figures for farm produce obtained by two methods, one by visiting the family continuously every four weeks and recording the figures for the previous four weeks, and the other by only one visit at the end of the agricultural year, were compared (Dhavale, 1963). It was found that in the reporting of food crops there occurred serious omissions in short-interval visits as compared to the estimated production at the end of the year. But when there were no such gaps in crop-reporting the production estimated on the basis of thirteen visits was higher than the production reported at the end of the year; it seems the latter then suffered from the memory lapse. On the other hand the reported figures for cash-crops by the two methods were more or less consistent. It appears that cash-receipts from sale are a help to memory both in four-weekly as well as annual interviews.

Two other points which came up in the investigation of farm-schedules are also worth mentioning. The first is about the inputs and outputs per acre which vary according to the size of plots. It was found that inputs per acre, especially labour inputs, were relatively high for smaller plots. Is this due to the fact that the natural over-reporting of inputs for these plots inflates considerably the ratios because of the small figures of size of plots in the denominator, or is it due to a natural tendency to work more intensively when the holding is small ? Probably it is the latter because the per-acre output was also somewhat larger for small plots. This has therefore to be investigated further. The second question is about the difficulties which have to be faced when it is intended to determine the general wage level in rural areas, in agriculture, from the data of farm studies and other household schedules. While there are the prevailing daily wage rates, different for men, women and children, for labour hired on daily basis, the general wage level may also have to take into account the contract labour which is prevalent for many seasonal operations and the labour of the permanently attached labour (or *Saldars*) employed by the big farmers. The problem of conversion into common units is thus quite complicated and both theoretical thought and systematic investigation are evidently necessary for tackling it. (Dhavale, 1964, 1965.)

To study systematically the quality of information which is collected by the interview method in village studies a very detailed and careful scrutiny was undertaken of the plot-wise schedules for the villages of the Agro-economic Surveys carried out for assessing village-change in a period of five years. (Kamat and Dhavale, 1964, 1965.) In the course of this scrutiny we had the benefit of extensive discussions with our colleagues, researchers in the field as well as field investigators. This led us to a number of interesting findings of which we shall mention only two or three in general terms. The data about actual observable items were mostly correct—they showed no apparent contradictions or internal inconsistencies; information about the size of the farm, areas under different crops, crop-rotations used, irrigation facilities, number of cattle, manures used and such other factual items. But when it came to the recall of operations, estima'ion of inputs, reporting of cash receipts or consumption expenditure, the data begin to display weaknesses and in respect of cash receipts (especially in the case of medium and big farmers) the information was either misleading or deliberately falsified In the case of labour and other inputs, (sometimes also in the case of crop outputs,) as also in the case of household consumption the reported figures were often a combination of the information volunteered by the householder and imputed figures calculated by the field investigator on the basis of assumed norms. It was also very disappointing to note that qualitative questions often evoked either very standardized replies or no replies at all.

Because of the emphasis which has been given here on the problems relating to the quality of statistical information and the examples cited here in their support, it is likely that some may carry the impression that it is futile to collect observational data about human affairs—it is so inaccurate. This sort of conclusion will be not only incorrect but also unjustified. This is not a cry of despair but a call for caution and consideration. It is my firm belief that the observation of human affairs is not only a most fascinating field of study but it is also one of the most worthwhile and useful activities. And it is precisely because of this that we should concern ourselves with the quality of material which we are collecting and devise methods to make it more and more precise and accurate.

REFERENCES

(Anonymous). (1966). Industrial Statistics. *Tata Quarterly*, Vol. 21, No. 2, 3, 4,pp. 35-57.

Bailer, Barbara A., (1968). Recent research in reinterview procedures. *J. Amer. Stat. Assoc.*, Vol. 63, pp. 41-64.

Belson, William, and Duncan, Judith A., (1962). A comparison of the check-list and open response questioning system. *Applied Statistics*, Vol. 11, pp. 120-132.

Census of India (1951), Vol. IV, Part 1, p. 137, and *Census of India* (1961), Vol. X, Part VIIIA, pp. 80, 109.

Chapekar, N.G. (1933). *Badalapur* (*Amacha Gaon*). Arya Sanskriti Press, Poona (in *Marathi*).

Coale, Ansley J., and Stephan, Frederick F., (1962). The case of Indians and teen-age widows. *J. Amer. Stat. Assoc.*, Vol. 57, pp. 338-347.

Counter Statistician, (1968). Why not discontinue the NSS crop survey? *Economic and Political Weekly*, Vol. 3, September 7, 1968.

Dandekar, V. M. (1953). *Report on the Poona Schedules of the National Sample Survey* 1950-51. Gokhale Institute of Politics and Economics, Publication 26.

Daming, W. Edwards, (1944). On errors in surveys. *Amer. Sociological Review*, Vol. 9, pp. 359-369.

Deming, W. Edward, (1963). On some of the contributions of interpenetrating network of samples. *Contributions to Statistics*, ed. C. R. Rao, Statistical Publishing Society, Calcutta, pp. 57-66.

Deutch, Karl W., (1963). *The Nerves of Government* (Models of political communication and control). The Free Press.

Dhavale, Shalini, (1962a). Analysis of the reported data on income and expenditure for some selected groups of families in Kolhapur. *Seminar paper at GIPE* (Gokhale Institute of Politics and Economics).

Dhavale, Shalini, (1962b). An analysis of reported cash receipts and expenditure of cultivators in Khandali village. *Artha Vijnana*, Vol. 4, pp. 210-226.

Dhavale, Shalini, (1963). Comparison of production figures collected by two different methods. *Artha Vijnana*, Vol. 5, pp. 52-62.

Dhavale, Shalini, (1964). Hired labour and wage-rates for farm operations in eleven selected rural centres in Maharashtra. *Artha Vijnana*, Vol. 6, pp. 127-141.

Dhavale, Shalini, (1965). An analysis of per-acre inputs and outputs for four major crops in Pusegaon and Visapur. *Seminar paper at GIPE.*

Divatia, M.V., (1963). Statistical requirements of agricultural planning, (being contribution to the Symposium on Current Status of Agricultural Statistics and Tasks Ahead). *J. Ind. Soc. Agri. Stats.*, Vol. 15, pp. 82-89.

Gales, Kathleen and Kendall, M.G. (1957). An inquiry concerning interviewer variability. *J. Roy. Stat. Soc. Series A*, Vol. 120, pp. 121-147.

Godel, Kurt, (1931). Uber formal unentscheidbare Satze der Principia Mathematica und verwandler] Systeme, 1 *Monatschefte fur Math und Phys.*, pp. 173-189.

Gray, Percy G., (1955). The memory factor in social surveys. *Journal Amer. Stat. Assoc.*, Vol. 50, pp. 344-363.

Hanson, Robert H., and Marks, Eli S., (1958). Influence of the interviewer on the accuracy of survey results. *J. Amer. Stat. Assoc.*, Vol. 53, pp. 635-655.

Hildebrand, F.B., (1956). *Introduction to Numerical Analysis* ; Chapter 10. Mac-Graw Hill, New York.

Jaeger, Carol M., and Pennock, Jean L., (1961). An analysis of consistency of response in household surveys. *J. Amer. Stat. Assos.*, Vol. 56, pp. 320. 327.

Kamat, A. R. and Dhavale, Shalini, (1964). Information obtained from the questionnaires for plot-wise schedules. (in *Marathi*). *Seminar paper at GIPE.*

Kamat, A. R., and Dhavale, Shalini, (1965). A study of plot-wise schedule of the village-surveys. *Seminar paper at GIPE.*

Kamat, A. R., (1967). Education in Gulumb. *Indian Education Review*, Vol. 2, pp. 13-35.

Kamat, A. R., (1968a). Some features of the growth of education in rural Maharashtra. *Indian Educational Review*, Vol. 3, pp. 1-56.

Kamat, A. R., (1968b). *Progress of Education in Rural Maharashtra (post-Independence Period)*. A forthcoming publication of the Gokhale Institute of Politics and Economics, Poona.

Kish, Leslie, (1962). Studies of interviewer variance for attitudinal variables *J. Amer. Stat. Assoc.*, Vol. 57, pp. 92–115.

Lahiri, D. B., (1958). Recent development in the use of techniques for assessment of errors in nation-wide surveys in India. *Bull. Inter. Stat. Inst.*, Vol. 36, Part 2, pp. 71-93.

Lange, Oscar, (1945-46). The scope and method of economic science. *Review of Economic Studies*, Vol. 13, pp. 20-21.

Mann, Harold H., (1917). *Land and Labour in a Deccan Village*. Oxford University Press.

Mahalonobis, P. C., (1944). On large-scale sample surveys. *Phil. Trans. Roy. Soc.*, Series B, Vol. 231, No. 584, pp. 329-451.

Mahalonobis, P. C., (1946). *Experiments in Statistical Sampling in the Indian Statistical Institute*. Indian Statistical Institute, Series No. 11. Reprinted from the paper in *J. Roy. Stat. Soc.*, Series A, Volume 109 (1946).

Mahalonobis, P.C. and Lahiri, D. B., (1961). Analysis of errors in census and surveys with special reference to experience in India. *Sankhya*, Series B, Vol. 23, pp. 325-358.

Mahalonobis, P.C. (1963). A preliminary note on the consumption of cereals in India. *Sankhya*, Series B, Vol. 25, pp. 217-236.

Mahalonobis, P. C., (1967). The sample census of the area under Jute in Bengal in 1940. *Sankhya*, Series B, Vol. 29, pp. 81-182.

Mauldin, W.P. and Marks, E.S., (1950). Problems of response in enumerative surveys. *Amer. Socio. Review*, Vol. 15, pp, 649-657.

Milne, W. E., (1949). *Numerical Calculus.* Princeton University Press.

Minhas, B. S., and Vaidyanathan, A., (1965). Growth of crop output in India 1951-54 to 1958-61. *J. Ind. Soc. Agri. Stats.*, Vol. 17, pp. 230-252.

Morgenstern, Oskar, (1963). *On the Accuracy of Economic Observations.* Princeton University Press, Second edition. (First edition in 1950).

Moser, C. A., (1951). Interview bias. *Review Inter. Stat. Inst.*, Vol. 19, pp. 28-40.

Mukherjee, M., (1963). Scientific approach to planning. *Essays on Econometries and Planning* (presented to Professor P.C. Mahalonobis on the occasion of his 70th birthday). Ed. C. R. Rao, Statistical Publishing Society, Calcutta, pp. 153-176.

Mukherjee, P. K., and Gupta, S. C., (1959). *B Pilot Survey of Fourteen Villages in U.P. and Punjab.* Asia.

Murthy, M. N., (1963a). Assessment and control of non-sampling errors in censuses and surveys. *Sankhya*, Series B, Vol. 25, pp. 263-282,

Murthy, M. N., (1963b). On Mahalonobis' contributions to the development of sample survey theory and practice. *Contributions to Statistics*, ed. C. R. Rao, Statistical Publishing Society, Calcutta, pp. 283-3.6.

Nandi, H. K., (1968). Rethinking on the application of statistical theory. (Presidential Address at the Statistics Section of the Fiftyfifth Science Congress.) *Calcutta Stat. Assoc. Bull.*, Vol. 17, pp. 1-14.

Naqui, S., Pillai, S. B., and Sah, R. P., (1968). Why discontinue the NSS crop surveys ? *Economic and Political Weekly*, Vol. 3, June 22, 1968.

Neale, Walter C., (1958) The limitations of Indian village survey data. *Journal of Asian Studies*, Vol. 17, pp. 383-402.

Neter, John and Waksberg, Joseph, (1964). A study of response errors in expenditure data from household interview. *J. Amer. Stat. Assoc.*, Vol. 59, pp. 18-55.

Panse, V. G., (1961). The national sample survey, agricultural statistics and planning in India, *Changing India*, Asia, pp. 207-232.

Popper, Karl, (1959). *The Logic of Scientific Discovery.* Hutchinson, London.

Rallis, Max, Suchman, Edward A. and Goldsen, Rose, K., (1958). Applicability of survey techniques in North India. *Public Opinion Quarterly*, Vol. 22, pp. 245-250.

Ritchie, A. D., (1923). *Scilntific Method.* Routledge and Kegan Paul, London.

Rudolph, Lloyd, and Rudolph, Susanne H., (1958). Surveys in India : field experience in Madras State. *Public Opinion Quarterly*, Vol. 22, pp. 235-244.

Sengupta, J. M., (1963). A study of field cost for the collection of household consumption data by an interview method. *Contribution to Statistics*, ed. C. R. Rao. Statistical Publishing Society, Calcutta, pp. 429-448.

Sengupta, J. M., (1966). On the validity of fertility data collected through interviews. *Sankhya*, Series B, Vol. 28, pp. 259–268.

Subramaniam, S., (1967). Vicissitudes of Indian agricutural statistics. *J. Ind. Soc.-Agri. Stats.*, Vol. 19, pp. 115-125.

Sukhatme, P. V., (1962). The food and nutrition situation in India : Part I, *J. Ind. Soc. Agri. Stats.*, Vol. 14, pp. 49-87.

Turner, Robert, (1961). Inter-week variations in expenditure recorded during a two-week survey of family expenditure. *Applied Statistics*, Vol. 10, pp. 136-146.

Von Neumann, John, (1947). The mathematician. *The Works of the Mind*, ed. Heywood and Nef, University of Chicago Press, pp. 180-196.

Wilson, Elmo C., and Armstrong, Lincoln, (1963). Interviews and interviewing in India. *International Social Science Jonrual* (*Unesco*), Vol. 15, pp. 48-58.

Wooten, Barbara, (1950). *Testament for Social Science*. George Allen and Unwin, London.

Zarkovich, S. S., (1966). *Quality of Statistical Data*. FAO, Rome.

REPORTS

The National Sample Survey. Reports No. 38, 73 and 106 of the 13th, 14th and 17th rounds, published in 1961, 1963 and 1966 respectively. Issued by Cabinet Secretariat, Government of India.

Report of the Technical Committee on Crop Estimates (1967), Government of India, Planning Commission, New Delhi.

A note on reliability of area and production estimates of major cereal crops based on the NSS and the Official State Series. (1968).