THE APPLIED STATISTICIAN[*]

T.V. Hanurav
ISI, Hyderabad

I feel honoured to be elected as the Sessional President of the 49th Annual Conference of the Indian Society for Agricultural Statistics. I think that this is more a testimony to the affection and goodwill that a large number of friends in the society have for me than to my own professional achievements. However, it seems that there is no such thing as free lunch as I am asked to prepare and deliver the technical address. Nevertheless I sincerely thank the society for the honour bestowed on me and shall proceed with my task.

At the outset I would like to dispel any notion (or apprehension) that my technical address is going to deal with mathematical and theoretical aspects of statistics. This is not because of any dislike on my part for theoretical and mathematical aspects. In fact I started my career with a Masters degree in pure mathematics (that too as a blue eyed boy of that legendary guru the late Prof. V. Ramaswamy of Andhra University) and my subsequent foray into statistics is accidental and almost against my wish. Apart from my love for pure mathematics this is probably more due to the wrong initial exposure I had to the subject of statistics. However, I am fortunate to be the student of some famous teachers of statistics at I.S.I. They imparted a balanced blend of both the theoretical as well as applied aspects of statistics. The result is that I enjoyed esoteric subjects like information theory, coding theory, ergodic theory, unified sampling theory etc. and with equal measure of enthusiasm spent long hours in late Prof. Haldane's laboratory experimenting on Drosophila (fruit flies) and conducted haemotological studies on the Santhals in deep forests of Midnapore district.

But to come back to my choice of a non-theoretical topic for the technical address, it is because

(a) many noble and some undoubtedly abler souls are regularly taking care of the theoretical statistics and more importantly

(b) applied statistics which was held in high respect when statistics was vigorously propagated by Mahalanobis, Sukhatme etc. is slowly getting looked down in academic institutions.

I think it is necessary to highlight and analyse this unhealthy drift before it is too late.

Statistics is not a branch of mathematics although like physics or economics, it uses some mathematics as a tool. It is to use Prof. Mahalanobis phrase a key technology needed in almost all branches of knowledge be they physical, natural or social sciences. Its main thrust has to be utilitarian, although some really able persons interested in pursuing the purely theoretical aspects should be encouraged to do so. Similarly the aim of a good teacher of statistics should not be just to produce good lecturers or professors of statistics. After all if everybody produces "statisticians" who will produce statistics?

I recall an interesting incident that occurred nearly thirty years ago when I was a lecturer at the university of Sheffield, U.K. in 1965-66. Prof. Joe Gani, a very dynamic organiser (besides a first class applied statistician), just then started the department. To spread interest in the subject he organised a half day meeting to which he invited a large number of high school students and their teachers. He and members of the faculty of the department gave short speeches on the importance of statistics and then the guests were invited to ask any questions they have in mind. After a few questions, one participant, a young girl asked the really crucial question "Can you guarantee that every one trained as a statistician will get job?" After a brief and embarassing silence and apologetic replies I answered the questioner that if a statistician is unemployed in a society it points out that the production of statisticians exceeded the demand for them and to rectify the situation and correct estimates of demand and supply you need a statistician's help. Hence, by induction every statistician will be employed!

To continue the statistician has a duty to be useful to the society at large (which foots the bill to train him) by providing reliable data for planning and for improvement of agricultural and industrial production. The Indian Society of Agricultural Statistics and the Indian Agricultural Statistics Research Institute are the foremost in upholding this principle in their training and research programmes and deserve a big "Thank you" from the people.

In this brief Technical address I would like to share with you with the help of some case studies, some of my experiences and thoughts as a teacher and applied statistician and my occasional interactions with nonacademic and governmental agencies. If some of my thoughts border on apprehensions and even nightmares, I beg to be excused for calling spade a bloody shovel.

When the discipline of statistics started blossoming in the west in the second quarter of this century mainly in U.K. and U.S.A. Indians were shrewd enough

to gauge its importance and potential and quickly jumped into the fray soon to emerge as a force to reckon with. Mahalanobis and Sukhatme concentrated on developing the applied aspects and methodologies of statistics. More importantly, they spent their energies in "Institution building" – the Indian Statistical Institute (ISI) and Institute of Agricultural Research Statistics (IARS) and oiling the wheels of the machinery and used their clout to get Government support to this branch of knowledge, leaving a number of brilliant minds to concentrate all their energies on teaching, training and research-both theoretical and practical.

Statistics was not taught in those days at undergraduate level in the universities. Even the few universities that have postgraduate programmes recruited mainly those with a masters degree in mathematics. As a consequence with a few exceptions they tended to teach statistics as if it is a branch of mathematics with questions in examinations like, "State and prove the theorem... Derive the distribution of..., What are the regularity conditions for... to hold good?" etc. the distinctive feature of the subject that it deals with the contingent world of reality (to use another pet phrase of Mahalanobis) is generally lost sight of.

However, at I.S.I. and at IARS, although the input remained the same (viz. post graduates in mathematics) the training had a healthy blend of applied aspects of the subject interlaced with theoretical foundations and excursions. The significant contributions made by the statisticians at these two premier institutions in the branches of Design of Experiments, Sample Surveys, Multivariate Analysis, Biometry etc. bear testimony to the research aspect while the large number of statisticians recruited by the industry and the research laboratories testify to the success of practical and applied approach imparted to the trainees.

Unfortunately inspite of a lapse of nearly thirty years after the introduction of statistics in the undergraduate curricula and the start of post graduate courses in almost all Indian Universities my own experience with a number of young graduates is that it is still being treated as a purely academic entity and as a branch of mathematics. This is true though to a lesser extent even among the graduates of I.S.I. nowadays. Even first class graduates can readily prove a theorem or derive a sampling distribution but can't draft a simple questionnaire or design an experiment to test a simple claim. Worse still, in a number of cases they find it infradig to indulge in such 'beivial' pursuits unworthy of budding mathematicians which they think they are in the last decade of his life Prof. Mahalanobis himself became acutely aware of this undesirable drift and used to be angry with the mathematicians as being responsible for looking down on applied statistics and project work. I contend, without any apologetic

tone that to be a good statistician one need not to be a big mathematician and can do, for most part, with simple graduate level mathematics. Even research workers in social sciences and natural science can acquire a moderate skill in statistics required for most of their purpose. Without going into questions of why and how I have been holding this view for the past thirty years although I am second to none in my love for pure and even abstract mathematics.

I would like to narrate an anecdote from none other than that great soul the late Prof. R.A. Fisher. This was in 1960-61 when Fisher visited I.S.I. for a month. At that time I was working in genetic statistics under Prof. Haldane. I was somewhat uneasy that in genetics we try to explain the phenomena of inheritance treating chromosomes as threads, with genes as beads on these threads and felt that we should try, as in the case of elementary particles of physics, to explain the phenomena at molecular or even submolecular level (By that time the famous helical model of DNA was not yet discovered). When I expressed my nagging doubts to Prof. Haldane he said that it would lead us deep into Biochemistry and that he was not ready for such an exploration. But when I mentioned the same doubt to Fisher his answer was characteristic and telling. He looked at me and said "See, Mr. Rao. If I am going on a bicycle to a nearby village and my bicycle breaks down on the way, what I need is a good cycle mechanic and not one who is an expert in Newtonian mechanics and dynamics of rigid bodies. I am sure all of you agree with the spirit behind that brilliant exposition. For example if on the basis of data available, a statistician is able to satisfactorily establish that prevalence of smoking and incidence of lung cancer are intricately and positively associated with each other he has already done a significant service to science and society. As to whether smoking is a cause and incidence of lung cancer is an effect and if so why should smoking cause lung cancer form a separate enquiry to be addressed to by the biologists and chemists where again the services of the statistician would be needed to test whether observed results are significant.

On the otherhand a statistician must have a basic understanding of the subject matter of investigation and shouldn't blindly apply his 'tools' and skills to data presented to him. He has to carefully scrutinise the data to polish it off. All of you are familiar with a number of real or artificial examples but I would like to mention just two examples which I repeatedly tried on a number of bright graduates and even postgraduates some in statistics too. They are :

Ex 1 : Heights of 6 college students in Hyderabad are as given below :

$$157.1, 163.2, 64.1, 178.3, 66.8 \text{ and } 176.4$$

Find their average height.

Ex 2 : Average water level (x) during the season and yield(y) per hectare of a variety of rice are given below for some values :

x (in cms)	0.5	1.0	1.5	2.0	2.5
y (in tonnes)	5.225	5.475	6.125	6.350	6.725

Estimate the value of y when x = 60 what is the standard error of the estimate?

Almost 90% of my respondents gave the answer to the first question as 134.3. Some gave accurate answers like 134.32 and 134.3167 etc. The second question put to statistics students after they were taught linear estimation and least square theory was also answered as per theory by most of the respondents while some had difficulty in giving the standard error of the "estimate". Interestingly, those who did not have much exposure to statistics tended to do better! My point is that a proper orientation to statistics is atleast as important as achieving mathematical rigour in the theory.

I would now like to share some of my thoughts about and experience with those who employ statisticians. Of these the State is the dominant one. As every one knows, since a long time government have been collecting data on land use, agricultural production etc. through their agencies, as a by product of revenue administration. Infact some say that the word "Statistics" originated from "state craft". Till recently in most states in India the village officer (V.O.) known variously as karnanm/ patel/patwari used to collect this information for proper collection of revenue in the form of land cess. In return for this part-time help he is paid a small consideration. By the very nature of the purpose, this information is needed accurately and on all land holdings. It was not much of a problem for him to do this job reasonably well. There are several reasons for this happy state of affairs.

(i)   The number of variables under consideration is relatively few.

(ii) In the stagnant and static agrarian society then existing there was not much of variation in the variables at hand from year to year and

(iii) The village officer post being heredity and with not much of social mobility the V.O. is rooted in the village and knows his domain very well.

However, this happy state of affairs no longer exists now. The government wants data on a huge number of socio-economic variables such as land use, cropping pattern, crop yield, literacy levels, immunization, public distribution and what not for the so called planning purposes. The complexion of the agrarian

society has become more and more complex with the introduction of various fertilisers, pesticides, varieties of crops, varieties of seeds of the same crop, types of irrigation facilities, marketing facilities, credit facilities (including the famous loan melas). And a serious blow to this data collection is the abolition for good or bad-of the post of the heredity part it me village officer and its replacement with a pucca government employee with all the trappings like various types of leave transfers every three or five years and one such V.A. for a group of villages.

Worse still, in my own state of Andhra Pradesh the heredity system was abruptly abolished without providing for the alternative system to step in its place. The earlier system, however imperfect, provided at least some data base for some key variables. As can be well imagined, the result of such hasty action is not much different from chaos. It is only a couple of years back the V.O.'s are appointed but they are unable to cope with the demands made on their time.

Compounding this crisis are two of the gravest errors that most Governments commit, viz. (i) Obsession for acquisition of all sorts of information on its subjects and (ii) Obsession for accuracy levels that are not at all needed. In some cases this arises out of pure innocense but you do come across the politician, and to a lesser extent a bureaucrat who arrogantly thinks that because the village officer is paid by the state he is at their beck and call to provide any information wanted and at any time. If the information is required for purposes of planning one does not require it correct to the last decimal place. If such are needed you need to employ accountants and arithmeticians and not statisticians.

The accuracy required of a data depends on the use to which that data is put to. For example if the state wants to make provision for enough groundnut it does not need to know the demand and production to a kilogram but probably only correct to a multiple of hundred tonnes, even if one wants to plan at district level.

When such chaos descended, a committee was set up by the A.P. Government in 1987 (of which this good self was coopted as an expert) to suggest methods of improving the statistical system in A.P. mainly for area and yield statistics. This committee after careful consideration and deliberations, recommended adopting a sample survey approach where by only one fifth of all land holdings, are surveyed on my rotation basis. The scheme recommended was a multistratified and multistage sampling with complete theoretical back

up for estimation of areas and yields for various crops and for obtaining estimates at even district level.

Like all useful things the report of this committee been kept in the state archives and will probably be put in a time capsule for excavation at a later date by another civilisation. Of course one can not blame any single individual politician or political party for this because the response remained static over all these years. Incidentally I understand that schemes similar to the one recommended by us are introduced in West Bengal and Orissa recently.

I would like to brief mention another case study which is even fresh.

As all of you are probably aware there are a number of castes communities classes (or more accurately self appointed leaders of these groups) all over the country competing for the coveted title of "backwardness". When one such agitation put pressure on the government of Andhra Pradesh, the later asked the A.P. Commission for Backward classes (a statutory body) to make suitable recommendation for inclusion or otherwise of that caste in the list of the O.B.C's which it was not hitherto. However, the commission expressed its helplessness in the absence of reliable data on the social, educational and economic profile of that caste. The Government then convened a meeting of a large number of academics (including this good self) to suggest a suitable method for collecting data on members of that caste. However, I argued (and for once successfully!) that nothing is good or bad, forward or backward without comparison and that we can't conduct a survey solely for a particular caste just because it staked a claim to be labeled as OBC. For otherwise we will be compelled to undertake such an exercise everytime some caste comes up with a similar claim. Instead one should conduct an omnibus survey covering all castes/ communities/ etc. with, if necessary, and extra emphasis on some of the groups on which we want to have more accurate estimates of socio economic indicators. This argument was well received by the Government which entrusted me with that stupendous task. An advisory group of eminent social scientists was also appointed but otherwise I am given a free hand to design and analyse the data. The services of the nearly 800 statistical officers spread over the whole A.P were kept at my disposal for part time use. A detailed but crisp four page schedule is being canvassed on nearly seventy five thousand households in A.P. The field work is almost complete and the scrutiny and coding are now going on. If all goes well my analysis will be submitted to the B.C. Commission in another four or five months. Which may then make its own recommendations. Without a blush, I can say that this would be the first time that recommendations on this important matter are going to be based on a scientifically collected and analysed data, that to on a large scale.

An important point needs to be stressed here. However sophisticated and efficient are our sampling design and statistical analyses they would be irrelevant if the basic data is of doubtful quality. Since data is being collected on sensitive socio economic variables it is important to get the informant's cooperation to the fullest. This is being achieved by

(i)   using seasoned investigators and giving them further training

(ii) careful scrutiny and constant supervision even at high level

(iii) assuring the informant in writing and through advertisements in T.V., radio and newspaper that the individual data will not be passed on to any government or other agency and that I will give only summary findings to the B.C. Commission, and last but not the least

(iv) we are not at all interested in any individuals data but only in summary figures for groups of individuals and hence are adopting a sample survey approach instead of a complete census. To drive home this aspect further we are telling the sampled households that if they stubbornly refuse to give the information asked for by us heavens are not going to fall on any body and we will simply replace that household by another random household. Needless to say we are getting a very satisfactory response and are rarely encountering stubborn non-responses.

Now comes an interesting development. While the field work of our survey is in progress, the State Government started revision of electoral rolls and issue of photo identity cards (PIC) for all its voters. For this it needs to collect name, age (in completed years) father's/mother's/husband's name and the residential address of all eligible voters. And then it occurred to some enthusiastic politicians and bureaucrats that when the investigator anyway goes to the house to get this information he might as well collect some basic data on the social and economic status of the individuals, at little or no extra cost. And this additional information includes the name of the head of the households, his religion and caste, type, ownership and even plinth area of the dwelling unit. Further for all members of the household, information is sought on name, sex, relationship to the head of the household, not just age in years but date, month and year of birth, occupation together with a 3 digit code, extent of land holding separately for wet, dry and fallow categories, details of urban property such as areas of houses, house sites, commercial and industrial establishments and total annual income from all sources. At the bottom of the schedule the informant has to sign a statement that the particulars given by him/her are true to the best of his/her knowledge. A distinctive number will then be given to

each person which will be printed on the reverse of the PIC which will then be called Multipurpose Social Security Identity Card (MSSID). The data collected in the multipurpose household schedule will be stored in computers and can be retrieved with the help of distinctive number. It is envisaged to collect in future more and more data on all the persons and add it on to the stored data to build a comprehensive data base. The MSSID is said to be useful to the card holder in obtaining caste certificate from revenue official participating in various government schemes (like IRDP) and obtain bank loans etc. Failure to give the data asked for would result, the people are told, in loosing access to these benefits besides not having the PIC required for voting. Nearly 80,000 school teachers are to collect data in a month's time as a crash programme taking time off their teaching duties. These investigators are given not even a perfunctory training.

I have taken the trouble to mention this example (i) Since many research workers get overawed by the sheer volume of a data it is necessary to warn them to carefully check the quality. (ii) Computers or other high tech gadgets with their associated yuppy jingo do not in anyway improve the quality of the basic data but can only speed up processing of the data. (iii) Just because computers are at hand one does not have to collect all sorts of data irrespective of its quality, and utility and most importantly (iv) Ultimately it is the poor applied statistician who will be found fault with for any lapse in the quality of the data. Even if he has to act under pressure and/or orders professional ethics demand that he put forth his candid opinion,if necessary in black and white and to dissuade the government of such wasteful and misleading ventures.

Some uneasy questions arise and these should not be brushed under the carpet. Can the Government compel a person to reveal his caste even if he does not intend to claim any benefit (like reservations, scholarships) therefrom? Is a citizen bound to reveal expect to the income tax authorities, his incomes and assets? Can the state do so under the pain of depriving him his right to vote or claim caste free benefits available to very poor people? Is a person bound by the information about him given by somebody else i.e. the informant although the later is a member or even head of the same house hold? Conversely, can a person get a bonafide caste certificate from revenue authorities just on the basis of the information given in the MSSID? (Although the authorities initially gave that impression wisdom dawned later and they hastily retreated.) Should the Government take as authentic the incomes/ assets declared in the MSSID?

Although these may appear to be more of a legal nature as one who delivers the data the applied statistician can not be impervious.

Recently it is a reported that in Kerala, which embarked on a complete census to elicit information on the caste and other related information, the processes had to be abruptly discontinued due to public's protests and non co-operation.

Recalling (iv) of para II, I can not think of a better example to prove the merits of a sample survey over a complete census.

I now wish to make a few comments on a recent fad involving statistics viz. opinion polls. Public opinion polls have originated in the U.S.A. With Gallup poll and Harris poll in the forties and fifties. While an opinion poll is meant to reflect public opinion on an issue it can easily be misused to influence and mould the opinion by creating a bandwagon mentality. And when those who conduct or commission such polls themselves have vested interest in the so called "estimates" given by them these should be taken not with a pinch but a ton of salt. Quite apart from the fact that our leaders have the capacity to corrupt even the most honest and upright individuals, the media itself is not above board?

Of course there is nothing wrong in propagating one's point of view. In fact I postulate that it is the duty of every intelligent person to brainwash others. However, I don't think it is ethical to create a bandwagon mentality.

One nowadays hears about the so called *exit polls also*. These are conducted on a sample of people after they have cast their vote. Assuming that by the time the results of the exist poll are known the pole process has ended (so that the published results do not influence the voting). I just do not understand the purpose of an exist poll, except as an idle curiosity (at the expense of others!) After all, the die had been cast and it makes no difference as to what others predict of an even that had already occurred.

But a bigger danger is the utter nonsense that goes in the name of opinion poll. While I am giving finishing touches to my manuscript I came across a prominent English Daily of Hyderabad which announced through its banner headline on front page the results of an opinion poll carried out by a well circulated English weekly from Kerala. They conducted by Fax, a poll in which 121 (yes, one hundred and twenty one) people all over India responded (i.e. faxed their answers) to its questionnaire. And the results are given as percentages correct to two decimal places.

Without bothering to go into questions about how representative is the sample or the magnitudes of sampling errors vis-a-vis the deceptive accuracy presented, I would say that any applied statistician with a minimum of competence and credibility should straight away condemn such absolute trash

and should even question the motives of those who conduct and publish such opinion.

An important area where the applied statistician has made a mark but is increasingly marginalised is that of Statistical Quality Control (S.Q.C.). From the mid-fifties the application of SQC to improve the quality of Industrial output has gained a lot of support and achieved good results. But of late I find that professional statistician is being slowly replaced by engineers (and even diploma holders) who are made into half backed statisticians with a few months of part time training in elementry statistics. I don't contend that there is no room for such elementry statisticians but would stress that they can't completely replace a fully professionally trained statisticians.

This is bad enough but worse has come. Even the part time courses are getting replaced by two/three day "courses" conducted in some star hotel. The prefix 'statistical' is avoided as if it is a dirty world and new fashionable terms like Quality Circles, Quality Management, Total Quality Management, ISO 9000 etc. have come in vogue. Slogans like "Our aim is to achieve the level of zero defectives" are coined. (In a world with zero defectives and completely homogenity the poor statistician will in fact be jobless ! Our very survival depends on the dictum that such a world just does not exist). Apparently statistician is marginalised by salesmen and management yuppies. To be fair, the statistician himself contributed not a little to this state of affairs by getting overawed by the glib talk and pays and perks of his clients.

And now for a googly. The rapid development of computer technology is undoubtedly a boon to the society. The replacement of bulky calculations machine by pocket calculators is definitely of help in the class room and even researchers. But we are witnessing the danger of the tail waging the dog. Student in the classroom do not want to use even pocket calculator and want to use software packages on a computer even to draw histograms, scatter diagrams and regression lines. They do not try to look into the data but blindly use a "package" resulting sometimes in howlers. (ef. para 6) And with the slightest twist in the problem they become helpless.

I would say that unless one is dealing with a voluminous data even the researcher can most often do with a pocket calculator. And great souls like Fisher and Haldane are with me in this: We should also educate the masses not to be misled by nonsensical statements like "the computer has chosen the candidates". "The computer predicted the result" etc. often used by newspapers, glossy journals and even by intelligent and well intentioned but ill-informed

persons. Afterall, the computer does what it is programmed to by the scientist, only it does the same faster.

Of course the computer is doing such disservice not only to the applied statistician but even to the theoretician. At the slightest hint of difficulty in finding theoretical solution to a problem, he runs to the computer to conduct simulation studies. Needless to say does not stimulate his theoretical prowess. And empirical results are never a substitute for theoretical proofs.

I now wish to call a halt to my expression of outrange and anguish at the state of affairs prevailing for statistics in general and applied statistics in particular.