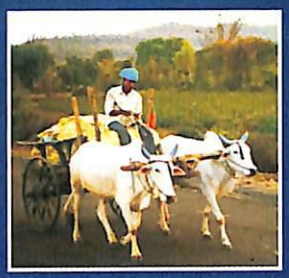
Statistics Everywhere Statistics Everywhere Statistics Everywhere Statistics
Statistics Everywhere Statistics Everywhere Statistics Everywhere Statistics
Statistics Everywhere Statistics Everywhere Statistics Everywhere Statistics
Statistics Everywhere Statistics Everywhere Statistics Everywhere Statistics

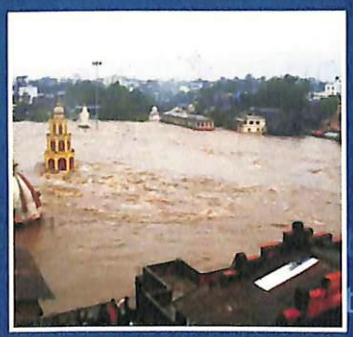
STATISTICS for EVERYONE

Anil Gore • Sharayu Paranjpe • Madhav Kulkarni


μ


σ


∞



χ^2


Statistics Everywhere Statistics Everywhere Statistics Everywhere Statistics
Statistics Everywhere Statistics Everywhere Statistics Everywhere Statistics

Share
Facilitate
Inform
Promote
SIPF ACADEMY

STATISTICS for EVERYONE

Second Edition

Anil Gore
Sharayu Paranjpe
Madhav Kulkarni

1 May 2012

STATISTICS for EVERYONE

Anil Gore, Sharayu Paranjpe, Madhav Kulkarni

First edition, 5 Sept, 2009

Second edition, 1 May , 2012

Published by SIPF ACADEMY, **Publishers and Consultants, Mrs. Madhuri M Kulkarni, Flat No 2, Rohini-Ravi Apt. Co-Op. Hsg. Soc., Canada Corner, Sharanpur Road, Nashik 422 002 (M.S.) India.**

E-mail: **madhurimk@gmail.com**
: **madhavbk@gmail.com**

Printed at M/s B. Y. Printing Press, Satpur, Nashik 422 007

Preface

Statistics is a subject that confuses many, scares some but excites few. 'After all, how thrilling can numbers or formulas be?' they seem to think. This situation is unfortunate and unnecessary. Unfortunate because citizens in all walks of life need to understand their own problems and problems of the society. This often involves appreciating statistics. It is unnecessary because heart of statistics is not numbers, nor formulas, but logic. And all of us can comprehend logic, especially if it is expressed in the context of our own problems. So, really there is no reason why we teachers of statistics cannot help laymen recognize the importance of our discipline.

Perhaps a major part of the blame for this failure can be attributed to the style of teaching of statistics in many countries, including India. We spend a lot of time on mathematical aspects of the subject, rigorous proofs of theorems and learning how to use formulas. On the other hand applied aspects of the subject get a short shrift. This is partly because most of the teachers of statistics remain aloof from practice of statistics. Since charity begins at home, we felt the urge to try to change this situation. Here we present our attempt to communicate the excitement and romance of the subject. In a sense we are well suited to the task. We three began our careers in the seventies under the tutelage of Professor P V Sukhatme, a father figure in the field of agricultural statistics in India. With his encouragement, we dirtied our hands with data collection through surveys and experiments, prior to analysis and interpretation. We learnt from him the importance of selecting a socially relevant problem and spending enough effort to become familiar with the domain of that problem. He told us that to learn statistical genetics, he began by growing fruit flies in milk bottles, not to outdo the biologists but to understand the nature of data. We have all been teachers of statistics in later life, but always getting involved in use of statistics in different fields. We propose to exploit this experience to write about statistics through examples and case studies.

As stated earlier, our audience is intelligent and curious men and women of any background. In fact, we began this venture with essays for the magazine 'Resonance' published by the Indian Academy of Science. Here the idea is to write for those who have completed high school education. We wrote seven essays and for various reasons, the project was shelved. We revive it now. All those seven essays are reproduced here, with the kind permission of the Academy. In addition we have written some pieces, based on our consulting experience. The collection does not follow any sequence of topics from a statistics book. It tries to narrate stories of interesting applications in many different domains. Statistical methods

used range from very elementary to fairly advanced. We give only minimal details of those methods. The stories are intended to stand-alone and the book need not be read in a sequence.

As will be clear from the remarks above, the book is not intended as a textbook of some course in statistics. It is likely to be useful as supplementary reading. We are by no means the first statisticians to write such a book. There are many fine predecessors. We will mention just a few. Statistics: a guide to the unknown (Judith Tanur, Editor), Statistics and Public Policy (Fairley and Mostellar), Statistics and the Law (Morris H. DeGroot , Stephen E. Fienberg and Joseph B. Kadane ;Editors). However, there are not many books of this type written in India. So, we hope our book will be particularly useful to Indian teachers and students. We have also included some items specifically aimed at classroom use by students and teachers. These include one hundred data sets for statistics education, a list of possible projects for statistics students etc.

We hope this material will be interesting reading for all and useful material for statistics students.

Anil Gore
Sharayu Paranjpe
Madhav Kulkarni

Table of Contents

Section I
Numeracy for Everyone: Series Introduction

Do you know about three R's in the traditional list of minimum skills to be acquired by every school going child? They are reading, (w)riting and (a)rithmetic. Perhaps arithmetic was needed mainly to carry out common transactions such as buying grocery, vegetables or paying rent.

In today's world of exploding information, it is not enough to know these three R's. It is also necessary to learn techniques to consolidate and interpret the continuously bombarding information. Statistics is the science of identification and art of interpreting patterns in numbers. Its distinctiveness lies in trying to understand uncertain events. Death is certain but age of death is not. We can predict rain but cannot guarantee it.

The aim of this series of articles is to introduce the reader to the basic ideas of statistics applied in many spheres of life. Thus we begin by emphasizing the need for quantification and illustrating how to represent the data using graphs. Then the concepts of probability and association/correlation are discussed.

As statistical consultants, our main interaction during the last decade and a half has been with ecologists/biologists. This has influenced our writing and many illustrations naturally come from the field of ecology. Some statistical techniques specially developed for applications in ecology are also discussed.

Any scientific experiment needs planning. Statistical techniques called design of experiments and sample surveys help a lot in such planning. An outline of these techniques is given.

Areas of biological sciences other than ecology, where statistics is used extensively, include health, epidemiology, clinical trials etc. Social sciences have a large component of uncertainty. Statistics is a science that searches for laws applicable to large groups of individuals (populations) in the presence of uncertainties. Its applications in social science lead to interesting findings, reviewed briefly in subsequent articles.

In the present liberalized and globalized economy, quality of an industrial product decides its fate in the market. Maintaining high quality of production, adjusting the production process to minimize the loss due to rejection of product are topics where statistical expertise can make a difference. We conclude the series by over viewing this area.

Numeracy for Everyone:
1. Why Quantification?

Introduction: India is a land blessed with beautiful and bountiful nature. We have the world's tallest mountains in the north and the heaviest rainfall in the east. There are dry deserts and long perennial rivers. Our forest wealth and wildlife are legendary. Western Ghats and north-eastern mountains are two of the world's mega biodiversity hotspots. No wonder so many of us love and cherish this treasure. It is to such nature lovers that we address the next few articles on numeracy, quantification and biostatistics.

We can anticipate the reaction of many. Some may say that as nature lovers they enjoy going out, trekking in national parks, watching birds and butterflies and photographing landscapes. Silence and serenity of the wild makes them feel recharged to face the humdrum life once again. They would prefer not to spoil their pleasure by getting into dreary details of numbers and statistics. Yet another group may argue that they took up study of biology, at least partly because they did not care for mathematics. And how can any one expect them to go through formulas and equations?

To these nature enthusiasts, we will say, hold your horses. While we are not apologetic about equations, formulas or numbers, this series is not about them. It is about nature and its conservation. Every detail of biostatistics to be discussed herein should be judged on the basis of its relevance to these matters. It is our plan to spare the reader of almost all technical details and give only the concepts necessary to understand processes of degradation and restoration of nature. The only assumption we make is that the reader is a serious nature enthusiast. Every such person routinely encounters news, views and comments that contain so much numerical information. Here are a few illustrations.

- Bharatpur is a Mecca for birdwatchers, especially those interested in winter migrants including ducks and waders. The jewel of them all is the Siberian crane. For years experts have been warning of a decline in the numbers of Siberian cranes returning to Bharatpur. If a bird species disappears altogether from an ecosystem, most birdwatchers will notice it promptly. But if there is a gradual decline in numbers, only regular counting will reveal the dangerous trend before it is too late.
- In the eighties, this man-made wetland was closed to cattle grazing. Subsequently *Paspalum* grass and other weeds have increased encroachment upon the water body which is getting choked. If the trend continues the wetland may be converted into a grassland.
- India's forest lands are supposed to be shrinking i.e. the area under forest must be declining over the years. Further, even the areas nominally under forest are

getting progressively degraded. This means that the number of trees per hectare must be going down.

- It was argued by ecologists a decade ago that if frog legs are exported on a large scale, the frog *Rana tigrina* may get endangered. What is the basis for this argument?
- Many wildlife enthusiasts have expressed reservations about the claims of forest departments that the number of tigers in Project Tiger areas is increasing continuously. People are very skeptical about the census methods adopted.
- Bird ringing programs of the Bombay Natural History Society (BNHS) have revealed that ducks migrate over great distances in very short times.

If you know the rationale behind all these statements, you probably will not see anything new in the sequel.

But many people are puzzled by these. When controversies and debates over environmental policies are heard, people are unable to make up their minds about which side is more reasonable. This is because they are unable to interpret quantitative information and arguments and to distinguish between valid evidence and junk. In that case you stand to gain something by continuing to read. We hope to discuss trends, harvesting strategies, diversity indices, census techniques, computer packages and a host of other things.
So are you willing to try a new adventure?

Why Statistics: Ours is, what has often been called, an age of information explosion. Newspapers, magazines, videos, TV, cinema, radio and god knows what else, keep feeding us information, much of it numerical. It comes in the form of tables, graphs, charts, estimates, predictions etc. It is all mind boggling for the uninitiated. Why cannot things be stated simply?

Take the case of our life span. How long does man live? It is difficult to answer this question precisely. Salim Ali, the father of Indian Ornithology, died at the ripe old age of 92 while his wife Tehmina died young. In India nearly every tenth child dies in its first year of life. So what is the length of human life? Well, it is a variable. It can range from zero (still birth) to a hundred years or even more. This sounds like utter confusion.

The discipline of statistics helps us sort it all out, make it simple and comprehensible. Once that is done, we can go deeper and consider comparison of life span patterns. Do tribals who are close to nature live longer than city dwellers? If not, why not? Perhaps tribal life spans are shorter because of poverty, malnutrition and lack of medical facilities, or in other words ineffective efforts for tribal welfare. If you think tribal welfare has nothing to do with conservation ask yourself the following. Can forests be protected while forest dwellers are hungry, miserable and desperate?

Statistics as a science was born out of the study of biological problems. Peculiarity of biological phenomena is variation. The mango tree is the same and so is the caretaker, but different fruits are not identical in weight. In a eucalyptus plantation, all trees are of the same age, but their heights are different.

In physics if the law of gravitation holds for one apple, it is true for every apple. In chemistry if one drop of an acid shows a property, it is true for the whole lot. In biology things are not that simple. Every individual has his own characteristics and yet the group as a whole follows a pattern (law). Statistics talks about these group properties or population laws. When we say subabul trees are faster growing and taller than tamarind, it is not meant that every subabul tree is taller than every tamarind tree of the same age. We mean the population of subabul trees as a whole is taller than the corresponding tamarind population. We will discuss later how this can be stated more precisely or convincingly.

Environmental impact assessment is a very popular phrase now-a-days. Developmental projects of many kinds, a dam, a factory, a highway can have an adverse effect on the surroundings. Our society has become very conscious of these aspects and there is an endless debate about them. If one side puts up impressive figures in support of the project, the other side has to be equally effective in their arguments. Consider river pollution caused by effluents of sugar factories. If the discharge is very heavy relative to the flow of water, it will kill most fish and other organisms. That is what you see in city wastewater disposal systems. Then the damage caused is obvious. Around many chemical factories, streams are brightly colored or full of foam. Here environmental impact is clear. But can we measure the damage at moderate level of pollution? It will be in the form of reduced numbers of fishes of common types, disappearance of some species, and increase in numbers of organisms typical of polluted water bodies etc. All these aspects have to be estimated statistically.

We believe therefore that familiarity with methods of quantification of nature is necessary even to understand conservation debates intelligently.

Figuratively Speaking

Information in bulk has to be condensed and summarized for effective communication. It has to be presented in such a manner that space needed is not unreasonable and the reader understands it easily. One of the ways of doing this is use of graphs, charts and figures. Take any issue of a good magazine like *India Today* or *Frontline*. You are very likely to encounter one or more of bar charts, pie charts, histograms and scatter plots. It is useful to know what they are.

Bar chart: Pollination is a crucial stage in regeneration of flowering plants. *Figure* 1 describes the relative importance of different pollinators of *Martynia annua*, devil's claw (Rao and Reddi (1994)). Notice that on the x-axis we have the climatic conditions. The corresponding bars give the percentage of visitors of three types of

pollinators. Clearly digger bees are important on a cloudy day or day after rain whereas on a sunny day carpenter bee plays an important role.

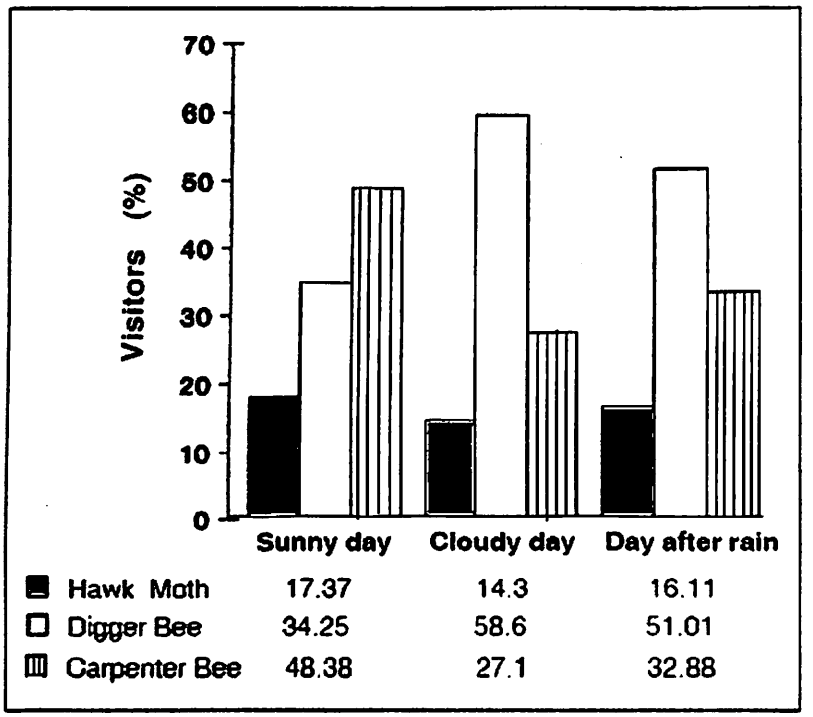


Figure 1. Pollinators of devil's claw.

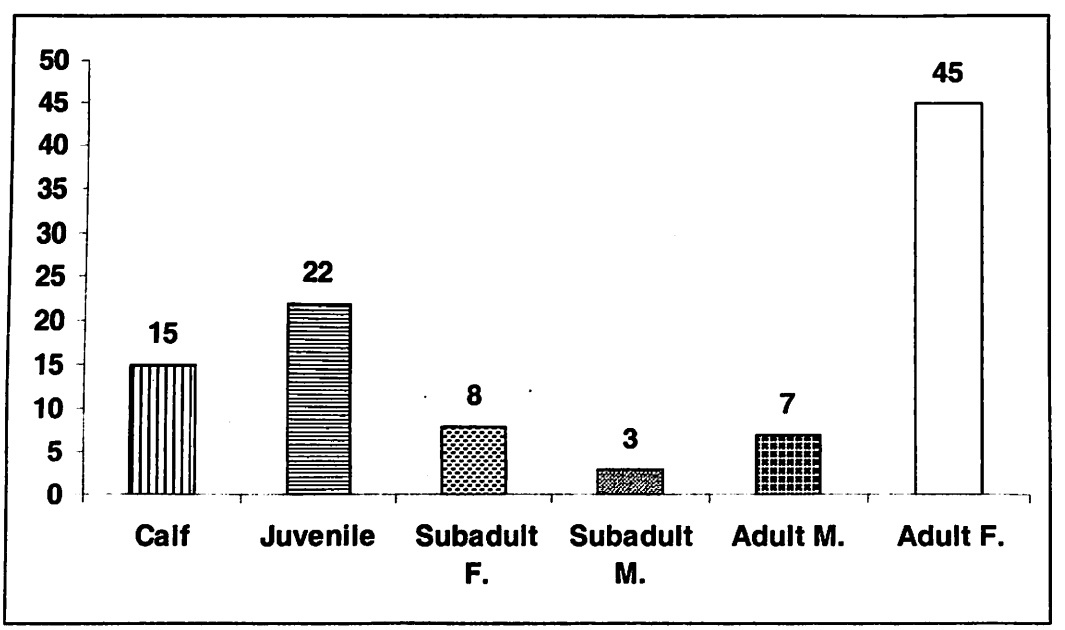


Figure 2. Age–sex distribution of elephant population in Parambikulam.

In *Figure* 2 we have age-sex composition of elephant herds in Parambikulam wildlife sanctuary (Easa and Balakrishnan (1995)). On x- axis we see the words Adult males, Adult females, sub-adult males, sub-adult females, juveniles and calves. The corresponding bar shows the percentage of males (or females) of each type. Clearly adult males are few. This may be because they are loners and missed in count. Or they are poached for ivory.

Can you imagine what a bar chart giving total annual rainfall in 1998 at Jaipur, Bombay and Cherapunji would look like?

Pie charts: Pie is a western baked food item that is circular in shape. It is cut into portions for serving just like a cake. *Figure* 3 shows the approximate composition of species of flora and fauna in India, described in scientific literature (Gadagkar (1992)). Clearly insect species are the largest in number. About 5000 mollusks are known. The number of bird species is close to 1200. Mammals, reptiles and amphibians together are about 900 species. Did you know that the insect species are so overwhelming in number? The famous British/Indian biologist J B S Haldane was a non-believer. He is supposed to have said once, "*I don't know if God exists. But if he does, he must love beetles. He created more kinds of beetles than any other animal!*" And beetles are just one kind of insect life!

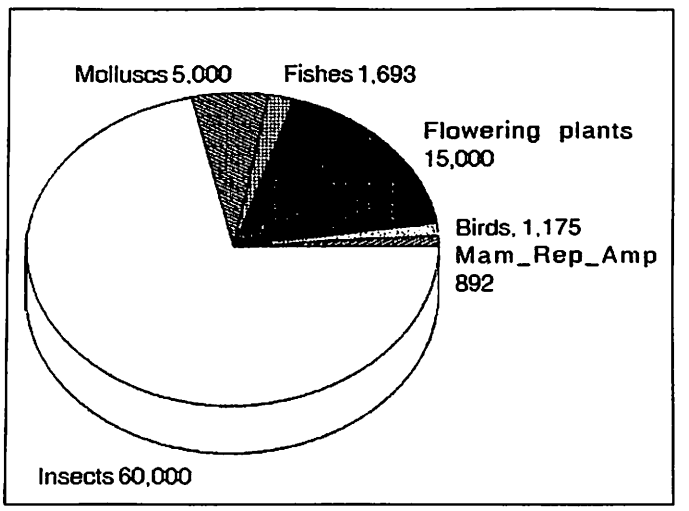
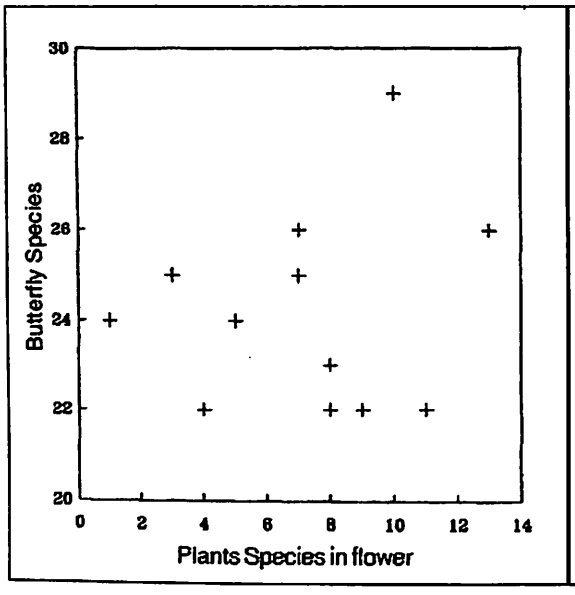


Figure 3. Described Indian species

A pie chart you will invariably see after the announcement of the government budget shows how a government earns income. The break-up gives percentage earning from income tax, excise and customs etc. Similar pie chart shows how a rupee is spent, i.e. in salaries, subsidies, development projects etc.

Scatter plot: Here is a puzzle: How are plants in flowering related to butterflies? That is trivial. Butterflies get nectar from flowers and plants get pollinated in return. Fine! Then how should number of plant species in flower be related to number of butterfly species present in an ecosystem? One guess is that greater the number of plant species in flower, greater should be the number of butterfly species present.



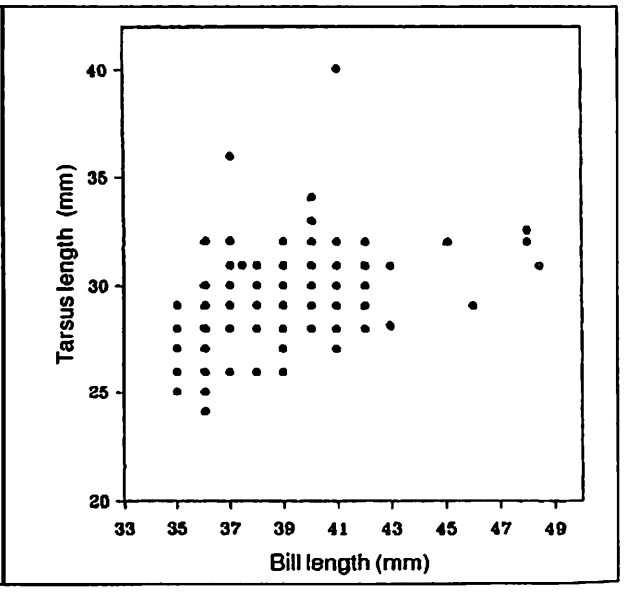


Figure 4. (left) Number of butterfly species and plant species in flower in Guindy National Park, Madras.

Figure 5. (right) Scatter plot of body measurements of Garganey ducks.

Figure 4 gives the relevant data for Guindy National Park, Madras for 1991 (Rajasekhar (1995)). One is struck by the fact that number of butterfly species occurring does not seem to be sensitive to number of flowering plant species. Perhaps the relation may hold for plant numbers and butterfly numbers, not their types. In that case a different scatter plot (i.e. two variables plotted against each other) may be needed. *Figure* 5 is an example of such a plot. It suggests that larger tarsus (i.e.leg) length is associated with larger bill length. (Remember this may not be true across species.)

We leave it to you to guess what a scatter plot of tusk length of a male elephant versus age may look like. Alternatively it could be shoulder height or pugmark circumference versus age. Can you imagine any use of this scatter plot for a field biologist?

One Number Says it All

Mean: It is true that a graph or chart can convey a message visually. But it is also essential to summarize data numerically. Consider the question of how long an organism lives. An *E. coli* bacterium perhaps lives for a few minutes or hours before dividing. An adult mosquito lives for a few days. Life span of a rice or a wheat plant is a few months. A cow or a dog lives for a few years. Man lives for a few decades. Banyan trees perhaps live for a few centuries. The longest living organisms on earth are redwood trees in California that are supposed to live a couple of thousand years. Now remember not all mosquitoes live for the same length of time, nor do all men. But a typical, representative figure is still very useful here for comparative purposes. If the need is more stringent, we may have to be more precise than above. If we wanted to compare the life span of men in India today with that of women, talking in terms of few decades will not do. Differences may be much finer. Actual arithmetic averages will have to be reported.

We all understand what an arithmetic average or mean is. But what are median and mode? Why are they needed? Mode is the most common value. Median on a road is a line that divides the road into two equal parts. In statistics median is the individual case such that half the group has smaller values than this and the other half bigger values. These are used because they are less volatile than a mean.

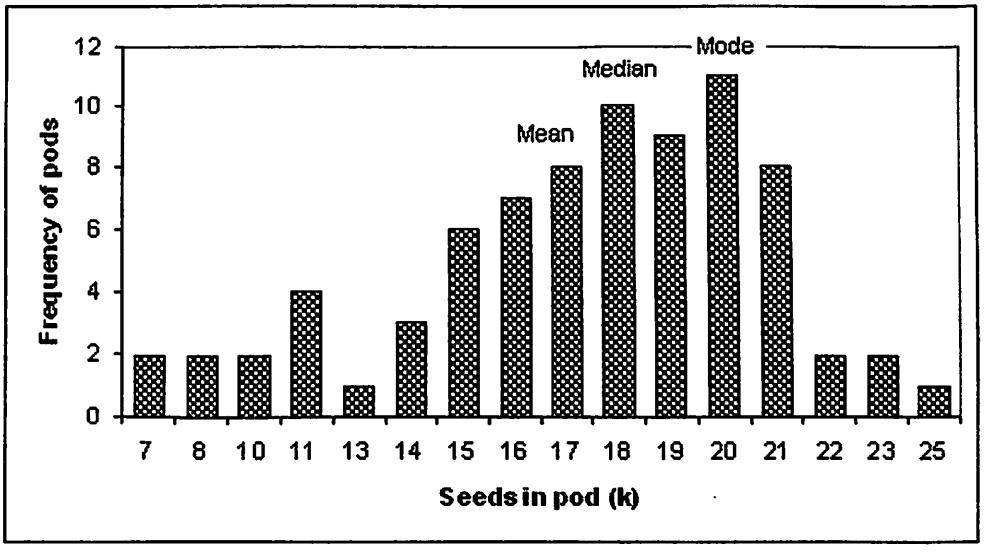


Figure 6. Frequency distribution of number of pods with k seeds/pod (Subabul).

Consider pods of subabul (*Lucena leucocephala*). The numbers of seeds in each of 78 pods were counted. The results are presented in *Figure* 6. This is called a histogram. It shows the frequencies of pods with different numbers of seeds. Mean, median and mode calculated for the data set are shown on the histogram.

Variation: One number that represents a whole set of values, like mean or median is called a measure of central tendency. But as noted earlier, variation is characteristic of biological data. Individual values are different from their average. In manufactured products, variation is undesirable. Variation shows lack of quality and lack of precision in manufacturing process. In nature, variation is the foundation of evolution. Because there is variation, some individuals are better adapted to environment. Hence natural selection favors them. If there were no variation, there would be no possibility of selecting better crop races/varieties. Variety is indeed the spice of life.

Poliomyelitis virus mutates very slowly and hence is stable. Vaccines have been developed successfully against it. Influenza virus mutates thousands of times faster. It has not been possible to develop a vaccine against it. Foot and mouth disease is a dreaded affliction of cattle. There is considerable variation in the genetic constitution of the pathogen. Hence different vaccines have to be injected for protection against different varieties.

So variation is important. Kelvin, the famous British scientist once remarked. '*If you can measure it then you understand it*', implying that if you cannot measure a thing, it is not understood well. How do we measure variation? One thing is range. Reddy and Reddi (1995) measured proboscis length of butterflies foraging on the flowers of *Clerodendrum infortunatum*. The lengths in millimeter ranged from 8 to 38.

Range is useful as far as it goes. But it does not tell us the whole story. It indicates only the extremes. Sometimes that is enough. To build cages for rabbits it is enough to know the largest size possible. For a domestic fish tank it is enough to know the range of lengths of fish species one wants to keep. On the other hand a fisherman is not satisfied to know only the range of sizes of a fish species, say Bombay duck, in his catch. He wants to know the number of fish in different size classes. The market price for each size class may be different. Range is the same if there is only one small fish and all others large or only one large fish and all others small. But the two cases are very different in terms of money they will fetch in the market. If we want to build a bridge over a river, we need to know the maximum water level. But in many other contexts extreme values are not enough. A store which sells shirts wants to stock various sizes. The manager needs to know how often each size will be asked for. In other words the entire distribution of shirt sizes has to be known.

In Bombay Natural History Society bird ringing project various body measurements were recorded for many birds. *Table* 1 shows the distribution of wing length of coots.

Table 1: Distribution of Wing Length (coots)

Length (mm)	Number of coots
≤ 190	111
190-200	1056
200-210	2926
210-220	1927
220-230	752
> 230	114
Total	6886
Mean	208.7
Mode	205.0
Median	208.9

This frequency distribution is graphically shown in *Figure* 7. We notice that histogram is just like the bar chart encountered earlier except that here on the x-axis we have wing length values, instead of just some names. The histogram is sometimes called distribution (of wing lengths in this case). The histogram shows that most birds have wing lengths near the average namely 209. As we go away from the average in either direction, the number of cases with that length declines.

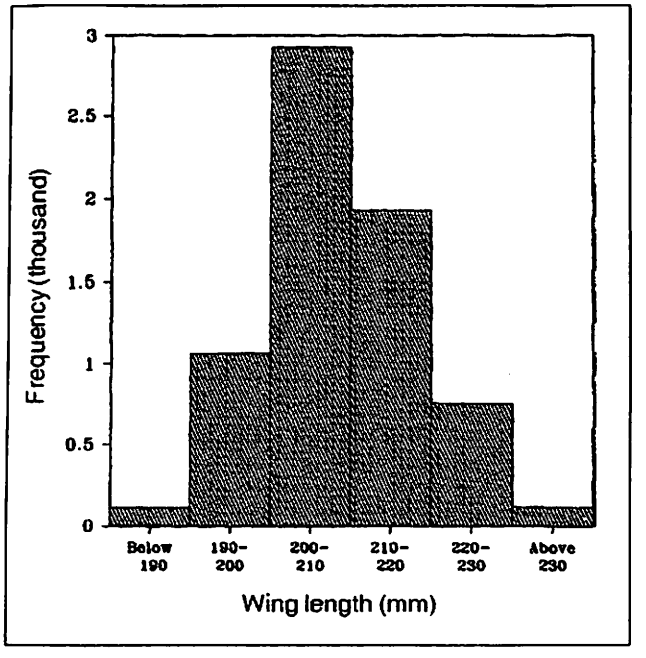


Figure 7. Frequency distribution of wing length of coots

These features are conveniently captured in a measure called variance. The formula is not critical here and can be found in any standard textbook of statistics. This measure checks how far an individual value is removed from the average, squares the difference (to remove negative signs) and averages all such squares. Standard deviation (s.d.) is the square root of variance. It is very common in ecological literature to report mean value of something together with s.d. e.g. for wing length data above s.d. is 7.9.

We will mention one more index encountered very frequently in ecological literature. That is the correlation coefficient.

Correlation Coefficient: Generally an older boy is heavier than a younger one. As ambient temperature gets warmer, plants grow faster. These are examples where two measurements are related. The intensity of relation is measured by the correlation coefficient *r*. It takes values between –1 and +1. A value of zero indicates absence of relation. A +ve value suggests that as one quantity goes up so does the other, as in the above two examples. Can the reverse occur? That is to say when one thing goes up the other declines. An obvious example is prices and demand. When a fruit is expensive, not many of us buy it. As it gets cheaper we flock to the market. In study of nature, consider activity time budget of an animal. Since total time is fixed, if time spent on one activity is increased, that on another goes down. Thus a blackbuck male, during rutting season, must concentrate on protecting his status, access to females etc. and must be constantly ready to fight with challengers. He has then no time to feed.

Figure 8. Time spent in brooding by G. cachinnans

Another example is time spent in brooding by Nilgiri laughing thrush. To help keep their chicks warm, Nilgiri laughing thrushes spend a lot of time on the nest brooding). As the chicks grow, the time spent in brooding declines as seen in *Figure* 8. The correlation coefficient between age of chicks and percent time spent by parents in brooding is –0.98.

Exercises

In one year's budget of the Government of India, the receipts and expenditure are divided into various categories as shown in *Table* 2.

Table 2. Percent contribution of various categories to receipts and expenditure

Receipts		Expenditure	
Category	%	Category	%
Excise	20	Interest payment	27
Customs	15	Defence	14
Income tax	8	States share	14
Corporate tax	9	Central plan	13
Other taxes	2	Nonplan expenditure	13
Nondebt capital receipts	6	State and Union Territories plan assistance	10
Nontax revenue	15	Subsidies	7
Borrowing	25	Nonplan assistance	2
Total	100	Total	100

1. Prepare two pie charts using Excel.

2. Prepare a pie chart of your personal activity time budget. Keep a diary of how you spend your time. Use a few simple categories like sleep, personal hygiene and nutrition, outdoor games, TV watching, interaction with other people outside place of work or education, reading etc. Draw pie charts for combined data as well as separate days of the week. Use Excel to draw the figure.

3. Measure heights of all your classmates and draw a histogram of that data (separately for males, females and combined). Do you see any differences between sexes?

4. Number of words in a sentence is a simple quantitative feature of writing style. It is said that biblical style involves short and simple sentences. Long, complex sentences are characteristic of legal pronouncements. Obtain frequency distributions of sentence lengths in different kinds of writing such as (a) editorials in newspapers or this magazine, (b) sports columns, (c) tender and employment notices etc.

5. Students in a class differ from each other in very many ways. In an English medium school, the mother tongues of students may be different, or the district from which each one hails or one's favorite color etc. Select an attribute and prepare a bar chart (using Excel).

Suggested Reading for the series

W B Fairley and F Mosteller (1977).
M O Finkelstein and B Levin (1990).
R Hooke (1983).
A J Jaffe and H F Spirer (1987).
J M Tanur, F Mosteller, WH Kruskal, R F Link, R S Pieters and G R Rising (Eds.), (1972).
H Zeisel (1957).
H Zeisel and D Kaye (1997).

2. Dice of Life

Lottery tickets are one of the hottest items on sale in India. Common people love to buy dreams of becoming millionaires overnight for a handful of rupees. All living beings play lotteries with nature. At stake are their own lives or lives of their offspring.

Gambling begins right at conception. In sexually reproducing species that are diploid, an offspring gets one chromosome out of every pair of homologous chromosomes from a parent. Which one? That is a lottery.

In human beings mothers have a pair of X chromosomes of which one is received by a son. Hemophilia is a disease in which blood outflow from a wound does not stop on its own. If a mother is a carrier of hemophilia i.e. has a hemophilia causing gene on one of the two X chromosomes, what is the chance that her son will get it and become hemophilic? 50%. Genetic disorder of color blindness has a similar situation.

Once born, an organism faces an uncertain environment. Where will it find food? Will it be sufficient? Should a bird in the hand always be preferred to two in the bush? Should an animal forage together with others of the same species or should it be a loner? Life is an unending string of decisions. Some prove to be right and others wrong. Together they decide the 'fitness' of the organism. This is the evolutionary paradigm.

Now let us try to understand this business of uncertainty and chance. What is chance? Well, the chance of an event (say success in a venture) is a number between zero and unity. If the chance (or probability) of something is zero it means the event is practically impossible. If the chance is one, it means that the event is certain. What is a chance of 50%? It means that in a long series of repetitions of an experiment, success is expected 50% of the times.

On the face of it, the phrase 'laws of chance' seems like a contradiction. Chance by definition is something that defies all rules. So how can there be laws of chance? The doubt and concern are legitimate but misplaced. Whether a particular conception will lead to a male or a female baby is indeed very difficult to guess. We presume to guess only the proportion of male births in a large set. When we discuss chance, we are not talking about a particular case but about a population. So the statement that the chance of a male birth is 50% means that in a large number of records the proportion of males should be close to 1/2.

Given this background, what is the chance of getting two sons in a row? The laws of chance say that the answer is multiplicative i.e. $1/2 \times 1/2 = 1/4$. There are 4 possible outcomes: son-son, son-daughter, daughter-son, daughter-daughter (all equally likely), and we are excluding the last three. It is because of this

multiplicative rule that on the race course, the correct guess of outcomes of several races becomes very difficult and hence it is entitled to the Jackpot. If on the other hand you bet that a horse will be at least in the second place, you are right if he is first or second. Here the chances get added, and your bet usually fetches lower rewards. If you toss a six faced die, the possible scores are 1, 2, 3, 4, 5, and 6. The chance of any particular score is 1/6. But the chance of getting an even number (i.e. 2 or 4 or 6) is 3/6. It is the sum of the chances of getting 2, 4 and 6. This is the addition law.

For insurance companies, understanding of chance is a matter of survival. Consider a motorcycle costing Rs. 10,000/-. It is to be insured against theft or loss. Suppose the chance of that is 1 in 1000. That means out of 1000 vehicles insured one will be lost (on an average) and the insurance company will have to pay Rs. 10,000/- to the insurer. So to recover this, the company will have to charge each of the 1000 insurers a premium of at least Rs. 10/-. Anything above Rs. 10 will go to cover expenses of the company and profit. If the calculation of the chance of theft goes wrong and many vehicles are stolen, the company will have to compensate all and its profits will plummet.

A foraging bird must constantly consider trade-offs between risks and gains. The saying 'no risk no gain' applies even for a merrily chirping bird. If the bird devotes itself single mindedly to eating seeds on the ground, it runs a high risk of being captured by a predator. So it must scan the surroundings and fly away if any danger becomes apparent. If the bird wants complete safety it must remain vigilant at every moment. But then it will die of hunger. So how does it manage? Do birds take a course on statistics and calculation of probabilities?

Of course not, natural selection eliminates birds which make bad choices and we are presumably left with those which make correct choices. So quantitative research in evolutionary ecology involves probability calculations to predict which type of bird should thrive in which type of environment.

We generally agree that the chance of the next human birth to be a male is 1/2. Consider families with 8 children. What is the chance that all are males? Using the multiplication law we get the answer $1/2 \times 1/2 \times \ldots$ 8 times i.e. $(1/2)8 = 1/256$. The chance of all 8 females is the same.

# Males	8	7	6	5	4	3	2	1	0
Chance	1/256	8/256	28/256	56/256	70/256	56/256	28/256	8/256	1/256

For other compositions the values are given in the table above. This is an example of the so called binomial distribution. Data on 53,680 families in England collected around 1920 had the proportions (Kunte and Jeffreys, 1992) given below.

Are these observed proportions close to what the chance calculations show? (This can be checked using a Chi-square test which is briefly discussed later).

	Observed Proportions								
# Males	8	7	6	5	4	3	2	1	0
Chance	342/53,680	2092/53,680	6678/53,680	11929/53,680	14959/53,680	10649/53,680	5331/53,680	1485/53,680	215/53,680

The answer is no. When a model does not match with data, we discard or modify the model. The data remains supreme. In this case the modification possible is that the chance of a male birth is not the same for all couples. Some couples are male prone and some are female prone. For exploring reasons of this kind one would have to go into details of human reproductive physiology.

Here is another example, this time from plant biology. In the species *Lablab niger* each flower contains up to 5 ovules. Suppose the chance that an ovule gets fertilized and becomes a seed is 0.7. Then out of 5 ovules how many seeds should we expect to get? How many seeds would there be in a pod of *Lablab niger*? Actually, there are no certainties. All ovules could fail (no. of seeds = 0) or all could succeed (no. of seeds = 5). A proper use of addition and multiplication laws mentioned above gives the chances of various numbers of seeds as given below.

No. of seeds	**1**	**2**	**3**	**4**	**5**
Chance	0.0261	0.1275	0.3058	0.3652	0.1754

The case of zero seeds is not given because it is not observable. This is an example of what is called a (zero truncated) binomial distribution (Prayag et al 1991). Two scientists from Bangalore, R Umaashankar and K N Ganeshaiah checked pods and found the following proportions:

No. of seeds	**1**	**2**	**3**	**4**	**5**
Proportion	1/69	4/69	22/69	41/69	1/69

In this case the observed proportions are quite close to the theoretical ones, after taking into account the fact that not all flowers have exactly five ovules. When you get such good models (showing agreement with observations) they can be used for prediction etc.

In genetics, a distinction is made between qualitative and quantitative traits. Eye and hair colors are qualitative traits while human height, milk-yield in cows, per hectare yield in case of rice or wheat are examples of quantitative traits. It is generally believed that quantitative traits are polygenic. A large number of genes, each with a small effect, together determine the phenotype or the actual value.

In such cases, the values in a population follow a bell shaped curve or a normal (Gaussian) distribution. In this distribution, the average and near average values are very common while very large or very small values are rare.

Is it just by Chance? In this section we discuss a rather subtle part of quantitative thinking. It is statistical inference. It becomes important because in ecology we often come across statements such as 'species A grows faster than species B' or 'in raptors females tend to be heavier than males'. How does one verify such claims? The obvious answer is 'by direct measurement'. Take a few cases of each type and check. This works in some cases. Suppose someone wanted to verify that humans are taller than bonnet monkeys. It turns out that every adult human being is taller than any adult bonnet monkey. So a direct measurement will also confirm this. But a statement 'men are taller than women' will fail such a test because some men are shorter than some women. A histogram of heights of men and women shows some overlap.

An alternative statement that is more acceptable is 'Average height of men is greater than average height of women'. How is it verified statistically? We measure the heights of a few men and of a few women and then compare the averages. Suppose the average for men is 162.5 cm and for women it is 162.1 cm. Is this enough evidence to claim that men are taller? Some may say that this difference can arise just by chance. If the difference was very large we would have believed the claim. But how large is 'very large'? Mathematical statisticians have laid down certain rules to decide this. Without explaining the technicalities we will illustrate the approach with an example.

Our interest is to check whether each of the ears in man is equally prone to develop hearing deficiencies (perhaps due to noise pollution). We check a series of case papers in an ear specialist's office. The first relevant case has a problem for the right ear. We say this could be just by chance. Our response is the same when the second and third cases are for the right ear. At the fourth case we get restless. At what point should we declare that the situation is unexpected? One convention is that a chance of 1 in 100 is small enough to say so. Let us calculate the chances of getting many cases of right ear problems in succession. We use the multiplication law and find:

Event	**Chance**	
First case is of right ear	1/2	
First two cases are of right ear	$1/2 \times 1/2$	$=1/4$
First three cases are of right ear	$1/4 \times 1/2$	$=1/8$
First four cases are of right ear	$1/8 \times 1/2$	$= 1/16$
First five cases are of right ear	$1/16 \times 1/2$	$= 1/32$

First six cases are of right ear	$1/32 \times 1/2$	$= 1/64$
First seven cases are of right ear	$1/64 \times 1/2$	$= 1/128$

So by the convention, if we get seven cases in a row of right ear problems, we can discard the idea that both ears are equally prone and conclude that for some reason the right ear seems more prone to developing hearing deficiency. The take home message is that it takes a rather substantial tilting of evidence towards one side before it is treated as 'statistically significant' and an old viewpoint is changed.

Where There is Smoke There is Fire: This old Sanskrit saying is a fine description of how we think. 'We' includes humans, birds and bees. A male moth will follow the smell of a sex pheromone of its species to a distance of hundreds of meters, hoping to meet a sexually receptive female at the end of the search. Ivan Pavlov, the Russian scientist who studied animal behavior, gave his dog some food after ringing a bell. Later, the dog would salivate at the ringing of the bell, expecting to get food right away. Deer in jungle prick their ears on hearing a monkey alarm call because the call may indicate the presence of a dangerous predator around. Something very akin to this has to be done in nature study also. We will take the example of age and girth of a tree.

What do we know about this? As trees grow old their girth increases.

We would like to guess the age of a tree by measuring its girth. That sounds interesting! But why bother? Here is a reason. If we measure girths of all trees of a species in a forest area and guess their ages, we get the age composition of the forest. Generally because of progressive mortality we expect to see fewer older trees and many younger ones. Suppose we find that five year old trees are very few. That means five years ago conditions were adverse for growth of new seedlings. This may be because all seeds were collected and removed or there was a severe drought or a big forest fire etc. If we find that proportions of young trees are consistently low, it is a cause for concern. There is inadequate regeneration. Is that species declining?

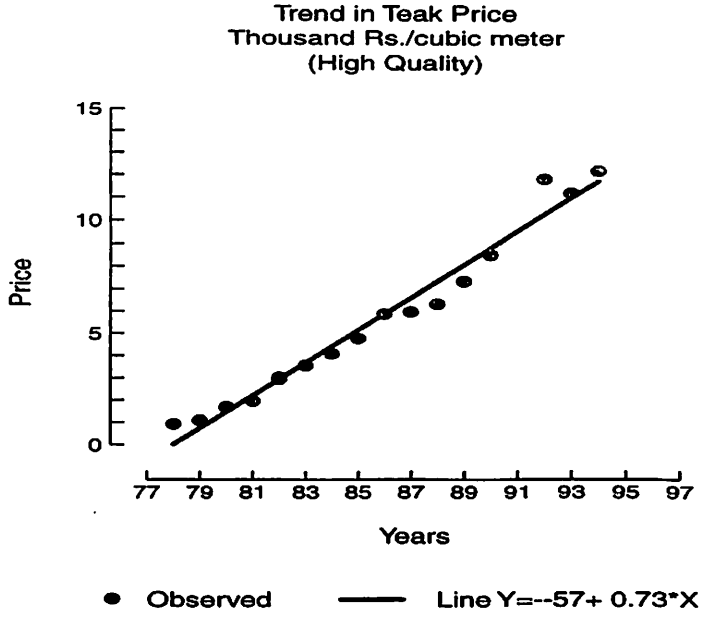
So it is useful to guess the age of a tree from its girth. Such a technique is called regression analysis. First we have to know somehow or the other, by direct record keeping if necessary, the ages of a few trees and their girths. A scatter plot has to be prepared. Finally a line or a smooth curve has to be drawn through the data points. Now if any new girth value is given, the graph can tell us the corresponding age estimate. This is only an estimate. For example some individual trees may get sick and remain thin even at an old age. But still such estimates are better than mere guesses.

If this technique is applied to tree species with very long life spans, then major events over hundreds of years can be guessed. Subash Chandran studied

forests of Western Ghats in Uttara Kannada district of Karnataka and calculated proportions of evergreen and deciduous trees in each age class (i.e. girth class) in an area. Whenever the proportion of deciduous trees was high, he argued, forest disturbance and tree felling must have been high. He was able to relate this to various historical events such as timber extraction by the British. In this sense, the history of a forest is recorded in the forest itself.

We can think of many pairs of variables similar to girth and age in terms of relationships. Annual rainfall and standing plant biomass, number of very old dead trees per unit area and number of owls or hornbills or such tree cavity dwelling birds, extent of open water (as opposed to water with lots of vegetation) in a lake and proportion of duck species that love open water, level of pollution in a lake and number of fish species present, moisture levels in a forest and density of tree frogs etc. are examples where a relationship will be revealed if data points are plotted on a graph.

One thing to remember about using regression lines/curves is the risk of extrapolation. There is always a temptation to extend the curve beyond the limits of observed data. That can give misleading results. Consider yield of wheat in one acre for a given dose of nitrogen fertilizer. As the dose level increases, so does the yield. So you plot points, draw a line passing through the points (as many as possible) and extend it freely to the right. It will mean that you can get as much wheat as you want from one hectare just by putting in adequate urea. So the whole country can be fed from one hectare of land which is obviously wrong. This has happened because the line was extended to a region where it was invalid. Increase in wheat yield occurs only at moderate doses. Beyond a point, not only does the yield not rise, it may actually come down.



Sometimes extrapolation becomes necessary in spite of the difficulties. It takes several decades for teak plantations to grow before they can be harvested. What would be the price of teak wood in the market then? *Figure* 1 shows the rising trend in these prices in recent years. One guess of future prices can be obtained by fitting a regression line to the data and extending it as shown in the figure. Separate regression lines will have to be fitted for each quality.

Associations: The techniques of correlation and regression described above are useful to study relationships between characteristics like height, weight, age, temperature, rainfall etc. These are called continuous variables.

But how does one study the relation between two attributes? By 'attribute' we mean a trait that is not quantified. Here are some examples of such traits: Group foraging or individual foraging; chase and hunt like a leopard or wait and catch like a spider; color of plumage –white like snowy owl or brown like spotted owlet. Let us briefly discuss the so called contingency tables used to study pairs of attributes.

Veena and Lokesha (1993) studied joint occurrence of drongos, common myna and jungle myna. They observed 177 feeding flocks of mynas and recorded the presence or absence of drongos.

Clearly for each type of flock, drongos are found to be around sometimes but not always. If drongos were always present (or absent) the matter would have been simple. Now the question is 'which kind of flock is more likely to have a drongo around?' In flocks of common mynas, drongos are present in (8/54=) 15% cases while in jungle myna flocks, the percentage is (21/34=) 62 %. In mixed flocks the value is still higher (70/89=) 79 %. The statistical question is whether these

differences are just due to chance or whether there are significant differences between flock types. This is answered using the so called Chi-square test. If there are no differences, then in each case the expected percentage is (99/177 = 56%).

Table 1. Co-occurrence of drongos and mynas.

Type of Flock	Drongo		Total
	Present	Absent	
Pure Flocks of Common Myna	8	46	54
Pure Flocks of Jungle Myna	21	13	34
Mixed Flocks	70	19	89
Total	99	78	177

The chi-square test checks whether the discrepancies between this expected percentage and the observed values are too large to be ascribed to chance. If the answer is yes, then the next question would be, why do drongos go with mixed flocks more often?

The answer seems to be that mixed flocks are larger and hence cause greater disturbance on the ground which causes more insects to scurry around giving greater foraging opportunities to a drongo. In that case a comparison of sizes of flocks in which drongos are present with those in which they are absent will reveal the same fact. But that will need data on flock sizes.

Exercises

1. Simulating family size under the rule 'stop reproduction as soon as a son is born': Almost every couple desires to have children. Most couples in India long for a son. Let us assume that the couple decides to stop having children once a son is born. We simulate a birth by tossing a coin. If it falls head, count it as birth of a son. If it is a tail, treat it as birth of a daughter. So stop tossing the coin as soon as a head is observed. Record the number of times you had to toss a coin (i.e. number of children). Repeat this experiment 50 times. Draw a histogram of number of children (the values will be 1,2,3...). Write a program to do this simulation and simulate family size 1000 times and prepare a histogram. How do the two histograms compare? As the number of simulations increases, the histogram smoothens out.

2. Simulating the dynamics of a farmer's cattle holding: Suppose a farmer has 2 bullocks and 2 cows. The bullocks will work well for the next five years. Will his cows generate their replacement? Many steps (uncertain) are involved. In any

year a cow must conceive (probability=.9), must carry the foetus to full term (prob.=.7), must deliver a calf and survive (prob.=.8). The calf must be a male (prob.=.5). The male calf must survive to two years (prob.=.7). Then it can be a new working bull. All probabilities given here are notional. Assume gestation period of 1 year, for simplicity. Simulation has to be separate for each cow. Check how probable it is that the farmer will get two young bullocks to replace his older animals. Can the farmer manage with only one cow? The simulation should throw light on the question of whether we have too many cows.

3. *Examining writing styles:* A person is supposed to have a literary style which is reflected in different pieces of writing by that person. One can ask whether the sentence length distribution remains the same for an author. Examine this issue with reference to a) various articles in this series, b) various editorials of *Resonance*, c) two of your favorite writers.

Suggested Reading

Kunte, S. and Jeffreys (1992)
Prayag, V. R., Paranjpe, S. A., and Gore, A. P. (1991).
Veena, T. and Lokesha, R. (1993).

Numeracy for Everyone:
3. Just for Ecologists

Most techniques in statistics are used in many disciplines besides ecology and have a wide range of applications from anthropology to zoology. They are used in industry, agriculture, social sciences and business. That is why we have called this series 'Numeracy for Everyone'. In the forthcoming parts, we will discuss applications of statistics in some of these fields. But some techniques are developed specially for ecological problems. Let us briefly look at a selection of those techniques.

Abundance Estimates or Wildlife Censuses: Let us suppose there are complaints that spotted deer in the wildlife reserve come out, destroy crops and cause damages. One possibility is that there are too many deer and culling is necessary. Then it is essential to estimate the size of the animal population.

Foresters claim that Project Tiger has been successful and tiger numbers have risen steadily. To confirm this, an estimation exercise, often called 'census' is essential. There are various approaches to 'census'. We will describe some briefly. Along with the method we also discuss limitations of these methods. Subject to these limitations they give estimates which are better in general than individual guess values.

(a) ***Water Hole Census:*** In summer, the number of waterholes in a protected wildlife area reduces to just a handful. Animals have to visit them to drink water. So observers are positioned to watch each waterhole continuously for 24 hours. All waterholes are covered simultaneously. Number of individual animals seen at each waterhole is added up. This is how lions at Gir were estimated to be around 240 in number (Hornbill, No. 4, 24-28, 1985). Two things have to be remembered here. Some animals like bison come to drink water twice a day. Secondly some animals could visit multiple waterholes the same day or some (e.g. blackbuck) could postpone drinking water till the observers disappear.

(b) ***Capture-Recapture:*** In this method some individuals are captured, tagged and released. After a suitable interval, capturing is done again. If the population is large, recapture (i.e. catching an animal that is already tagged) is rare. If the population is small, many previously captured animals reappear in the second capturing exercise. Relevant formulae have to be used to estimate population sizes using capture and recapture figures. Suppose 100 marked fish are released in a pond. Next time, 200 fish are caught, out of which 20 turn out to be already marked. So we guess that in

the pond 10% fish are marked. In other words total fish population must be 10 times the number marked. Thus our estimate of fish number is 1000.

While simple as a concept, implementation of capture-recapture method can be quite tricky. How are animals marked? In case of fish, the fin may be pierced and a wire ring may be attached. In case of birds, a thin metal or plastic strip is tied around a leg. Insects are marked using paint. Marks must not be lost or else estimates get inflated. Capturing must not hurt the animals. Some animals are captured with baited traps, using a suitable food item as bait. Some animals become trap attracted and others become shy. This introduces bias. How does one capture tigers and lions? Lions in Gir forest of Gujarat are not secretive and can be photographed in daylight. It turns out that their whisker patterns are unique and so capture as well as recapture is done notionally by photographs.

(c) ***Line Transect Sampling:*** In this method an observer walks along a chosen path and records animal sightings and approximate perpendicular distances of animals seen from the path. If animal density is high, sightings are many. If visibility is high, animals are seen even at great distances. Calculation of animal densities from such data is a rather complicated matter.

Even while using this method, one must never forget the ecological ground realities. In one study in a small national park, the number of female chitals was found to be very high in the wet season whereas the count of males was high in the dry season.

Clearly sex ratios cannot fluctuate wildly within such a short span of time. The explanation of the anomaly lies in changing male behaviour. In rutting (breeding) season males become bolder and can be observed at a shorter distance from a transect which pushes up the estimated number.

(d) ***Nearest Individual Distances:*** This method is used mainly for static objects like trees, nests, anthills, etc. Points are selected in a forest area and distance from each point to the nearest individual (say a neem tree, if we are measuring density of neem trees) is measured. If density is high, nearest distance is short. If measured distances are long, it suggests low density.

(e) ***Indirect Census Through Dung Piles:*** Here an attempt is made to estimate the number of dung piles (of say elephants or chital) per unit area. From direct observations one can estimate how many dung piles are produced by one individual per day. Also it is possible to estimate the number of days for which a dung pile lasts on the forest floor (after which it merges with the soil). Dividing the number of dung piles per unit area by the life of the dung pile and also by per day per individual output, we get an estimate of animal density. This approach has been

used to estimate elephant density in Mudumalai forests of Tamil Nadu. (See Dekker et al 1991).

(f) ***Tiger Census Using Pugmarks:*** Some foresters believe that pug mark of a tiger in the soil is essentially a signature of an individual animal. So pugmarks are located, traced on paper and compared. Similar tracings indicate the same animal. Distinct tracings provide an estimate of the number of tigers.

Many ecologists treat the above method with great skepticism. A recent innovation by Ullas Karanth (1995) is identification of individual tigers from stripes using photographs taken automatically as a free roaming tiger triggers the camera. In general, estimation of abundance of animals has proved to be a rather difficult task.

Measurement of Biodiversity: A remarkable feature of our biosphere is diversity. Animals and plants come in such a great variety as to amaze even the most casual observer. Introductory biology courses often devote a great deal of time and effort to teach taxonomy or the system of classification. In animals we have vertebrates and invertebrates. Vertebrates are further divided into fishes, amphibians, reptiles, birds and mammals. (In plants there are flowering and non-flowering types etc). The splitting goes on till we reach species (which may be further divided into races, varieties, etc.). For the purpose of our discussion, let us stay with species. The number of species of mammals or birds is known to a reasonable degree of accuracy. Discovery of a new mammal species is rare and becomes scientific news. On the other hand the number of species of insects present on earth is not known even approximately. The diversity of insects is enormous.

It is now widely recognized that biodiversity of all life forms is a resource for human beings. Plant and animal species provide us food, medicine, fodder, clothing and many other things. Secondly diversity is a nonrenewable resource. If a mango tree is killed, another can be grown from a seed. If a species is wiped out, we have no way of recreating it. From algae to angiosperms and from bacteria to baboons, the whole range of organisms is of potential use for mankind. Hence loss of a species of any kind is a definite loss of resource. In 1992, leaders of many nations in the world gathered together at Rio de Janeiro in Brazil and signed a convention in that Earth Summit. The convention asks each country to prepare an inventory of its biodiversity and monitor any changes in it. So we must know how to measure biodiversity.

Thinking about biodiversity and its measurement can be difficult for some traditional biologists. Typically they concentrate attention on one taxon or sometimes even a single species. Contemplating the whole biosphere quantitatively may seem rather bewildering. But one can start with simple steps. Some countries

have greater biodiversity than others. India is a mega biodiversity country. So are Indonesia and Brazil. In contrast temperate countries have low biodiversity. Some specific geographic areas have very high levels of biodiversity and are called hot spots. Western Ghats and North Eastern Himalayas are two hotspots in India. On the other hand deserts and mountaintops have very low diversity. From such broad generalities we can come down to smaller localities, a district or a tehsil or even a watershed. Also we can narrow down the discussion to a taxon of interest e.g. trees. Measurement of biodiversity has two features – species richness of an ecosystem and evenness. Greater the number of species, richer is the ecosystem. Also, greater the evenness of species, greater is the biodiversity.

If our target is a small area, say a sacred grove, we can prepare a list of tree species. Birdwatchers often have such checklists for areas of interest to them. If the area is large, a complete and thorough check is not possible. So the method of sample checking has to be adopted.

Species Individual Curve: Suppose we select trees randomly and identify the species (for which good experience is necessary). Then as the number of individuals checked goes on increasing, the number of species encountered also goes on increasing. Initially, it is easy to find a new species. But gradually, as the accumulation progresses, it becomes harder and harder to come up with a new species. A graph of the number of individuals seen versus the number of species recorded is called a species individual curve. This curve gives us a reasonable idea of the total number of species.

Other things remaining constant, greater the number of species, greater is the biodiversity. The remarkable degree of variation in species richness can be illustrated by a study of all plants in 50 hectares of forestland in Malaysia, India and Panama. The number of individual plants and number of species are given in Table.

Table 1. Tree count in 50 ha. plots in 3 countries.

Country	Malaysia	India	Panama
# speies	816	71	303
# plants (living stems above 1 cm diameter)	3,35,000	27,000	2,44,000

(If we observe 10 times as many trees in India by going over greater area if necessary, will we come across as many species as in Malaysia? The answer appears to be in the negative.)

Now two places with the same number of plant (or animal) species are equally species rich. But should they be regarded as equally diverse? Suppose forest patch A has 99 ber (*Zizyphus* sp.) tree and 1 babul (*Acacia* sp.) tree. Patch B has 50

ber trees and 50 babul trees. Species richness is the same. But patch A seems vulnerable. Loss of 1 babul tree will make it a monoculture of ber. Its diversity is less. If number of species is the same, greater the evenness of numbers of individuals, greater the diversity. Hence the issue of relative abundances becomes relevant. In a typical community, there are a few abundant species and many other species occurring at low level of abundance. When we talk of a bamboo forest or sal forest we mean that the concerned species is very abundant.

Table 2 gives the summary of data presented by Parthasarathy and Karthikeyan (1997). The number of woody tree species is given by abundance. Thus out of 482 trees recorded, about 1/6th are just one species, 50% are covered by 4 species. 23 species are represented by just one individual each.

Table 2: Species Abundance Distribution

Abundance	Number of Species	Abundance	Number of Species
79	1	10	1
72	1	8	1
57	1	7	4
40	1	6	3
32	1	5	2
14	1	4	3
13	1	3	3
12	1	2	6
11	3	1	23

If our study area has a small number of species, we can list the species and mention abundance level of each. If the number of species is large, this becomes impractical. Instead, it is customary to use certain indices, which summarize this information. We will mention two common indices.

Simpson's index: If there are k species with relative abundances p_1, p_2 ... p_k, Simpson's index of diversity is given by $1-\sum_{i=1}^{k} p_i^2$. It represents the chance that two individuals selected randomly will belong to different species. For the above data, the value is 0.9277.

Shannon–Wiener index: Its formula is $H=-\sum_{i=1}^{k} p_i \ell n(p_i)$ where *l*n is the natural logarithm. For the above data, the index has a value of 3.3523.

These indices have higher values if different species have similar abundances. If one or a few species dominate the scene, the indices come down. These indices are said to measure species diversity or α diversity. By β diversity we mean dissimilarity between two sites. Similarity can be measured by the number of species common to the two sites (suitably normalized by total number of species). A popular index of similarity is,

Jaccard Index $= \frac{n_c}{n_1 + n_2 - n_c} * 100$; where n_1, n_2 are number of species in two sites, n_c is the number of species common in two sites. If no species is common in two sites, similarity is zero i.e β diversity is 100%. On the other hand if all the species are present in both the sites, similarity is 100% or in other words β diversity is zero.

Diversity indices are useful mainly for comparison between ecosystems over space and/or time. See *Figure* 1 which gives a plot of species richness (one possible diversity index) for all birds ringed by Bombay Natural History Society bird ringing team (Gore and Paranjpe (1995)). The graph has a peak at 1981 and a decline up to 1987 and again a rise afterwards. The cause of this pattern needs to be explored.

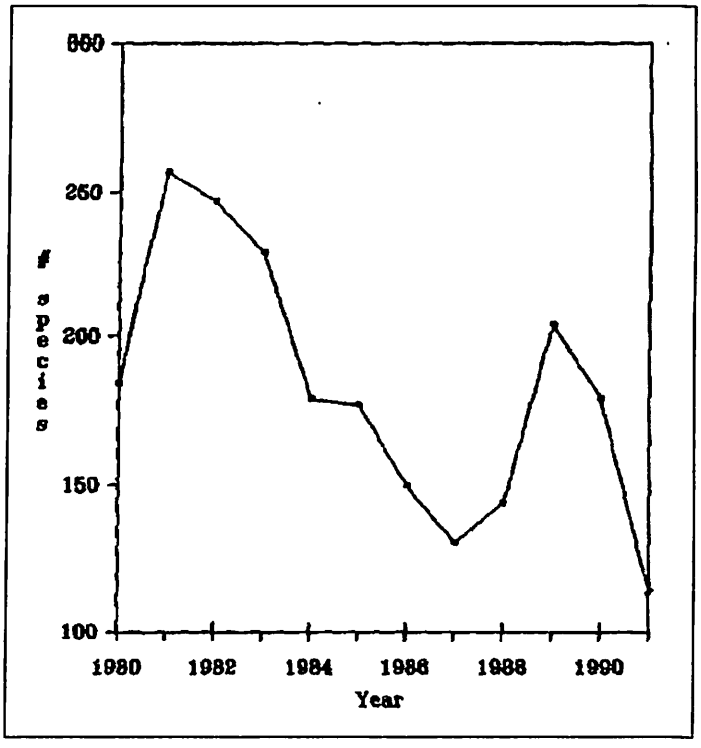


Figure 1. Yearly species richness in BNHS bird ringing data.

We talk about diversity in a particular group of organisms (birds, trees, bacteria etc). We could also talk about diversity of intestinal parasites or worms. Watve and Sukumar (1995) studied diversity of intestinal worms of wild animals. They measured egg density in dung. They proposed various possibilities regarding patterns in this diversity. Elephant gut should have higher diversity than hare or porcupine. Gregarious animals (e.g. chital) should harbor a greater variety of parasites than solitary animals (e.g. tiger). Animals with multiple stomachs should have greater diversity than animals with single stomach. Carnivores should show in their faeces greater variety of worms than herbivores. As you can see, this is an excellent way of relating host with parasite. Their results show that the only relationship confirmed by empirical evidence is between predation pressure and parasite load. As the predation pressure increases, prey animals with higher parasite loads, being weaker, get eliminated.

Sustainable Harvesting: This is a crucial concept for utilization of any renewable biological resource such as pasture, timber, fisheries and so on. We do need to extract biomass for food, fuel, fodder, fertilizer, fiber, etc. But we must do it in such a manner that the resource is not exhausted and we can get more when we need it.

Consider a tree that provides leaves for animals. We can lop all its branches and get lots of fodder at one time. But then the growth of fresh leaves will be very slow. Instead it is better to lop it less heavily. This simple principle of moderation has to be applied to get a sustainable yield.

Consider grass. This is an important source of fodder. In traditional arrangements, each village has a grazing area. As the monsoon rains begin, new growth of grass makes the grazing land emerald green. In summer cattle are often starved of greens. So there is a great temptation to let the cattle graze there right away. This is an unwise practice. Growth of grass tends to be sigmoidal as shown in *Figure* 2. Very roughly speaking, growth is very slow initially in June. It picks up in July and tapers off in August. Flowering is in September. Maximum benefit is gained if we harvest grass just after the period of fast growth i.e. roughly say end of July. This may be called the optimal harvesting policy. Given a sigmoidal growth, how to identify the best time for harvest is a technical matter, which we shall skip.

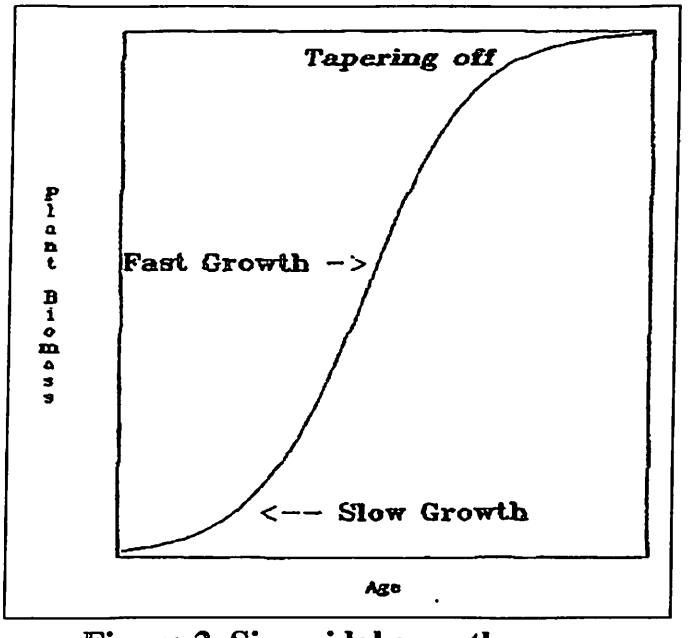


Figure 2. Sigmoidal growth curve

In case of fisheries it has been found that managing the resource sustainably is very difficult. One traditional practice that helps greatly the conservation of fish resource is that of having a refugium, i.e. an area where there is no harvesting. This forms a place for breeding, restocking, etc. Sacred groves act as the plant refugia in forest areas.

It is very important for nature lovers to know traditional practices important from conservation view point, to evaluate their efficacy and to promote them if found useful.

Animal Behavior Models: While it is pure joy to observe animal behavior, understanding it is a tough scientific task that has been taken up with gusto in recent decades. The evolutionary approach to study of animal behavior assumes that animals optimize. They extract the highest benefits out of any activity. If not, a mutant that works better will replace them.

(a) ***Clutch Size:*** Usually clutch size or number of eggs laid by a bird in a breeding season is fixed. This is optimum in the sense if more eggs are laid, due to food shortage or other reasons, mortality rate in chicks goes up and in the end fewer offspring survive.

(b) ***Prey Choice:*** Fish eaters know that some fish are full of small bones that have to be separated with effort, patience and skill before a morsel of tasty meat becomes available. Other fish are easy to handle and clean. Predators generally encounter

some prey types that are tedious to handle, clean, open, etc. (bad) and other types that are easy (good). Predators are very smart and choose the prey so that they get high reward with least effort. If both good and bad types are available aplenty, they choose only the good ones. If good ones become scarce, they take whatever comes their way. All these ideas can be made very precise. They lead to quantitative expectations verifiable through careful experiments.

(c) ***Central Place Foraging:*** A foraging bird has to search food and bring it back to the nest to feed the chicks. It is necessary to work such that feeding rate of chicks is commensurate with their phenomenal growth rates. To optimize food gathering effort, birds adjust the time spent on a patch searching food to the distance of the patch from the nest. If the patch is near the nest, birds make short, quick trips bringing small amounts each time. If the patch is far, they spend much more time gathering food so that the extra travel time is not wasted. Is it not similar to what we do? We don't mind a trip to the corner store even for a single matchbox. If we go to a far away place for shopping we make elaborate preparations, shopping list, etc. That makes the long journey worth while.

Exercises

1. Make your cat or dog walk on wet and dry soil. Try to trace outlines of pugmarks by keeping a piece of glass on top and using a marker pen. Try to quantify the shape of the pugmark. Assess visually if there are differences between front and hind pugmarks or between left and right pugmarks. Check if replicate pugmarks of the same foot (say rear left) are equal or just similar.

2. Fill a bottle with a known number of beads of a color (say white). The aim is to estimate this number using capture-recapture technique. Add a small number (known) of beads of different color. These are our 'marked' beads. Now shake the bottle well and take out a sample of the mixture. Find the proportion of marked beads. Hence, by rule of three, estimate the number of white beads. Repeat this a few times to see the extent of variability in statistical estimates.

3. Take a pack of playing cards. Suppose the four suits Spade (S), Heart (H), Diamond (D) and Club (C) represent four species. We know that each species is in the same proportion (13/52 = 1/4). Calculate Simpson's index. Now shuffle the cards and take a sample of say 12 cards. Count the number of cards of each suit and hence calculate the relative 'abundance' of each 'species'. Now calculate Simpson's index for this sample. Repeat a few times and see how close these estimates are to the 'true' value. Can you write a computer program to simulate this procedure?

Numeracy for Everyone:

4. Numeracy in Research Planning: How to Check if Rice is Cooked

We take a pinch of grains of rice and press to check if the core is soft. If yes, we declare that rice (we mean all grains) is cooked. How do you know? The unstated assumption is that all grains of rice in that pot are alike (in an identical state of being cooked). If this assumption is not true, our conclusion can go wrong. This predicament indeed does occur if the heat of the stove is too strong while cooking. Then we get multi-tiered performance. The bottom tier is charred and black. The middle layer is edible while the top layer is still raw. Why don't we check every grain to rule out errors? Simply because, it is cumbersome, time consuming and usually, unnecessary!

All these considerations are equally relevant in many studies. We take water samples, air samples and soil samples. Studies of sample plots in forests are common. By sampling we hope to know about a huge entity at low cost. This is a valid and successful approach. It is only necessary to be careful about assumptions. Is the entire forest area of interest uniform? River banks, eastern and western slopes of hills and rocky outcrops may have very different vegetation. So it is best to divide the forest area into strata or divisions such that each division is relatively homogeneous but there are differences between divisions. Some sampling has to be done in every stratum. If a stratum has a lot of variation in it, more effort has to be spent on studying it. If a stratum is very uniform a few observations will suffice to show what is in it.

How large should a sample be? Or how many individuals should be examined to estimate reliably some feature of the population? Generally, larger the sample size, greater is the precision of an estimate. But we cannot go on observing because resources in terms of time, money, personnel, etc. are limited. While the question about sample size is very important, it is also very difficult to answer in general. If we want to know the number of mammary glands in different mammals, observing one or two adult females in a species suffices. This is because the trait is almost without variation. The same is true of clutch size in birds. On the other hand if our interest is number of eggs laid by frogs or size of crowns in tree species, many more observations would be necessary. If a population is small, like trees in a few hectares of forest, 5% sampling is usually adequate. If the target population (not necessarily human) size is very large, this rule becomes inappropriate. Instead it is the actual sample size that becomes crucial. When election patterns are predicted, though the size of a parliament constituency is of the order of 10 lakhs, a few hundred individuals selected carefully and interviewed skillfully can indicate the result.

Ecologists often want to measure diversity in an ecosystem. It appears that here also sampling a few hundred individuals does the job quite well. A good familiarity with the population to be sampled is very essential for effective sampling. If we want to estimate sex ratio in black bucks or langurs (the black faced monkeys) or elephants and we examine some breeding groups, we will find that adult males are very few. We have to know that large number of males live separately as bachelor herds or just loners. Otherwise we will get a totally misleading answer even though we use correct statistical procedures.

Having discussed broad outlines of sample survey methods, let us turn to some specifics. In sampling we talk about a 'frame'. It is the list of all individuals in the population of interest. Such a list is available in case of some studies but not always. Electoral roll of a constituency is the frame for sampling of voters. List of houses in municipal records is another example of a frame. In an educational institution, students enrolled are all listed. Given that a frame is available, given that a sample size (number of individuals to be checked) is decided, the important task is to identify in the frame, the individuals to be interviewed or subjected to some measurement. The basic method for this is 'simple random sampling'. Let us see how to select such a sample.

Simple random sampling: Suppose we have a list of one thousand individual students, houses, shops, families or whatever and a sample of 50 is to be selected randomly. We number the cases from 000 to 999. Now we refer to a table of random numbers and select 50 triplets. Suppose the first triplet in the random number table is 178. Then the individual corresponding to that number is located and relevant observations are recorded. If instead of a printed random number table you have a computer, it can generate random fractions. Multiply the fraction by 1000 and use the integer part. There can be small problems in any such procedure. Suppose our frame contains only 500 cases. Then the above procedure may have a hitch. What if the random number selected happens to be above 500? (It will never exceed 1000 since we are multiplying a fraction by 1000). Well, if the number chosen exceeds 500, one can divide the number by two and ignore any fractional part. The idea is somehow to ensure that each case has the same chance of being selected. We have mentioned earlier that a population should be split into homogeneous subgroups called strata (to ensure higher efficiency). In such 'stratified random sampling', sample size is decided for each stratum and a simple random sample is drawn from it.

Systematic Sampling: One difficulty about simple random sampling is that often the selected individual is hard to locate, or if selection is to be done by field staff they may get confused. Hence sometimes a modified selection procedure is recommended. In our problem of selecting 50 cases out of 1000, we can start by selecting a random number from the set 1 to 20. Say 13. Then we go on adding 20

to this number. Thus 33, 53, 73, 93 become the selected set from the first hundred. 113, 133, etc. are selected from the next hundred and so on. This is systematic sampling with a random start. While there is convenience in it, a word of caution is in order. If there is a rhythmic pattern in the list and we start with a high value, we may get all high values and hence inflated estimates. In American cities, residential areas are divided into blocks of roughly equal size. To choose a sample of houses we choose one from a block and select the corresponding house in each block. But suppose our random start gives us a corner house. Then all cases in the sample may be corner houses. If corner houses have larger plot size or higher market value, we have a sample that is not representative. Eliminating such biases becomes possible with experience of studying the same or similar populations again and again.

Cluster Sampling: Suppose we wish to study agricultural practices of farmers in a district. Suppose there are 1000 villages and about 200 farmers in each village. We plan to interview about 1000 farmers. There are various ways to select these. Select one farmer from each village. Select a simple random sample of 1000 from a frame of two lakh farmers. A drawback of these strategies is that travel time increases immensely. In the first option the interviewer has to go to every village. In the second option traveling is less but still considerable. A strategy to avoid this is to select five villages randomly and interview all farmers in those villages. For this strategy(with very low cost of traveling) to be satisfactory we need that each cluster should be representative of the whole population of interest. Variation from village to village should be low. Variation between farmers in the same village should be high. In the district of Pune, where we live, the assumption is not valid. Western part of the Pune district is hilly and has heavy rainfall while eastern part is in plains and with scanty rainfall. In such a case, sampling should be done separately in two parts.

Mail Questionnaire Surveys: One common method of sample survey is to send questionnaires to a random sample of respondents. This saves cost and effort of travel. It allows the respondent to answer the questions leisurely after thinking about them instead of giving impromptu responses under pressure. So far so good! There is one major drawback of these surveys. It is that of non-response.

Many people simply ignore the letter requesting them to fill a form and return it (even when postage is prepaid). There are many reasons for this, lack of time or interest, diffidence about the questions, fear of losing confidential information or even sheer laziness. Someone will argue that we should estimate the degree of non-response and compensate for it. If we need a sample of 500, and only half the people are expected to respond, send out 1000 letters. This is correct. However there is one additional problem. Non-respondents often tend to be different from respondents.

If questions are about one's personal income, people with high incomes are reluctant to answer (or let us assume that it is so). In that case only low income individuals will respond. If we use this information to estimate average income, we may get a serious underestimate. This problem is genuine.

In a democracy, a vociferous minority can create an atmosphere in favor of this or that political decision. But we must never forget the silent majority. A truly democratic decision should take serious note of what the majority believes or feels.

So what can we do about the problem of non-response? One pragmatic view is that an attempt should be made to interview at least a few non-respondents by extra effort. This information should be added to whatever else is available before pronouncing the final estimates. This brings us to the issue of questionnaire construction. It involves expertise of marketing specialists and psychologists. However a few generalities can be clear to anyone. Firstly a questionnaire has to elicit answers from a person over whom you have no (or marginal) control. If the control is significant, the response loses its credibility anyway (Who will (should) believe it if an employee praises his boss?) So it is most important to humor the respondent and to avoid irritating her/him. Hence the questionnaire should be as brief as possible. Answering the questions should not require any complicated calculation or thinking. Yet the question must not be suggestive. "Do you use toothpaste A?" is not a good question. It is better to ask "which brand of toothpaste do you use?" Sometimes there is a difference between the intent of the surveyor and the understanding of respondent.

If you ask "what is the cause of ozone depletion over Australia?" and the answer is "I do not know but I closed all gas cylinders," you are sure of miscommunication. To avoid this confusion, it is customary to conduct a pilot/trial survey and make any corrections in the survey strategy. There is one category of questions that is quite important but difficult to ask. We know that some students cheat in final examinations. We want to estimate the proportion of such students. It is pointless to ask "did you cheat in last year's final examination?" Everyone will say, "No, I did not". No one wants to be caught confessing to such misdemeanor. Statisticians have found a way to avoid invasion of privacy and yet to get estimates needed. The method involved is called 'randomized response'. In this approach we offer two questions with yes or no as possible answers. Here is an example.

Q1. Did you copy or cheat in the last examination?

Q2. Were you completely honest in the last examination?

Everyone knows these questions ahead of time. Then a respondent throws a die. If the score is 1 or 2 he is to answer Q1. Otherwise he answers Q2. No one but the respondent knows which question he has to answer. Hence he can give a candid reply without fear. Suppose out of 100 students 40 say yes and 60 say no. What can we conclude about degree of cheating? The estimate of proportion of cheaters is

given by the equation $0.40 = \frac{1}{3}p + \frac{2}{3}(1-p)$. Here p is the unknown proportion. (Try to see how this equation arose). Solving it we get $p = 0.8$.

Another variant of the method involves asking a different question as Q. 2. For example, 'were you born in the first trimester of the year (i.e. Jan - Feb - March)?' We know that roughly the probability of an affirmative answer is a quarter. We will let the readers think about how to write an equation to get p.

Sample surveys are a powerful tool in every walk of life. Experience in sample surveys can be useful if you seek a job in marketing. Even otherwise, you will come across many reports based on sample survey and knowing about the technique will help you interpret the reported material better.

Designing an Experiment:

Preliminaries: Experiments are the heart of science. Scientists believe what they observe and what their experiments confirm. A good experiment is one which yields an unequivocal conclusion. Every conceivable objection is raised before a conclusion gets accepted by the scientific community. This skepticism is crucial in scientific enquiry. We need to design an experiment in such a way that there is minimal room for skepticism. It is widely held that overgrazing by cattle affects grassland adversely. Can we design an experiment to check this? Yes. Go to an area where cattle graze. Build a fence around a small portion of the area, to exclude cattle. Come back after say 4 weeks. If condition of grass inside is different from that outside, then cattle must be the reason. If there is no difference, cattle have no impact. Does this sound reasonable?

Wait a minute. What if the experiment was done in May and there were no summer showers? Then all grass would be yellowed and dead even before the experiment begins and no perceptible change may occur. So perhaps it should be done in August. To consider different ways of plugging loop holes, let us discuss a simpler experiment that you can perform at home. There are two companies A and B, each selling a soap to wash clothes. We wish to know which soap is better. So we design an experiment.

Attempt 1. You wash your shirt with soap A and your friend washes his shirt with soap B. Check which shirt is cleaner. If your shirt is cleaner, declare soap A as better. Remember such a conclusion if widely known and believed, will reduce the sale of soap B. Hence that company will quickly raise objections to discredit the conclusion. Here is one objection. Anything can happen in one case. You should try soap B many times. Only then will the truth come out. This is a valid objection. We need adequate 'replication'.

Attempt 2. You and your friend wash five clothes each and a cleanliness score is given to each piece. If your total is greater, soap A wins. Now defenders of soap B

raise another objection. You wash more diligently and for a longer time. Hence your score is better. Not because soap is better. This objection could be wrong. But the experimental design did not take care of it explicitly.

Attempt 3. Each person devotes precisely the same amount of effort. Suppose soap A still fares better. Then there can be yet another objection. Your friend is an athlete while you are a sit at- home type person. So your clothes are not as dirty as your friend's. Well this objection can be taken care of using only your clothes to test both soaps. This is the principle of 'local control'. Comparisons are valid if experimental units are similar.

Even this precaution may not be enough because the objection may be that soap B is really better for dirty clothes while you tried it only on relatively clean ones. Remember, the side that loses, will leave no stone unturned in the attempt to discredit the experiment. So you will have to run parallel trials with exercise clothes and normal clothes. But the objections will not end. The opponent may attempt hair splitting by arguing that even among your clothes some are cleaner than others and they went to soap A. How do we ensure absolutely 'level playing field'? Perhaps the best thing would be to wash one half of a shirt by soap A and the other half by soap B. Even then there can be an objection that soap A was used to wash left side while soap B to wash right. And it is the right armpit that gets more sweat. The last resort in designing a fair experiment is to allot sides of shirts to soaps by tossing coins. This is the principle of 'randomization'. When we know ahead of time, about any systematic differences in experimental units, we take care and balance things out. Randomization is the tool to even out factors that may be present but unknown.

Replication, local control and randomization are the three cannons of the art of designing good experiments. In science there are usually no favorite sides and no vested interests but we take good care all the same so as not to be misled.

Let us consider one experiment carried out a few years ago by a Swedish ornithologist named Malt Anderson, to see how he handled such problems. In the theory of evolution, there is a concept called 'sexual selection'. This concept suggests that males develop extraordinary traits (e.g. peacock's plumage) in response to female attraction for those traits. Anderson wanted to verify this proposition in the case of an African bird species called blackbird. In this species, males possess very long tails. If this is due to female preference, Anderson argued, then males with short tails should have lower reproductive success in terms of number of females that lay eggs in the territories of the short tailed males (and number of eggs laid). So, analogous to soap A are the ordinary males with long tails. Analogous to soap B we should have short tailed males. Now in nature all males have long tails. (Presumably short tailed cousins have been wiped out over time). To overcome this problem, Anderson cut off a portion of the tails of a few birds. If the experiment is done in the same area and in the same breeding season

local control is taken care of. With effort one can ensure enough number of males of each type (precisely how many is a difficult question and can be answered only with experience).

Is this design adequate? Are all loopholes plugged? If males with long tails get greater number of females in their territories and greater number of eggs laid (presumably fathered by them) is the hypothesis of sexual selection to be treated as confirmed? Anderson anticipated two possible objections.

Firstly, shortening of the tail may affect flying, fighting or foraging abilities of males and females may avoid them for that reason, and not because short tail is ugly. This objection was taken care of, by actually measuring various abilities of the short tailed males. Anderson could confirm that there was no adverse effect on those males.

The second objection is more subtle. Perhaps the act of cutting its tail may have affected the general behavior of the male in some manner noticeable only by females but not by scientists. If so the attractiveness or otherwise is not in tail length but that behavioral feature. To protect against this possibility, Anderson introduced in the design a third group of male birds. These had their tails shortened but the portion cut out was pasted back. So the group suffered the act of cutting without actually getting a shortened tail. We shall return to this kind of arrangement when we discuss clinical trials and the concept of 'placebo'.

At the end of the experiment Anderson found that short tailed males had lower rate of reproductive success whereas the other two groups had higher rate. So it turned out that the act of cutting does not have any impact but shortened length of tail does create a handicap for a male bird. In this sense, females seem to have promoted long tails.

Having generally outlined the approach to design of experiments, let us mention a few basic terms used commonly.

Completely Randomized Design (CRD): Suppose we want to compare several hormone treatments on flowering and fruiting of teak. We identify several trees of the same variety, similar age, in similar soil condition and allocate trees to each hormone treatment randomly. Presumably there are no differences among individual trees and any observed differences in fruiting are attributable to hormones.

Randomized Complete Block Design (RBD): If we have several varieties of teak, we should compare the hormones for each variety. Block is a group of homogeneous experimental units. Here all teak trees of a given variety form one block. Within each block we allocate trees to hormones randomly.

Incomplete Block Design: Here the group of similar individuals is very small. Suppose we wish to check if bird songs are inherited. Our experiment involves rearing in isolation and in different levels (low, medium and high) of

exposure to other birds of the same species. Thus we will be comparing four 'treatments'. For this we need littermates which will constitute a natural block. If litter size is only 2, we get into a problem. We cannot have complete blocks in the sense that within a litter we can only try two out of four treatments. This can be handled by trying different treatment pairs in different blocks.

Factorial Experiments: Traditionally scientists adopted a 'change one thing at a time' approach in designing experiments. This means all other conditions were held constant and only one 'factor' was changed to study its effect. Suppose we wish to check if it is economical to use higher doses of fertilizer in growing rice. So we keep all other things constant and use less fertilizer in some fields and more in others. This is quite legitimate. But it has two drawbacks.

Often the number of factors is large. For a biscuit maker the following factors may be important in determining biscuit quality: (a) type of wheat, (b) type of oil and its quantity, (c) oven temperature, (d) amounts of sugar and salt. As the number of factors goes up, the number of trials needed increases and the experiment becomes more and more expensive.

Second problem is that of interaction between factors. Suppose two factors of interest are nitrogenous fertilizer and irrigation in growing wheat. It turns out that the level of fertilizer use that is best depends on the amount of irrigation available. If irrigation is plentiful, high dose of fertilizer may be good. But if irrigation is not available, high dose of fertilizer may not be useful. So we have to study combinations of factors. Sir Ronald Fisher developed the technique of designing experiments in which many factors are changed simultaneously and yet valid inferences about each can be drawn. For recent account of Fisher's contribution see Krishnan (1997).

Exercises

1. Design a survey strategy to estimate the market share of different daily newspapers in your city/town. How will you obtain the profile of a typical customer of each newspaper?

2. Identify four surveys reported recently in the newspaper you read. Check whether there are (can be) some loopholes in these surveys.

3. How will you generate a triplet of integers 0 to 9 using only a coin (or a six faced die)? Can you show that your procedure ensures equal probabilities for all cases?

4. Aim: to check effect of three treatments on quality of chapati. The three treatments are: after preparing the dough: a. immediately make chapattis, b. Keep dough for some time and then make chapattis, c. Keep dough in refrigerator for a couple of hours and then make chapatis. Design a suitable experiment.

Numeracy for Everyone:
5. Numeracy in Medicine and Public Health

Epidemic is occurrence of a large number of cases of a disease in a locality at a time. Similar phenomenon in animals is technically called an epizootic. But one is often not particular about this. Epidemiology is the field in which scientists play detectives and try to identify 'reasons' behind the event. This is done in many ways. Let us consider some examples.

Cholera: In 19th century London, epidemics of cholera were not rare. In an attempt to identify any commonalities among families with cholera patients, it was found that there may be some connection with the company from which the family sourced their water supply. There were several companies in the business of city water supply. Occurrence of disease seemed to go hand in hand with water supply from some companies and not others. This seemed curious since all companies pumped their water from river Thames. Further inquiry showed some companies, pumped water quite upstream and others downstream below where city waste water was released into the river. Clients of latter companies were more prone to getting cholera. Given this diagnosis, corrective measures seem clear. Is it not? Of course scientists can go deeper into the question of what in the waste water may be the cause and this may further help prevention as well as cure. Preventive and social medicine (PSM) is a subject included in standard medical curricula in India. However, it seems to suffer from low status and neglect. (We recommend that the relevant map of London available on wikipedia should be examined. In fact one hand pump is identifiable that is surrounded by many cholera cases.)

Heart disease: Epidemiologists talk about factors that influence the chance of an individual getting heart attack. The chance increases with age. Higher the cholesterol content of blood, greater is the risk. So age is called a risk factor. Habits such as smoking, drinking alcohol make an individual prone to heart disease. On the other hand regular moderate exercise, balanced and just adequate diet, absence of obesity make it less likely for a person to be affected. Other things remaining constant, males may be more vulnerable than women. A fast and stressful life style can make achievers more prone while a sedate and regular life style may make one less prone to heart attack. All such conclusions are based on statistical analysis of sample surveys.

Sexually transmitted diseases: Now a days, we are all very concerned about AIDS. But for a very long time in the past , public health experts have been concerned about something different, viz. spread of syphilis and gonorrhea, known to man for centuries. Epidemiologists have tried to recognize patterns in the number of cases recorded by hospitals and other facilities. Rather elaborate differential equation models are developed to understand the patterns. Let us

cursorily consider one such model called the S-I-R model. S stands for susceptible, the population of individuals who can get the disease. 'I' stands for infective. Susceptible has to come in contact with these to get disease. R stands for removed. These are cases which are cured or immune or dead or physically isolated. Numbers in each category change depending upon the interaction among individuals.

Attempts to fit such models to data on gonorrhea in USA led analysts to realize that there were two kinds of infective individuals. The core cases were those with promiscuous habits. They tended to acquire the disease quickly and to transmit it to several susceptible before taking treatment and getting cured. The opinion therefore was that protecting the core type through regular check up and other preventive action was essential for control of disease. The non-core patients had become infected by a stray event and were not a major threat. As a consequence, an elaborate interview was introduced in case of each patient taking antibiotic treatment. The interview helped social workers identify the core type cases which were then kept under careful follow up.

Smoking and cancer: This has been one of the most controversial areas in epidemiology. As you probably know, tobacco is a new world (i.e. American continent) plant and was introduced into Europe and then Asia after discovery and conquest of America. Its use spread like wildfire. In 1940's medical experts began to suspect that use of tobacco was responsible for occurrence of cancer of respiratory organs. Lung cancer was earlier a rare disease. In 50's it became more common. This was attributed to smoking.

How does one check this? One can take lung cancer cases and look into the habits of those individuals. Suppose we find that a large percentage of cancer patients are heavy smokers. Can we say that smoking 'caused' cancer? Here is a counter argument.

Smoking is a very common habit. Since percentage of smokers in society is large, it will be large among cancer patients also.

This argument will not stand if percentage of smokers among patients is much larger than in the population. If 50% people smoke and cancer occurs randomly, then we expect about 50% patients to be smokers. In a study on lung cancer and smoking habits among British male doctors, it was observed that out of 1,79,277 doctors, 1,09,368 (61%) were smokers. 201 doctors were detected of having lung cancer. Out of these 195 (97%) were smokers. So we have strong reason to suspect that smoking has something to do with cancer.

Another method is to examine longevity of life among smokers and non-smokers. Thus in another study on death rates from coronary disease among British male doctors, it was observed that among non-smokers, about 55% doctors survived the age 75 years while among smokers the figure is only 44%. One can get such

figures for various ages and plot what is called a survivorship curve. The curve for smokers turns out to be below that for non smokers (see Fig. 1). One can of course raise doubts about how individuals in the study were selected.

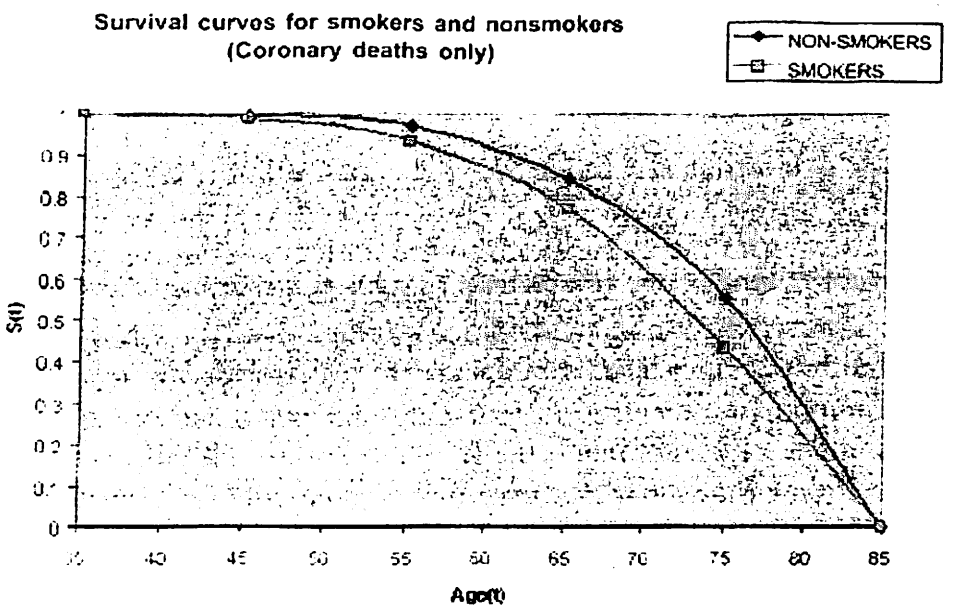


Figure 1: Survival curves for smokers and non-smokers.

Later in US and UK nearly 60,000 individuals were tracked for years to check if percentage of cases developing lung cancer is similar or different among smokers and nonsmokers. It turned out that, heavy smoking increases the chance of lung cancer over twenty times. But even this and subsequent studies in fifties with massive sample sizes could not finally clinch the issue. The reasons were basically logical. One argument in defense of tobacco the 'vile weed' was that among smokers, not only were deaths due to lung cancer more common, but also deaths due to many other diseases. Tobacco simply could not have been responsible for all these diseases. So perhaps there were some hidden biases in sample selection. Perhaps people with inclination to smoking also had a hereditary tendency to catch the diseases. If such were the case, giving up smoking does not change the genetic constitution and hence may not change the tendency to get diseases.

The only way to answer such doubts is to carry out a properly designed randomized clinical trial. In that trial randomly selected volunteers will have to agree to take up smoking at an adequately high level and continue it for years. You can see why such experiments are not possible. However, experiments were indeed done on animals and it was found that tobacco does cause lung cancer.

In mid sixties, the Surgeon General of USA accepted a report of an Advisory Committee on tobacco and declared that smoking is injurious to health.

The current wisdom is that tobacco consumption in any form increases the chance of getting cancer. In India today, gutkha, a mixture of condiments and tobacco has become very popular. Doctors everywhere are advising that it can cause problems in oral health rather quickly and the habit needs to be resisted.

Hygiene and Death: Have you heard the name Florence Nightingale? She was called the lady with the lamp. She brought solace to wounded soldiers of the British army which suffered very heavy casualties in the Crimean war against Russia in 19th century. In addition to nursing soldiers, Sister Nightingale did something most extraordinary. She dwelt over all the information about deaths of soldiers and discovered a fact which everyone found very striking. More British soldiers died away from the battle field, in the barracks, than on the battleground. She repeated such a study in India and the same was true here too. More redcoats died sitting at home (in barracks) than while fighting. They generally succumbed to diseases caused by filth and lack of hygiene. The report recommended that proper and clean accommodation with healthy surroundings is a must for soldiers. This led to the development of cantonments in India. Florence Nightingale was made a Fellow of the Royal Statistical Society for her creative use of statistical data to influence public policy.

To illustrate how epidemiological summary of information regarding a disease helps, let us see the following paragraph about rotavirus.

Rotavirus: is the most common cause of severe diarrhea in children all over the world. Children get this infection early in life (below 2 years) and develop antibodies. That is why adults are generally immune to it. In developing countries 0.75 million children die of rotavirus each year. This is 6% of all deaths among children under 5 years. In India about a quarter of all hospitalized cases of diarrhea are due to rotaviruses. Males seem more susceptible to it than females. Winter is the season favorable to rotavirus attack.

Antibiotics useful in other types of diarrhea are of no use in rotavirus infection. But it is difficult to distinguish between the two infection types? Misuse or overuse of antibiotics can be harmful. The only way to identify the diarrhea type is a biochemical test (ELISA). It can be expensive and time consuming. However, given age, sex of the patient and month of occurrence of diarrhea, it is possible to assess the chance of the case being rotavirus. If the chance is low, antibiotic can be used. Otherwise an ELISA can be asked.

Breast Cancer and Mammography: Perhaps you may know that women are vulnerable to breast cancer. As in any disease, early detection increases effectiveness of the treatment. A test for detecting this condition is called mammography. But who should take this test? If the test is recommended indiscriminately, most will be negative and one may be accused of colluding with test laboratory. It is perhaps better to consider various risk factors such as age, age

at birth of first child, number of children born, dietary habits etc. So every general practitioner should have a simple numerical formula to calculate risk that a given patient may get a breast cancer. He/She should also have a threshold value. Mammography can be recommended in cases which cross threshold. To our knowledge such values are not available for various Indian groups. Bodies such as Indian Medical association can take a lead in development of this formula.

Many such numerical aids in diagnostics are possible but at least in India, they are not used commonly. Part of the reason is lack of numeracy among doctors.

Clinical Trials: Human population is exploding on earth. This is because of lowered mortality rates due to wonder drugs, vaccines and other medical innovations. Diseases like cholera, typhoid, tuberculosis, malaria, pneumonia, etc. are fully and quickly curable, thanks to medicines. In fact with vaccines we have eradicated small pox, yellow fever and other epidemic diseases. The story of how these wonder drugs are discovered is an enthralling saga of modern science.

Statistics plays a key role when it comes to testing of new drugs. In the fifties, Jonas Salk developed a vaccine against poliomyelitis. Its efficacy had to be checked before extensive use. Now polio is a disease which afflicts children and leaves a very tragic impact by way of paralysis of a limb or even death. The epidemic occurs mainly in summer. American President during Second World War, Franklin Roosevelt was himself a victim of polio. This was one reason behind the great impetus to research in prevention as well as cure of polio.

Efficacy of polio vaccine was put to test in 1954. One million children participated in the test. Why so many? Polio is not a common disease. Even at its best (or worst?) its attack rate is only say 1 in thousand, and vaccine is expected to cause reduction in that value. Rare events are in general difficult to assess and hence the large 'sample size'. Deciding how many individual subjects are needed in test to make it adequately sensitive is one of the most important (and difficult) questions.

The test used placebo controls. This means some children were given vaccine and others only salt solution. The latter provide a valid benchmark for comparison. Incidence rate among inoculated children must be significantly lower than the rate among children getting placebos. Simultaneous application of vaccine and placebo ensures that comparison is in the same season and under similar conditions. If the comparison were made with incidence record in a previous year, a skeptic could have argued that any vaccine can be proved to be effective by comparison with a really bad year. Do you know the trick of shortening a line segment without touching it? Draw a longer line segment beside.

Volunteer children were assigned randomly to vaccine or salt solution. This ensures that no unknown bias is allowed to vitiate the experiment. It was also a double blind trial.

Now what is this? If two drugs are being tried and everyone knows who is getting which drug, then it is an open trial. This has some risks. A patient knowing that he/she is given experimental (unproven) drug may feel stressed and this may affect the response. To protect against such risk, patients are kept blind i.e. they are not given information about which drug they get. This is a single blind trial. In this system, the doctor who administers the drug knows who gets what. The doctor normally would have some inclination (favorable or unfavorable) to the drug under trial. This may affect the response. A double blind trial is one in which neither the patient nor the doctor knows who gets which treatment. In a clinical trial in which we participated as statisticians, the experiment was over, the data were in. Yet we were only told that group 1 was given drug R and group 2 was given drug S. We protested. That was for some reasons. Firstly, any bias on our part cannot affect the data. Secondly, if it is found that we are biased, analysis could be done by someone else. Lastly, some finer aspects of analysis need knowledge of which is the new drug and which is well established. If the drug being tested has a response not as good as an established drug, we will discard the new drug. But even if the new drug is just slightly better, we will not favor it. Demands on it are stringent. It must turn out to be 'significantly' better than the established drug. This requires a so called "one tailed test". But let us not get too technical.

One more interesting aspect of the trial design deserves mention. This trial was with volunteers. If you were to volunteer, would you volunteer to take a placebo or a trial drug? Some may argue that they do not want to expose their children to an untested entity. Others may feel that giving placebo to children deprives them of potential benefit. Question of ethics also become important. Can drug be tested on unsuspecting innocents? Suppose their consent is sought and obtained. Is it informed consent? Do the subjects really understand what they are consenting to? Often the custom is to constitute an ethics committee and get experimental protocol approved by the committee. Usually new drugs are first tried on animals. As you know, now there are animal rights groups also. They can and do ask that inflicting of pain should be kept to a minimal and unavoidable level. We hope you see that the business of drug trials is fraught with many tricky angles.

Drug screening: This refers to scrutiny of a large number of chemicals hoping to encounter one or a few useful drugs. This is nearly as difficult as searching a needle in a haystack or searching a lost boat in an ocean. Often there are hundreds or even thousands of compounds which, according to scientists, have some prospect of being useful against a disease. Success rate in searching for a useful drug from among potentially useful compounds is frustratingly low. One estimate puts it around one per ten thousand. In a search for anticancer drugs, laboratory mice, in which some cancerous growth has been induced, may be used, a handful getting each compound on test. A relatively large control group is also

maintained. At the end of the experiment, cancer tumors may be removed from each mouse and weighed. Lower weight of tumors as compared to those in the control group would be taken as suggestive of usefulness of the compound. A drug on trial is either rejected i.e. further research on it is discontinued or it is included in a shortlist for further investigation.

One can of course make errors. A useful drug may get discarded early on or a useless drug may be allowed to tag along causing wastage of resources. If a drug seems promising it is subjected to a more detailed quantitative study. Its dose is varied and effect of this variation on the disease is examined. The number of animals used at this second level experiment is usually large. If only 3 or 4 mice may have been treated with a drug in the earlier round, now the number may be raised by one order of magnitude to 30 or 40 say. So a useless compound that sneaks into this level, uses up resources to test 10 compounds at the first level. So it is a tricky thing to decide the criterion for selection in round 1. Statisticians constantly strive to make the choice of decision rule more and more efficient.

One way to improve screening efficiency is to introduce what is called a sequential procedure. In the procedure described above, in the round 1, a drug is either discarded or selected for round two. In a two stage procedure we have three options. Discard the drug, retain it for round two or thirdly take some more observations on it before making up your mind. This increases the cost of testing in round 1 but reduces the chance of erroneous decision.

We have to bear in mind that no matter how intelligent the screening procedure may be, in the final analysis, its success depends on how effective the 'good' drugs are. If they are only marginally better than placebos, not much success can be expected.

Exercises

1. Prepare a questionnaire to check relationship, if any, between food habits (vegetarian/non vegetarian), number of times a person brushes teeth, use of cold drinks/ soft drinks/ ice cream and dental health. Collect information on 50 children in a suitable age group say below 10 years and prepare cross tables of dental health and other variables.

Numeracy for Everyone

6:Numeracy in Social Sciences

Art and Science of Quantifying Amorphous Concepts, Impressions and non-measurable entities

Matters in the sphere of politics, economics, sociology, law etc. affect lives of all. Here too numerical information is often the basis of many decisions, policies and actions. As intelligent citizens, participating actively in social life, we need to understand the logic behind several such matters.

Many students and parents hold the wrong notion that social sciences are totally descriptive and any aptitude for quantification is quite unnecessary in study of Social Sciences. While good amount of work can be done in social sciences without numbers, measurement always strengthens any argument and can be essential in some situations. Here is a simple example. Large sums of money are currently spent on increasing awareness of AIDS. Funding agencies are keen to know if the work has any impact. So we may record interviews among people vulnerable to AIDS (e.g. truck drivers) before and after the education programs such as awareness camps. The difference if any will be a numerical measure of effect on society.

The word measurement need not always imply a thermometer or measuring tape etc. Even if we simply give names and classify, it is a step in the right direction. Sex (male/ female), profession (teacher, scientist, lawyer, social worker etc.), religion (Hindu, Muslim) are the cases of naming. Sometimes it is called 'nominal' measurement. If there is some hierarchy, that is yet another step in measurement called ordinal. Let us consider such a case.

Abodes of Gods: If you study religion and rituals of tribal groups, you may describe their deities and sacred groves, various sacrifices etc. None of this will require numeracy skills. But consider the following proposition: Higher the degree of elaborateness in a cult or worship spot, greater is the degree of destruction in the surrounding forest. If true this will mean that more primitive tribal groups will live in better conserved surroundings. But how does one quantify either extent of conservation or of degradation?

A deity in the open consisting of a rock covered with red lead can be regarded as the most primitive. A roof, walls and floor make the temple more elaborate. If the deity is beautifully carved, elaborateness is greater. So we have three ordered categories.

Now order the surrounding forest by the tree density i.e. number of trees per unit area and decide on suitable intervals. Then there can be categories such as very dense, moderately dense and disturbed. So, a contingency table can now be designed as shown in Table 1.

Table 1: Association Between Deity Elaborateness and Forest Status

Status of Surrounding Forest	Elaborateness Level of Deity		
	Low	Medium	High
Disturbed	e		c
Moderately Dense		b	
Dense	a		d

Here we can actually measure association between the two. If all cases fall in cells a, b, c and other cells are empty then the proposition is strongly supported. If, on the other hand, all cases fall in cells d, b, e, the conclusion is reversed in that elaborate temples seem to harbor dense vegetation around etc. In practice there may be non zero counts in all/most of the 9 cells and a suitable measure such as chi-square may have to be calculated to quantify intermediate level of association. We hope the point is clear.

Gallup polls: Ours is a democratic constitution. Periodic elections are the means by which we, the people, exert our supreme right to choose our rulers. Every general election is almost an upheaval. Thousands of candidates seek the blessing of the electorate. Therefore any means of assessing inclination of voters is of great interest. Consequently, sample surveys of voters evoke tremendous excitement. Such surveys seem to have started in early nineteenth century in United States. For a general description of Sample Surveys see article 4 of this series.

Literary Digest, a prestigious magazine in USA conducted in 1930's a mail questionnaire survey of about 2.4 million voters and estimated that Alfred E. Landon will win over Franklin D. Roosevelt in the presidential race. One advantage to the public (and disadvantage to the pollsters) in pre-election polls is that the day of reckoning is right there. In this instance, actual result was the opposite of the prediction. Roosevelt won handsomely. A post facto analysis of what went wrong in the survey was inevitable. The sample size was absolutely massive and the survey could not be faulted on that score. However, selection of voters turned out to be biased. The survey used lists of car and telephone owners to send the query forms. These were all individuals in the average and above average income groups. In this section of society, the Republican candidate was popular. But Democrats had support in lower income groups which were more populous.

This signal failure of electoral poll emphasized the importance of proper selection of a truly representative sample. This was not the only election survey at that time. Several others were carried out too. One of them was designed by George Gallup. It ensured better representation of the electorate and did forecast Democratic victory with a much smaller sample size. Ever since, name of George Gallup has become associated with election surveys. In fact often they are called Gallup polls. Since thirties, methodology of such surveys has been improved considerably, and their credibility today is high. In the American Presidential

elections in 2004, pollsters had predicted a close fight between George Bush and Al Gore and indeed it turned out to be so. In 2008 Barrack Obama won the election with a comfortable margin as predicted.

Apart from anticipating the winner there can be other aspects of elections which need a quantitative study. Of late television debates have become an important part of American Presidential elections. It is of interest to see if such debates make voters change their mind and in which direction. Suppose in a hypothetical study 100 voters expressed their preference before and after a debate and the result is given in Table 2 below.

Table 2: Preference Before and After Debate

Before	**After**		
	Democrat	Republican	Total
Democrat	50	20	70
Republican	5	25	30
Total	55	45	100

Here 25 out of 100 voters have changed their minds. 20 people switched from Democrat to a Republican candidate while 5 switched the other way round. Should the two numbers be regarded as essentially equivalent or have Republicans benefited more? Without going into technicalities we will say that the difference squared and divided by the number who switched i.e. $(20-5)^2/25$ is a good measure of relative effect of debate. If this quantity is large we may regard the debate to have helped Republicans significantly. In technical jargon this is called McNemar test after the psychologist who invented it.

Surveys are useful not only to study elections but in fact to study virtually any aspect of society. Kenneth Boulding, a famous social scientist termed the survey method 'telescope of the social sciences'. This is a very apt simile because surveys help researchers scan social skies. In Arabian stories, Harun Al Rashid, the Caliph of Baghdad had to walk the streets of the city in disguise at night to understand ground reality. This subterfuge is no longer needed. Sample surveys can obtain information about almost any feature of social life. Many news papers, including Times of India, conduct frequent opinion polls about current issues and publish results. These include surveys on internet. 'Times of India' puts out a question every day and publishes the reactions received. A word of caution is in order. The responses come over internet. So at best they represent those Indians who have access to Internet. That is a tiny fraction. Secondly these are responses from volunteers. Often volunteers are not representative of the whole group. Of course, to be fare to Times of India it must be admitted that they make no bones of these limitations. All schools and colleges can undertake the exercise of conducting

such surveys locally on questions of local and broader interest. National Sample Survey Organization (N.S.S.O.), a professional group set up by the Government of India conducts periodic 'rounds' of expenditure surveys. These reveal the felt need of people in urban and rural areas. Results of such surveys constitute major inputs into the national planning processes.

Measuring the non-measurable:

Economics is the science and art of using limited resources to satisfy our unlimited wants. We bring home a salary; spend the money on food, housing, clothes, education, entertainment, health and what not. Many people get a basic salary plus a dearness allowance (or cost of living allowance / COLA). What is cost of living?

Economics is of course concerned with prices, demand and supply. When price falls demand rises (case of negative correlation). Cost is another name of price. We understand what is cost of milk or bread. But what is cost of living as a whole? Clearly, we need an overall figure measuring this concept. Here we encounter a peculiar situation. The entity to be measured does not really have physical existence. It exists only in our heads. It is an amorphous concept. In this sense we are trying to measure the unmeasurable. Cost of an item can easily be measured through market inquiry. But to measure (overall) dearness level, we have to make this concept operational first before measurement is possible. Such a measure is consumer price index. This takes us into the art of index construction. An index is constructed when direct measurement is not possible. Human height can be measured in cm and fruit yield can be measured in kg and no index is necessary. On the other hand, intelligence, health, poverty etc. cannot be measured so simply. Here a formal calculation procedure is spelt out and once agreed upon, it is followed by all. This general agreement is the strength behind any particular index.

Consider the case of measuring intelligence. We can say that Mr. X is more intelligent in mathematics than Mr. Y since latter got fewer marks than X. But that is only one aspect of intelligence. There are others such as verbal skill, recognizing geometric patterns, judging weights, having a ready wit and so on. But how do we combine such measures? Psychologists have answered this question by means of a formula we call the IQ or intelligence quotient. Once such an index is devised, authors have to help others to understand it. In case of IQ we know that average humans have IQ of around 100, geniuses have values say above 140 while value below 70 probably suggests an imbecile. Secondly, in certain cases, when people know that the individual is a very bright person, the index value should come out quite high. Such agreement increases credibility of an index. Once people see that where they are sure, the index agrees with them, they feel comfortable about using that index in situations which are ambivalent.

Coming back to dearness level, construction of a 'Consumer price index' involves 3 steps. First, prepare a list of items that a typical consumer buys. Importance of each item is assessed and a weight is given accordingly. Next we calculate a weighted sum of prices.

Lastly the figure is expressed as a multiple of the corresponding value for a base or reference year. Index value for base year is 100. The corresponding value for current year is usually much higher. Since the Second World War, by and large, there is continuous rise in the index. There are very few items in case of which prices are on the decline. Computers are among such articles of purchase. But otherwise, the overall index is found to creep up throughout inexorably. This rise is termed as inflation. Generally we express inflation rate as % increase in consumer price index. In developed countries, rate of inflation is very low, say a couple of percent per year. In India it is often higher (though for a short period in 2009, it became negative). When inflation rate enters double digits, all Indians feel concerned (or such is the assessment of some analysts). The population tends to become more resentful of the rulers. But double digit inflation in India is nothing compared to the so called runaway inflation sometimes experienced by countries in Latin America. In this situation, with rate of inflation one (or even two) order of magnitude higher, everyone rushes to convert money into goods. Suppose you have 100 rupees, enough to buy a shirt. But tomorrow, the same shirt will cost 200 rupees. Then it is best not to keep your money idle but quickly exchange it for goods.

Another economic nightmare is grinding poverty.

Recently Professor Amartya Sen made all Indians proud by winning the Nobel Prize in economics. Among other things he is known for his research on measurement of poverty. Sen's index of poverty (or minor variations thereof) is the standard method of poverty measurement at least in research literature. So, how do we measure poverty? Here a crucial term is poverty line.

This is the level of income needed to satisfy a person's basic needs. Make a list of all items that are essential (e.g. food items, clothing, medicine etc.). Decide the quantity of each item needed. Obtain information on price of each item. Hence find the amount of money needed to satisfy basic needs. Anyone with income below this line is regarded as poor.

At a conceptual level, this sounds very reasonable. But making a list of essential commodities is not a trivial task. Some people consider meat / fish essential at least occasionally. Others abhor meat. South Indians cannot survive without rice and north Indians have to have wheat. Alcohol is an essential commodity for many tribal groups. It is part of all religious rituals. Some people consider tobacco and alcohol as essential.

Even if items are agreed upon, their quantities are not easy to decide. We may then refer to nutrition experts who may know how much food a man needs. You may find it surprising, but estimates of nutritional requirement are constantly revised, often downwards. So how reliable can estimated poverty line be at any moment? Thus it is very difficult to come up with a really good evaluation of poverty line. Suppose we do have one reasonable value to be used. Still measurement of poverty does not follow. There can be various ways of using poverty line. 'Head count' is simply the number (or proportion) of people whose income falls below poverty line. This index is not too bad, except for one hitch. Among people who have incomes below poverty line, some incomes are just below while others are far below. This gap is ignored by 'Head Count'. Also there is the angle of 'relative deprivation'. If every one shares the same suffering, there is no relative deprivation and poverty hurts less in such case. Inequality makes the impact of poverty more severe. Amartya Sen showed how all these considerations can be given due weight in constructing a 'good' poverty index. This index P is defined as $P = H \{I + (1-I)*G\}$. Here H is head count or proportion of people with income below poverty line. I is the income gap or average deficit for poor people defined as $I = (z - \overline{X})/z$. z is the poverty line. $\overline{X}$ is the average income among the poor. G is the Gini coefficient of inequality among poor. given by $G = \dfrac{\sum_{i=1}^{q}\sum_{j=1}^{q}|X_i - X_j|}{2q^2\overline{X}}$.

If there is no inequality among poor G is zero (can you see why?) and Sen's index is simply H*I.

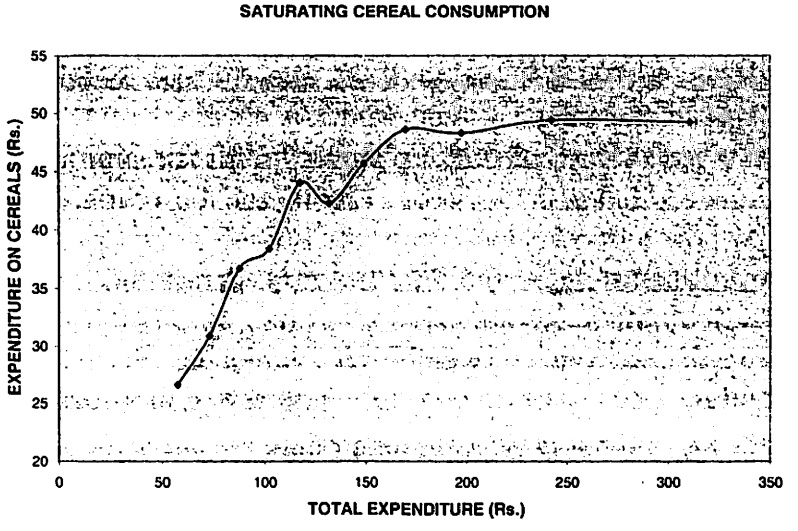


Figure 1: NSS 45th Round Rural (1989-90) Data

(Expenditure/Capita/Month)

The question of avoiding use of a seemingly arbitrary poverty line has foxed many an expert. Here is one possible alternative.

If we plot income level on X- axis and money spent on cereals (rice / wheat etc.) on Y-axis, the data seem to follow a rising but saturating curve. See Figure 1. This is a well known phenomenon. As you spend more on cereals, you get satiated and any further income is spent elsewhere. The level to which this curve approaches can be regarded as requirement. We can say that people reveal requirement through their expenditure pattern. Poverty (in cereal consumption) can be assessed by comparing actual consumption with this requirement.

Index construction is only one area in economics. But quantitative inputs are needed in many. To give a couple of examples, consider forecasting share prices or tax collection or export earning. Quantitative methods are now in use by administrators. Take the example of the Customs and Central Excise Department of the Government of India. In addition to revenue forecast, models are used to guess under or over invoicing to avoid taxes. First a certain set of cases is examined thoroughly to decide which cases involve malpractice. Then lessons learnt from this exercise are applied to other cases to predict malpractice without detailed examination. The method in use is called logistic regression.

If you read a good newspaper or magazine, you will invariably find an economics section. We leave it as an exercise for you to read such sections and identify matters involving numeracy.

Let us now turn to a different area in social science, viz. law and judiciary.

Uncertainties in Law: 'Oh come off it! What has statistics to do with law? Law may be an ass. But it is firm and rigid, nothing uncertain about it'. Or so you may think. But the truth of the matter is that in many criminal cases there is not enough evidence for the judge to be absolutely sure of guilt (or innocence) of the accused. So any judgment is prone to error. Errors can be of two types. An innocent person may be wrongly pronounced guilty. Or a guilty person may be acquitted. It is virtually impossible to avoid both errors. Then the tendency is to avoid more serious error. Attitude in India (inherited from the British) is that it is more important to insure that no innocent person gets punished. So a judge tries to see if guilt is established beyond reasonable doubt. You see, doubt and uncertainty are part of jurisprudence.

But statistics can come handy in the legal field in many ways. We will briefly describe a few cases documented in literature.

The first case concerns a workers' union in an aluminum smelting company in Canada. The workers during a strike, disconnected electric power supply to the

aluminum plant. Now aluminum smelting is a continuous process in which ore is heated to very high temperature in crucibles. These crucibles became cold and cracked after power outage. The company sued the union for damages, demanding that the union should pay for purchase of new crucibles. Union lawyers argued that the crucibles had been in use for a long time and needed replacement anyway. So the matter boiled down to how much longer the crucibles would have served had it not been for the union mischief. In other words what was the expected residual life of the crucible? Any reasonable compensation should be commensurate with this duration.

Estimating residual life is a statistical problem. If there was a fixed period beyond which a crucible becomes dysfunctional, the matter would have been very simple. But it is not so. Techniques of 'survival analysis' have to be deployed to complete the calculation. Details of that would take us way beyond the scope of this writing.

The second case is a dispute between a municipal corporation of a major city in USA and a company contracted to collect coins from parking meters. These are metal devices with clocks, mounted on pipes and hoisted near each car parking slot in public places such as roads. Users of the parking place insert coins in meters and are eligible to use the slot for time proportional to the amount of money inserted. The contracting company was to open meters periodically and collect coins from thousands of such meters and to deliver all money to the city authority. It was suspected that there was considerable pilferage. So, as a first step, authorities decided to estimate the total amount of money deposited in parking meters. They used the capture – recapture technique (see part 3 of this series). They secretly marked some coins and inserted them in different meters. Only some of these coins came back in the bags representing the day's collection. It became clear that a substantial portion of the collected money was siphoned off. (Let us say D dollars). Hence employees of the company were followed on a day, their activities were secretly videotaped and then a case was filed. The company pleaded guilty to pilferage on the dates for which videotaped evidence was available. Municipal Corporation demanded that the company should pay back D dollars per day for the entire period of contract. A counter argument made was that the collections were much lower in previous periods. This necessitated estimation of true collection in previous periods. Such estimation requires fairly sophisticated statistical methods.

We hope examples given above convince you of usefulness of numeracy in social sphere.

Exercises:

1. Plot changes in share prices of some companies in the field of information technology (e.g. Wipro, Infosys) over the last 8 weeks.
2. Seek opinions of your classmates about a TV serial. Then seek opinions of about equal number of people of your parents' age group about same program. Is the percentage of people who like the program same in both groups?
3. Following data are on the last 12 parliament elections in India.

Election	Total Voters (T)(crores)	% votes cast (P)	Election	Total Voters (T)(crores)	% votes cast (P)
1	17	39	7	36	43
2	19	39	8	40	36
3	22	45	9	50	38
4	25	39	10	51	39
5	27	45	11	59	43
6	32	40	12	61	38

Plot graphs of T and P over time and comment on them. Try to fit simple linear regression and check which slope is steeper.

4. Examine the possibility of formulating indices for following subjects:

(i) Farm mechanization
(ii) Food security
(iii) Soil fertility
(iv) Land degradation
(v) Forest degradation

Try following steps:

(a) Identify region of interest such as district or state.
(b) List variables to be taken into account.
(c)Search for sources of information.
(d) Decide on relative weights.
(e) Try validation of your formula.

Numeracy for Everyone

7. Numeracy in Industry:

Statistical skills to improve quality and maximize profits

If we ask a typical student of standard twelve about what her/his choice of higher education is, a very likely answer would be 'Medicine' or 'Engineering'. This is understandable in view of high employability, prestige and remuneration associated with these professions. There is a common belief that while students going to medicine need not be well versed in mathematics those opting for engineering have to be. So it may come as a surprise to many that mathematics as practiced in a typical engineering course does not prepare students in two crucial quantitative aspects of industrial environment viz. quality and profit. The purpose of this article is to highlight mathematical/ statistical tools commonly used in quality improvement and optimization.

Quality Improvement: Have you heard the terms globalization and liberalization? Over the last decade or so, India has opened the gates of her economy and industrialists all over the world are now welcome to manufacture and sell their products here. From very sophisticated and essential (say IBM or Volvo) to very unessential (say Pepsi or McDonalds) the whole spectrum of companies has setup shops here. Why do customers throng to (at least some of) them? Partly it is image building and partly quality. People feel that Maruti cars are better than the old Premier (and worth the price), hence the tilt. Quality has therefore become a key factor in industrial production in India. It is a life and death matter. Premier Automobiles has closed its business.

For our limited purpose let us define quality as uniformity. This is exactly the opposite of the world of art in which uniqueness is at a premium. Uniform Ganesh idols can easily be made from a mould. However, those made one at a time cost more. Latter are regarded as of higher quality. So notion of quality is different. When it comes to industrial products, consumer expects high class performance. Whatever the specifications in terms of performance, each item (or batch or package) produced should fulfill the specifications and all should essentially be alike. This is quality. We take such uniformity in products for granted. When a spark plug of a scooter needs replacement, we buy one and return without a moment's thought as to whether it will fit 'our' vehicle. We assume that there is no individuality in scooters (of the same company and brand) or in spark plugs. In fact things are not quite 'the same'. There is variation. But it is within tolerable limits. Suppose the plug intended for say a Honda Activa scooter cannot be fitted to a vehicle because say the OD (outer diameter) of the threaded portion of the plug is too large. Then we will say that it is beyond tolerance limits. Items which differ

from specifications beyond tolerance limits are rejected by customers. They have to be reworked if possible or discarded as scrap. So a basic duality is that consumer wants assured good quality products while manufacturer wants to avoid rejections. Quality improvement protects interests of both sides. Understanding nature and causes of variability is the foundation for attaining high quality.

Each production process has its own intrinsic variability. Many factors contribute to it. Raw materials or components supplied by vendors may not be uniform. Performance of different operators or machines may be different. Things may vary from day shift to night shift and from one day to another. In a popular novel by Irving Wallace named 'wheels', about automobile industry in USA, , an insider recommends to a friend that one should buy a car produced on Wednesday. The reason for this is supposed to be that as weekend approaches (Thursday / Friday), workers become negligent. Reason for avoiding production on Monday / Tuesday is holiday hangover. The point is that some variation in products may remain in spite of all precautions. Attempts to insure quality can be made at each of three major stages; product design, production process and final inspection. It turns out that statistical tools useful in these three stages were developed over time in the reverse order.

Acceptance Sampling: If you are not a producer but a (bulk) consumer you wish to ensure that supplies received are of requisite quality. It is not practical to check every item. Sometimes checking is destructive. (Suppose you wish to test if explosion of a fire cracker is loud enough to be used at 'Deepawali'. You cannot test one and have it too). Hence a sample can be taken instead and the number of defective items in the sample counted. If this count is too high, the lot is rejected. Otherwise it is accepted. The key question is `how many items should I inspect?' The aim is to be able to reject all bad lots and accept all good ones. It is not possible to avoid both errors but by choosing a sufficiently large sample size one can keep the probabilities of making these errors reasonably low. The actual calculations have to be based on the so called hypergeometric distribution. Let us consider a simple hypothetical example.

Suppose we have to buy a box of 200 lead pencils. We propose to check a random sample of 32 pencils. Our rule is to accept the lot if sample contains no more than 3 defective pencils. Suppose the lot contains 40 bad pencils and deserves to be rejected. Then what is the probability that our decision rule will fail to do so? This is the so called consumer's risk.

P (the lot will be accepted) = P (# of defectives in sample ≤ 3)

$$= \sum_{i=0}^{3} \frac{\binom{160}{32-i}\binom{40}{i}}{\binom{200}{32}}.$$

This turns out to be approximately 7.5 %.

Now suppose the lot contains only 9 defective pencils and deserves acceptance. Under our rule now the probability of accepting the lot is,

P (the lot will be accepted) = P (# of defectives in sample ≤ 3)

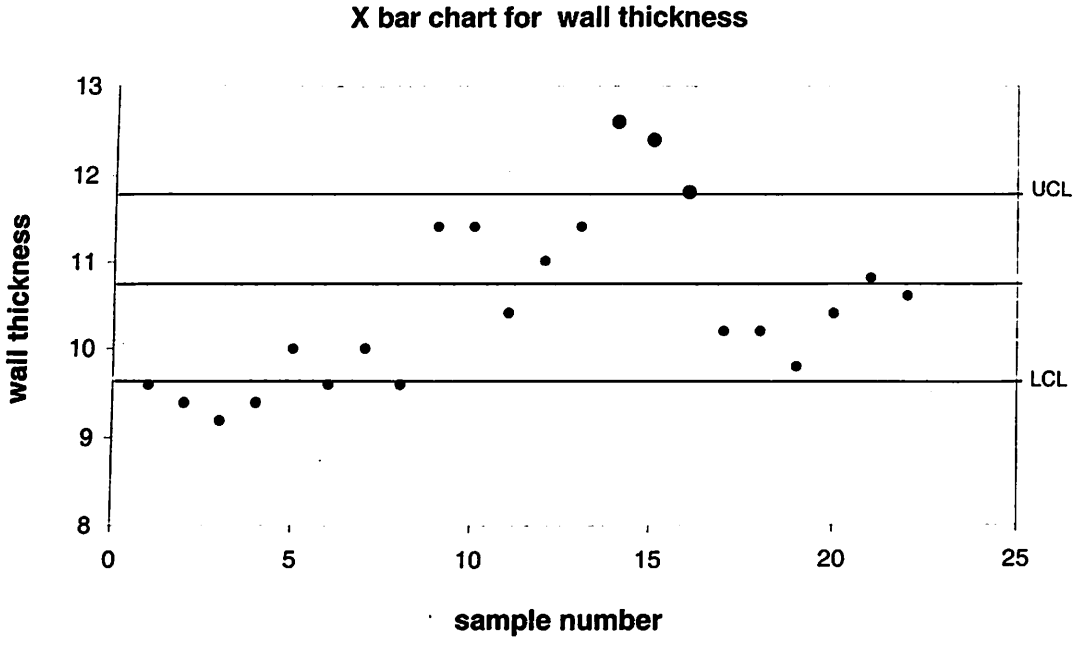
$$= \sum_{i=0}^{3} \frac{\binom{191}{32-i}\binom{9}{i}}{\binom{200}{32}}.$$

This is approximately 96%. Here producer's risk is only 4 %. These calculations can be verified using EXCEL. In actual practice one has to select sample size and decision rule so as to ensure that consumer's and producer's risks are at pre-specified levels. Such a selection is called an 'acceptance sampling plan'. Such plans for commonly used risk levels are available in standard manuals.

Control Charts: Acceptance sampling plan reduces the risk of a consumer getting a substandard item. But rejected lots constitute a major loss to the producer. As in health, prevention is much better than cure. Therefore it is necessary to provide for midcourse correction. Opportunity for such correction is provided by control charts. They involve monitoring production flow, taking periodic samples; measuring relevant traits (e.g. OD of a spark plug) and plotting, say, the means on an appropriate graph. Each control chart has a centre line representing the desired value of the trait and upper/ lower control limits (say at a distance of three sigma). If a sample value falls outside the control limits, it is treated as a warning of trouble in the production line. Let us consider an example from mechanical engineering. The data, collected by a student working on an industrial assignment in a factory in Pune, are about cylindrical bearings being made for an IC (internal combustion) engine. Wall thickness is the criterion of interest here. A sample of five pieces is selected every 15 minutes and average wall thickness is plotted. If $\bar{\bar{X}}$ is the grand average i.e. mean of sample means and S is the standard deviation of these sample means then the control limits are calculated as $\bar{\bar{X}} \pm 3S$. Figure 1 shows the center line and upper and lower control limits (UCL, LCL). Each dot represents mean of one sample.

Notice that dots corresponding to sample numbers 14 and 15 are above UCL while that for 16 is on the UCL. This suggests that the process had gone out of control. In such a situation the thing to do is to search for an assignable cause responsible for the process going out of control and to eliminate that cause. In this case inquiry showed that a cutting tool used had worn out and had to be replaced. Dots corresponding to samples 11, 12 and 13 indicate a rising trend giving an early warning which was not heeded. Of course one may ask why three rising points all within control limits should be regarded as a sign of trouble.

Figure 1



The objection is valid unless there is circumstantial evidence to suggest that trouble may be brewing. A thumb rule followed by some is to wait for occurrence of 7 successive points each higher than the previous one, before raising an alarm. The basis for this thumb rule is that probability of a point being higher than its predecessor is ½. Two successive points each being higher than previous one is an event with probability 0.25. Proceeding thus we note that at 7 the probability falls below 0.01 (a commonly accepted threshold) for the first time. One may use charts not only for means but also for ranges, proportions etc.

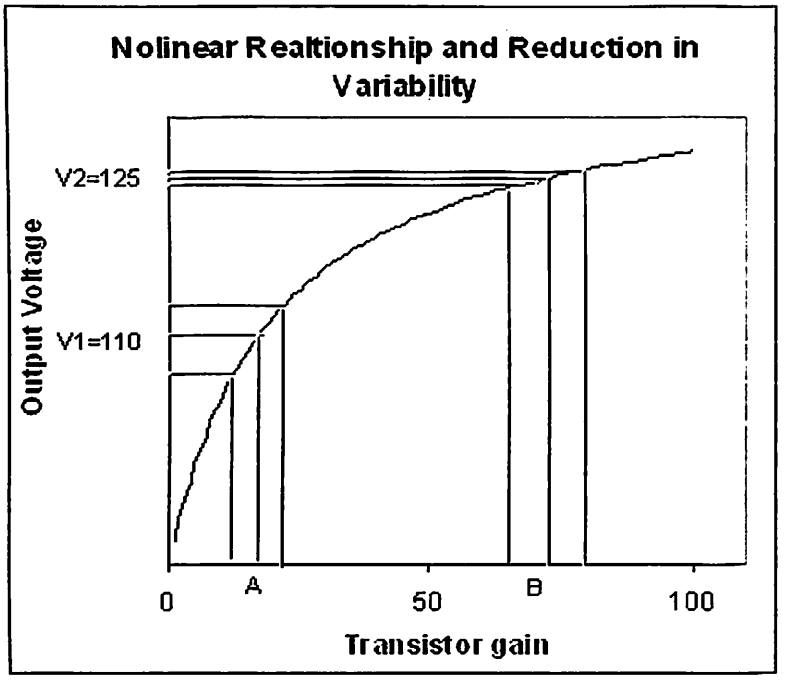
I.S.O. 9000: Control charts are a technique for anticipating and detecting a failure at a particular stage in production. To ensure quality, a manufacturing unit

needs to follow appropriate techniques in all activities. So international norms e.g. ISO 9000, have been laid down as guiding principles to be implemented by a company. These include statistical methods as well. There are organizations which act as quality auditors. Many countries now a days import goods only from companies which conform to such procedural standards. This forces exporters to follow the norms seriously. You may have seen advertisements of companies regarding ISO 9000 certificate earned by them.

Taguchi Approach: So far we explored methods concerned with final inspection and production process. Now we turn to the third stage namely product design. Here the approach is to ensure a product design such that output quality will be good in spite of variation in inputs. This approach is attributable to Taguchi. In the last couple of decades name of Taguchi has become synonymous with consideration of quality and productivity. This Japanese engineer introduced a novel approach to problems of quality in manufacturing. He argued that some variation in inputs is unavoidable and it is better to require the product design itself to be robust. In other words, design should be such that quality is invariant (and good) even when inputs are not uniform. Madhav Phadake (1989) gives the example of a power supply circuit to illustrate the underlying concept. We can use nonlinearity of a relationship to reduce effect of variation in input on product. Output voltage y is an increasing function of x, say transistor gain. If we seek output voltage 110 (used in USA), the appropriate x value is A. But here small change in x causes big change in y. On the other hand output voltage y=125 can be gained at x=B and here change in x does not cause major change in y. (Figure 2).

Hence a robust design will involve x=B. But now the stable value of y available is wrong. It has to be shifted to a desired level by introducing some resistors. Of course this is a rather simple example. In practice many factors are involved and identifying a robust combination can be rather difficult.

Figure 2



Starting from inspection after production we have ended up with the idea of robust product design for quality. This is the spectrum of statistical quality control ideas today. We may note that basic ideas in this field were developed mostly by statisticians in the west. Perhaps the most famous name among these is that of W. E. Deming. However, it was the Japanese industry that enthusiastically embraced the faith and harvested bumper profits. In seventies this was recognized all over and western industries caught up with the Japanese. We in India have a long way to go in this respect.

Optimization: The second major concern in industry is profit maximization. This can be achieved by optimization in resource use. Numerical techniques can play an important role here too. Let us begin with a simple yet very real example.

In a large automobile manufacturing plant hundreds and thousands of gauges are used regularly. These must be accurate or else measurements go wrong. For this, they have to be checked periodically and recalibrated. How often should this be done? In one factory in India the convention was to check each gauge every two weeks. There was a suspicion that this method did not work satisfactorily. Hence a simple study was launched. Gauges were classified by frequency of usage.

Frequency	percentage of gauges with such use
100 times per day or more	20
10 to 99 times per day	25
Up to 9 times a day	55

Consultants suggested that calibration frequency should be commensurate with frequency of use. Where there is heavy use (i.e. category 1) there should be weekly check as error in these gauges affects many measurements. Moderately used gauges (category 2) can continue to be checked once a fortnight. The infrequently used ones (category 3) need be checked once a month only. It turns out that the above simple modification leads to a reduction in measurement errors without any increase in calibration effort.

Of course there are many methods in mathematics that come handy in solving a variety of optimization problems in industry. Operations Research (OR) is the name given to such a tool box. This discipline crystallized from military related research done during the Second World War. There are many aspects of OR; Linear, nonlinear and dynamic programming, queues and inventories, modeling and simulation etc. After the war, the methods came to be used by people at large. We will take a brief look at some of these techniques.

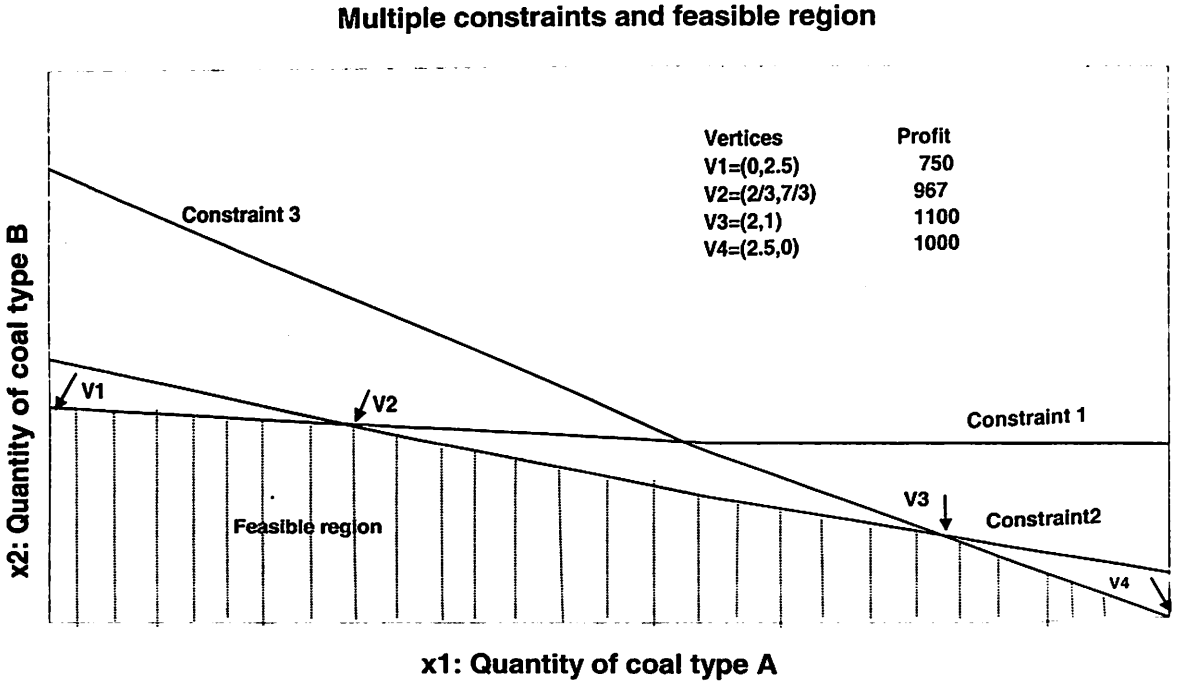
Linear Programming: Here the aim is to maximize a linear function of variables which are subject to linear restrictions. Let us understand this technique through a simple example. A fictitious coal mining company Bangalore Coal Fields produces two kinds of coal, type A and type B. Profit from selling one ton of type A is Rs.400/- and a corresponding figure for type B is Rs.300/-. Decision to be made is quantity x_1 of coal type A and x_2 of type B to be produced in a day so as to maximize the total profit $P= 400*x_1 + 300*x_2$. One cannot select very high values for both x_1 and x_2. This is because of restrictions on time availability of cutting machine (10 hours in a day), screens (9 hours) and washing plant (10 hours). Time requirement for producing one ton of coal type A is 1, 3, and 4 hours of cutting machine, screens and washing plant respectively. Corresponding values for coal type B are 4, 3 and 2. Hence mathematically the problem is to maximize total profit P subject to three restrictions viz.

$$x_1 + 4*x_2 \leq 10$$
$$3*x_1 + 3*x_2 \leq 9$$
$$4*x_1 + 2*x_2 \leq 10$$

These constraints can be satisfied by several pairs of values of decision variables (x_1, x_2). This set is called feasible region because each of the choices has the possibility of implementing. Among these we have to choose the one which

gives maximum profit. A mathematical theorem simplifies the choice by asserting that the best combination has to be at one of the vertices of the feasible region. Figure 3 shows the feasible region with four vertices indicated by arrows. Profit at each vertex is also shown. Clearly the optimal choice is at $x_1 = 2$ and $x_2 = 1$, which will give a profit of Rs. 1100/- per day.

Figure 3



As usual reality is more complex than a conveniently selected example like this. As the number of decision variables goes beyond 2, graphical presentation becomes difficult. For moderate number of decision variables and constraints, computer programs can be used to implement a procedure called simplex algorithm to arrive at the best choice. As the problem size goes beyond a point, even this approach begins to falter. For such super complex problems a new algorithm was devised by an Indian mathematician, Prof. Narendra Karmarkar who became an instant celebrity.

Transportation Problem: Have you noticed that there are large godowns that store up industrial products (cement, steel, paper, tyres and what not), just outside the municipal limits of most cities? This is mainly to avoid payment of

octroi. But in any case every major manufacturer has to store production in warehouses and then send shipments to widely distributed wholesale and retail vendors or customers. The problem here is to satisfy requirements of all customers while keeping the transport cost at a minimum. This is a special case of above linear programming problem but receives independent attention. The following example will illustrate the nature of this problem.

A company has three factories with production capacities 20, 15 and 10 units respectively. There are three warehouses which can store 5, 20 and 20 units. The problem is that of shipping products from factories to warehouses such that cost of transportation is minimized. Following is the matrix of transportation costs:

Cost of Shipping One Unit from Factory to warehouse

From factory	To warehouse		
	G1	G2	G3
F1	9	10	10
F2	10	14	8
F3	13	10	8

A manager not trained in OR had adopted following scheme: Send 5 unit from factory F1 to warehouse G1 and 10 units to G2 and remaining 5 to G3. Send from F2 10 units to G2 and the rest 5 to G3. Send from F3 everything to G3. Cost of this scheme is Rs.455/-. Such a scheme which satisfies all the restrictions is called a feasible solution. However, the least cost solution may be different. To improve on a feasible solution, one identifies the most expensive element in the cost matrix and tries to reduce burden on that route. In our case such an element is F2 to G2 (costs14/-). Hence we reduce the burden on this element from 10 to 5 and send these 5 units to G3. As compensation we cancel shipment from F1 to G3 and send those 5 units to G2 instead. So the revised scheme is as follows:

Ship 5 units from F1 to G1
Ship 15 units from F1 to G2
Ship 5 units from F2 to G2
Ship 10 units from F2 to G3
Ship 10 units from F3 to G3

The cost of this solution is Rs.425/- which is an improvement over earlier one. But even this is not the final answer.
The least cost solution turns out to be as follows:

Ship 5 units from F1 to G1
Ship 15 units from F1 to G2
Ship 15 units from F2 to G3
Ship 5 units from F3 to G2
Ship 5 units from F3 to G3

The cost of this solution is Rs.405/-.

Let us remember that this is a simple problem and real problems are much more complex.

Queues: These are an inevitable feature of modern life. Queues are for a service. If you go to a barber shop on a Sunday you have to wait for long. Why aren't there more service points? You wonder. Even aircrafts have to form a queue for landing or take off. Why not build one more runway? The cost of adding an extra service point is often very substantial. Will the benefits outweigh costs? Experimentation is not possible. But we can do statistical simulation.

This is the method of creating a mathematical reality, analogous to virtual reality created in computers. Arrivals of aircrafts are mimicked using random numbers from say an exponential distribution with a suitable arrival rate. This rate can be extracted from traffic data available. Time for which an arriving aircraft keeps airport facilities busy (service time) can be modeled using another suitable distribution.

If the model is good, it will generate waiting times and queue lengths comparable to the actual experience. This is validation of the model.

Now imagine that one extra runway is added. This will mean that service time distribution will have to be suitably changed. The revised model can be run for as many days as needed by using more and more random numbers to represent more and more arrivals. Changes in waiting time and queue length can then be estimated. They will form the basis of any rational judgment as to whether the cost of a new runway is commensurate with benefit. Similar exercises can be done about reservation counters, postal service counters, new berths in a port etc.
For more examples of simulation see Kunte (2000).

Inventory Control: Mass manufacturing is an ongoing process. It requires continuous inputs of raw materials, bought out components etc. You may have seen pictures of car assembly lines. As the car moves down the line, different parts are fitted to it. At each station there is a supply of a specific part, ready at hand. A worker at a station does his job and then the car moves on. All day worker at a station does the same job repeatedly. In the famous film 'City Lights', Charley

Chaplin has humorously depicted psychological effects of such monotonous assembly line tasks on workers.

For smooth functioning of this set up, there must be an uninterrupted supply of relevant parts. This means that parts must be stocked up. Such a stock is called an inventory. An important decision is the quantity to be stocked. Larger the stock lower is the chance of interruption in production due to non availability. But larger stocks also imply that greater amount of capital remains locked up which involves a cost. A stores manager must take into account the time it takes to send an order and the delivery time before deciding the volume and frequency of order. If all systems are really fine tuned, it is possible to keep nearly zero stock level. The deliveries then have to be 'just in time'. This keeps inventory cost to a minimum.

One can see that there is lot of room for innovation in industrial management using these techniques and many more. In fact their use need not be restricted to industry and can easily be extended to all walks of life.

Exercises

1. Consider the problem of gauge calibration discussed in text. Verify that the saving in the revised calibration scheme is 7.5 %. Assume that cost of calibrating any gauge is the same. Try to work out the extent of reduction in measurement error.

2. A baker bakes two types of cakes each day, chocolate and banana. He makes a profit of Rs.7.50/- on one chocolate cake and Rs. 6.00 on one banana cake. A chocolate cake requires 4 units of flour and 2 units of margarine and a banana cake requires 6 units of flour and 1 unit of margarine. However, only 96 units of flour and 24 units of margarine are available on each day. How many cakes of each type should the baker make on a day so as to maximize the profit?

3. In the study of bearings (see section on Control Charts), for each sample of five bearings maximum and minimum wall thickness was recorded. Data are given below.

Sample No.	Max	Min	Sample No.	Max	Min
1	11	8	12	13	10
2	11	8	13	13	10
3	11	8	14	13	12
4	10	9	15	13	11
5	11	8	16	13	10
6	11	8	17	11	9

7	11	9	18	12	9
8	11	9	19	11	9
9	12	11	20	11	10
10	13	10	21	11	10
11	11	10	22	12	9

The difference between maximum and minimum value in a sample is called sample range (R). A wide range indicates a poor quality. Calculate range for each sample and plot a control chart for range using the value R = 2.23, LCL=0 and UCL=4.7. Check if any points fall outside the control limits.

Epilogue

We bring this series of articles (section I of this book) to a conclusion now. Our attempt has been to describe briefly, applications of Statistics in as many areas as possible. The intention was to paint with just a few bold strokes of the brush but almost totally devoid of details. We have referred to a large number of techniques, but only sketchily. We want to communicate to the readers our conviction that the core of Statistics is not formulas but logical ideas. Given a sound idea, a suitable formula can always be devised to implement it. The vision of Statistics essential for an intelligent layman is neither one of piled up numbers nor the third kind of lie in the famous Mark Twain quote but search for elegant patterns in a confusion of data. Perhaps readers will agree that this is also a good description of science itself.

Lastly, readers interested in dialogue about application of numeracy to any real life problems are welcome to write to us by e-mail or snail mail.

This series was written several years ago. Our interaction with readers since that time suggests that it would be useful to move further and describe actual case studies in application of statistics. We shall do this in the next section of the book.

Section II
Excursions in applications of statistics

Statistics has been called a new technology of the twentieth century. That means the subject focuses on solving specific and immediate problems arising in different walks of life. In fact history of statistics is replete with examples of how new developments in methodology were triggered by problems in this or that are of application. Regression analysis arose in trying to understand heredity. Design of experiments was meant to help agricultural research with an economical way to conduct studies to identify the best varieties or agronomic practices. Mahalanobis developed his D^2 statistic to solve some problems in anthropology. Abraham Wald launched sequential analysis to help increase efficiency in production of ammunition. The list can go on and on. So, applying statistics is (or should be) an integral part of the culture of statistics. But Indian students do not get enough opportunity to learn of applications first hand. They can only read about things done in the west. We have tried to remedy the situation to some extent by writing brief essays about live examples of application. These are our own experiences and they all arise in Indian context. We hope that they will turn out to be more understandable to the audience of Indian students and laymen. The exposition is informal. We make no attempt to show details of the problems; they can be complicated and rather daunting. Similarly, statistical methods used are described very cursorily. In many cases details are available in technical papers published. References to such work are given. Our objective is to make the subject come alive for students.

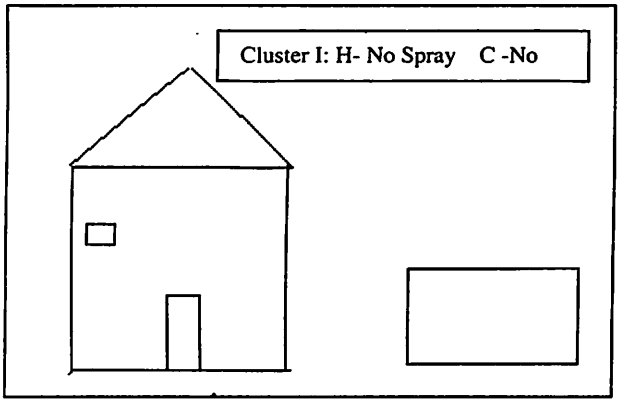
8. Mosquito, Malaria and Men

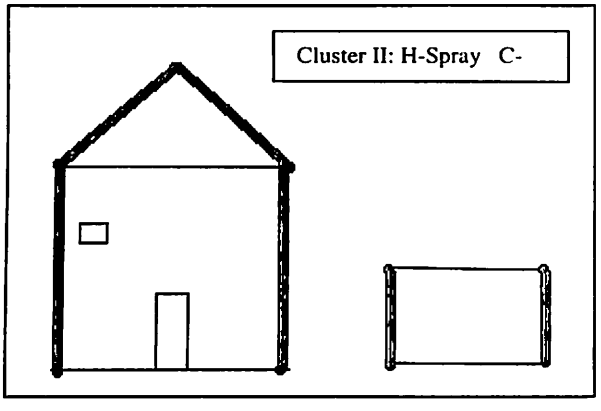
Situation: Malaria affects about two million Indians every year. At the time of independence there were an estimated 75 million cases per year and number of deaths was about 800, 000. Our National Malaria Eradication Program was so successful that the number of cases reduced to 100, 000 and deaths due to malaria were essentially eliminated by 1966. That was a remarkable achievement. Alas, the success was short-lived. Within ten years the number of cases went up through the roof to cross 6 million. So the control efforts were revived. At the beginning of the new millennium, we are still saddled with over two million cases and over 1000 deaths per year.

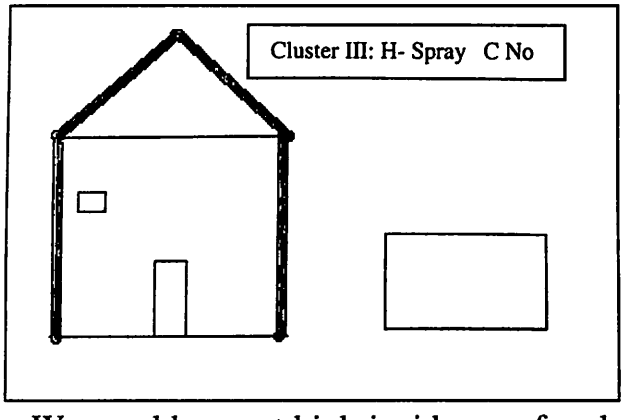
Why should government intervene to control malaria? It may seem like a preposterous question. But sometimes it is good to ask such questions. One obvious answer is that life, health and happiness of citizens is the responsibility of the state and hence making all attempts to keep the population healthy is part of the duty of a welfare state. There are financial implications too. Can we put some value on being free from malaria? How much is the cost to the society due to infestation of malaria? Such calculation can be enlightening. It will suggest the kind of financial outlays from exchequer that are justifiable. It is not easy to estimate such cost (sometimes called 'burden'). About 2 million cases of malaria occur per year and the distribution is very uneven. Orissa has 3.5% of India's population but 25% cases and over 30 % of malarial deaths in the country. According to some scientists, each rupee spent on malaria control gives a benefit of about 20 rupees to the society. At the global level, about 400 million people fall sick with malaria every year, 90% of them in Africa. No wonder Bill Gates foundation assigns a large chunk of its grants to malaria related work!

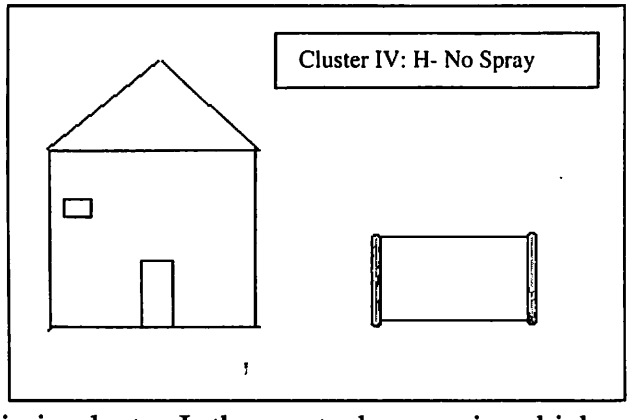
Coming back to Indian scenario, what was the key to success up to 1966 and what caused the return of the curse? Our success was due to mosquito control using DDT. This wonder chemical seemed to work every where. Areas with very high incidence of malaria such as Himalayan foothills (Tarai) were uninhabitable. But with control of malaria, these areas became livable and people migrated into such areas. Then, slowly, mosquitoes resistant to DDT raised their heads. The old trick of DDT use could not do the job well. So, yet another trick was used to control mosquitoes. The new trick was use of a bacterium named 'Bacillus thuringensis'. It was able to infect mosquito larvae and restrict population growth. However in just a few years, resistance to this control agent also developed. This is the so called arms race between humans and mosquitoes. We were involved in two studies concerning malaria, which addressed two important questions about malaria incidence. One study was in a tribal region of Orissa and the other in the metropolitan region Mumbai.

Can DDT increase malaria? DDT being effective or not effective in reducing malaria incidence is understandable. But what if someone says that DDT can increase the malaria? A study in a tribal region of Orissa was designed to answer this question. In Keonjhar district of Orissa, DDT was used in different ways in different village clusters. In fact there were four types of clusters. Cluster I- This can be called the control group. There was no spraying of DDT at all in these villages. Cluster II- Here, it was the other extreme. DDT was sprayed on walls of residential huts as a repellant. Similarly it was sprayed on walls of cattle sheds. Cluster III- Here, cattle sheds did not get any spray, but residential huts did. Cluster IV- In this last group, walls of cattle sheds were sprayed but not residential huts. The populations of all four clusters were under surveillance. This can be viewed as a 2X2 experiment. Two factors are spray of DDT on the walls of residential hut (yes/no) and on walls of cattle shed (yes/no). Following sketches describe the four clusters. Dark/ broad borders indicate sprayed area. H represents 'Hut' and C represents 'cow shed'.









We would expect high incidence of malaria in cluster I, the control group in which there was no use of DDT. However, it turns out that the cluster 4 in which cattle shed got the spray but not the huts gave the largest incidence. Solution to this

puzzle is that DDT does repel mosquitoes and they all turn to the residential hut (which lacked the repellant). This can indeed increase malaria incidence. In this sense, DDT, if wrongly used, can indeed increase malaria. Further, spraying the walls of residential hut does not seem to make a difference. Is this in contradiction with the previous claim that DDT did repel mosquitoes in cattle shed? Why should things be different in huts? Again the answer to the puzzle is simple. Tribal families have a practice of plastering walls of huts with cow dung slurry. This ensures longer life of walls and cleanliness. Unfortunately, dung covers the DDT sprayed and eliminates its effect.
So the conclusion is that, improper use of DDT can increase malaria incidence.

Are women immune to malaria? It is often not known that genetically speaking, males are the weaker gender among humans. The small y chromosome characteristic of males, fails to give them protection against some recessive genes. Thus, color blindness is more common among males than females. What about malaria? Many studies have noted that incidence of malaria is higher in men compared to women. But is this a genetic phenomenon or cultural? Men work outside or sleep in open, where the chance of getting bitten by a mosquito is higher. Perhaps that is why men get malaria more often. A different answer seems to be on the horizon. We were involved in a study of fever patients and others admitted to public hospitals in Mumbai. Study was done by the Tata Institute of Fundamental Studies. Our role was that of analyzing and interpreting data. We used chi-square tests and logistic regression analysis in order to understand association between various factors and malaria incidence. Here is what we found.

Mumbai is an area with relatively low incidence of malaria compared to the so called hyper-endemic regions like Orissa. If there is lots of malaria, it hits children more than adults. As children become adults, they acquire immunity to malaria because of repeated exposure. This does not happen in Mumbai. Hence we expected similar rates of malaria in all age groups and both sexes. In fact incidence is highest among adult males. Risk of malaria is similar among children and adult women in reproductive age. The mosquito responsible for malaria in Mumbai is of the species *Anopheles stephensi.* It bites throughout day and night, indoors and outdoors, and attacks men and women without fear or favour. So, any differences in incidence cannot be attributed to the mosquito. Further, when blood samples were checked, it turned out that proportion of samples with malarial parasite (Plasmodium vivax and Plasmodium falciparum) is similar for all ages and for both sexes. So, children and women get the same dose of the disease causing bugs. But they do not develop malaria as often as men do. Why are men more vulnerable than women? Statistics cannot answer such a question. It can help. Numbers showed that proportion of men sick with malaria rises with age. Effect on male children is no

different from female children. But once beyond puberty, results are different. Proportion affected goes on increasing in men while in women it remains low and flat. As men grow old, the proportion declines and becomes similar to proportion in women in the same age group. We know that testosterone level goes on increasing in males till the age of 40 after which it declines by about 2% every year. All this suggests that presence of male sex hormone testosterone may be the cause. Perhaps female sex hormone estrogen helps women fight malaria. The exact mechanism of this sex hormone-linked immunity/susceptibility to malaria is not clear yet. Perhaps further research in immunology will throw more light on the phenomenon. TIFR study seemed to confirm existence of the effect in a hypo-endemic area (where malaria is present but in small proportion).

What does this result mean for policy? This finding has important implications in the development of health strategies. If we test a new drug/vaccine for malaria, we should check if it works well with men. Today as it is, millions of people live in conditions that are suited to malaria. In addition, there is lot of discussion about global warming. One likely effect of it is spread of malaria to new geographic regions where adults not immune to the disease will be exposed to the parasite. Prediction from this study is that males are more likely to suffer from malaria though both males and females are equally likely to have malaria germs in blood.

9. Paper Wasps: The case of forgone fertility

Insects are a major part of our environment. Some of them are a great nuisance such as lice, bed bugs and house flies. Some are a threat e.g. locusts. Others such as silk worm or honey bees are very useful. Some others are curiosities like beetles, fireflies or butterflies. Some insects live in large communities, e.g. ants.

Scientists have always been curious about insects. One curiosity is about the social life of insects. Think of a beehive or an ant colony. Thousands of individuals seem to work together with perfect unison. How does this come about? One very basic idea in Darwinian Theory of evolution is that each individual tries to maximize the number of offspring by every possible means. This tenet seems to be violated in case of these social insects. In these societies, only one individual, the so called Queen, does the job of reproduction. All others help in feeding and rearing this next generation. In fact a multitude of tasks necessary for smooth running of the colony are all carried out by specific groups of individuals. So, there are soldiers, workers etc. These have been called 'castes'. It turns out that genetics decides the caste of an individual and once decided it leaves little room for the individual to choose.

Paper wasp is an insect that is somewhat different. In this case colonies are much smaller. Beehives are made of wax. Paper wasps make their colonies from paper! The raw material for making paper is wood scraping and saliva. Unlike bees and ants, all females, not just one, among paper wasps seem to have the anatomical machinery to produce eggs. And yet, in a colony, at any point of time, only one individual female (Queen) produces eggs while all others devote their time to non-reproductive activities. Professor Raghavendra Gadagkar, a distinguished entomologist, studied paper wasps for years to decipher the nature of sociality. He carried out a series of simple but brilliant experiments to understand the wheels within wheels in a paper wasp colony. We will discuss one early experiment in the series conducted.

A group of newly born females from a colony were put in individual cages (tiny plastic bottles). They were provided food (certain worms), water and small piece of wood. All the females scraped the wood, made some paper and started building a nest. Each was observed till death or till an egg was laid in the paper cell prepared (which ever was earlier). It turned out that nearly half the females laid an egg. The other half did not. This confirmed the idea that a large fraction of females are capable of laying eggs. Now the question was, which wasps have the ability and which wasps do not?

Clearly, we had two groups of female wasps. One that died without laying an egg and the other that laid an egg each. How did they differ? Perhaps there were

physical differences. Did larger wasps lay eggs and smaller did not? Size of a wasp could be measured in many ways. Each became a random variable of interest. For example total body length, length of thorax (stomach), wing length etc. So, we carried out two sample t-tests to compare egg layers with non-layers. Unfortunately, all the tests gave a conclusion that differences were not statistically significant. This seemed like a dead end. We could not identify any difference between the two groups.

A new direction of analysis was suggested by Ashok Shanubhogue, a research student in the University of Pune at that time. He pointed out that the response observed was dichotomous, success or failure to lay egg. Hence we could try a general linear model, or what is now popularly known as logistic regression. Now this method does not give any closed form solutions. We cannot estimate regression coefficients using some formulas. The approach has to be iterative. In other words, we have to start with a trial solution and improve it step by step. This is computer intensive. There were no ready programs available. So, he wrote necessary programs. He did one more thing. In addition to body measurements (that varied from one wasp to another) he also proposed use of colony measurements (that were the same for all wasps from the same colony though different for two wasps from two different colonies). These included (a) number of eggs in the colony, number of larvae, number of pupae, and number of adults (b) number of cells in the colony, number of occupied cells and number of empty cells. In all there were more than a dozen potential explanatory variables or regressors. For each regressor, the null hypothesis of interest was that the regression coefficient is equal to zero. Testing of these hypotheses began. Again it seemed that we may get nothing since in most cases we failed to reject the null hypothesis.

At last there was light. One null hypothesis was rejected. Regression coefficient of the variable 'number of empty cells in the colony' was significantly different from zero. We wrote a report to Professor Gadgkar and waited. The reply was quick (as quick as it could be in the pre-internet times with almost no access to telephones). He said it made a lot of sense. Normally, the egg laying capacity of a queen is much larger than the number of cells available. So, as soon as a cell becomes empty (the pupa opens and an adult emerges and workers clean the cell) it is instantly refilled with a fresh egg. Thus, normally there are no empty cells. Presence of empty cells indicates decline in vitality of the queen. Perhaps she is growing old. So what? Well, when the queen is young, she is sure of laying many eggs and she wants many workers who will take care of this expanding brood. Once she grows old, she may want to produce a successor. Hence, assuming that manipulation by queen is involved in enabling females to lay eggs, a queen should produce workers in the early part of her reign and egg layers in the later part of her reign.

Wow! We had never thought that one rejected hypothesis in regression could mean all that. We felt good. But the last part of the letter was a pin prick to our bloated egos. Professor Gadagkar went on with his reaction to our analysis. He said he drew a simple graph. He separated wasps into groups, one group representing one colony. Now among wasps from a colony he calculated the proportion of egg layers. So for each colony he had two values (1) number of empty cells and (2) proportion of egg layers. On x axis, he took number of empty cells in a colony and on y axis he took proportion of egg layers among females from that colony. Lo and behold, there was a clear rising and linear relationship. So, said he, what was all that brouhaha about general linear model and iterative procedure and so on? We wondered too. Were we making much ado about nothing? It was a disturbing thought. Then, slowly it emerged. Our complicated analysis was needed to identify the one regressor out of a dozen that seemed to have an impact on the response of interest namely egg laying. As soon as it (number of empty cells in the colony) was identified, a simple graph could be drawn to demonstrate what that variable was doing. It was a simple graph indeed, but it was an after thought. We did not think of drawing that graph until the result of regression was out. So, perhaps Ashok Shanubhogue's analysis was relevant after all!

Applied statistics is a venture in which interactive work is a very crucial part of each project. Statisticians have to understand the questions that scientists wish to answer. Results of analysis have to be interpreted jointly. If there is no interesting interpretation, then analysis is not of great value, no matter how fancy it may be. Statisticians have to be prepared to learn a thing or two from subject matter experts. If the attitude is right, all this can be a lot of fun.

(For details see Gadagkar et al 1988).

10. Do Birds Think?

Some people believe that humans are different from other animals in that they think. I think, therefore I am! Is this really true? Is it possible that birds and the bees think too? How would we ever know? Dr Milind Watve, a brilliant biologist from Abasaheb Garware College in Pune, did a simple experiment and showed that probably birds do think. We were lucky to be associated with the work. In fact one major benefit of being an applied statistician is having access to other people's research. It is a very privileged access because you see it before the work gets published. Most people come to know interesting scientific findings only after they become stale news in science and are written about in popular science magazines.

Let us begin with a classical story. A generation or two ago, milk was delivered in glass bottles with seals of aluminum foil. In early morning deliveries, bottles were often left at the door step of the customer so as not to disturb sleep. It turns out that some birds in England learnt to break the seals. Then the birds drank milk or ate the cream floating at the top of the bottle. Birds could recognize milk in transparent glass bottles. They probably thought that an unguarded bottle was a soft target and attacked it and reaped good benefit. Is it not thinking? It could be. But we should accept new explanations only if conventional explanations are not adequate. In this case, one could argue that a bird will always approach an edible object. It will always hit a seed or a bug or whatever the object may be, as a precursor to eating. All this can happen by instinct. There is no compelling evidence that the bird thought about how to steal milk. In this instance there was the additional evidence that milk bottle raiding began in one area and gradually spread to other cities. So, the proposed explanation was that birds in one area discovered the possibility of stealing cream and milk and then gradually other birds learned by copying. Thus the MO (modus operandi) spread like an ink blot. This is very fascinating, but not entirely convincing. A simple and controlled experiment would be a much stronger proof.

So, here is a planned experiment. A common bee eater (a migratory bird that builds nests and reproduces in Pune area), was under study. A bird spends a large part of a day in feeding the chicks which grow very fast. So, the bird makes repeated trips to and from the nest. It catches a morsel of food, returns to the nest and puts the food item in the mouth of the chick. However, the return to the nest with food is not quite direct. The bird perches on a branch or wire nearby, waits a while and then flies to the nest. Presumably this is a defensive strategy. The bird looks around for any potential predator. The nest is usually well camouflaged, but going straight to the nest may reveal its location and the chick would face grave danger as the parent goes away again in search of more food. So, the parent makes

sure that no one is looking and then moves from the temporary perch to the nest. This is our idea of how the bird THINKS.

So, what would be the difference in the behavior of the bird in the presence and absence of a human observer? Perhaps the bird will linger on the perch longer when a human is present. If this logic is correct, the nature of study is clear. Observe a bird from a hiding place and measure how long it waits at the temporary perch before flying into the nest. This is the random variable. The bird will keep making trips all day. So, we can collect X_1, X_2, X_3 ... ,X_m. Now observe the same bird from open where you can be seen and generate Y_1, Y_2, Y_3, ..., Y_n. You have two random samples from two distributions. Null hypothesis says the two are identical. Alternative hypothesis says mean waiting time is longer when a human observer is present and visible. You can apply a t test or a non-parametric test, whichever is suitable.

So far things are smooth. But they do not tell us convincingly about any thinking by a bird. A significant difference in the means of X and Y could be attributed to an instinctive reaction to presence of a human being. It takes a more ingenuous experiment to yield that kind of evidence. Here is what Dr. Watve did. In his third trial, the bird, the nest and the human observer are all there. But there is one more thing. A tall wall is created between the human observer and the nest. How does it change matters? Now the bird sees the human being. So, if any delay in going into the nest is instinctive, the bird should wait for long even when there is a wall between the observer and the nest. So, we observe Z_1, Z_2, ...,Z_p. We can compare Z values with Y values. If they are similar, it will show that raising a tall wall between the human observer and the nest did not affect behavior of the parent bird. What if Z values are in general smaller than Y values? The only explanation for it would be that the bird recognizes that the potential predator (human observer) is present but (due to the wall) cannot see the nest. Hence it is safe to go in and feed the hungry baby. If this line of argument is right, then we cannot escape the conclusion that the bird thinks and then acts.

For details see Thakar et al 2002.

11. Will Frog Leg Feasting Finish the Species?

Situation: In the 1980s when Indian exports were not phenomenal, there was one item of export that added a tiny amount to country's earnings of foreign exchange. The item was frog legs. This meat item was eaten with relish in countries like Japan and USA.

This was good for exporters. It was also good for tribal workers who caught these frogs in paddy fields at night. They got some wages for this work. But not everyone was happy. Ecologists were concerned that such continuous catching of frogs (of species *Rana tigrina*) may eliminate that population altogether. Society for Prevention of Cruelty to Animals was upset because rear legs of frogs were chopped off and then rest of the animal was left to die. Agriculture experts were also worried in their turn. That was about the impact of frog harvesting on paddy crop. Their argument was that frogs eat insects. This is a natural form of pest control. Populations of insects that attack paddy may explode if there are no frogs to control the insects. If this happens, yield of rice may decline. In view of such opinions, export of frog legs was banned. Our interest was to scrutinize the basis of this decision.

Population size: We began by examining the population of frogs being harvested. A capture-recapture experiment was done. Captured frogs were marked by clipping a particular toe with a nail cutter. A drop of anti-biotic was applied and then the frog was released. Now, standard textbook formulas are available to estimate population from such data. But the formulas assume that there is a closed population (no one leaves, no one comes in). It turned out that frogs sometimes ran away after their toe was clipped. (You can hardly blame them for that!) . This meant that formulas had to be modified. In the end we came up with a modified formula. It gave an estimate of about 100 salable frogs per hectare. What do we mean by salable? Only large frogs (100 grams or more in body weight) are salable.

Population dynamics: Now we had to worry about how frog populations increase or decrease when left alone. For this we needed to estimate birth and death rates and rates of growth. To our surprise, though frog was used for dissection in colleges for many years, there was lack of basic data. For instance no one seemed to know the age at which a frog female lays eggs for the first time. In the end, out of frustration, one well wisher reared frogs and kept checking about age at egg laying. To his surprise, a female laid a clutch of eggs at a tender age of 6 months, much earlier than expected. Two things happen to frogs as they grow older. First is that their size increases. But numbers decline. At birth (i.e. when a tad pole sheds its tail, ends aquatic life and comes to land) a frog may weigh less than one gram. He may cross one kilogram at the age of one year. One can fit a suitable growth curve so that a rough estimate of weight is available at all ages up to one year. We know

the number of eggs laid by a female. But we do not know death rates from egg stage onwards. Mathematical models come handy in some situations of this kind. We examined the proportions of frogs of different sizes in a field and assumed that it is the equilibrium state. This gave us estimates of death rates.

Harvesting: If you kill some frogs, the population is sure to decline. Or so it seems. But there is some capacity to rebound. This is true of many biological systems. When we donate blood, the stock in our body gets depleted. But there is enough capacity in human body to generate fresh blood as replacement. Of course if we overdo it, the person may die of blood loss. It is most important to know how much blood can be taken out safely and at what interval of time we can do it. Similar considerations apply to the problem of harvesting any renewable biological resource. Ecologists talk of an optimal harvesting policy. It ensures good yield year after year. Such yield level is called MSY (maximum sustainable yield). Mathematical methods exist for calculating such strategy. We applied the methods and found that sustainable yield is possible. To our great surprise, the optimal strategy thus calculated indicated harvesting frogs above 100 grams in weight. This was precisely what frog collectors were taught to do. Not because the traders knew the optimal policy, but rather because smaller frogs had no market. So, we were disappointed somewhat, since we did not discover something totally unknown. However, we did find that the harvesting practice was sustainable. In other words, the risk of wiping out the population was not high. So, it seemed that the concern of ecologists was not justified. Of course our findings were based on mathematical models and some assumptions about birth/death and growth rates. Hence the findings have to be taken with caution. In general it is prudent to observe nature for a reasonable duration (in the present case, a few years) and be alert. But subject to this caveat, we find that there was no need to ban export.

Insect control: Next we turned our attention to the issue of impact on rice crop. Our decision was to seek evidence for any claim. So, we looked for research reports on frog diet. Again we were surprised by the scanty evidence available. We could lay our hands on a paper published in the Bombay Natural History Society that gave tabular information on diet composition of frog at different ages. It turned out that frogs eat insects and also small crabs. The proportion of insects in frog diet decreases as the size of the frog (i.e. age) increases. Now harvesting reduces the proportion of large frogs and increases the proportion of small frogs. This means consumption of crabs goes down but consumption of insects GOES UP. Hence we concluded that harvesting leads to more stringent insect pest control. Agricultural experts were way off the mark!

We did not have much to say about cruelty. However, on other points, our study suggested that ban on exports of frog legs was NOT justified.

See Prayag and Gore (1993) for more technical details of this study.

12. Diffusion of two innovations: Cross-bred Goats and solar cookers

Innovation is a new idea or technology or method of doing something and doing it better than before. Innovations can improve quality, productivity and/or reduce difficulties. For a society to benefit from innovations, it is necessary that the new ideas are adopted quickly by people. In other words, innovations have to diffuse in the society. In the absence of diffusion, an innovation remains unused. In fact this is the fate of many innovations. Hence it is of interest to study the process of diffusion in a society. We shall discuss in this essay the process of diffusion of two specific innovations.

Situation: India is a country in which majority of citizens live in villages. Poverty is the most pressing problem of villagers (mostly farmers). Modern science and technology can play a big role in relieving farmers from their dire situation. One such technology is that of cross breeding of animals using artificial insemination. This has been tried successfully in case of cows. We were involved in study of dissemination of a similar attempt in case of goats.

Goats are termed poor man's cows. They cost less, can be managed by women, old people or even children. They are hardy and live off the land. They have a ready market for meat and hence can help a family tide over financial difficulty. However, Indian goats usually give very little milk. One NGO operating in Narayangaon near Pune, imported a small herd of goats from Israel. These animals were of a breed called 'Saanen'. They are fine milk animals. The NGO offered breeding service to farmers and kept a date wise record of fees collected. This new idea or innovation has to spread among poor families and has to benefit them.

Modeling diffusion: It is of interest to understand the process of spread of this innovation. Who adopts such new ideas? What are their expectations? Are they fulfilled? How do they come to know about the innovation? Do they face difficulties? How widely will the innovation spread? These questions are of great importance not only for this innovation but also for rural development in general. Building a model for the diffusion of a new idea/method/technology etc can be helpful. Such models are widely used in market research. A common model is the one called logistic growth. On x axis we have date/time and y axis we have number/ proportion of people who have adopted the innovation up to that date. We found that the model did a fine job in the sense that data points were very close to the fitted curve. The model estimates the so called saturation level i.e. the maximum proportion of families that will adopt the innovation. It also suggests how long it will take before this saturation level will be reached. This answers some of the

questions about diffusion of this innovation but not all. For other questions we carried out a sample survey of farmers who adopted sannen goats.

Survey: We conducted a survey in 1986 and asked farmers who had adopted the new breed about their motivation and experience. They were mostly motivated by the example of friend or a relative. Media or political leaders did not play any role in their decision. The expectation was that of getting more milk, early maturity of females and a good price in the market. By and large the expectations were fulfilled. We asked some non-adopters about why they did not adopt. The reasons were many. Goat keeping does not have prestige. Richer farmers prefer to keep cross bred cows. Some families may not have a person free to look after goats. We found confirmation that diffusion has been through word of mouth.

Surrounding villages: Having examined diffusion in Narayangaon itself, we turned to villages nearby. There were very few adopters in these villages. Nevertheless, we went ahead and fitted the logistic curve to that data as well. The fit was good but the saturation level (the proportion of families that will eventually adopt) was very low (6%). This came as a surprise. Why should the behavior of families in surrounding villages be so different from families in Narayangaon where the saturation level was 40%? To find out the answer, we conducted a small survey in the surrounding area. The reason was as follows- Poor farmers had to take their female goats to the breeding service center on foot. If an animal in heat is made to walk the distance of 5-6 kilometers, it loses its heat. Hence the distance puts a barrier. It can be overcome if breeding service is available locally.

Solar cookers: Fuel availability is a major problem for poor families in India today. In fact one possibility is that for some poor people food may be available but not the means to cook it. Fossil fuels are expensive, wood sources are depleted. Sun is one source that is unlimited. So, any innovation that enables use of sun is most welcome. Solar cookers are an innovation that lets us take advantage of the plentiful sunlight available in India. Hence in one class in 1989 we decided to study diffusion of this new gadget in the households of Pune City using a sample survey. It requires a frame (a list of all individuals in the population of interest). We were lucky to meet the distributor of solar cookers. He had to keep a record of all sales because he had to claim a subsidy from government. He shared this information with us. We selected a stratified random sample of addresses of adopters of solar cookers. Strata were the different parts of the city, known to represent different socio-economic classes. We prepared a simple list of questions and student volunteers visited the sampled households to interview the family members (generally the lady of the house). Responses were summarized in tables.

We found that by and large users were happy with the cookers. Many people explained that food cooked in solar cookers tastes better. There were some suggestions about heaviness of the box and lack of wheels and aesthetic

appearance. People had read about the product in local newspaper and otherwise heard from friends. Political and social leadership did not seem to play any role in adoption of solar cookers. Two aspects of the findings were puzzling. One is that hardly any poor people bought solar cookers. Perhaps as a corollary, owners were not particularly interested in the savings made by using the cooker. They were either interested in the convenience or in being eco-conscious. It seemed that a subsidy was not necessary for the gadget. Why did poor people not adopt this innovation? The answer was simple. Use of solar cooker needs safe and secure open space. A slum dweller has hardly any place open to sky where the cooker can be kept securely for hours to cook the food. Clearly the gadget is unsuited for urban poor. Marketing recommendations emerging from the survey were 1) to present it as a device that promotes nature conservation. 2) to improve the appearance of the gadget and to reduce its weight.

Students enjoyed participation in these studies. Field work was regarded as fun and visit to villages was treated as trips.

Postscript 1: Over the years, consciousness about benefits of solar energy has increased in our society. Solar water heaters can be seen on terraces of many buildings. Solar lanterns are also more common. This is good progress. However, possibilities of use of solar power are far from exhausted.

Postscript 2. Sample surveys are an important part of statistical too box. Conducting a small survey relating to a problem of current social relevance/interest is a very fine way to give practical training to students and also a way to promote good citizenship. Every college can conduct such surveys without any significant additional resources.

The material relevant for goat innovation diffusion was published in following articles:

Lavraj and Gore (1987 a)

Lavraj and Gore (1987 b)

Material on solar cooker was published in Gore et al (1990)

13. How to count Wild Tigers?

Situation: 'Project Tiger' is a special activity of the Ministry of Environment and Forests, Government of India. It was launched over 35 years ago. Under this project, many areas in forests were declared as 'Project Tiger' areas and received special attention, extra funding and personnel etc. The purpose was to protect tigers. This was considered necessary because of two reasons. One is that the number of tigers in India was estimated to have declined by say 90% from about 50,000 at the beginning of the twentieth century to less than 5000 by 1975. Tiger is considered a flagship species among Indian wildlife and India is home to most of the tigers in the world. Hence many countries were interested in conservation of tigers in India. Within India, awareness of ecological problems was on the rise and there was considerable popular support for government initiatives in nature conservation.

As work on Project Tiger progressed, a public debate arose about the effectiveness of the efforts. Officials in charge of the project carried out censuses of tiger populations in different project areas and reported area wise tiger counts. These appeared to increase steadily. If these counts were accurate, the project had fulfilled its mission. Many scientists in general and field ecologists in particular, attacked these numbers and claimed that they were unreliable. Increases were too smooth. There were no ups and downs, as expected in natural populations. Counts of cubs (baby tigers) were too low. There was no validation in terms of prey base. It was argued that the method of counting followed by forest department was not accurate. Hence there was a need for independent review of the methodology of counting. This is where some of us came into picture. We decided to examine the practices of forest department and suggest modifications.

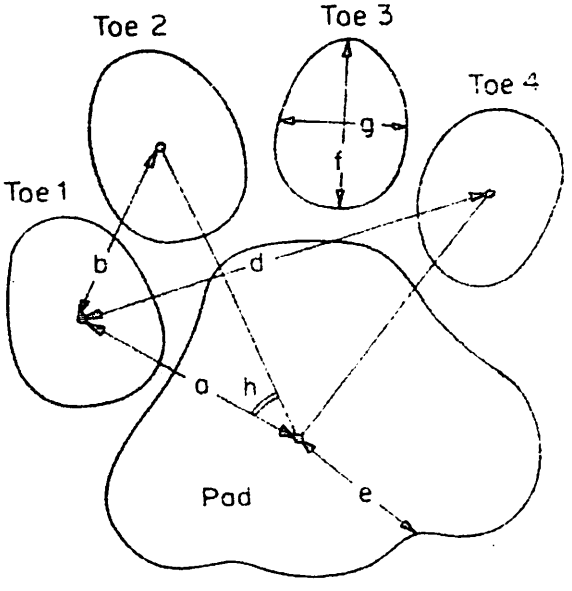
What was the method followed by forest officials? Observing tigers is difficult and dangerous. Hence a new method was devised. It was based on pugmarks. During a specified summer week, the forest area is scanned and all pugmarks located on ground are recorded on paper using a simple glass tracing method. These tracings are then compared. If two tracings match, they represent the same animal. So, one of them is discarded. The number of pugmark tracings left behind in the end represented distinct animals and hence gave the animal count.

Our strategy: We felt that basically the procedure seemed intuitive and reasonable. Its weakness lay in the fact that there were too many judgments involved which could make the method arbitrary. So, it seemed necessary to make the method objective and based on quantitative analysis. We began by measuring various areas and distances in a pugmark. This collection of measurements would replace a picture. An illustration is shown in the figure below. This one step introduces quantification. Once that is done, all analysis can be computerized and

judgments can be minimized. But did we capture all features of the pugmark in our set of measurements? We tried one check.

Foresters believed that female pugmarks are elongated while male pugmarks are bigger and square. So we collected pugmarks of a few males and females from a zoo in Pune and used the data to set up a formula based on measurements. This formula estimated the probability that the animal was a male. If the probability was high, then we guessed that the animal was a male. In this training data set we knew the truth. Hence accuracy of our formula could be checked. It was not 100%. But more like 90%. That is not too bad.

Quantification of Shape of Tiger Pugmark



Next our doubt was whether the measurements would vary depending on who drew the tracing or what the soil was like. We carried out two experiments to check this. In one experiment we located a trail of pugmarks (a series of prints produced when a tiger walks along a forest path). All these prints come from the same animal. So, any differences are due to chance alone. We selected a few pugmarks and asked several foresters to trace each of those pugmarks. Any differences in tracings were now entirely due to differences among operators. We checked this using analysis of variance (ANOVA) and eliminated any measurements that were too sensitive. In most measurements, differences from operator to operator were minor.

Next experiment was comparisons of substrates. We needed pugmarks of the same animal on fine soil, sand and mud (three different substrates). This seemed

too difficult to get in nature. So, we requested an NGO in Coimbatore to help us. Their caged animals were made to walk on substrates prepared to our specification. Again the resulting measurements were analyzed and it turned out that effect of substrate on measurements was not too significant. We were satisfied that the measurements were a reasonable depiction of the pugmark and robust to operators and substrate.

The last step was to set up a method for comparing pugmarks and deciding if they represented the same or different animals. For this purpose, we set up an algorithm for classification/clustering. The logic of the algorithm was as follows: Two pugmarks differ in measurements, even when they are of the same individual animal. If the difference is excessive, then very likely, they represent two distinct individuals. So, we obtain a good estimate of the extent to which pugmarks of the same tiger differ. This is called intra-individual variation. Put pugmarks in the same or different clusters depending on how their difference compares with intra-individual variation.

Capture-recapture: One standard method of estimating size of animal populations is the so called capture-mark-release-recapture (Capture-recapture for short) method. In this method, animals are captured and tagged (metal tag, spot of paint, clipping a nail etc.) and then released. The logic is the rule of three in arithmetic. Suppose we capture 10 fish in a pond and mark them and let go. Next time again we capture 10 fish and find that half of them are already marked. Then we judge that half of all the fish in the pond are marked. But we know that number to be 10. So the total has to be 20. Of course this method cannot be applied to tigers since they would not be kind to anyone trying that. Instead, cameras are set up in the forest and tigers get photographed repeatedly. This method is now practiced in some forest areas.

Of course nature is full of surprises. Sunderban area in west Bengal, which is home to a sizable population of tigers, seems to defy either of these two methods. Such challenges are the great stimulants for development of other innovations. Indeed history of statistics is full of cases in which an attempt to solve a difficult real life problem led to discovery of exciting new statistical methods.

14. Harvesting strategy for Eucalyptus

India is a land of villages and farmers. Poverty is their most pressing problem. Green revolution has helped one section of farmers come out of poverty. But that needs irrigation, fertile land, and use of inputs such as hybrid seeds, fertilizers, and insecticides. What does a farmer in a rain-fed region do particularly when his land is of poor quality? In fact traditionally, farmers did not adapt their land use to suit the topography and soil type. In the USA, there is land classification into various types and prescribed land use for each type. Land that has lost fertility may have to be given rest for fertility to get restored. Could we try something similar here? Classifying land is not difficult from a technological viewpoint. The issue is whether farmers can afford the luxury of following the recommended use. As an example they cannot let a piece of land remain fallow for restoration of fertility. They have to put it under plough each year or they will go hungry. Thus plots on hill slopes also get ploughed and the loosened soil is washed away by rains causing permanent loss of topsoil and fertility.

One via media for low quality land is agro-forestry. This means growing trees instead of annual crops. This has the advantage that labor requirement is low, green cover on soil is maintained continuously and there is a steadily rising price level for wood and wood products. One drawback is that the payoff comes after many years. So, a farmer has to worry about cash flow.

How profitable is agro-forestry in fact? What kind of income can a farmer expect? This is an important question and we wanted to get answers based on hard facts. So, a project was launched. We were lucky to get co-operation of Nasik Zillah Nilgiri Utpadak Sahakari Sanstha Ltd - a group of farmers in the district of Nashik in north Maharashtra. This group had taken up planting of eucalyptus trees in their land. Nilgiri, i.e. the eucalyptus is one of the straightest trees in the world, but the debate among ecologists on its suitability has been far from straightforward (Agrwal A. and Narayan S. 1985). There are five main uses of forest resources. They are: food, fodder, fuel, fiber and fertilizer. Eucalyptus trees are no good as source of food, or fodder. Their main use is for making wood pulp for paper and to some extent in construction industry. Members of the co-operative had undertaken to plant, grow and harvest eucalyptus wood and had accepted the project of establishing a pulp and paper mill. In Nasik district about 2311 farmers planted eucalyptus on approximately 10,000 acres of land.

We began with interviews of farmers who were members of the group. It turned out that the fields in which this species was planted were of different types (in terms of fertility as well as access to irrigation). Clearly then, the earning had to be different in such diverse groups. So, we decided to do calculations separately for each group. Every year the tree grows bigger. Hence its worth in the market also

goes up. In that case what is the age at which the tree should be harvested? This becomes an optimization problem. Should a farmer wait indefinitely? That does not sound right. There is some cost associated with waiting. So, economists consider discounted present values. Suppose interest rate is 10%. Then 100 rupees deposited in a bank today will increase to 110 in one year. In other words, getting 110 rupees after one year is similar to getting 100 today. If we can get more than 110 after one year, it may be better to wait. This logic was to be applied to the farmer's dilemma.

Here is the plan underlying the solution of the optimization problem. Estimate the yield of wood per unit area at different ages of the plantation. Estimate future prices and hence total gross income. This will be higher, the longer we wait since volume of wood increases with time and so do prices. Now find out the discounted present value in each case. Select the age of harvest that gives maximum present value. In order to implement this idea, we needed to know the volume of timber expected at different ages of the plantation and also prices in the current year and in future. This is where statistics can be very helpful. Our agenda was development of a statistical model that will predict growth of eucalyptus in different types of fields at different ages. To build such a model, we needed data on growth. Luckily there were farmers who had planted trees at different times. So, we could visit plantations with different degrees of fertility etc and with different ages. Now we had to measure growth. This is not as simple as it may seem. We need to measure the volume of wood of a standing tree. The only way volume can be measured accurately is by felling the tree, cutting the bole (main stem) into smaller/manageable logs and measuring volume of each log. But we could not fell the trees. So we had to estimate the volume of a standing tree. Here we note that a bole tapers from ground up. So a tree can be thought of as a cone. Volume of a cone is given by area of the base multiplied by height divided by three (as we learn in high school geometry). Area at the base is easy to measure. We can use a tape and measure the circumference. We know that circumference equals diameter multiplied by π (= 22/7). So we can get the diameter and hence radius. Further we know that area is $\pi.r^2$. All this is fine. But how do we measure height of a tree? That is not very easy. An indirect way is to measure the length of the shadow of the tree and then estimate height using properties of a right angle triangle. But this is also not practicable in farms where density of trees is as large as 1452 per acre as was the case in Nasik district. So we decided to use age-old method of using a rod as suggested by forest officials. The suggestion was to prepare a measuring scale with the help of light weight G I pipes. So we purchased a few pipes, did some fabrication work and prepared measuring scale of adjustable length. Then a measuring tape was hooked to it and using the same we could measure height of eucalyptus trees. Fortunately plantation was not more than 4 year old and trees were not too tall, so our trick worked.

Suppose we know the height of a tree. How do we get volume? The formula would be height multiplied by $\pi.r^2$. This looks very precise. But do not get fooled. Natural entities like trees do not obey such mathematical rules too well. They are not quite the shape of a cone. It is an approximation. It is what scientists call a model. So we have to get some idea of how crude an approximation we have. This can only be checked if we have, in a few cases, true values as well as values given by the formula. Again we were lucky. There were a few trees cut for market and we did get true values of volume. The two (true volume and volume calculated assuming a conical shape) did not quite match. There was approximation. So, we decided to predict volume from height and basal area using data on true values. The statistical tool employed was 'multiple regressions'. This is a standard practice in forest management (Maslekar, A. R. 1981).

Table 1: Estimated total volume per acre in different plantations

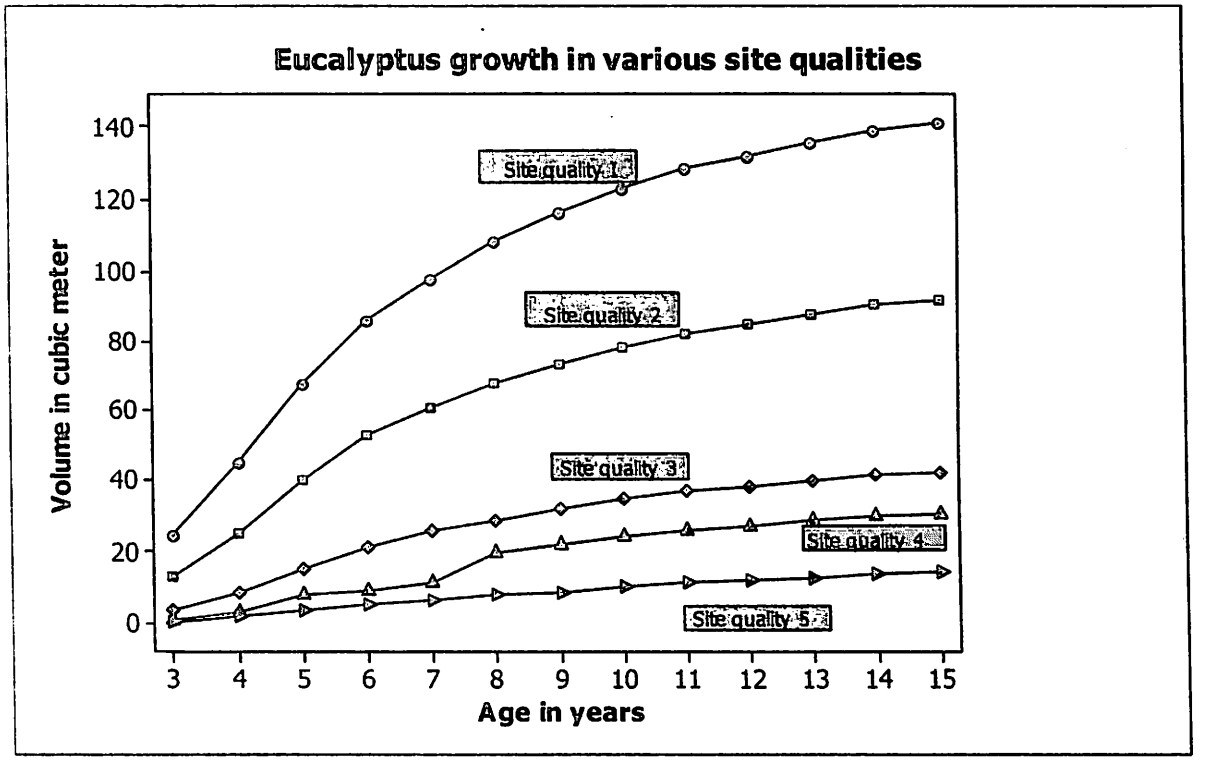
Age(in years)	Soil Type	Water availability	Volume (in m^3) Per acre	Income (in Rs) Prices of 1989
2	Poor	No	2.67	1760
	Poor	Medium	5.63	2750
	Medium	No	4.56	1895
	Medium	Medium	16.85	8750
	Medium *	Good	43.02	9125
	Good	Medium	3.95	1435
3	Poor	No	1.01	570
	Poor	Medium	16.12	6810
	Poor	Good	18.86	7755
	Medium	No	1.54	610
	Medium	Medium	12.21	5375
	Medium	Good	1.85	800
	Good	No	13.81	4875
	Good	Medium	12.68	5875
	Good *	Good	70.37	17180
4	Poor	No	10.57	6500
	Poor	Medium	14.68	7875
	Medium	Medium	7.48	3750
	Good	No	11.28	5250
	Good	Medium	39.78	17125
	Good	Good	25.63	10875

*: Plantation density was almost double in these farms

One question arose in our mind. Can we improve this method of estimation? That would require putting in more information. We decided that one more measurement about the shape of the bole should be tried. That would be girth, not at the breast height but observation on girth at one meter above breast height. We checked out this idea and found that such input (which is easy to measure) does improve the estimation of volume. This was published as a paper in a forestry journal (Kulkarni and Gore 1988).

One more thing. We collected data on 21 farms of various site qualities and also studied the data published by Chaturvedi A. N. (1983). See Table 1 on volume of timber by age of plantation in sampled farms.

We can fit a suitable model to the available data on eucalyptus growth and use it to estimate the volume of trees per acre at all different ages. Here is the data on volume of eucalyptus hybrid in five site qualities. Logistic growth model turns out to be appropriate for these data. This exercise enables us to estimate volume at all ages



All right, let us take a review. Where do we stand? We can estimate volume of trees from a few measurements. Using data on different ages and different types of fields, we can predict the volume of wood produced at different ages. Now, if we estimate market prices in future, we can calculate income for a farmer and also its

discounted present value. All this was done. What was the outcome? We discovered that the right age of plantation at which it should be cut and marketed is about 6 years for "good" farms (site quality 1, 2 and 3) and was 7 or 8 years for "poor" quality farms (site quality 4 and 5). What would be the income of a typical farmer? The conservative estimated income, after the discounting the total amount at the rate of 10 percent is reported in Table 2).

Table 2:
Expected Income* per year from 1 acre of Eucalyptus plantations in different site qualities

Site quality → Harvesting year ↓	1	2	3	4	5
4	2472	1373	709	313	88
5	2799	1657	893	399	137
6	**2878**	**1755**	**963**	445	167
7	2632	1640	930	**454**	168
8	2434	1519	863	435	**171**
figures in **bold** indicate highest income for corresponding site quality * price level in year 1989					

We were rather disappointed. It did not give a very rosy picture. A farmer was not going to get a roof of gold (to use an old Marathi expression). Income was going to be very modest.

We took this report and met a leader of farmers involved in agro-forestry. We explained our work and how we were rather unhappy about the results. His reaction was quite different. He said to us that in fact we had proved that the venture has a good potential of making a reasonable profit. In that sense we had shown that it was a proposition that deserved loan from a commercial bank. We were delighted to see that after all our endeavors were not in vain.

15. Adaptive Sampling

Estimating Number of Species in an Ecosystem

Situation: Human livelihood is based on use of renewable biological resources. We need crops, forest produce, dairy and fishery for survival. We depend on a large number of species to satisfy our requirements. By and large, poorer sections of our society depend more critically on fruits of nature. Against this background we have to see the fact that the diversity of life forms (now- a -days called 'biodiversity') is on the decline. Every year many species of plants and animals disappear. We cannot re-create a species once it is lost. Set of all existing species is a non-renewable resource. Of course since time immemorial, species have disappeared from the face of earth. One well known example is the entire range of dinosaurs. They became extinct long before human beings appeared on the scene. However, of late, under the impact of industrialization, the rate of disappearance of species has increased a thousand fold. This has alarmed scientists and planners everywhere.

Slowly but surely threat to biodiversity is appreciated by political leaders as well. Leaders from all over the world gathered in Rio de Janeiro (Brazil) and signed the earth convention in 1992. The convention recognized biodiversity as an important asset of the people and laid down the objectives of preparing an inventory of species and also monitoring status of biodiversity over time. These are immense tasks and cannot be completed without participation of people at the grass roots. Clearly, there is an urgent need to plan for monitoring biodiversity. To put it simply, we need to know how to estimate in an area, number of species in different taxa (animal/plant groups such as birds, trees etc.) Following table gives approximate number of species for some selected taxa for the whole country.

Animal/Plant group	Approximate number of species in India
Flowering plants	15000
Fishes	1700
Birds	1200
Insects	60,000

However, our interest is in estimating such counts for much smaller areas such as a national park or a village forest. To study these areas, we need to plan and prepare budgets. So we should be able to say how many person days of work will be needed. Field biologists may be able to judge how many trees or birds they will see and identify in a day. So it is convenient to think in terms of count of individuals as the effort put in. In this sense how much effort do we need to put in

to get good estimates? No one knows. But it is known to be a difficult task. An easier task is to measure some index of diversity. We use indices in many contexts. Cost of living index is one that is familiar to many. When stock markets are booming, every reader of the business page of newspaper knows about Sensex or Nifty (two indices of stock prices). Similarly, some indices have been constructed to measure biodiversity. Two most popular indices are Simpson's index and Shannon-Wiener index. These are spelt out in the third essay in section 1.

Effort needed to measure biodiversity: Typically, field biologists survey forest areas and record the species of each individual animal/plant seen. So, they build a list of types and tokens (list of species and how many individuals seen). We need to know for how long this survey should go on. Budgetary constraints are such that we have to keep the demand to a minimum. To answer this question we adopted a simulation approach. What does that mean? We imagine that a biologist is taking a survey and she sees a tree. What species is it? We have a reference list of species with abundances (how frequent are the trees of that species). One tree is selected randomly from this list. That is the one seen. This process of surveying (on paper) is continued and we accumulate the results.

We will illustrate this idea with a fictitious example of drawing a sample of trees from a list. Suppose our ecosystem has 10 different types of trees with frequencies as given in the table below.

Serial number	Name of the tree	count	Cumulative count	Cumulative proportion
1	mango	50	50	0.50
2	jackfruit	20	70	0.70
3	bamboo	10	80	0.80
4	neem	5	85	0.85
5	tamarind	4	89	0.89
6	Teak	4	93	0.93
7	sal	2	95	0.95
8	champak	2	97	0.97
9	coconut	2	99	0.99
10	arecanut	1	100	1.00

We want to draw a random sample (with replacement) of 3 trees from this ecosystem. I draw 3 random numbers from uniform distribution over the range zero to one, (in simple terms, three fractions). The numbers come out to be .98, .31 and .06. The first case puts me in class 9 (the fraction is between 0.97 and 0.99). So the tree selected is coconut. Next number is in class 1 (the number is below 0.50). So the tree selected is mango. Last number is also in class 1 and again a mango tree is

selected. So, in my sample I have 2 species. It is clear that the ecosystem under study has at least two species.

We know how big the species list is. How many random numbers do we have to draw to "see" all the species? That can be a too demanding. Rare species are hard to see. The effort needed may get out of hand. So, we set up a more reasonable target such as seeing 80% of all species. Even this is very hard. In that case our next alternative is to estimate a diversity index. Our simulation study suggested that continuing a survey till 1000 individuals have been observed should lead to a fairly reliable estimate of the diversity index. On the other hand if we insist on estimating the total number of species, the effort needed is one order of magnitude higher, in other words, we have to observe 10,000 individuals.

Sampling: Suppose we have decided about the effort to be invested. The next question concerns how that effort should be distributed among the different types of areas under study. One forest area we studied had three ecotypes. They are: Teak plantation, evergreen belt and moist deciduous forest (in which trees shed their leaves every year). So, it is better to go about systematically, sampling these different forest types. We discovered that a very efficient way of sampling is to go cyclically. Say spend a day in each forest type in the first round, and then repeat the whole exercise. In every round you will see unseen species. If that count falls too low, drop that type. Such cycle sampling can cut the cost by half. Why call this adaptive? Because, the sampling strategy is flexible. If one area gives a lot of species, then we spend more effort there. If an area yields poor results, we drop it.

Conclusion: Through a simulation study, we found an answer to what may be an important question about planning a species count in a locality. About 1000 individuals have to be checked to get a good estimate of a diversity index and about 10,000 individuals have to be checked to estimate the number of species. Whatever the level of resources available, it is better to divide the study area into different ecological types and spend the resources in different parts in a cyclical manner. Effort should be continued in forest types that show many species and discontinued in areas that fail to show many species. This insight seems to be new to ecology.

16. Green Revolution, Evergreen Revolution and Statistics

Now a days all over the world, when the name India is mentioned, people think of IT, BPO, KPO, high growth rates, urbanization etc. But the truth is that India is still very much an agrarian society. Admittedly, the proportion of GDP coming from agriculture is declining (now about 25% from nearly thrice as much at the time of independence). But the proportion of the population dependent on agriculture is still very high (now about 60%). So, we must never lose sight of problems in agriculture. Also, so much of modern statistics can trace its origin to problems in agriculture.

Problems in agriculture! What problems? Many are not even aware. So, let us recall. People of this country have lived through periods of food shortages. In the sixties, food aid from the USA (under the so called Public Law 480) bridged a critical gap and any major crisis like the Bengal Famine of forty's was averted. Yet signs of stress were there for anyone to see. In the sixties milk and sugar were scarce and rationed. It was illegal to invite large crowds for feasts even in marriages.

Fortunately, all this is history. Indian population today is far greater than what it was in sixties. And yet, there is plenty of food to go around. Credit for this goes to the green revolution. Farmers, scientists and government machinery together worked wonders. Hybrid seeds, irrigation, chemical fertilizers and pesticides; these were the main drivers of the paradigm shift from shortage to plenty.

Today, as we continue to enjoy the fruits of green revolution, new concerns are being voiced. Is this era of plenty sustainable? Our population is still growing at a rate well above 1%. Planners fear that food production may not be able to keep up. It is getting harder to improve seeds any further. We are running out of places to build dams. Irrigation projects like Tehri in Uttaranchal and Narmada in MP and Gujarat have faced stiff opposition. Pests and pathogens are found to develop resistance to chemicals and farmers have to use more of the same or more poisonous chemicals to protect crops, with associated problems of pollution and health. Chemical fertilizers are expensive. They increase our dependence on imported hydrocarbons. Fertilizer subsidies are crippling our public finance. These are the storm clouds on the horizon. Will we be able to sustain even a modest growth rate in farm production? Will our land and water resources continue to be squandered or will we preserve them? A sustainable growth in agriculture can be termed 'evergreen revolution'.

Converting our green revolution into an evergreen revolution is a major challenge facing India. Some lines of action are obvious. Biodiversity prospecting- we must examine wild flora and fauna to locate new genes that will enhance

productivity of crops. IPM- Farmers have to adopt integrated pest management systems (use of insect traps, use of insects to control pests, herbal pesticides etc.). Crop planning- better crop rotation, mixed cropping. Water/soil conservation – drip irrigation, better drainage etc are means to achieve water conservation. These are some options.

All this is fine. But what role can statisticians play? We believe that statistics teachers and students can play an important and creative role. We will try to elaborate.

Statistics is mainly concerned with understanding variation/uncertainty and with exploiting that understanding. Is there uncertainty in agriculture? Yes, of course. The most important example of it is weather. Weather can make or break a farmer. Too much or too little rain, too high or too low temperature, too many days with cloud cover and many other events can have adverse impact on crops. Farmers deal with these uncertainties in many ways. It is never certain when monsoon rains will begin. But farmers prepare fields ahead of the rains. Next decision is choice of sowing date. You cannot sow until there is enough moisture in the soil. But the longer you wait the greater may be the chance of missing the bus in some other way e.g. number of sunshine hours may become too small in the last phase of the crop. If rains are too late, farmers may give up on kharif crop altogether and wait out the whole season. They can then sow the seeds after rains are over and hope to get a rabi(winter) crop using the accumulated moisture plus the morning dew. How are all these decisions taken? This is an interesting field in its own right. We will just mention some names. Amos Tversky and Daniel Kahneman who received Nobel Prize in 2002 for their work in the area of behavioral economics have addressed a general theme of this kind. We recommend reading about their work. It may give ideas for student projects or even new research. At least the Indian farmer can be compared with western populations in terms of decision making methods. Who knows, traditional Indian decision process may be better. It seems fair to say that humans have considerable capacity to manage uncertainty, perhaps through accumulated knowledge. But the capacity is not unlimited. People fumble if things change too quickly. Here is one example.

Traditional farming is based on a limited range of crop varieties. If a new variety is brought in, farmers may not understand associated idiosyncrasies. In one locality, a new variety of groundnut was enthusiastically adopted for its higher yield. But over the years there grew some frustration because too often there would be rains at the time of harvest. Too many pods are left behind in soil when moisture level is high. A systematic study showed that the difficulty was going to persist. The new variety had a shorter time to maturity and rains in this locality did not finish by then. The variety was more suited to the rainfall pattern at the research station where it was developed. This simple fact was lost sight of by scientists as

well as farmers. This is avoidable. Statistical skills can be harnessed to overcome some of these problems. We propose that statisticians can study uncertainty and build decision support systems useful for farmers. A similar problem described by a horticulturist concerns choice of pruning date for an orchard. Generally fruit trees are pruned once a year to promote new growth. Take the example of zizyphus (ber) tree. These are pruned in summer and flowers blossom in about 90 days. How can we select the pruning date so as to minimize the probability of rainfall while the tree is in bloom? If it rains at that crucial time, there is flower loss and fruit formation is reduced. So, there is interest in avoiding rain just as there is interest in receiving it.

Teachers and students can try their hand at some such problems, once or twice, and get lots of enjoyment in the exercise. In one such attempt, we found that a groundnut farmer needs to decide whether/when to spray pesticide on an insect called 'leaf miner'. If it rains right after the spray, it washes out the pesticide. On the other hand if there is a good shower, that itself can act as a pest control. We used rainfall data at the locality (for over 80 years) and lots of assumptions about the biological processes involved and came up with the rule 'After the pest attack is noticed, wait for 6 days. If it does not rain, then use pesticide on 7^{th} day.' We called it a 'Wait and see strategy' (Gore and Paranjpe 1998). In another case we found that we could anticipate the intensity of attack of a fungus species on groundnut crop, several weeks ahead, using weather data for that season. Such anticipation can be used to give the crop extra nutrient inputs thus making it better prepared to fight the onslaught (Mayee et al 1998). We want readers to note some attractive features of such work. (1) It is interesting. (2) It is publishable. (3) It takes us closer to the reality of our society. One can get a sense of participation in ongoing action. (4) It makes attractive seminar material. Students and other audiences are far more interested in such work than purely theoretical work. (5) There is always a possibility that the work may suggest some theoretical investigation. If that happens, you will be on strong turf in terms of defending the relevance of your theoretical piece. (6) There is endless scope here. Problems differ from crop to crop and locality to locality. Likelihood that someone else would anticipate you is almost zero. In fact there is enough work here to keep a whole generation of statistics teachers and students busy. (7) Remember that no one expects any spectacular results. Small incremental gains are quite enough for the work to be considered worthwhile. All we need is an increase of a couple of percentage points in productivity to keep up with the rise in population!

There are two kinds of inputs necessary for such work. One is weather data. It is available from India Meteorology Department. They charge some fee. But for academic work it is low. Some data for Karnataka are available at the Indian Institute of Science website (check the page for Center for Atmospheric Sciences).

You may want to get data for the locality of interest to you and hence extra effort may be needed. The second input needed is information about crops. For this you must liaise with the nearest agricultural college or university or perhaps some botanists near you may also want to help/collaborate. Many people show interest in such work but do not know which problem to pursue. This can be recognized only through interaction with farmers and agricultural scientists.

17. Modeling Intense Rain

Situation: Rainfall is one of the most important natural phenomena for all of us. When rain gods smile, there is prosperity. If they get angry, there is suffering. Whether it is too much or too little, abnormal rainfall causes lot of damage to society. And the key feature is that no one knows ahead of time how things will pan out in a year. This uncertainty prompts use of statistics to predict rainfall. This prediction is then used to guide action. We shall describe one such case.

Dams are a powerful means by which man harnesses the energy of water. Dams are useful for irrigating farms, generating electricity, supplying drinking water to towns and cities and also to control floods. Dams are very expensive to build. Also, they create a large reservoir of water and extensive land gets submerged under this manmade body of water. Safety and security of dams is of great importance both because of the cost and also because of the potential havoc downstream if the dam breaks. To prevent its collapse, a dam must be strong enough to withstand pressure of water collected by the wall of the dam. This pressure depends on the quantity of water collected, which in turn depends on the rainfall in the catchment area. If there is intense rain, then large quantities of water accumulate in a short time. This can cause, even after opening all the outlets to release as much water as possible, high stress on the wall of the dam. If stress crosses a limit, the wall can cave in. That is a catastrophe! So, we want the wall to be strong enough to withstand effects of intense rain. The key question is 'how intense?' Engineers focus attention on rainfall in a day. If we check daily rainfall at a location, we can find the largest value in a year. Dam must be strong enough to face such a rainy day. But the rainfall of that day varies from year to year. So it is necessary to think of how high the rainfall in a day can be. This is given in terms of PMP or probable maximum precipitation. We have to be pretty conservative here. It means we should take a fairly large value of daily rainfall as reference in designing a dam. Value selected should be such that chance of daily rain exceeding that value should be slim. This prescription is reasonable but still vague. So, it is made more specific. In case of small dams, select the value that is exceeded with chance less than one in hundred. In case of large dams go still higher and take a value exceeded with a chance of one in thousand (or even ten thousand). The key question is "How do we estimate PMP for a particular place?"

Data availability: Before we answer this question, we need to know what kinds of data are available. In India, weather data (including rainfall) are gathered by IMD (India Meteorology Department), an arm of the Government of India. It was created and nurtured by the British rulers and it has continued to prosper after independence. Its workforce gathers data at nearly 400 locations in India. Data for about 100 years is available in the archives of the Department. Students and

teachers should indeed acquire (for a fee) data on their locality and study it. In any case, we have about 100 observations on the variable 'maximum daily rainfall in a year in a specified locality'.

Estimating PMP: If we have 100 'x' values, and we want to estimate population mean, we can simply use the sample mean (i.e. mean of the 100 values available). That trick does not work for estimating an extreme value. If I want PMP with 1% probability of being exceeded, perhaps largest among 100 values may work. But for major dams we want probability of one in thousand or even one in ten thousand. This needs more sophisticated tools.

Modeling maximum daily rainfall: To understand the behavior of this variable, we should prepare a histogram. It turns out to be, not symmetric bell shaped but positively skewed. That means there are a few very large values in the set. This is typical of extreme variables. So, we had nearly 400 data sets corresponding to as many physical locations within India. We tried to fit a skew distribution to each set. Different models turned out to suit different sets. Gamma, lognormal, Gumbel are some of the distributions used. Once a model gives a reasonably good fit, it can be used to estimate the rainfall value such that area to the right of it is only 0.001 (or 0.0001 depending upon our choice).

This is a valid method. However, mathematical statisticians were concerned that the estimates so obtained may be unstable. They may have a large variance. Such estimates are not dependable. To check stability, we did a simulation study. Random samples of 100 observations were generated from a given skew distribution and estimate was calculated from each sample. Such repeated calculations can show us whether the estimate is stable. Luckily, we found that there was no problem. So, we did get estimates of PMP for use by irrigation engineers. It also turned out that these estimates were better than the estimates in use. (for detail see Deshapande and Gore (1999) and Gore et al (2001)).

Hourly maximum rainfall: We discovered that the method we had developed was useful in designing storm water drainage systems in cities. Here the reference unit of time is not a day but an hour. When it rains cats and dogs, if the water does not get drained, it accumulates in city roads. That can cause disruption of business and industrial activity. Such disruption can be costly. So, drainage system must have adequate capacity to take care of extreme storms and intense rains. What should be the capacity of the drainage system? Here we face the usual conflict. If capacity is low there can be disruptions. But if we want a large capacity then cost of building it goes up. So, we need a dependable estimate. Now remember that the data needed is hourly rainfall for many years. This is a more stringent demand. For the city of Pune we had such data for 30 years. However, the mathematics of the problem remains the same. So we were able to obtain required estimates which were given to the design engineers.

18. Weather Insurance

[This essay is inspired by an article in The Indian Express (Monday March 20th, 2006). Author of the article is Ms. Sucheta Dalal. (e-mail address-suchetadalal@yahoo.com) She was a regular columnist and wrote about economic/financial matters. Her emphasis was on good governance and social well-being. In 2007 the Government of India recognized her yeoman's service to the profession of journalism with an award of 'Padmashri'. Her website is worth visiting.]

Statistics has played an important role in development of agriculture in post-independence India. Perhaps foremost use is in estimation of area under different crops and estimation/forecasting of yield. This is very important for policy making, though it is of limited interest to individual farmers.

The second use is in designing experiments for selecting the best agronomic practices. Third field of application is plant/animal genetics and selection of improved varieties/breeds. A new and emerging field in which statistics can play a big role is crop insurance.

There are about 100 million farmers in India. They seem to work the hardest and yet go through a life of poverty and misery. Among other things, they seem to have to face very high levels of risk in their enterprise. Any government that is committed to welfare of farmers has to think about reducing the risk in farming. One major risk in agriculture is due to vagaries of weather. These can make or break a farmer's fortune. If rains fail, crops fail. That is obvious. If rains come at wrong times, then too result is the same namely crop/income failure. Weather can help or hinder growth of insect pests or fungal disease. Hence a college of agriculture always has a department of meteorology. Its job is to teach how to use weather forecast for better crop management. The advice is sometimes rather simple-minded. If rain is expected, don't irrigate. If a frost is expected, give some warmth to your orchard. If cloudy weather is expected, harvest your cauliflower quickly or else it will become loose and will fetch a poor price in the market. All this is useful but not enough.

How do we, common people, manage risks? How do we prepare to face accidents, health emergencies, fire and other disasters? By purchasing an insurance policy! So, we can ask (like Professor Higgins in 'My fair lady') why can't the farmers in India be like us? Perhaps they do not know about availability of insurance or it may indeed not be available. But that is not quite true.

In India, a pioneering attempt at crop insurance was made in 1870 in the erstwhile Mysore state but the idea did not spread. In the post independent India, there were many committees that deliberated on the feasibility of crop insurance. But one man who promoted the idea very strongly and successfully was late Prof. V M Dandekar. Under his intellectual leadership, the Government of India launched a

pilot crop insurance scheme in 1979. It was administered by the General Insurance Corporation and limited to some areas of Maharashtra and Gujarat. In this pilot phase a modest premium amount of Rs. 1.65 crores was collected and claims were paid to the extent of Rs. 1.35 crores.

In the simplest terms the scheme collects insurance premium money and compensates farmers for loss of yield. How is the loss assessed? For that we have to know expected yield and actual yield. Expected yield is a 3 to 5 year (prior to year of interest) moving average of actual yield. Actual yield is assessed by crop cutting experiments. Estimates are obtained, not on any individual farmer's field, but for a large 'homogeneous area' such as a Tehsil /*Taluka.* If estimated yield is below expected yield some indemnity is paid. It is calculated as {[expected yield-actual yield]/expected yield} multiplied by sum assured. All this was done through the banking system. Sum assured was always the crop loan taken by the farmer. The bank paid premium and indemnity was paid into the loan account. It took about a year for payment. Farmers really did not know what was going on. It seems like a scheme to cover the risks, not of farmers, but of banks giving loans to farmers. No wonder the impact on farmers has been minimal. Perhaps there are difficulties in running a crop insurance operation. Estimating crop loss due to an unexpected weather event must be very difficult. In fact estimation of potential yield and actual yield are both difficult (scientifically as well as administratively).

Of late one reads about farmers' suicides in different parts of our country. Sometimes the cause may be market fluctuation. Thus for example, some areca nut farmers in coastal Karnataka had to face a virtual meltdown when very cheap supplies from Southeast Asian countries reached our markets after the import barriers were lifted. But otherwise, the most common cause of suicide is inability to cope with crop failures. Current insurance schemes have done very little for these hapless farmers.

It seems that an excellent alternative exists which can mitigate effects of crop failure. That alternative is weather insurance. You pay a premium (similar to car accident insurance) that the insurance company keeps if weather remains normal. If weather turns bad, insurance company pays you an agreed multiple of the premium amount. This eliminates the need to estimate crop loss. Hence the method seems much more practicable.

In order for such an insurance policy to be fair to both sides, we need good risk assessment. This is where statisticians can play a useful role. Firstly we need more data on weather since good insurance policies must be based on reliable data; about rain and about yield. There are nearly 500 rain gauge stations operative in our country while there should be ten times as many. A simple and yet very useful activity will be to measure rain every day in one's locality. You can get a measuring cylinder with suitable markings and place it in a suitable place and keep record.

This has to be done for a long period of time (many years). Data collection is a good student project in the first term of an academic year and its analysis can become a project for the second term. In fact, not just collages but even high-schools as well can do this data collection. Collecting data on other weather parameters can also be contemplated. Assuming availability of weather data, we can think of risk assessment.

Let us consider a hailstorm that can wipe out citrus fruit crop in eastern Maharashtra. If the risk of a hailstorm were 1%, what would be a reasonable trade off? A farmer should be paid 100 times the premium if the disaster occurs. [Of course there has to be some provision to cover the administrative expenses and profit of the insurance company.].

How big is this task of risk assessment? It is simply enormous if our desire is to cover the entire country. Here is one off-the-cuff estimate. There are about 30 districts per state (i.e. about 750 districts in the country). A typical district may have two agro climatic zones. District of Pune in Maharashtra has an eastern half that is a low rainfall zone while the western half is a high rainfall zone. Thus there may be about 1500 agro climatic zones. Each zone may have say two crops of interest to begin with. In eastern part of Pune we can consider grapes and figs. Thus there will be 3000 separate entities to be studied. In each case a model is to be arrived at to relate rainfall to crop yield. (Our sneaking suspicion is that we may have underestimated the size of the task by one order of magnitude since there are many crops in need of insurance and each crop has many varieties and each variety may have distinct weather requirements.)

It seems that every college in India offering a degree in statistics can take up a distinct study that may be of interest to the local economy. There can be an "All India Coordinated Network" [or AICON] specially developed for this purpose. Perhaps insurance companies interested in expanding activity in this sector may sponsor such work. Agricultural Insurance Company of India made a beginning by offering insurance of the type we have described. (Please do visit their website for more information). Perhaps some private insurance companies such as ICICILombard also have some policies for this purpose.

Under the program we visualize, a typical project will involve the following steps:

1. Select a local crop (say sorghum and mango)
2. For each crop select one (or more) major risk(s). As an example, in case of Jowar (sorghum) if it rains just before harvest, grains turn dark and fetch a lower price. In case of mango, if it rains during the flowering season in spring, the yield goes down. [For identifying these aspects, liaison with botanists, agricultural colleges and practicing farmers is necessary].

3. Collect weather data relevant to your problem (from India Meteorology Department or local agricultural meteorologist etc.). Analyze the data and estimate the risks of events of interest.

4. Design a model insurance policy. [Liaison with Commerce College or Business Management College will be useful.]

This is our proposal. It can be implemented as a research project with funding from the government or sponsorship from a commercial organization or a co-curricular activity. If the community of statistics teachers is, on the whole, supportive, this action plan can easily turn into reality.

Of course we should also know the steps that have already been taken by different players. Government has now launched a new scheme called 'Varsha Bima' which operates differently from the older scheme described above. It is motivated by the fact that over half the variability in yield is due to rainfall variation and that over two thirds of all farm yield is in kharif season. Hence the scheme tries to evaluate elasticity of yield in response to rainfall. It is calculated using the regression equation

$$\text{Ln}(y_t)=a + bt + c\{\ln[\text{actual rain/normal rain}]\}$$

'y_t' is yield. 'c' is the elasticity parameter. 't' is time. This equation is used to compute expected yield. Estimated reduction in yield multiplied by the minimum support price of the crop gives the amount of indemnity. (If they do not know it, students will have to find out what minimum support price is). Thus only data on rain is enough to compute the amount payable if any (assuming that the model is ready).

Two more policies have been drawn up. One is called a 'sowing failure policy'. If you purchase this policy and rains fail during pre-specified sowing season, costs of sowing are reimbursed. Again statistics students who are not familiar with agriculture may wonder about cost of sowing. It can be large. In case of groundnut, sowing cost can be as high as a quarter of total cost. The other policy recognizes that total rainfall is not adequate to predict yield. It distribution during the season matters a lot. So, rain requirements in different growth phases are considered and a composite rainfall distribution index is constructed. Indemnity is based on this index. This particular product has received a cool response from the market. No one wants to buy this policy, perhaps because it is complicated and evaluation of the index is not as transparent.

There are some more schemes in pipeline. One is excess rainfall insurance. The other is drought proofing for district collectors. If there is a drought, the collector will be paid a sum to cover relief operations. It is proposed to float 'varying interest bonds' to be sold in the capital market with interest rate tied to rainfall. If rain is good, interest rate is high and if rains fail, the interest rate is low.

Clearly, many innovative things are happening in this field. Only sky is the limit. If you know your local situation better, very likely you can design a policy better suited to your area/crop. Would it not be wonderful to apply statistical skill for such purpose? We hope our professional colleagues would get quite excited about the possibility of playing an active role in helping our distressed rural brethren.

19. Poverty

Situation: Perhaps the most serious problem faced by the Indian nation is that of poverty. Millions of citizens of this country have to survive with meager income. Poverty leads to mal nutrition, sickness, short life span and in general deprivation in all aspects of life. Poverty reduction, if not elimination, is a main objective for any sensitive government. India has always had poverty alleviation programs throughout the post independence period. Perhaps the most famous of them is the "Garibi Hatav" slogan of Indira Gandhi. How can we judge whether any such program has succeeded or not? How can we compare poverty situation in different regions or states? One way of comparing poverty between regions or between time points is by using some index of poverty. This is how a statistician's skill becomes useful in the context of poverty. We shall discuss here, how an index of poverty can be constructed. This essay has some overlap with one earlier essay (article 6 in section I). However, we have kept repetition to a minimum. We have added new material here.

Head Count: The simplest way of measuring poverty is by calculating proportion of people in the country, who are poor. The head count of the poor people in the numerator and total population of the country in the denominator. How do we find out the head count of poor? This is done by using the so called "poverty line". It is the level of income so low that any one earning less than it is obviously poor.

This approach has the weakness of not taking into account the actual deficit.

Income Gap: To eliminate the drawback of head count a different index of poverty is proposed. Here income of each poor person is subtracted from poverty line to get income deficit. Average of such deficit value is calculated.

However, it does not pay attention to inequality among poor people. According to some economists, that is an important aspect. Hence some improvement is necessary.

Amartya Sen's axiomatic approach: The famous Nobel prize winning Indian Economist Amartya Sen developed ideas about how a poverty index should be constructed. He proved a mathematical theorem which showed that his own index is the one and only formula that fulfills all requirements that a good index should satisfy. Here is the index constructed by Amartya Sen.

$P = H\{I + (1-I)\ G\}$ where; G is the Gini coefficient (index of income inequality), I is a measure of the distribution of income, both computed only for individuals below the poverty line and H is the head count.

Where is Poverty Line? As you must have noticed, all these poverty indices depend upon a poverty line. Therefore, we should ask a question namely, "How do we find the poverty line?" A simple answer has been given by economists. Poverty line is the income needed to buy all goods and services

necessary for life. This description indirectly tells us the method of calculating poverty line. Make a list of items needed, write down the quantity of each item needed. Note the market prices of different items. Multiply price by quantity and add over all items. This gives you the poverty line. A very reasonable procedure indeed!

There is, however, one difficulty. How do we get a list of essential items? Such a list will necessarily vary from one group to another. In Maharashtra, coconut is not an essential commodity. In Kerala, in contrast, it is a part of everyday cooking. For a vegetarian, meat and fish are not essential. For a non-vegetarian they are. Is alcohol essential to anybody? Some people will say no. Others will point out that alcohol is a very important ingredient of food and religion in many tribal communities.

Let us ignore these difficulties for a moment. Suppose we do have a standard list of essential items. How do we get the quantities needed? If we talk about food, perhaps experts in the field of nutrition may know how much we need. In fact, nutrition books do tell us the minimum requirement of various food constituent such as starch, proteins, oil, minerals and vitamins. We learn about these things in science text books. However, the science of nutrition also has its own soft underbelly. It turns out that historically nutritionists have often revised their estimates of minimum quantities needed. This by itself should not worry us. It shows that the science has a self correcting culture. Unfortunately, almost all revisions are downward. To put it differently nutritionists have always overestimated requirements. It seems therefore that calculating the poverty line is very difficult task. We feel that it can be simplified greatly by observing what people do. Our philosophy is that in nature all animals know what to eat and how much of it to eat. The same should be true for people also. Therefore, we should examine behavior of poor people to estimate poverty line. This is where statistics plays a crucial role.

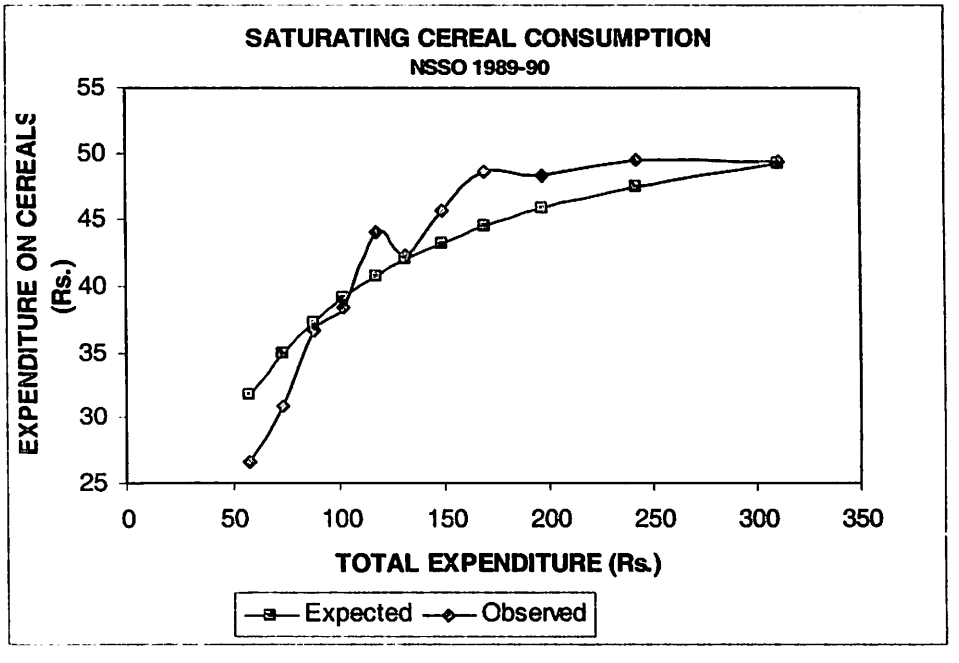
Saturating Function Model: What do we mean by observing poor people? We try to see how much of their limited income, they spend on any given commodity. Let us be more specific. Perhaps the most essential item is cereals. So we should monitor expenditure on cereals. In fact such monitoring is done regularly by the National Sample Survey of India. Their periodic surveys (called rounds) give information on the income of a family and money spent on cereals. (Statistics students should procure reports on some rounds and browse through them). We shall plot income on the X axis and expenditure on cereals on Y axis (a small point of detail is that instead of income it is the total expenditure on all commodities that is recorded. This substitution is needed because total expenditure is easier to record compared to income and is regarded as a good surrogate). It turns out that in the initial stages as income increases so does expenditure on cereals. However, after a

while, there is saturation. In other words, increasing income does not lead to increase in expenditure on cereals. When this happens we shall say that expenditure on cereals has saturated. In other words, the poor people are telling us that now they prefer to spend extra money on items other than cereals. This saturation level can be interpreted as the required quantity of cereal. The wonderful thing about this approach is that we depend on nothing else but the people themselves. There are no external experts dictating how much the requirement should be. A mathematical function that is very suitable for modeling saturating curves is the rectangular hyperbola given below.

$$Y = \frac{Vx}{(K+x)}, \quad x>0$$

(We invite students of mathematical statistics to verify the following properties of this function: This is a differentiable function. First derivative is positive and the second derivative is negative. The function has an asymptote at y=V. We will call it the saturating value. At x=K, the curve attains y=V/2. Hence K is called the half saturation constant.)

Here Y is the expenditure on cereals made by family with income x. V and K are parameters of the model. The model can be fitted to data given by National Sample Survey Organization (NSSO) using methods of regression analysis (The NSSO data are freely available. Hence we recommend to students and teachers of Statistics that they carry out an exercise in regression analysis using data for their own states/regions). The fit of the model to actual data is quite good. One illustrative case is shown below.



New Index of Poverty: So we have now discovered how to calculate the requirement of cereals. We will use this value and obtain average deficit for the population. That will be our measure of poverty. The mathematical formula is given below.

$P = \sum f_i \left[\frac{K}{(K + xi)}\right]$; the sum is over all expenditure classes.

Poverty Scene in India: Whatever be the formula for a poverty index selected by experts, a common man would want to know the extent of poverty. Have there been any changes in poverty level? If we do necessary calculations then a clear picture seems to emerge. Over the years there has been a gradual but consistent decline in poverty in India. Secondly, poverty in urban areas is slightly lower than poverty in rural areas. Of course, many people will be skeptical of these number games. They would want to see concrete evidence of decline in poverty. Some of the indicators that can be used for this purpose are as follows:

(a) Do we see many people begging? How often do we come across a person wearing tattered cloths?
(b) How many people are homeless? How many have homes made of very poor quality material?
(c) What is the extent of unemployment? Do people come to famine relief works etc.?

Answers to such questions can give concrete evidence of change in poverty. The answers can be based on every day experience or we can carry out statistical surveys. That is yet another opportunity to use statistics in study of poverty.

(For details see Sitaramam et al 1996)

20. Market research

One area of application of statistics that is becoming an attractive field for employment is market research. Students of statistics will benefit a lot if they try to become familiar with way statistics is used in this field. We are going to describe experiences gained while collaborating with some experts in the field. Here is what wikipedia (the web-based free encyclopedia) says about this field.

Market research is the process of systematically gathering, recording and analyzing data and information about customers, competitors and the market. (Does this not sound familiar? This is very close to some people's definition of statistics. The only difference is that here we are focusing attention on customers.) *Market research can be used to determine which portion of the population will purchase a product/service, based on variables like age, gender, location and income level.*

Market research is generally either primary or secondary. In secondary research, the study team uses information compiled from other sources applicable to the product of interest. The advantages of secondary research are that it is relatively cheap and easily accessible. Disadvantages of secondary research are that it is often not specific to your area of interest and the data used can be biased and is difficult to validate. Primary market research involves surveys, field tests, interviews or observation, conducted or tailored specifically to that product.

Two terms are used a lot in this field that we should be familiar with. We give below a description for each of them again from wikipedia.

Market segmentation: *Market segmentation is the division of the market or population into subgroups with similar motivations.* (This seems familiar too. It sounds like classification problems in multivariate analysis.) *Widely used bases for segmenting include geographic differences, personality differences, demographic differences, use of product differences etc.*

Market trends: *The upward or downward movements of a market, during a period of time are a topic of considerable interest to marketers.* (Notice that time series analysis is a standard topic for statistics).

We hope you see that statistics is a discipline that is very useful in this commercial field. Therefore some experience in market research can enhance employability of statistics students to a great extent. Students and teachers should check out websites of some market research companies as a part of statistics training. Here are some examples.

www.imrbint.com , www.acnielsen.co.in , www.metricconsultancy.com

We are going to write about two cases. One involves a bank and the other involves an oil company.

The bank: Here the question of interest was 'why customers change banks'. In particular, the desire was to study the nature of customer dissatisfaction with services of the bank. Our plan was to take a random sample of account holders and ask them their level of satisfaction (on a scale of 1 to 10) about individual services and estimate average satisfaction level.

The first hurdle was sample size. How many account holders should we interview? This calculation is very important since cost of the study depends on it. We had to make a series of assumptions. Assume that the responses will follow a discrete uniform distribution over the set 1-10. Then we know the standard deviation of the response. Let us assume that we want the sample size to be enough to ensure that the estimate has a confidence interval (confidence level 90% say) with a half width of 1. We came up with an answer of 36. There is often some problem of drop outs. So, we started with a value of n=40.

Next step was identifying the account holders to be interviewed. Manager of the branch of the bank was ready to cooperate and we advised him to select 40 account holders randomly. He did the job. The people were interviewed. It was time to analyze the data. We began with a summary. The first set of questions was about personal background. It turned out that average age of the 40 people in our sample was more than 70 years. We were shocked! How could this be? Does this bank cater to old people only? A brief observation of the operations of the branch did not give that impression. So what went wrong? Perhaps our sample was not representative. Did we not draw a random sample? It suddenly dawned on us that we had committed a mistake by depending on someone else for this very critical step. It turned out that the manager had delegated the task to a clerk who faithfully opened the list of all account holders and took the first 40 cases. These were the people who had opened account in the branch when it was inaugurated 20 years ago. No wonder their ages were all on the higher side!

We had to start all over again. This time we did not repeat the error. With a correct draw of the random sample we could show the bank what priorities customers had among the various services given by the bank.

The next exercise involved risk modeling. Here the bank had in its archives, records of loans/advances given to different customers and the repayment pattern in each case. The question of interest was whether it is possible to guess before sanctioning the loan whether the person will be punctual in repayment. We treated this as a problem in logistic regression with a binary response (repaid on time/ defaulted) and all variables (about the borrower) with information in the loan application as covariates. This helped us test hypotheses about each regression coefficient. If the null hypothesis that the regression coefficient is zero is accepted, then the relevant variable is not useful for predicting default. Final model selected

can be used in case of a new applicant. If the estimated probability of default is high, the application can be rejected.

Oil Company: As you know one major concern of vehicle owners in our country is adulteration in fuel (hence the slogan 'Pure for sure'). Retailers may illegally add cheap substitutes like kerosene to petrol for making quick profit. We all want this to be stopped. One way is to take samples from the fuel being sold and send them to laboratories for analysis. That can be time consuming. Can there be a short cut? A possibility considered by some oil companies is addition of a marker. Have you seen blue kerosene? A dye is added to the kerosene on sale in ration shops. The purpose is to let everyone know that the material is subsidized.

Purpose in the present case is different. Hence the same trick will not be suitable. The alternative system will work as follows: A very small quantity (2.5mg/liter) of an invisible marker chemical is added to the fuel as it flows in pipes from refinery into storage tanks. If a dealer adulterates petrol, the proportion of this marker will go down. A portable machine is taken to petrol outlets. A sample of fuel being sold is checked by this machine which measures proportion of the marker. If the proportion of the marker in the petrol is the same as in supplier's storage tanks, the machine gives a reading of 100. If the dealer adds kerosene, the proportion of the marker is reduced. If proportion has gone down the machine reading also goes down. That is proof of adulteration and quick action can then be taken.

Well, this seems like a tamperproof system. But it is better to exercise abundant caution. So, we examined data of various kinds to check feasibility of this proposal. The main reason for caution is variation. It turns out that the proportion of the marker in the parent stock may be just one value, but as petrol is pumped into trucks to be transported to retailers, samples from the trucks give rather variable values of the proportion of marker. How do we interpret this result? (If you want to be applied statisticians, you should close the book at this point and think of some reasonable interpretation. You can then compare it with our interpretation. This is a very important step. With availability of software and computers, calculation is no longer a mark of expertise in statistics. Interpretation, on the other hand, cannot be delegated to a computer.) We thought of two possible reasons for this variation. One is that in fact in spite of all care, the marker does not get mixed well with petrol. The other possibility is that the machine that measures the proportion of the marker is itself the source of variation. Which of these is easily testable? We believe reproducibility of the portable machine is the property that can be tested easily. How? (again, we recommend that you give your own answer and then compare it with ours). We can take many samples from the same source and check measurements for all of them. SD or CV of such values is a good indicator of how consistent the machine is. It turned out that the machine is not perfect. So the idea

was to avoid insistence on getting a reading of 100 and accept anything from 98 and above. But even this relaxation was not enough. We discovered that in almost 5% cases, a sample of pure petrol gave a machine reading below 97!

Having absorbed this information, we did some calculations. Since the machine may not give the true value of the proportion of the marker in the fuel at the pump, we have to allow for some margin of error. If the test measurement is close to 100, we should say the product is pure. If the value is much lower, we suspect tampering. Remember we are talking about a one sided test. Adulteration is indicated by lowering of the proportion of marker. If we use such a test, we are faced with two types of errors. We may wrongly declare an honest dealer as dishonest or miss on someone who is involved in adulteration. We have to choose our decision rule such that both these undesirable events have low probability. What recommendation did we give? We prepared a table of different decision rules and consequences. Decision rule specifies the machine reading that should lead to closure of the retail outlet and prosecution. The table lists a range of choices and probability of type I error for each choice. We have given power (i.e. 1- the probability of type II error). It depends on the degree of adulteration/impurity. Power will increase as the adulteration level goes up. If we decide that a reading of 96 or below will lead to closure of the pump, we will do so wrongly in 1% cases (this is α) . Such a rule will detect 10% adulteration with almost 99.9 % chance. That is what the table tells you.

Comparing alternative rules
Power at various impurity levels

Cut off (error rate)→ (α) % impurity ↓	97 (.05)	96 (.01)	95 (.005)	94 (.001)
1	.13	.04	.02	.01
5	.85	.63	.53	.33
10	.9999	.9985	.9968	.9864
15	1	1	1	1

This is as far as a statistician can go. Next step is choosing one of the options and implementing it. It is a very difficult step. Just one example of that difficulty is that such a procedure may need a change in law. Second difficulty is that implementation of such a law can lead to direct confrontation with organized criminals and can be dangerous to the implementers. But discussion of these matters, though very important, is beyond the scope of this article.

21. Cosmetics

Scenario: We all want to look attractive! We dress well, wear ornaments and otherwise decorate ourselves. But of course there are limitations to how far that can go. After all we cannot change the basic components namely our body and our face. Or can we? There is plastic surgery to correct any deficiencies in the shape and size of various body parts. But that is a bit much for many. Something less drastic is more in order. We use cosmetics (face powder, lipstick etc.) But these are short lived, typically just a few hours. Many of us are unhappy with our skin complexion. Often we wish to look fairer than what we are. Plastic surgeons cannot arrange this for us. Cosmetics have to be applied every so often and can be recognized as such. If only the skin complexion could be made more fair! Nearly a generation ago, some chemists from Hindustan Lever Limited, came up with a new product called 'Fair and Lovely'. This product, they discovered, could give a skin lightening effect. The product became a phenomenon. A generation later, it continues to fetch crores of rupees of revenue. A slew of competitors have entered the market now. Initially, the target population for this product was young females. Now the sex discrimination seems to be on the way out in this field too. Advertisements, of late, invite young men also to spend money on such products to get a fairer face.

We three have been statistics consultants to the skin products group of Hindustan Lever Ltd for over fifteen years. This essay is a description of use of some statistical tools in the development of products in the 'cosmetics' stable.

A statistical design for a skin lightening experiment: How would we plan an experiment to check efficacy of a formulation for making you look fairer? First we decide the target population. Let us say young women in the age group 18 to 30 years in a defined region, say metro area of Hyderabad. Then we select a random sample from this population. This is not easy. We can draw samples from electoral roles. But contacting these individuals will be tedious and many of them may not be interested in participating in a trial. So, it may be more feasible to advertise and offer some monetary incentive for participation. It means that the applicants may be from lower income groups. Does that vitiate the trial? The answer is 'no, if the response to be studied is not dependent on income levels'. Some judgment will be involved here and also some experience. We believe that volunteers are indeed suitable as subjects.

Now the selected participants have to be trained to use the formulation in a particular and uniform way and also to maintain a roughly uniform lifestyle. If some women work for long hours in intense sun, it may have an adverse effect on their skin. This needs to be avoided. Etc. Now comes the question of measuring response. Some machines are available for this. They are called mexameter and chromameter. However, it turns out that assessment of the degree of darkness of

skin by an expert is regarded as a better way. Why? Are the machines not more consistent and objective? They may be. But some how, the aggregate effect of various aspects of skin such as color, shine, dryness etc is judged better by an expert. After all user of the formulation wants her skin to LOOK more beautiful. The assessment is generally a grade or rank. So, we have a case of ordinal data. Suppose we make the assessment once a week. Then change in skin color can be recorded as difference between current measurement and previous measurements. If we are looking for change in one particular week, we take a difference between two successive measurements. If our interest is effect of the treatment as a whole, we take the difference between pre-treatment measure and current measure. Using these values, we can test the null hypothesis that application of the product does not cause any change in the degree of darkness of the skin. Perhaps the test suitable for this situation may be Wilcoxon's signed rank test. If the difference is statistically significant, we may want an estimate of the change. We can use the Hodges-Lehman estimator of the median here. If advertiser wants to claim that the formulation works in say 5 weeks, then we should check the difference between reading after 5 weeks and initial reading.

This is just the basic test of efficacy. There are many more aspects of interest. Perhaps the effect of the formulation depends on the initial level of darkness of skin. If so, we have two options. We can use analysis of covariance and treat initial level as a covariate. This will yield a regression coefficient and will tell us how much the effect is. Another possibility is to divide the population into different sub-groups by level of darkness (low, medium, high) and compare the effect of the formulation across the groups. Further, markets in different countries may be different in terms of what the customer is looking for. Hence we may want to conduct distinct trials in different countries of interest. Now laws in different countries may be different and we may have to adjust with these differences. Here is one example. A good design of experiment seeks to use homogeneous blocks. In case of cosmetics, the ultimate block is two parts of a person's face. We apply the formulation to left half of the face and leave the right half untreated. This is done for one group. For another group, we reverse and apply the active formulation to the right side. This is very good from a statistical point of view. But such 'half face trial' is not allowed in India. We can imagine the reason. If the formulation is effective, the person may end up with one half of the face much fairer than the other! Would you want that?

The next step is that of drafting a claim. It has many aspects. How much improvement? One possibility is to quote the estimated median difference from initial value. But this means that half the population will get a benefit less than the value quoted. Many people do not realize this implication. Perhaps you think it is too much that half the customers will get a benefit that is less than the estimate

published. We can take care of that concern by quoting something smaller, say 10^{th} percentile (or first decile). Some others may regard this as too conservative. How can we be bold and correct? Can we claim that our product is twice as good as the competing one? Marketing people think that such a claim can do wonders to a product. Will such a claim be challenged in court? Can it be defended? We worked on this problem and published what could be the beginning of a methodology for making a justifiable 'times better' claim. Some tricky statistical arguments have to be used to arrive at such a claim.

If a half face trial is allowed, two formulations can be compared on one person. What if there are more than two formulations? We can use a balanced incomplete block design with bloc size 2. What can we do in India where half face trial is illegal? We can use arms instead of face. Here too we have only two options, left and right. However, we can accommodate multiple formulations on each arm. Oh, but an arm is not homogeneous. Underside of the arm (often called volar) is light while the lateral side (away from body) is generally darker. So, we have to treat volar side as a block and make sure that all formulations are tried on volar side. Lateral side becomes another block.

In any serious trial the endpoint is a conclusion about one or more products. Statisticians are expected to give a crisply stated conclusion in the form of a certificate. If there is any court case about claims of a company, statisticians may have to appear before a court of law as expert witness and give a technical deposition. Now there is always a lot of discussion about the exact wording of a certificate. If you check 'sale' advertisements, you will notice wording like 'up to 75% off'. What does it mean? Is it fair to the customer? Draw your own conclusion. But remember, there is always a tussle. Businessman wants as attractive a claim as possible. Scientist wants as accurate and scientifically defendable a claim as possible. A via media has to be found.

So, the field of cosmetics involves rather significant amount of statistical thinking and analysis.

22. Can we measure writing style?

Statisticians are in the business of interpreting data from any field. What about literature? That sounds like a case of over extending oneself. After all, creative writing is so intuitive that any quantification of it is beyond the realm of possibility. Or so it seems. However, truth of the matter is that many attempts have indeed been made to quantify literary style and some of them are well known in the field. In fact a name has been coined for this subfield. It is 'stylometry'. It seeks to 'measure' in some ways, the style of an author. You can try a Google search with this key word. You are sure to get many references.

The main idea here is to quantify style. This may sound like a tall order. But one can try some simple steps. Let us begin with a couple of things used in the early stage of research. One is word length. If a text has n words in it, we can denote by X_i, the number of letters in the i^{th} word. We can think of distribution of word length. It is possible to do a class exercise and obtain word length data on a text of interest. Once we have a frequency distribution, the next question is whether any standard probability distribution model can be used to describe the data. Generally, the data are positively skewed.

Another variable of interest is sentence length. Again, it is easy to generate sample data from any text. Again, the data are typically asymmetric. There can be a few sentences that are too long. For both word length and sentence length, a lognormal distribution can be a good candidate.

Apart from general curiosity, will this exercise of fitting distributions be of any immediate practical use? One answer is that the models may be helpful in resolving authorship disputes. What are these? In literature, sometimes we do not know who the author of an essay, a book or any such piece of writing is. Or, there may be differences of opinion among experts. For example, there have been debates in the England about whether all the plays of Shakespeare were written by him or there were other authors (for example Francis Bacon and Christopher Marlow) who used the same name. Similarly, it is not clear if Bhagavadgeeta, the great Indian philosophical discourse was work of one author. In more modern times, there were many essays written during the American struggle for independence. Some of them are unsigned and they could have been written by Alexander Hamilton or Thomas Jefferson or John Adams. Two statisticians (Mosteller and Wallace) examined these materials to decide authorship. Their book on the subject is now famous.

How can stylometry help in resolving such issues? The resolution is based on some premises. 1. An author's style remains uniform in all material written by that author (or at least the material under investigation). 2. If a text with disputed authorship is subjected to stylometry analysis, then it can be compared to literary pieces of known authorship (by one or more candidate writers). The item under

investigation can then be attributed to the author whose writing is most similar/very close to the disputed material. Nearness or distance in terms of Style can be characterized by patterns such as word/sentence length distributions.

Are these assumptions correct? One of us studied some essays in Marathi with authorship in dispute. Likely authors of this material in 19th century were Lokamanya Tilak and Gopal Ganesh Agarkar. We checked the hypothesis that for the same author the word length distribution remains the same across essays (where authorship is not in dispute). The answer was disappointing in that the null hypothesis was rejected. Similar is the story about distribution of sentence length. So, we arrived at a rather negative result that word length and sentence length distributions may not be very helpful to decide who the more likely author is. It turned out that usage of commas and some typical words was a possible discriminator (see Gore et al 1979).

In fact Mosteller and Wallace proposed that word use frequency may, in general, be a better discriminator between authors. What kind of words? Well some words are contextual. An essay on diabetes is sure to have the word 'sugar' used often. But that is not really a reflection of author's style. On the other hand the so called function words (words like a, an, and, but, if etc.) and other non-contextual words are better indicative of style. Hence frequency of such word should be used for comparing authors. Such study naturally involves counting words in texts. This can be a tiring job (Try it and you will see).

Here is one description of a statistical method based on word use frequencies. It is given in wikipedia.

In one such method, the text is analyzed to find the 50 most common words. The text is then broken into 5,000 word chunks and each of the chunks is analyzed to find the frequency of those 50 words in that chunk. This generates a unique 50-number identifier for each chunk. These numbers place each chunk of text into a point in a 50-dimensional space. This 50-dimensional space is flattened into a plane using principal components analysis (PCA). This results in a display of points that correspond to an author's style. If two literary works are placed on the same plane, the resulting pattern may show if both works were by the same author or different authors.

Please remember that the job of implementing such a method can be quite demanding and on top of that, there is no guarantee of success.

A recent item in this field is a news in Indian Express dated October 21, 2008. It describes James Pennebaker, a Professor of Psychology in the University of Texas who specializes in analysis of communications from terrorists. His website http://wordwatchers.wordpress.com/ is worth visiting.

The particular analysis of interest here is that of speeches of contestants (McCain and Obama) in the election for President of USA.

"We have recently analyzed the nomination acceptance speeches of candidates to perform deeper computer analyses of language..."

"We have also found evidence to suggest that McCain and Obama have different thinking styles. Whereas McCain tends to be more categorical in his thinking, Obama is more fluid or contextual in the ways he approaches problems..."

There were also a few departures in language use by the two candidates compared to their earlier debates. Obama, for example, used more 1st person singular pronouns than his opponent for the first time in any debate we've analyzed. This may be due, in part, to the fact that McCain used his "my friends" only once. Obama also used more achievement words than McCain which has typically been a reliably high marker for McCain.

Using the LIWC computer program, the differences in language usage between the categories in the third debate were as follows:

Category	Examples	McCain	Obama	Interpretation
Word count		6596	7339	Obama talks more
Words per sentence		13.83	18.39	Obama longer sentences
Big words (> 6 letters)		17.77	18.72	Obama bigger words
Personal pronouns		10.22	9.22	McCain more personal
2nd person	You, yours	1.91	1.39	McCain more pointed
3rd person singular	He, she, her	1.33	0.63	McCain more reference to others
Indefinite pronouns	It, those	6.67	7.67	Obama more vague
Articles	A, the	6.76	6.24	McCain more categorical thinking
Past tense	Was, gave	3.35	2.68	McCain talks about things in the past
Present tense	Am, is	10.01	12.06	Obama more present oriented
Future tense	will	1.39	0.91	McCain more future oriented
Social references	Friend, we, talk	11.75	10.19	McCain more references to others
Overall emotion words	Happy, hurt, kill	5.43	5.01	McCain more emotional
Causal	Because, reason	1.43	2.13	Obama more causal reasoning
Tentative	Maybe, perhaps	1.52	2.15	Obama perspective difference
Work	Job, paycheck	3.99	4.86	Obama more references to work
Achievement	Try, succeed	1.82	2.58	Obama higher in achievement words

We hope you will note how differences in style are likely to be interpreted by others. There will be questions of statistical significance of observed differences. But that apart, the above example is intended to show you the possibilities of using stylometry for very contemporary purposes.

Section III
Introduction to 'Statistics education'

A very important section of intended audience for this book is students and teachers of statistics. We the authors are ourselves teachers. Hence it is natural that we have some opinions about how this subject should be taught and related matters. We think that there is much about statistics education in India that could be improved. So, we have taken upon us the task of suggesting changes in syllabus, practical exercises, software, illustrative data sets, books and what not. These are not just details. We believe that we are trying to propose a change in perspective. We view statistics as a tool for solving societal problems and not as a branch of mathematics. Hence emphasis has to be on ability to understand problems and translate them into statistical questions, on good communication with non-statisticians and on modifying standard tools to suit particular situations. We suspect that absence of this perspective is responsible for what we regard as a decline of the role of statistics in our societal life. Our feeling is that the suggestions presented are quite feasible since they are all tried and tested. We hope our readers are able to use some of the tools we describe.

23. Why statistics has lost its central role in societal affairs in India?

All three of us started our careers in 1970s as assistants to Professor P V Sukhatme. He was perhaps the first Indian to get formal advance education in statistics. He finished his Ph D in statistics in University of London in 1935. He made a deep impression on us. One of his precepts was that society at large is far more interested in general issues and difficulties faced by the public than in any technical aspect of statistics. It responds warmly if there is any attempt to analyze societal issues and to devise solutions for them. His famous book was titled 'How to feed India's growing millions'. Now who can resist the temptation to find out just how? In contrast, the ambience in a typical statistics course in India or Statistics department in an Indian University is far removed from any kind of reality, let alone a current issue.

Mahalanobis and Sukhatme, two most distinguished names in statistics in our country are associated with efforts to develop, not just statistical theory, but answers to vexing problems of the country. This commitment not only brought them laurels but also gave them immense prestige among decision makers. Our impression is that current generation of academicians in statistics does not have this perspective.

One natural reaction to this observation is that those were giants, exceptional individuals with vision and mission backed by dedication, ability and influence! These unique episodes cannot and will not be replicated!

Obviously there is some truth in such sentiment. To the extent that we are lesser men/women, some of the decline in prestige of statistics since their time, may be inevitable. And yet, we cannot absolve ourselves of the responsibility to do whatever is possible. Can we say, in all honesty, that we have tried our best? We would like to submit humbly, that there is room for a lot of work.

Theoretical statistics work emanating from Indian universities during the last twenty five years has had, at best, a marginal impact on the field. The situation in applied statistics is no better (and in fact may be worse). There are two essential features of a good work in applied statistics. They are- (i) strong relevance to domain of application and (ii) innovation. They require an intimate interaction with researchers in fields of application (something rare in Indian educational institutes). In fact there is often an undercurrent that applied statistics is for those who find theoretical statistics rather tough. Nothing could be farther from truth. There are, no doubt, some 'concessions' available while working in applied statistics. If the result obtained is very important in the domain of application, then the paper becomes famous even if the statistical methods used are elementary. Secondly if a new method is successfully used to solve a crucial problem, then a very rigorous proof of all results can be postponed to a future date. But we see very little of serious

applied statistics in Indian Universities. Our thought therefore is the following: since work in applied statistics is likely to have a greater immediate impact on the society and since theoretical work being done is not very effective, would it not be better to increase the effort put in the area of applied statistics? Unfortunately, we seem to have a mental block when it comes to applied statistics. Here is a revealing quote from T V Hanurav, a distinguished statistician, from his article titled "The Applied Statistician" (J. Ind. Soc. Agr. Stat.(1995) Vol 48 No 1 p-1-12.)

Applied statistics, which was held in high respect when Mahalanobis, Sukhatme etc., vigorously propagated statistics, is slowly getting looked down in academic institutions. With a few exceptions, University Professors tend to teach statistics as if it is a branch of mathematics with questions in examination like 'State and prove the theorem...', 'Derive the distribution of...'. Distinctive feature of the subject that it deals with reality is lost sight of. R A Fisher once said '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 an expert in Newtonian mechanics or dynamics of rigid bodies'.

One can only hope that the coming generation of statisticians will keep this in mind.

Professor B R Bhat in his lecture at the 25th annual conference of the Indian Society for Probability and Statistics (ISPS) at Bangalore made some interesting remarks. 'One of the intentions of ISPS was that it should be a body liaising with government and other institutions in statistical matters as a mouthpiece of statistical community. But this desire does not seem to be fulfilled and we have to work harder in this direction.' We believe that Professor Bhat's observation is correct. That is not to say that statisticians are not invited to interact with government bodies. But the interaction is of a limited nature. Professors are called upon to set question papers and examine candidates (for examinations conducted by the Union Public Service Commission, Council for Scientific and Industrial Research etc.). They also sit on committees to scrutinize research-funding proposals received by Department of Science and Technology and other bodies. But that is true of all fields. It is nothing special about statistics. And we all believe that our discipline has an extraordinary role to play in societal and governmental affairs. The central government conducts many surveys and there is a felt need for expert advice. Generally some experts from Indian Statistical Institute get invited. If the survey is in the field of agriculture (which is often the case) someone from the Indian Agricultural Statistics Research Institute (IASRI) is also invited. Participation from universities is rare. This is partly because ISI and IASRI are premier institutes and have a long tradition of research in sample survey theory and practice. Unfortunately, the number of available experts is declining. If expertise were developed elsewhere, opportunities would follow. But sample survey is not a

fashionable topic today. You do not hear lectures on it in scientific conferences. Very junior faculty often teaches it. In reality, one will find exciting challenges in this field if we look in the right direction.

We would like to point out a few areas in the field of sample surveys that need our attention. In the area of ecology and conservation, there is considerable current interest in measuring biological diversity. Economic and efficient sample survey designs to estimate diversity need to be developed. Then there is the area of using satellite images to carry out farm surveys to estimate area under a crop and total yield. Physicists at the Space Applications Center Ahemadabad, seem to have done more work on this topic than statisticians. Senior officers in the ministry of agriculture often express frustration that areas under crops such as sugar cane are too fuzzy. We should strive to come up with some smart method of getting precise estimates. So, would it be fair to say that we need the right kind of expertise before others will listen to us? Our recommendation is that university departments need to pay serious attention to teaching of and research in sample survey design (and such other relevant topics). This activity needs to combine theory with practice. Teachers and students have to be in touch with surveys being conducted at local, state and national levels. It is not enough to teach formulas for variances of estimates under different designs. Familiarity with conduct of surveys is also essential.

So, let us say we agree that neglect of sample survey methodology is one reason for reduced interaction between academic statisticians and government. Another such area is demography. There are several issues being debated in India and indeed in the whole world, which are in the field of demography. Let us mention only two. First is the number of missing women. Censuses tell us that the proportion of women is on the decline. This is partly due to selective abortions and partly due to discriminatory treatment of the girl child. What is the best way to estimate these contributions? Another issue of great concern is infant mortality regardless of gender. Lowering of this parameter is now universally accepted as an indicator of development. Good and current estimates for small regions are not available. But again, demography is out of fashion in statistics departments in universities. So, we are not able to contribute to this discussion.

Similar is the story with weather forecasting. It is key information for many walks of life including agriculture, health and defense. Physicists regard the phenomena to be so complex that they willingly give space to statistics if only people would come forward to play a role. Our impression is that university statisticians are either unaware of the opportunities or unwilling to take up the challenge.

Consider the field of insurance. There is agreement that statistics plays a basic role in developing new products in this field. The field has now been opened

up for non-government companies and there are many new entrants. They all need trained manpower in the field of actuarial statistics. But the supply is a trickle. There is no major effort by statistics departments to provide training, even though employment prospects are excellent. A new opportunity that requires combination of insurance and weather prediction is opening up in our country. It concerns crop insurance. Till very recently, the only crop insurance available was to farmers in a 'homogeneous area'. Expected and actual yield of a crop were estimated for the whole area through historical records and actual crop cutting. All this was very time consuming. Now (in the year 2007) a new scheme has been introduced in which compensation for loss of yield is estimated based on rainfall data only. Measurement is therefore easy and quick. In fact the policy specifies that compensation if any is calculated based on official record of rain. Such policies need a good statistical model that relates rain (and other weather parameters) to yield. This work has to be done for different crops for different areas. Will the community of statisticians rise to the occasion?

Talking about employment opportunities, it is necessary to recognize how the scene is changing. In the nineties, it was news if a student was recruited by a well-known company. In the new millennium, it is news if someone is left without a job offer. This is true of Pune, about which we know in detail. It is of course true of Indian Statistical Institute. The case of Indian Agricultural Statistics Research Institute is peculiar in so far as their intake for M.Sc. is only five while the number of scientists in the institute is nearly one hundred. But all their graduates also get excellent employment. Perhaps it is not true everywhere. That only shows the great opportunity the teachers have. They can modify the training given and make their graduates attractive to industry and business. As an example, if students know SAS, even at a beginner's level, their employment prospects brighten. Companies in many fields employ SAS programmers. One field particularly interested in SAS programmers is Pharmacy.

The situation in the field of pharmaceutical research has changed dramatically in India. Since 2005, the patent regime in our country has changed. Our earlier practice of recognizing process patents instead of product patents, is now gone. Hence, multinational companies feel more secure. (Also Indian pharmaceutical companies are going multinational). Consequently, there is a multifold increase in clinical trials conducted in India. Statisticians have a role in clinical trials similar to that of Chartered Accountants in business. Good statistical design and analysis is a pre-requisite to acceptance of claims about new drugs. And all the statistical analysis is done using SAS. So, is it not time to take full advantage of this situation?

Dr N. Unnikrishnan Nair in his Presidential address at the 25th annual conference of Indian Society for Probability and Statistics held in Bangalore made

some interesting remarks. 'Many universities in India insisted on undergraduation in mathematics as an essential qualification for higher studies in statistics, interpreted statistics as a part of mathematics in forming departments. Also the curriculum contained a large percentage of mathematics ... Journals showed preference for articles that contained sophisticated mathematics...(There is an) impression among industry, technology and other sciences that many statisticians have very little to offer to resolve their problems.' We think Dr. Nair has hit the nail on its head!

So, to summarize, statisticians are respected for what they give to the society in terms of immediate payoffs. These come through innovative applied statistics that focuses on crucial social problems. Lack of this focus has hurt the status of statistics as a discipline in India. If this diagnosis is correct, the implications for teachers and researchers are clear.

24. Statistics in India Today – Past Perfect, Future Tense!

Today we are all used to the idea of young Indian techies traveling all across the world providing IT solutions and services. But hardly anyone knows that a similar trend started 50 years ago. What? You ask incredulously. But there were no computers then, let alone PCs. Very true. Those techies were statisticians, helping many nations, some free and others still in colonial bondage, in their attempts at development. How did this come about? To answer such a question, we have to trace a little bit of history.

Statistics is a young discipline, barely a hundred years old. It came to India as Professor P. C. Mahalanobis traveled home from England after the First World War and discovered that a journal named Biometrika of this fledgling discipline had many things potentially useful for India. Perhaps the first Indian to get formal education in this subject was P. V. Sukhatme, who completed his Ph. D. in statistics from the University of London in 1936. He settled down in Delhi as a founder leader of a small statistics group within the Imperial (later, Indian) Council for Agricultural Research. Soon his work became well known and in 1956 he was appointed Director, Statistics Division, Food and Agriculture Organization, UN, Rome. There was a general awareness that progress in agriculture needed research, which in turn needed statistics. But a trained statistician was a rare breed. The new chief knew of one source of manpower. ICAR, Delhi. The rest, as they say, is history.

1933, the year in which young Sukhatme joined the University of London was also the year in which Mahalanobis founded the Indian Statistical Institute (ISI). He launched a series of statistical researches concerning critical social issues of the time such as management of floods, assessment of yields of jute and other crops, anthropometrical analysis of ethnic elements of Indian society. ISI attracted an outstanding pool of talent and soon became (by an act of Parliament) an institute of national importance. Associates of Mahalanobis such as S. N. Roy, R. C. Bose and C. R. Rao dazzled the world by their contributions to theory of statistics. By 1960, India began to be counted in the top league of nations in the field of statistical theory and practice. The league consisted of UK (where modern statistics flowered), USA and Russia.

Today, the discipline of statistics in India can boast of a separate ministry in the central government (Ministry of Statistics and Programme Implementation), a separate arm of bureaucracy (Indian Statistical Service), a world class set up for information gathering called National Sample Survey Organization, several specialized research institutes (ISI, Indian Agricultural Statistics Research Institute, Institute for Research in Medical Statistics), nearly a dozen journals, about 100 educational centers that offer training at Master's and Ph D level and perhaps a

thousand or more colleges with degree programs. In this age of outsourcing, many pharmaceutical industries have established centres for statistical analysis of their data, in Mumbai. Similar centres for business intelligence use statisticians for analytics (stock market, retail sales management, etc.). This is a very creditable list for any country.

And yet, all is not well. Is the statistics community able to face challenges arising from problems of development? Regrettably, the answer is far from a resounding yes. There are many problems. Some of these may be common with other fields of science while some may be peculiar to statistics. Among the common problems, we face low demand for seats and low quality of entrants, small number of Ph D awards, a shortage of researchers and a great variation in the quality of training and output. In addition, one major problem I see is failure to respond to current needs such as in insurance and drug research. Considerable activity in these fields has led to greater demand for statisticians but universities have paid little attention.

Not only has the educational establishment been unresponsive to changes in the business world, there is a lack of response to the changing technology scene as well. We need to develop new statistical tools to analyze satellite imagery as well as medical images such as MRI. To my knowledge only a couple of statisticians work on problems of interpreting such images. Use of micro array chips in genetics and related fields is on the rise. Again very peculiar statistical problems are involved in interpretation of these measurements. In the absence of good statistics, conclusions drawn (e.g. genes selected for further work) may go wrong more often. Again no thrust has been developed in this area. Lack of purpose and focus is apparent in the fact that even in areas of traditional strength, fact that even in areas of traditional strength, our grip is loose. A good example is sample surveys.

Mahalanobis and Sukhatme – two founding fathers of statistics in India – both recognized and emphasized the role of statistical sample surveys. India became well known for achievements in theory and practice of sample surveys. The situation has changed a lot since then. Most experts in the field have retired and the number of new entrants is miniscule. There is hardly any serious research being done in the country on survey methodology. The Ministry of Statistics and Program Implementation is running out of names to be included in advisory committees.

What is the situation on the institutional front? The Indian Statistical Institute, the jewel in the crown, continues to get full support from the Central Government. It is still looked upon as the main source of expertise and advice. But one sometimes wonders if the institution is gradually losing its focus. For the last many years, scientists who are not statisticians hold Directorship of the institute! In fact the position should always be held by an eminent statistician. New academic programs started by the institute are in areas other than statistics. The number of

statistics teachers/scientists in the institute is on the decline. Lastly, a record of the institute working with industry is not satisfactory.

The Indian Agricultural Statistics Research Institute is another apex body. It leads all statistics groups in agricultural universities and ICAR institutes. This entire community is rather insular. They have their own society, their own journal, and their own conferences. They rarely write in other journals or attend other conferences. I suspect that it leads to isolation. The principle 'larger the lake, larger the fish' works against such exclusive clubs.

Research institutes such as IISc, CCMB and NCL need to develop their own statistics groups. There is lot of discussion about drug discovery and drive to get patents. But many CSIR institutes (with some exceptions such as CDRI Lucknow) carry on merrily with hardly any inputs from statisticians. This is in stark contrast with the fact that in the three major medicine markets in the world (USA, Europe and Japan) a new drug approval application without strong statistics component is inconceivable.

However, there are some indications of growth. The number of places offering postgraduate training has gone up gradually and today there may be around one hundred such places. But the number of teachers at a place may be small. Indeed I know some Universities where the number is just one! There is of course great variability in terms of the quality of training given. I have already pointed out the need to keep syllabi abreast of current societal requirements. As far as style of teaching is concerned, it is dominated by formalism, theorems and proofs. Emphasis on independent work, reading of journals, problem solving, giving seminars, etc. is low. There is hardly any connect with real life problems and current practice of statistics. A course on design of experiments involves no experiments and a course on sample survey theory requires no first hand experience of participation in a survey. The so called 'practicals' are anything but that. There is little interaction with users and hence low appreciation of practical difficulties or inadequacies of current theory. The culture of collaborating with colleagues in other disciplines needs to be nurtured among students and teachers. If people do not go to their colleagues next door, going to an outside industrial concern seems even more unlikely. My guess is that most statistics teachers have never seen the inside of a factory as a statistician. Of course one may argue that it takes two to tango. Indian industrial concerns have shown precious little enthusiasm for use of statistical tools to enhance quality or productivity. Their profits came, not from better production but from adroit management of licenses and permits. There was little incentive in a seller's market to improve performance. There is more than a proverbial grain of truth in this argument. But now our frontiers have opened up. There is a lot of international competition. Shortages have disappeared. Some companies face a crunch because cheaper and better imports are available. So, survival instinct forces

such units in distress to look for help. Other companies have become ambitious and want to sell their produce in the world bazaar. But buyers in the West often lay down stringent quality specifications and statistical norms. All this should be music to the ears of a statistician. This turn of the tide in favor of statistics remains to be exploited but gives hope for future.

One area where we see clear growth is publication of statistics journals. One or more journals get published from Bengal, Assam, Uttar Pradesh, Gujarat, Maharashtra, Kerala, etc. Department of Statistics, Kolkata University is home to three journals. This surfeit of numbers has expected consequences. The number of submissions is low as is rejection rate and quality. Some of these publications serve only one purpose namely making many people editors and associate editors. Major consolidation is needed here.

One can only hope that the community of statisticians in India will shed its isolation and sloth, rise to the challenge of the new era and in the process bring success and prosperity to their parent institutions and students. Posterity will judge them harshly otherwise.

25. Clinical Trials

This is a topic that can have a significant impact on all aspects of statistics; education, research, professional opportunities etc. in India. We have touched upon this topic in the fifth essay in section I. We will elaborate upon it here.

Since the nineties, India seems to have broken out of the 'Hindu rate of growth', which hovered around 4% for decades, and now there is talk of even double digit economic growth rates. This is mainly due to performance in the manufacturing and service sectors of the economy. Our blue-eyed boys are the software people. But a quiet revolution is taking place in the health sector also that may have largely gone unnoticed by many including us statisticians.

One feature of the new paradigm is medical tourism. We have some fine hospitals and expert Doctors and Surgeons. Treatment is first class. And all this comes at a cost that is remarkably low. So, getting treated in India makes a lot of economic sense to people in many countries. Health insurance companies in the west are sure to take advantage of this high-quality-low-cost service. It will enhance their profits and competitiveness. Of course the service provider has to have excellent reputation. 'Apollo Hospitals' is one company often mentioned in this context. In our own city of Pune, 'Jehangir Nursing Home' (an affiliate of Apollo Hospitals) and 'Ruby Hall' are two hospitals already in the game. In addition to treatment, there will be a plethora of remote services like diagnostics and expert advice.

There are two other major areas in the field of health that may see great expansion. They are drug manufacturing and drug discovery/testing. Our pharmaceutical companies such as Ranbaxi, Dr. Reddy's Lab., CIPLA and others have already become global players. Earlier their important business was the so-called generics. These are medicines that are out of the patent restrictions. In other words, the medicines have been around long enough that proprietary rights of discoverers are over and the items can be manufactured by anyone. Cost of production of our companies is low and hence their competitiveness is high. In exceptional circumstances, patent rights may be relaxed for humanitarian considerations. You may have heard that some drugs for HIV-AIDS are in this category. Indian companies may manufacture these medicines to make them available to victims in Africa, Asia etc at an affordable price.

For all these years, activity in the area of drug testing was minimal in India. One possible reason was our patent regime. It was different from that in the west. We allowed production of a formulation by people other than the innovator as long as the process of production was different. Hence huge expense for trials did not seem worthwhile. Another effect of the regime was that the MNCs were very cautious in introducing a new drug in India for fear that it may be copied and

hijacked. Now, the law has changed and is more in line with the west. So, there is much greater interest in shifting the work of developing and testing new drugs to India. Under the new law, MNCs seem to feel safe enough to outsource drug-testing work to India. This has opened up new opportunities for statistics.

The opportunity is to participate in conduct and analysis of experiments (called clinical trials) to confirm safety and efficacy of a new drug. Such trials are carefully regulated by concerned government departments. As an example you may want to visit the website www.fda.gov of the U.S. Food and Drug Administration. Statistics plays an important role in this work.

How big is this activity? A few numbers may give some idea of the enormity of work going on in this field of clinical trials. All trials have to be registered before starting (so that we do not have the problem of failed trials being suppressed). The website www.ClinicalTrials.gov lists thousands of trials done since the registry began (August 2008). Time gap between starting a trial and its successful completion and beginning of sale of the drug may be of the order of a decade and the cost in US may be several hundred million dollars. According to one estimate, total annual outlay on clinical trials in the world may be about 40 billion dollars (about 2 lakh crore rupees, or market capitalization of Infosys and TCS together). As you can imagine, this involves many hospitals, doctors and countless patients. Data collected are huge and the work of analysis is also very considerable. About a decade ago, Pfizer, a giant MNC in pharmaceuticals, opened a center in Mumbai to handle part of the work of data management and statistical analysis of clinical trials done by (or commissioned by) them. Novartis, another MNC in the same field has also followed suit. Similarly Glaxo-Smith-Kline and Bristle-Meyer-Squibb have set up work groups. Others may come too. It must have been relatively easy to outsource statistical analysis. The big step is conducting the trial itself in India. This is also happening. Companies farm out part of the work to other companies called CRO (contract research organizations). Many such CROs (newly started as well as old and established ones like 'Quintiles' www.quintiles.com) are now operating in India. According to one estimate total turnover of CROs in India will rise by two orders of magnitude in five years to reach 10,000 crore rupees per year. (Also visit www.informa-ls.com/ct-ceeasia for details of a conference on issues concerning clinical trials to be conducted in India and other countries)

What are the implications for the field of statistics education and research in India?

1. Availability of jobs for holders of M. Sc. (Statistics) degrees. Many pharmaceutical companies and CROs recruit such students. Campus interviews have become a regular feature in Pune and some other universities. Neither teachers nor students have experience in this respect. Teachers should get to know how students of MBA prepare for interviews. Aspects such as group discussion, essay

writing, case studies are quite neglected in traditional teaching of M. Sc. statistics. Some correction is overdue.

2. Holders of M. Phil. and Ph.D. degrees in statistics are also needed. But supply is short. These individuals will become supervisors and group leaders and hence will get better emoluments. We should make students aware of this. Any perception that time spent in research is 'wasted' has to be corrected.

3. Training in commercial software packages for statistical calculations needs to be emphasized. In particular SAS programming is a widely used skill in analysis of clinical trials. Fresh recruits can be trained but those who already know it will have a competitive edge. Hence launching of short programs in statistical computing (SAS and SPSS along with some other packages) would be worthwhile. (Do think about organizing an evening program of two hours per day for four to six weeks. You may find takers in many disciplines including economics and management. The program can easily become self-supporting.)

4. It would be useful to ensure that topics in statistics that are directly relevant to clinical trials are duly emphasized in teaching; for example crossover designs, (group) sequential tests, bio-equivalence testing, multiple testing and issues arising from there. Further, specific attention should be paid to increase skills in report writing. Our students are often weak in written and oral communication, which becomes a handicap.

5. An elective course on statistical methods in clinical trials may be launched. This will require extra effort by faculty. Teacher training programs can be thought of as aid in initial stages. Experienced academicians and scientists from companies can be invited to speak.

6. Small research projects of relevance to clinical trials can be taken up. We will give below an example of the kind of research that may be directly relevant and can be taken up by any statistics group in a college or university.

7. Students of M.Phil. can be asked to review selected recent material on clinical trials in journals like Biometrics.

8. Statisticians employed in such work can be encouraged to work towards a research degree such as M.Phil. and Ph.D. They will bring to the table domain knowledge and clearer view of problems of interest to the industry. Faculty can complement them with expertise in theory.

9. Companies are keenly aware of the important role that universities will play in development of this sector. After all we are talking about knowledge-based activities. Companies are therefore favorably disposed towards any ideas of collaboration. Teachers of statistics should be bold in approaching pharmaceutical companies in the vicinity. Any interaction is bound to be of mutual interest.

10. In allopathic field, while MNCs are clear about need for statistics, many government laboratories lack awareness and skill in statistics. Several laboratories of the CSIR are seriously involved in drug related research. But awareness of statistics is inadequate (to say the least). Expertise in statistics is almost non-existent. Hence extensive statistics training for chemists, biologists and other scientists in these drug research laboratories is essential. Hence everyone should think about exploring the possibility of collaboration with institutions like the Central Drug Research Institute, Lucknow.

A sample project: As stated earlier, clinical trials are extremely expensive. Hence even a small reduction in cost is of interest to industry. There are three prongs of a clinical trial: time needed, cost involved and power to detect efficacy of a drug. So, any research that can reduce time and/or cost or increase power is of interest. {Time is very precious in clinical trials because duration of the trial directly cuts into time available to market the drug under patent protection.} One reason for delay in a trial is non-availability of enough patients. Planners have some idea of how many patients they need and in what period. Very often, hospitals have no idea about the number of patients that can be recruited and they give crude (and sometimes inflated) estimates. If patient recruitment falls short, trial just has to wait. So, we take up a hospital, select a disease (asthma, heart attack, hypertension, cancer of different types, diabetes etc.), study records of patient arrivals and develop a suitable model for patient arrival rates (stochastic process, queue, time series etc). In addition to giving a better estimate of patient arrival, such a model may also throw new light on the epidemiology of the disease.

Breadth of the field: If drug testing in allopathic research needs statistics, the need is equal or more in naturopathy, ayurved, yunani, siddha, homeopathy (abbreviated as AYUSH) and other alternative medicine systems. Awareness of the need for scientific trials is increasing among followers of such alternative systems. But unfortunately, awareness of statistics may be much less. Statistics teachers need to adopt a missionary spirit and to proactively connect with such people. This is a big opportunity if we decide to go out and grab it. It is an ideal situation because we can expect material gains while doing something for social good. We would like to draw your attention to a very relevant guest editorial in Current Science (January 10, 2006) by Prof. M S Valiathan.

26. Innovations in Statistics Teaching

This essay is addressed to teachers and students of statistics in India. On the one hand we see around us lot of prosperity derived from knowledge-based businesses. On the other hand, the standing of a typical college or university, in the minds of citizens, is pathetically low. People, by and large, believe that the only institutions that give worthwhile education are IITs and IIMs. There is endless discussion about reservations in IIT and IIM. But with or without reservations, these institutions will provide only a laughably small number of the army of trained individuals needed to service worldwide demand for Indian expertise. Our knowledge industry has an insatiable appetite for good human resource. It can be satisfied only if AVERAGE quality of education goes up. So, each of us has to worry about improving quality of education in our own little corner. One aspect of this pursuit of quality is 'better teaching'. Hence we need to think continuously about teaching innovations. These could be new topics, new format of examination, new method of teaching etc.

In what follows, we have given a few examples of possible innovations. These are organized into three broad categories namely, new syllabus, practicals and competitions.

I. New syllabus: University Grants Commission of India periodically sets up expert committees to draft model syllabus and we are all invited to consider the document for adoption. There is no compulsion, but many of us think favorably about UGC's recommendations. If you have not done this, it may be useful for you to look into these model syllabi. In addition, we would like to share some of our ideas. Our suggestions below are not necessarily new, but reflect our opinion about what is needed for the society and also for enhancing employment prospects of students.

1. Sample survey theory and practice: This was a standard topic in statistics teaching in India but has been losing its prestige. Not many teachers do research in this area. But the topic is very important for the society. Further, its teaching should not be devoid of practice and hence students need to participate in some surveys. Students and teachers should collaborate with those who conduct surveys. In a reasonably active university campus, at any point of time, there are many surveys in progress, mainly in social science. Those who plan and conduct them are usually grateful for any assistance. Our advice is, take the initiative and join that group. You will certainly learn some new things.

2. Statistical ecology: This is our favorite topic. Many issues such as irrigation dams and displacement, forest resources and their rational use, conservation of wildlife etc are often discussed in media. It shows that society's interest in

ecological policies is quite high. Hence a course on statistical ecology is of current relevance. It can be offered at undergraduate level and/or post graduate level. It is now taught at both levels in Pune. Books and teaching material are also available.

3. **Insurance** – Our government has opened this sector to commercial organizations and there are many new entrants now. They all need people with statistical skills and training in insurance related work. Opportunities are huge and supply of personnel limited. We should train our students so that they can appear for examinations of professional bodies like Institute of Actuaries. To become 'Associate of the Institute of Actuaries' one has to pass 9 papers roughly divided among mathematics, statistics and economics. To pursue this line, students have to learn some economics. Teachers can help students by introducing a course on mathematical economics/econometrics and also allowing them to take some courses in economics as a part of the syllabus of statistics.

4. Genetics-A basic course in statistical genetics can be taught at undergraduate level without difficulty. A new area emerging in this field is analysis of the so-called micro-array data. These data are obtained from experiments based on micro-array chips. They enable a scientist to study response of thousands of genes to stimuli of interest. These experiments have many applications including drug discovery. Scientists need statistician's assistance to interpret the data so generated. People with such skill will find opportunities in research institutes, pharmaceutical companies etc in all parts of the world. Again, a beginning has been made in Pune.

5. **Statistics of remotely sensed data**- Satellites have become a source of vast amount of information about surface of earth. Interpretation of this information is based on statistical models of reality. Users generally depend upon standard commercial packages without knowing the basis of conclusions drawn. Statisticians with the necessary skill are not available in our country. Universities will have to tie up with relevant organizations such as 'Indian Institute of Remote Sensing', Dehra Dun and 'National Remote Sensing Agency', Hyderabad. A related area is medical imaging.

6. Clinical trials- The patent regime in India has changed from process patent to product patent. For this and other reasons, activity in the area of clinical trials and their analysis has increased considerably. This includes pharmaceutical companies and also the so called CRO (contract research organizations) Their business in the field of clinical trials is expected to reach stratospheric level of Rs.10,000 crores per year. This has opened up new avenue of employment for statistics students. A course on statistical methods used in clinical trials will therefore be useful. Again, such a course is taught at Pune.

7. R Language: In years gone by, there used to be courses on fortran, cobol, C++ etc. Instead, it would be very useful to have courses on the comprehensive,

popular, open source software package R down loadable from www.r-project.org Ready programs in R can be used to carry out practicals.

We should all think about these and other possibilities to offer new courses.

II. Practicals: Here we have a legacy of the pre-computer era when computational skill and patience had to be learned by doing. There was a lot of drudgery and only very small data sets could be analyzed. All these things have become obsolete. With modern PCs our computational capacity has gone sky high. We can take large real life data sets for exercise. But teachers may not know where to locate such data sets. One international site for data is www.statsci.org . We have prepared a CD with over 100 real life data sets together with brief descriptions and suggestions for analysis. These are free to download. And can be found at http://stats.unipune.ernet.in/Databook/ or http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html.

In addition to taking up 'real' data, we need to incorporate an element of decision-making in these exercises. Students seriously lack skills in comprehending a real life problem and selecting a strategy of analysis. So, in a practical exercise, we must not tell the student which tool to use. He/she must read the description of the situation and decide: [Americans call these 'word problems']. Lastly, we do not ask students to write any comments/discussion on the results. In fact hardly any attention is paid to report writing. This must change. Someone employed as a statistician has to explain the findings to superiors who may or may not be familiar with statistical methods. Perhaps the most suitable format for such work is projects. We should reserve a part (say 50%) of the time allotted to practicals, for doing a project. It can be done in teams. We should not make too much ado about difficulties in giving marks to individuals when work is done jointly. Such assessment is done routinely in professional courses. In Pune University, the Board of Studies has now approved this pattern for T.Y.B.Sc. In fact, a competition is held annually to select the best projects done. We will give one example of a project that received a prize in such a competition.

The problem selected concerned quality of food in hostels. There was a feeling that mess contractor perhaps buys the cheapest vegetables available in the market on any day and that may be one cause of student dissatisfaction. So, the aim was to test the claim that the vegetable served in the mess is the cheapest available on that day. Students gathered data on vegetable prices in local market (and had difficulties in doing this). They were surprised that the claim was not justified.

This has all the elements of an ideal project. A real life problem, some data collection, some analysis and a conclusion that is unexpected. Students can learn a lot from such projects.

Here is a short list of some other interesting projects done by students of TYBSc or MSc in Pune University. We have only given titles as provided by students. That is at best only marginally helpful. We are very aware of this limitation. But the whole idea is to make students think independently and the list gives some indication of the theme. How to delineate the boundaries of the project, where to look for data, what analysis to do and how to report the findings are all matters to be resolved by the students planning to write a project report. Teachers can be approached for advice, but major responsibility rests on the students' shoulders.

1) District wise production of different crops over a span of twelve years.
2) Statistical study of drip irrigation
3) A Study of dry spell during rainy Season at Jalgaon (1946-1964)
4) Modelling relationship of weather with wheat crop germination
5) Comparison of Two Wheat Varieties (Parabhani Station.1944-1964)
6) Modeling Growth (Height) of Wheat plant (Parabhani station 1944-73)
7) A Statistical analysis of Rainfall in Jalgaon
8) Analysis of wind velocity of Dharvad city.
9) Study of Rainfall
10) Air pollution in Pune city
11) Do parents Influence their Children's Educational Decisions?
12) An enquiry into educational status of primary school children in Jawhar Taluka (a tribal area)
13) A Study of Municipal transport (PMT) Service on Selected Routes
14) Life Style of People in age group 35-60 Years with special reference to health consciousness.
15) Health Study of Various Age Groups
16) Call-Center.
17) Comparative Study of Four Wheelers.
18) Expenditure on petrol in use of two wheelers
19) Analysis of market survey of wristwatches.
20) Statistical Analysis of mode of payment & Occupancy of rooms in Hotel Swaroop
21) Statistical Analysis of Crime Data of Pune City
22) Analysis of Life Insurance Policies Using Statistical Tools
23) Statistical Survey of Library Services
24) Use of Library by Teachers
25) Statistical analysis of ATM

26) Sample Survey Of Telephone Users With Special Reference to Mobile Holders
27) A Study of Exports of Garware-Wall Ropes Limited
28) Statistical Survey of Household Computer Users
29) Process of manufacturing 'Half Bearings' at Mahila Udyog Ltd.
30) Modeling and forecasting share prices of Bajaj Auto Ltd.
31) Single sampling plan for Unitech Industries.
32) Sample survey on the lifestyle of college youth.
33) Ideal Restaurant
34) Study Of Beauty Parlors
35) A Statistical Study of News Channel Viewers
36) Distribution of number of girls of two children families.
37) Blood glucose levels.
38) Accident proneness among children.
39) Birth conditions in Pune city
40) A Statistical Survey About Exercise
41) Old Age Problems
42) Effect of Music On Mental Stress
43) A Project Report on Means of Recreation
44) Credit Cards-A necessity or Status Symbol
45) Analysis Of T.V. Viewing Habits Of People
46) A Project Report on Reading Newspapers
47) Effect of Advertisements on People
48) Statistical Study of "Astrological Influence on Human Life"
49) Analysis of Banking Sector
50) Analysis of Blood Bank Using Statistical Tools
51) New Born Babies in Hospital
52) A Project Report on GYM Users
53) ROTA analysis for MICO, Nashik
54) Female Age at Marriage
55) A Study of vegetables in College Mess (How good how bad!)
56) Travel Time
57) Study of status of a women in Nashik
58) Study of internet users in Pune city
59) Vehicle preference among Pune residents
60) Sample survey of beauty consciousness in Pune
61) Service time distribution at Pune railway station
62) Is your mind a suitable source for random numbers?

We are convinced from the experience so far that project assignments are feasible in undergraduate teaching. Projects are quite essential in post graduate courses. They will sharpen the analytic skills in students. This will improve their performance in job interviews.

Presentations: There must be occasions when students have to give presentations. They can be part of an examination or competition or social functions. Use of modern audio-visual aids is part of the skill set needed in the business world. Language skills get tested and also problems of stage fright are reduced if students are made to present their work.

Industrial assignments: If we regard a statistics degree as a professional course, a component of practical training is essential. It could be in industry or business or government office or research institute. Working, even if only for a brief period, in a non-academic environment broadens the perspective of a student. This can be done as a vocational non-credit assignment. Such a component involves substantial effort by teachers. They must approach various organizations for summer placement of students. It is not as hard as it may sound. Management schools do this routinely. Current scenario in our society is quite favorable to such activity, provided of course teachers get actively involved. Students at Pune University have done projects in (a) large private sector companies like Bajaj Auto (b) Government organizations like ordnance factories and Institute for Armament Technology and (c) small companies that are suppliers of parts to large ones.

Innovations suitable for a specific college or university may have to be devised to implement such a program. Perhaps a local hospital will accept a few students or a body such as agricultural produce marketing committee may have some interesting data to analyze (one college teacher here did a project on prediction of onion prices). No one formula will suit all. The only commonality is interaction between statistics and society to facilitate training of students (and teachers).

III. Competitions: We feel strongly that examinations should not be allowed to dictate the content and style of a teaching program. A good well-rounded learning experience will shape the personality of a student much better than rote learning for conventional examinations. Competitions can be used to promote balanced development of students. We briefly describe the kind of competitions held for statistics students in Pune University during the last few years.

First Year: The number of students at this level is large and any elaborate competition is not manageable. Hence a simple quiz lasting one hour is administered. All questions are on elementary statistics and are 'multiple choice' type. A token fee is charged for participation. Questions are designed to test understanding and analytic ability instead of memory. This can be real fun for

students. A few top scorers are invited to attend competitions for SY and TY to see the performance of senior students

The test is followed by a general lecture on some aspect of statistics application. Teachers from one college visit another college and give this lecture. As a result, interaction between teachers in different colleges has gone up considerably.

Second Year: Here also a test with 'multiple choice' questions is given to all interested students. Top scorers are selected for a second round that involves multiple activities. These include essay writing, group discussion, giving a short prepared lecture on topics of general interest, reading a description of a real life situation and deciding the right kind of statistical analysis, poetry appreciation etc. This round involves considerable preparation by participants. We found that even parents like to attend this round to see how their wards perform. We invited about 25 students for this round and found that it lasts several hours. These items also prepare the students for typical job selection routines.

Third (final) Year: We have already described the competition of project presentations. In addition, we organized annually a leadership development camp of 3-4 days. Lectures of statisticians and eminent people in different walks of life were arranged. Lectures on improving ability of spoken English and also communication skills were very popular. Teachers too benefited from the event.

M.Sc: For the last few years, Pune University has organized an inter-university competition under the umbrella of Indian Society for Probability and Statistics (ISPS). Financial support is given by International Indian Statisticians Association (IISA). Here too, students put in a lot of efforts and benefit from participation regardless of the result of the competition. Since 2008 the activity has been taken up by the Department of Statistics, Shivaji University. A brief description of the competition follows.

A Project topic is announced and letters confirming intention to participate are invited. Competing team is a group of M.Sc. Year I students. The group works on the project and submits a report (say about 10 typed pages.) to the organizers. A team of reviewers scrutinizes submissions and 3-4 top reports are short-listed. Two students from each of these four teams are invited to attend the next annual ISPS conference. One 2-hour session is organized and each team is given 30 minutes to present their work. Referees then rank the teams and suitable cash prizes are given. Here are the project topics of previous years:

1. Comparison of performance of state and private passenger bus transport services, from the viewpoint of users.

2. In my city- who reads which newspaper and why?

3. How I used 'Design of experiments' to improve a traditional cooking recipe.

4. Measurement of fuel consumption of my two-wheeler.

5. Predicting height of the tallest Indian male in 2020.

6. Preparing a test report and help file for a public domain software package in statistics

7.Which newspaper analyst predicts stock prices better?

We will end with description of some activities organized in colleges in the city of Nagpur.

Stat-storm: A group of college teachers in Nagpur conduct a statistics quiz for college students. They have named it Stat-storm. The effect of this storm on students is astonishing. They enjoy it and want more of the same and want to take up statistics as a subject of further study etc. This team has conducted quiz competitions in many cities. We have prepared a video on stat-storm and also presentations by MSc students as a part of the project competition in ISPS annual conference. This CD is available for showing to students.

Stat Maze: This is a cross word puzzle in which all words are from statistics.

Stat Smart: This is a quick and dirty competition in which a topic is given and each competing group has to collect data, analyze it and prepare a report in very short time (say 2 hours).

These are some ideas. We are sure other teachers will come up with different and better ideas. The key is to take a proactive stance and go beyond the current culture of just completing the syllabus and preparing for examinations.

27. 100 DATA SETS FOR STATISTICS EDUCATION

This is an essay on problems of quality in Statistics education in India. We spell out our view of the flaws in current practice and propose that creative analysis of interesting, real life (preferably Indian) data sets is one possible step to improve quality. We describe a "data book" prepared by us to help teachers who may wish to practice the ideas presented here. The 'book' is available for free download from following addresses or alternatively, a CD can also be made available for a nominal charge.

(i)http://stats.unipune.ernet.in/Databook/

(ii)http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html

Background: Statistics is a relatively young discipline and most of the modern developments in the subject have occurred in the 20th century. Research in statistics began in India quite early, 1933 being the year in which Indian Statistical Institute (ISI) was founded. Work done in ISI was superb and within a couple of decades India became a frontline country in research and application of statistics, comparable in stature with UK, USA and USSR. Statistics research carried out in India was a good blend of theoretical and applied work.

This heritage naturally gets reflected in education of statistics, or that is what one expects. Reality however, is somewhat different. Purpose of this article is to bring out the divergence between a historic perspective of statistics and actual teaching of the subject in India today. We will conclude with some suggestions for change and offer help in that direction.

Statistics education in India: This can be divided in three groups. First group is 10+2 i.e. high school and junior college education. Second group is college and university students taking one or two courses of statistics as a supplement to the principal area of study being pursued. Lastly, we have students who take up statistics as their principal subject for a bachelor or master's degree program.

At the 10+2 level, time devoted to statistics is quite limited and as a consequence, coverage is also minimal. This is understandable since there are many subjects and topics competing for attention. The material on statistics is included under the broader umbrella of mathematics. The contents include summarizing of data and some elementary calculations of probability. Students get to know terms like mean, median, mode and variance. Perhaps because of such exposure, many people equate the subject of statistics with the troika of mean-median-mode. A few may even remember relevant formulas. What no one remembers or recalls is learning something new about the data set for which mean etc are calculated. What

did we discover about the data? Was there anything surprising/ unexpected? Such questions are not even asked. This need not be so.

There are many books on elementary statistics which show, without fancy mathematics, how statistics can play a very active role in day-to-day life. These may not be conventional texts but certainly are excellent introductions to the subject. We will mention two of them. One is 'How to lie with statistics' by Huff, Darrell(1954) and the other is 'Say it with figures' by Zeisel, Hans (1985) .

On the subject of averages, Huff points out that one has to be careful in interpreting them. 'Otherwise you are as blind as a man choosing a camp site from a report of mean temperature alone'. Oklahoma City has an average temperature of 60.2^{0} F which seems very pleasant but hides the fact that it can get as cold as -17^{0} F and as hot as 113^{0} F. So the important message is that average values can misguide unless accompanied by some indication of variability. This message is completely lost if the focus of attention is all the formulas for grouped and ungrouped data etc. Hence in our classes, either the students are lost in the maze of rules or even if they survive, acquire no skill to put the knowledge to use.

Zeisel does a wonderful job with simple cross tables. In a data set of 14030 car drivers, 62% have never had any accident while the remaining 38% had at least one accident while driving. What factors may characterize accident free drivers, he asks. One possible factor is gender and a cross table answers the point.

	Men %	**Women %**
Never had an accident while driving	56	68
Had at least one accident while driving	44	32
Total	100	100
Number of cases	7080	6950

It appears that women are safer drivers than men. Zeisel then recommends that before closing the case we should consider introducing additional factors, which may modify the conclusion or at least give some warning. He proceeds to include a factor namely 'amount of driving'.

	Men		**Women**	
Distance drove (000 miles /year)	>10	≤ 10	> 10	≤ 10
% of persons in that group with at least one accident	52	25	52	25
Number of persons	5010	2070	1915	5035

He introduces two categories, those who drive more than 10k miles annually and those who drive less. Now we can compare genders within each of these sub-categories. The result is surprising. It appears that there is hardly any difference between genders. The earlier apparent difference is attributable to the fact that women generally tend to drive less than men.

Now this involved no boring calculations or slippery concepts and the lesson is clear as a day light. This can easily be taught to high school students. It is meaningful and useful. Why can we not use and write such books? Why do we make the subject mechanical and devoid of human interest? We will get back to this issue later.

There is ample opportunity to touch a curious mind while teaching statistics. Consider the following question: When do you call someone in your class very fat (or tall or skinny or smart)? A simpleminded answer is the person with maximum weight. Some discussion may persuade students to use not the absolute weight but weight/height ratio. The real difficulty, however, lies in the fact that there is always a maximum value in any group, but the person need not be 'fat'. A fancy answer is to use some 'standard'. But that only begs the question of how to decide the standard. This is often not clear. Sometimes, a readily available standard developed for a different society is used. In any case the logic of a standard seems unclear. In fact, elementary statistics offers an easy and elegant way out. We can generate a yardstick from within the data themselves. [This is indeed the main strength of statistical methods. Standards are developed anew for each data set. They reflect the variability within the data.] A graphical tool that can get a quick answer is the so called 'box plot'. It identifies outliers in the data set. A person should be called fat if the value for this person is an outlier on the higher side within the reference group. Let us not mystify the matter. By current convention, a value that is far above the 3^{rd} quartile (75^{th} percentile) is regarded as an outlier (more than 1.5 times the inter-quartile range).

Our contention is that if the subject is to be attractive and study is to be rewarding, students must get involved in situations and data that relate to their daily life. It is not the mechanics but the emerging message that needs to be emphasized. Here is the opinion of an expert group in UK: introductory teaching should 'not have a heavy emphasis on theory nor should it emphasize computational methods but should make basic concepts clear and show how these are used in the interpretation of experimental and observational data. Children should become aware of and appreciate the role of statistics in society; i.e. they should know about the many and varied fields in which statistical ideas are used including the place of statistical thinking in other academic subjects.... They should know the sort of questions that an intelligent use of statistics can answer, and understand the power and limitations of statistical thought' Holmes P. (2003).

This idea seems equally applicable to the second level of teaching, namely service courses for students of other disciplines. Why are students from disciplines like biology or social science required to take a course in statistical methods? Because model syllabi given by the University Grants Commission include such a course! What decides the topics that go into it? Very often the detailed syllabus is nothing but a list of chapter headings from some book with the title 'Statistics for ...'. How does one decide relative importance of various topics? One doesn't bother! In fact it is an easy matter. Things that are needed more often are more important. But if we apply this criterion, perhaps nothing is important. For, in most cases no teacher uses statistics in his/her research (if there is ongoing research activity!). This leads to some very warped situations. No one in his/her right mind memorizes all formulae in statistics. Yet students are expected to do that. Why? We do not have the system of open book tests. In fact now with availability of software packages, the ability to use the formulas for realistically large data sets is also not crucial. It suffices if students can apply the formulas to tiny data sets to demonstrate that they understand what the formulae mean. So giving an examination in which students have to rush through computations on a big data set is also unnecessary. In real life they have all the time in the world and will very likely use a computer and a software package. Then what is crucial? Diagnostics, decision-making and interpretation! What is the nature of the data or the problem? How to take some exploratory steps to arrive at good understanding of the problem? How to select a suitable statistical tool? These are matters in which it is hard to find help. Our courses mostly remain clueless in this regard. If teachers in the domain of application are not practitioners of statistics, they are helpless. What about teachers of statistics? They are not of much help either. Why? Because often they are not practitioners either! It is our estimate that more than 95 % of college and university teachers of statistics in India, go through their entire professional life without a single occasion to put statistics into practice. It is a vicious circle. Teachers from fields like biology and social science do not use statistics and hence do not call upon teachers of statistics for help. Since no one expects help from statistics teachers, they continue to reproduce material from textbooks on black boards. Hence their skills as consultants remain infantile {untested}. Since their discourses are not hampered by their lack of familiarity with domains of application, they never wonder how a method is going to be applied in a particular situation.

The last level of statistics teaching is for students who specialize in statistics at B.Sc. and M. Sc. Here too, it is possible to exist without any meaningful application of the tools of trade. At this level also our education is passive. Here is a triplet of sayings that describes limitations of passive learning. 'I hear and I forget. I see and I remember. I do and I understand.' We will add a fourth item or perhaps

modify the third. 'I slip, I fall, I get a bloody nose and then the lesson is permanent'.

Relating Statistics to other subjects: Subject of statistics has two faces, 'Statistics as information' and 'statistics as inference'. Admittedly statistics as a scientific methodology (grammar of science) refers to the second aspect. But total neglect of the first aspect, as is happening in Indian colleges and universities today, is detrimental to statistics education. Firstly, some pieces of information should be known to all students, those of statistics or otherwise. This is just for becoming a good citizen. These include per capita income, level of poverty, shortage or surplus of food, size of population, sex and age composition etc. Secondly, some knowledge of the domain in which to apply statistics is essential to understand statistical methods. We teach theory of sample surveys without any participation in a survey. Students learn concepts of design of experiments without designing a single experiment. A lot of time is spent on regression analysis including interpretation of residuals. The related 'practicals' include the mechanical steps but fail to give a flavor of decision-making since neither students nor teachers are 'involved' in the data being analyzed.

Even from a purely pragmatic viewpoint of enhancing employability, paying attention to statistics as information can make a big difference. Consider the Reserve Bank of India. This huge organization has a separate division for statistical analysis and many officers are recruited at master's and Ph.D. level in the field of statistics. Here, a competent statistician without any knowledge of economics may find herself/himself at a disadvantage compared to an economist with some statistical training. The statistician often has no idea about the functions of RBI and role of statistics in it. It is not widely known in the academic statistics community in India that many Nobel prizes in economics have gone to economic statisticians such as Simon Kuznets, Wassily Leontif and Laurence Klein. The neglect of this subject is going on for so long that econometrics has almost disappeared as a statistics specialization in Indian Universities. This is because hardly any professors of statistics have any serious interest in economics. And interface between statistics and economics cannot be taught without some familiarity with basics of economics.

Many years ago at the Inter-science/pre-professional year, everyone studying mathematics (the so called A group) at Universities in Maharashtra, also had to study economics and English literature. Further in University of Bombay, for many years, students taking statistics at B.Sc. were required to study economics as the so-called 'subsidiary' subject. In addition, topics like econometrics and psychometrics were part of the statistics syllabus. This made a lasting impression on students. It is not essential to study these very subjects or topics to complement study of statistics. Any discipline, which benefits a lot from application of statistical

methods, will provide a good background. These subjects include botany, zoology, microbiology, geology, atmospheric science, anthropology, education etc. To our knowledge, combining study of such subjects with statistics (in college) is virtually impossible in India. Why is that so? A college that offers statistics as a subject usually does have degree programs in one or more of the other subjects mentioned above. So why should a combination not be possible?

The difficulties are of many kinds, mostly of our own making and mostly ludicrous. First is rigidity of structure. Statistics is usually classified as a 'science' subject and economics and psychology are classified as 'arts' subjects. And it is heretical to want to take arts subjects if you are in science. Our structure often does not allow taking up topics across 'faculties'. This can be easily remedied at the university level by listing economics and psychology under arts as well as science, just as mathematics, statistics and geography are. This is rarely done. In fact we know only one example. It is that of North Maharashtra University at Jalgaon in Maharashtra. It's first Vice-Chancellor, Dr. N. K. Thakare, a mathematician, ensured that all necessary resolutions are passed in appropriate bodies. [Sadly, however, not a single college in that University took advantage of it]. The second difficulty is the divide within 'science'. Mathematical sciences and biosciences allow no truck with each other. Our biology students are required to (and are happy to) keep away from mathematics and vice versa. Since statistics falls in the mathematics group, it is unimaginable to combine statistics with biology. This is totally contrary to modern trends and also antithetical to the spirit of interdisciplinary research. The current trend is to acknowledge that societal problems do not recognize artificial partitions created by academicians. You have to use concepts, methods and expertise in many disciplines to solve them. But we seem to pay only lip service to the notion of interdisciplinary study. Third difficulty is administrative. Our colleges are overloaded with students and have inadequate laboratory and class room space. Hence scheduling classes and practicals is a difficult task. An easy way out is to rule out unconventional subject combinations for study. So, we find most Principals of colleges following the path of least resistance. The last reason is deep rooted and perhaps the most difficult to rectify. Our universities are autonomous to a considerable extent as far as syllabi are concerned. So 'Board of Studies' designs statistics syllabi. College teachers often dominate this committee. In many universities, statistics teachers have ensured that students taking statistics as a principal subject take only statistics papers and nothing else. So the backdoor entry of other topics such as econometrics or psychometrics is also not possible. Why so? Sometimes this is due to lack of awareness. But often it is deliberate. The purpose is very down to earth. It increases the 'work load' of statistics teachers and hence leads to more appointments in statistics. An outright parochial trade union attitude! We find the attitude

outrageous since it puts academic interests of students on the back burner. All this ends up with students taking physics and/or chemistry as subsidiary subjects along with statistics. Physics may be a very exciting subject in its own right. But it is not one in which statistics is used routinely.

A possible solution: Of late college recruitment scene has changed dramatically. Many state governments are in a dire financial situation. With an empty treasury, they are very reluctant to make new appointments in colleges. In fact vacancies created by retirements often remain unfilled. Programs are being rolled back. There is almost a sense of panic among college teachers. So, those tactics for expanding statistics pool are not workable any more. But that does not mean change in syllabi and systems. Under the circumstances, a pragmatic approach would be to appeal to those interested in enhancing quality in statistics education to adopt new ways and to offer tools for the same.

We offer here a 'book' [freely available at the link given here http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html] with over hundred data sets to be used in statistics teaching. In case of any difficulty in downloading the free material, we invite the interested reader to get in touch with us for a CD.

We emphasize again that this study tool will contribute something only if teachers participate actively in the process of learning through discovery. At present this element is virtually absent in education, which has become a process of rote and mechanical 'learning'. Why should it be so? The reasons lie at least in part in our examination system. There is nothing wrong in trying to evaluate learning. But ours is a case of tail wagging the dog. All attention is focused on examination. Students only want to know what is important and what is very important (i.e. quite likely to be asked as a question in the final examination) and they limit their effort to memorizing answers to these potential questions. This is the reason for the popularity of 'guides' over scholarly texts. An extreme situation is one in which students of literature would not even read the novel meant to be studied, but stick to only a compendium of questions and answers about the novel. Even senior teachers feel weighed down by the examination system. A proposal for a change in syllabus may be ruled out because it is difficult to set questions on the new topics. This approach ignores the crucial fact that education is a total experience. Lectures, discussions, practicals, fieldwork, tests, co-curricular activities, hobbies, they all contribute to the shaping of the young mind. Once we make examination the sole focus to the exclusion of all other aspects, academic life loses much of its meaning. Why attend classes? Just read the guide. So you find more students outside the lecture halls than inside. Why study the whole syllabus? Only half of the questions in the paper need to be attempted. So skip study of half the topics. Why solve

problems? Restating the theory will suffice to answer half of each question. This is a downhill path with zero attention to quality in education. We need to move away from this course. Our contention is that correct use of data sets can be helpful in restoring, to some extent, quality in statistics education.

A book with data sets is not a new idea in statistics. One such book is already available by Hand et al (1994) ('A handbook of small data sets'). Another book which gives many data sets and develops statistical methodology around them is 'The Statistical Sleuth: A course in methods of data analysis' by Ramsey F.L. and Schafer D.W. (1997). Title of this book reflects an approach to statistical analysis that we endorse fully. Aim of analysis is better understanding of a situation and finally discovery of that lynchpin which makes clear the whole story. Another analogy is a logjam. When a lot of stuff flows down a river, occasionally it creates a block or jam. Shifting one piece in a critical place restores normal flow. To discover such a piece is the climax in analysis. All kinds of forces have to be mustered towards that one end.

If books exist in multiplicity, why write one more? Our reasons are the usual ones and perhaps well known. First reason is access. The above books are expensive and hard to get for a typical college and certainly for an individual student/teacher in India. We bring the material within easy reach of any net surfer. Second reason is familiarity with the background. If problems are based on Indian reality, students are more likely to relate to them. If we wish to study time series of share prices, it is better not to use the example of Intel Corporation. Instead, Infosys Technologies may prove to be more useful. A student may encounter a news item about this company or may see a TV program connected with it. Or some one around may be familiar with history of the company and may know the reason behind some dip (or surge) in price. All these bits constitute important inputs in analysis and interpretation and learning. If such information is not readily available, it is possible to ask students to search for it and there is a reasonable chance that it can be done. So, a majority of data sets here are connected with the Indian scene. Third reason is that we have tried to highlight interactive research in which we have been involved.

How do we expect readers to use the data sets? We offer a brief description of the problem and its background. Some discussion on these features (and perhaps even some reading or internet surfing) among teachers and students will be useful. This should lead to a precise and meaningful formulation of the problem and hence a smart choice of the tools to be used. A plan of statistical analysis should then be evolved. It implies discussion of available tools and their suitability, assumptions implied in different methods and their applicability. We suspect that some interaction with teachers in the domain of application will be not only useful but even necessary. That is a good beginning of a genuinely interdisciplinary study. A

short report should be written about the whole endeavor once it is decided that work on a data set is over. The report should be intelligible to the person from domain of application consulted during discussions.

Many of the data sets are suitable for undergraduate students. In some cases the statistical methods appropriate are generally studied only in M.Sc. courses. Those can be used to give students a glimpse of what lies ahead. Often people lack any comprehension of the boundaries of knowledge and hence do not know what constitutes research. The tendency is then to assume that as you go up the ladder, it is more of the same. We can alleviate such misunderstanding by showing problems for which answers do not exist in text books at a lower level.

We have tried to include problems, which will require the whole range of statistical methods available. We have also attempted to include examples from many different domains of application. However, there is a dominance of examples from the area in which we are active, namely biostatistics (and in particular, ecology). But this is perhaps inevitable. It would be easy to launch a discussion group and expand the collection using expertise of others who are willing to help. We invite interested readers to send us more data sets (in our format, i.e. data in EXCEL and description in MS word) for possible inclusion in the 'data book'.

Happy learning!

28. e-learning

This last section of the book is devoted to statistics education. Perhaps it is relevant here to discuss various options available in terms of technology of education. Traditionally, there was only one standard way of teaching. It was chalk and talk. There was of course the microphone, if the class was too large. But that was rare. Then came overhead projection. So teachers used hand-written transparencies in lectures and seminars. Transparencies could be prepared in advance, could be made colorful and artistic and they worked for large audiences as well. Next, with personal computers everywhere, their use in teaching became popular. With ability to project material from computer displays to wall size screens, presentations using power point and other programs became ever more 'real' and artistic. All this helped teachers and students.

A major limitation of these direct methods was that teacher and taught had to be in the same place. This was unsuitable for a large group of potential students who were already working and/or residing in a place far from where teaching was done. So, the idea of distance learning came up. Initially, teachers would send by mail lecture notes and other learning materials. Students would work with material mailed and also write back queries. This system was formalized through starting of 'open universities'. In India we have Indira Gandhi National Open University based in Delhi. In Maharashtra we have the YeshwantRao Chavan Maharashtra Open University (YCMOU) located in Nashik and many other correspondence courses. Programs of such organizations often have two parts, one that is at a 'distance' and the other that is face to face. In the latter part, students and teachers meet occasionally by appointment for a brief period to sort out difficulties. To our knowledge, these universities do not offer any major programs of teaching statistics. We know more about affairs at YCMOU since we are directly involved in planning some activities there. While there are no students enrolled for a degree in statistics in YCMOU, there are hundreds of students registered for M. Phil in various social science subjects. They have to write dissertations often based on some empirical work. Hence many of them will have to use statistical tools. However, skill level in statistics is meager and hence there is a significant need for counseling. So, perhaps a panel of experts can be constituted, members of which can receive consulting requests from students and can respond by email. New technology has made it possible to bring experts in touch with students in this manner. It also shows that statisticians can play a significant role in higher education in traditional as well as non-traditional forms of organizations.

Our main purpose in this essay is to discuss briefly the use of internet as the latest technological innovation in education and its application to statistics education in particular. Internet cuts most barriers of space and time. Teacher can be where she/he is and students can be where they are. Each one works when

convenient. And yet a lively interaction becomes possible. All three of us have participated in such teaching activity. This is how it worked for us. We became instructors/tutors for courses run on the website www.statistics.com. This program is conducted by a small company based in Washington DC, USA. It runs over 50 different courses on different topics in statistics. The levels of courses range from elementary to highly specialized ones. Students enroll and get access to the chosen course. Instructor places lecture notes on the website. A specified text book is used. Weekly lessons are also posted that specify which parts of the text book to be read and which problems to be solved and so on. There is no physical class room. But there is what is called, a white board. Students post comments and questions on it. Teacher posts responses. Sometimes other students in the class may also offer answers. So, a fairly intimate interaction develops among the class participants. Of course one can choose to be silent. But for getting maximum benefit one has to be proactive. There is home work assignment and grading. Students submit their solutions of assignments. These are graded. This is where tutors come in. The above program has a back office in Pune (the company is known as TechKnit IT enabled services Pvt. Ltd.) statisticians in this back office do the work of checking the solutions submitted by students. Many teachers and students of the Departments of Statistics at Pune University and Shivaji University have participated in the activities described here.

Beneficiaries of this form of training in statistics are mostly from North America and Western Europe. This need not be so. India is today a leading provider of ITES. (IT enabled services). We have a very wide network of internet café's. So the physical infrastructure needed is available. It means we can provide e-learning without any glitch.

29. Book writing

Books are the principal tools of the profession of education. We often take their availability for granted. This was not always so. {As an aside, before the advent of printing, books had to be copied laboriously by hand and were high value items. In the middle ages, one of the duties of church priests was to copy bible and other religious texts. Only a few decades back, in India in the days before Xerox machines, if one could not afford to buy a book the only other option was making a copy by hand. At Pune we cherish one such copy in our library, made by Late Prof. M. N. Vartak during his student days). In fact there were few books available on statistics in the fifties. Even in the sixties, the number of books on statistics could be counted on fingertips. It is only in recent times that there are many publishing houses each with a long list of titles in theoretical and applied statistics. However, we in India face a problem. Most books on statistics (in English) are produced in the west and tend to cost a fortune. Hence they are beyond reach of typical students. If each student has one copy of a textbook for a course in his/her possession, quality of teaching/learning can improve significantly. Instead of spending time in writing and copying standard results and proofs, it would be possible to spend time on thinking, solving problems etc.

So, it is our opinion that availability of good and inexpensive books is crucial for improving quality of education in statistics (and many other fields). The only way to ensure such availability is for Indian statisticians to write and publish books in India. Publishers will invest money if they believe that items published will sell. How do we assess what is available and what is needed? By the direct route of asking teachers! So, teachers can use their vast experience and make two lists. List one should contain names of topics and corresponding books in use and judged to be satisfactory. Second list should contain names of topics that are taught (or ought to be taught) but no textbook written by an Indian/foreign author and published in India at a modest price is available. If there are many topics for which books need to be published in India, some action can be contemplated. Many possibilities exist. We can make publishers aware of the situation and they can activate their channels to seek out willing authors. Or we can appeal to members of the community of teachers, who can then take the matter in their hands, form groups and prepare an outline and approach a publisher.

While on the subject of books, it would be nice to think of the range of books on statistics. While everyone would have her/his idiosyncratic categories, here is one classification of some use. Statistics books seem to come in seven varieties.

The first is the so-called 'popular' type. It is intended for the lay audience. It has limited coverage but a lively style. Two classics in our field are 'Facts from

figures' by Morony and 'How to lie with statistics' by Darrel Huff. Such books are important though not used in day-to-day work. Each of us should try to write material in a similar vein. We owe it to the society at large. The present book is our attempt in this direction. In many fields of science, some reputed institutions have a tradition that every good Ph. D. student writes a popular book on an area of interest. You may have seen many fine books on animal behavior. If that is regarded as too far from our subject, think of the popular books written by Professor Jayant Naralikar on relativity and astrophysics. So, why can't we do it in statistics? One can begin by writing a short popular article. Every statistics teacher should consider writing an article for the magazine Resonance.

The second variety is the so-called 'cook books'. R .A. Fisher wrote the first book of this genre, (Statistical Methods for Research Workers) which went into several editions. Our favorite one in this category is 'Statistical Methods' by Snedecor and Cochran. These books enable potential users of statistics to apply methods to their problems. Again we feel that each teacher of statistics should try to write one book in a relevant area. Have you seen such a book about stochastic processes? There may be other areas of statistics for which a cook book is yet to be written.

The third variety is a textbook for students of statistics. This is one type where many teachers try their hand.

The forth variety is a monograph. Admittedly, there can be a lot of overlap between this and earlier type. This is likely to be written after substantial publication activity in a narrow area of research. Perhaps a rather senior person should write it. Another formula is a team that combines experience with enthusiasm. A bright Ph. D. student goes deep into a narrow topic. So a logical possibility is that the student collaborates with guide to write a review of that topic. This is not beyond the realm of feasibility. But our mindset in India seems to be one of diffidence. We should be able to overcome it.

Fifth type of book is a case study collection. An excellent example is a book edited by Judith Tanur titled 'Statistics: A Guide to the Unknown', published by the American Statistical Association. This material is highly educative for students (and teachers) as co-curricular reading. Every teacher of statistics should have a list of such books and should point them out to students

Sixth variety is a book providing data sets. Again there are a handful of these published in the west. It is high time we write something that is just right for us. We have made one attempt (100 data sets for statistics education) and the data sets are available for free down load.

Seventh variety is the type that we would like to avoid. It is a book with a title such as 'Statistical Inference for semester 7 paper 3 of ABC University'. A book written just to cover syllabus of one examination is likely to be too narrow. It

begins somewhere and stops somewhere, both ends dictated by the syllabus. Usually it pays scant attention to the history of the topic, its linkages with other topics and contributions by prominent workers, in short to the wholeness of the topic. Such writing enhances neither studiousness nor scholarship.

But book writing seems like a daunting task, taking several years and lot of efforts. If this is the reservation, an obvious solution is forming a team and splitting the effort. Such team work will lead to further gains in terms of higher productivity and more camaraderie. Finally, it is better to have fallen in love and failed than not to have loved at all! So, every teacher should resolve to write at least one book in the career and pay dues to the next generation.

30. Industrial Consultancy

We all know that as University/college teachers of statistics, our brief is three fold. It is teaching, research and extension. The last item, namely extension, may mean help to our own colleagues in other departments within the university or to scientists in research institutes elsewhere or to potential users in the society at large. The last group includes users in industry and business. Helping them is sometimes called consultancy and is treated as something different since it implies receiving money. We want to point out that the interaction involved is very similar to scientists in other disciplines except for this monetary aspect. Our judgment is that the proportion of our fraternity that is active in consulting with business and industry is too small. A significant increase is desirable and possible. Such participation has many attractive features. Teachers can get interesting research problems for work. Students can get exposure to current issues in industry and business and thereby become more attractive as employees in these fields. Our experience is that participation in consulting has a favorable impact on quality of day to day teaching of virtually any topic in statistics.

Perhaps we should point out that UGC (www.ugc.ac.in) and CSIR (www.csir.res.in), two major bodies of the Central Government have come out in favor of scientists offering their expertise/services to business and industry and earning revenues for their parent institutions (and themselves). We give below some excerpts from their websites. National Policy on Education (on the UGC website) states "*Institutions will be encouraged to generate resources using their capacities to provide services to the community and industry.*" It further lays down categorically that "*Active interaction between technical or management institutions and industry will be promoted in program planning and implementation, exchange of personnel, training facilities and resources, research and consultancy and other areas of mutual interest.*" In fact there are guidelines available on the CSIR website regarding sharing of fees earned. '*The revised pattern of distribution of honorarium (max. 2/3 of intellectual fee) for consultancy work taken up after 1st April 1990 will be as follows: Team of consultants 65%, other S &T staff 15%, supporting staff 15%, CSIR Welfare Fund 5%*'. Clearly, we have blessings from the government for doing work for business and industry.

Our proposition is that a new thrust to increase consultancy for business and industry is quite in order. We call upon all our colleagues to launch jointly a major effort to promote such consultancy.

We know that many teachers will smile and say that it takes two to tango. How can consultancy flourish unless users show interest in seeking it? That is right. While our economy was highly regulated, there seemed to be little interest in using statistics (and other) expertise available in universities. However, our perception is

that since opening of the economy and reforms initiated by Dr. Man Mohan Singh, the situation has changed dramatically. We see definite signs of favorable trends.

There are two main reasons for the new interest in use of statistics. One is survival and second is expansion. Some industrial units that were quite entrenched in our system now face new competition. A textile factory that we have been working with, finds that imports from South Korea and Taiwan are cheaper and have better quality. Hence bottom line of this company is under threat. So, they are seeking help from every possible quarter (including statisticians) to improve their forty-year-old process. This is survival instinct. Another company that has approached us produces components of certain high voltage electrical assemblies. Here too the process is old. But production of such items is being phased out in the west. Hence companies in Europe are looking for new vendors. This scenario has vetted the appetite of the Indian company. They see a big opportunity out there, waiting to be exploited. But there is one hitch. These European customers are sticklers for quality. Their queries are all based on statistical specifications regarding sampling schemes, sample sizes, process quality indices etc. The company needs help from statisticians to face this challenge. A third company in personal products is also faced with stringent sampling rules of the European Union. The company felt that the sample size specified by the EU was unnecessarily large and wanted us to offer probability-based arguments to support their stance.

The examples cited above are typical. If this assertion is correct, there must be many more companies facing similar situations and sooner or later they all will look for help in statistics. Hence our perception is that there is a large potential clientele for statistics in India's manufacturing sector.

Some of you will raise the doubt that if there is need for statisticians, companies will employ people of requisite background. Well, things are not that simple. We find that employees often get bogged down in generating routine reports and also lack contact with methodological bases of procedures. What is needed is not repetitive application of standard procedures but some 'out of the box' thinking. Troubleshooting is the name of the game. If a textbook solution does not exist, one has to design a new solution. In short, one has to operate in a research mode. Teachers who are familiar with a wide range of statistical tools, who do research as a regular activity, are singularly qualified for the assignment. Another advantage is that teachers have a nation wide presence. For an industrial unit looking for services, the nearest university/college department of statistics can be the easiest source.

We know from experience that one feels diffident in taking up such challenging assignments, especially if that is for the first time. We believe there are members of the teaching fraternity that are experienced in industrial consultancy

and first timers can call upon them to help. It should be possible, if necessary, to constitute a small cell in the fraternity that can act as a think tank and resource center. Perhaps someone from the cell can attend the first couple of meetings with officials of the industrial unit to get the ball rolling and then local statisticians can take help through email etc. We believe it will be an exciting venture. Participants will get tremendous satisfaction for involvement in the process of wealth creation: for make no mistake, when we help in enhancement of quality, we create wealth and do a national service. The simplest way to begin would be to hold discussions on this topic among local colleagues and friends and share thoughts with the wider community during conferences and gatherings.

References:

Agrawal, A and Narayan, S (1985). *The State of India's Environment 1984-85.* The Second Citizen's Report, Centre for Science and Environment, New Delhi.

Chaturvedi A. N. (1983). *Eucalypts for Farming.* U.P. Forest Bulletin, No. 48

Crossley,M.L.(2000) *The Desk Reference of Statistical Quality Methods*, American Society for Quality, Quality Press, Milwaukee, Wisconsin.

Darrell, H. (1954) How to Lie with Statistics, W. W. Norton, NY.

Dekker, A. J. F. M., Dawson, S. and Desai, A. (1991) An indirect method for counting Asian elephants in forests, In Censusing Elephants in Forests, Proceedings of an international workshop, Asian Elephant Conservation Centre of IUCN/SSC Asian Elephant Specialist Group, *Technical Report* No. 2.

Delampady, M. and Padmawar, V.R. (1996) Sampling, Probability Models and Statistical Reasoning, *Resonance* , 1(5) 49-58.

Deshapande, N.R, and Gore, A.P. (1999) On Computation of probable maximum precipitation, *Calcutta Statistical Association Bulletin*, Vol. 49, pp 113-123. (Winner of Bose Nandi Award for the best paper in applied Statistics).

Easa, P.S. and Balakrishnan, M. (1995) The Population Density and Structure of Asian Elephants in Parambikulam Wildlife Sanctuary, Kerala, India, *Journal of Bombay Natural History Society*, Vol.92, No.2, 225–229.

Fairley, W. B. and Mosteller, F. (1977) *Statistics and Public Policy*, Addison-Wesley, London.

Finkelstein, M. O. and Levin, (1990) B. *Statistics for Lawyers*, Springer, NY.

Gadagkar, R. (1992) World's Biodiversity Needs to Be Preserved, *Down to Earth*, Vol.1, No.11, 43–44.

Gadagkar, R., Vinutha, C., Gore, A.P. and A. Shanubhogue (1988) Pre-imaginal biasing of caste in a primitively eusocial insect. Proceedings of the Royal Society of London, b 233, p.175-189.

Gore, A.P., Gokhale, N, K., Joshi, S.B. (1979) On disputed authorship of editorials in Kesari. *Indian Linguistics*. 40, pp. 283-293. Summarized in Mosteller, F. and Wallace, D.(1984) Applied Bayesian and Classical Inference, The case of Federalist Papers, Springer.

Gore, A. P., Paranjpe, S.A. Geeta Rajgopalan, Watve, M.G., Gogate, M.G., Kharshikar, A.V. and Joshi, N.V. (1993) Tiger census: role of quantification. *Current Science*, Vol. 64, No. 10, p. 711-714.

Gore, A.P. and Paranjpe, S.A.(1995) *Tabulation and Analysis of Bird Ringing Data from BNHS*, Tech. Report, Department of Statistics, University of Pune.

Gore, A.P. and Paranjpe, S.A. (1998) Wait and see strategy for control of leaf miner in rainfed groundnut, *Current Science*, Vol. 74 No.4, p. 296-299.

Gore, A.P., Paranjpe, S. A., Pandit, S. J. and Prayag, V. R. (1990) Why solar cookers don't sell? *Changing Villages,* Vol. 9. No. 4. Oct. pp. 219-229.

Gore, A.P., Paranjpe, S.A. and Deshapande, N. R. (2001), On Estimation of Extreme Quantiles From Limited Data (2001) *Journal of Indian Statistical Association, Special IssueDedicated to Prof. Vasant P. Bhapkar,* pp 251-261.

Hand, D. J.,F. Dal, A.D. Lunn, K.J. McConway and E. Ostrowski, (1994) 'A handbook of small data sets' Chapman and Hall , London.

Holmes P.(2003) 50 years of statistics teaching in English schools: some milestones (with discussion) Journal of the Royal Statistical Society series D (The Statistician) v.52 no. 4 p.439-474.

Hooke, R. (1983) *How To Tell The Liars From The Statisticians*, Marcel Dekker, NY.

Jaffe, A. J. and Spirer, H. F. (1987). *Misused Statistics: Straight Talk For Twisted Numbers*, Marcel Dekker, NY.

Karandikar, R.L. and Basu, A. (1999) Opinion Polls and Statistical Surveys: What They Really Tell Us, *Resonance*, 4(7) 49-58.

Karanth, K.U. (1995) Estimating Tiger, *Panthera tigris*, Populations from Camera – Trap Data Using Capture Recapture Models, *Biological Conservation*, Vol. 71, 333-338.

Krishnan, T. (1997) Fisher's Contributions to Statistics, *Resonance*, 2(9) 32-37.

Kulkarni M B. (1987). *Modeling Growth and Harvesting Strategy of Eucalyptus*, M. Phil Dissertation Dept of Statistics, University of Pune

Kulkarni, M. B. and Gore, A. P. (1988). *"Improvement in Volume Estimationby Adding One More Girth Observation"*, inVan Vigyan, Vol 26, Nos 1 & 2.

Kunte, S. (1999) Statistical Computing 1. Understanding Randomness and Random Numbers, *Resonance*, 4(10) 16-21.

Kunte,S.(2000) Technique of Statistical Simulation, *Resonance*, Vol. 5, no. 4, PP 19-27.

Kunte, S. and Jeffreys, (1992) Lindley Paradox and a Related Problem. In *Bayesian Analysis in Statistics and Econometrics*, Ed. P K Goel and N S Iyengar, Lecture Notes in Statistics, 75, Springer Verlag, pp.249–255.

Lavraj, U.A. and Gore A.P. (1987 a) Innovation diffusion in a heterogeneous population. *Technological Forecasting and Social Change,* Vol. 32, p. 163-168.

Lavraj, U.A. and Gore A.P. (1987 b) Cross-bred goat adoption in a rural community. *Indian Journal of Agricultural Economics*. V.XLII, No.4, Oct.-Dec., p. 587-594.

Masalekar A. R. (1983). *Managing The Forests*. Jugal Kishore and Company, Dehra Dun

Mayee, C.D., Shah, A., Paranjpe, S.A. and Gore, A. P.(1998) Modeling progress of fungal attack on groundnut', *J. Ind. Soc. Agr. Stat. (P.V. Sukhatme memorial volume)* LI, August-December, pp293-302.

Paranjpe, S.A., Gore, A.P., Gogate, M.G. (1993) Application of Bhattacharya Technique in Sex Determination and Sex Ratio Estimation of Tigers from Pugmarks. *Indian Forester*, Vol.119, No. 10, pp. 793-797.

Parthasarathy N. and Karthikeyan, R. (1997) Biodiversity and Population Density of Woody Species in a Tropical Evergreen Forest, in Courtallam Reserve Forest, Western Ghats, India, *Tropical Ecology*, Vol. 38 (2), 297-306.

Phadake M.S.(1989) *Quality Engineering Using Robust Design*, Prentice Hall, New Delhi.

Prayag V.R. and Gore A.P. (1993). Should export of frog-legs be banned? *Chapter 20 in "Environmental Problems and Prospects in India"* Ed. M. Balkrishnan, Oxford and IBH Publishers, New Delhi.

Prayag, V.R., Paranjpe, S.A., and Gore, A.P. (1991) Mixture Models for Distribution of Number of Seeds per Pod in some Multiovulated Plants, in *Recent Advances in Agricultural Statistics Research*, Ed. Prem Narain, V K Sharma, O P Kathuria, Prajneshu, Wiley Eastern, pp.462–466.

Rajasekhar, B. (1995) A Study on Butterfly Populations at Guindy National Park, Madras, *Journal of Bombay Natural History Society*, Vol.92, No.2, 275–278, 1995.

Ramsey F.L. and Schafer D.W. (1997), The Statistical Sleuth : A course in methods of data analysis, Duxbury Press.

Rao, C.B. and Reddi, C.S. (1994) Pollination Ecology of *Martynia Annua L.*, *Journal of Bombay Natural History Society*, Vol.91, No.2, 187–193.

Reddy, T.B. and Reddi, C.S. (1995) Butterfly Pollination of *Clerodendruminfortunatum* (Verbenaceae), *Journal of Bombay Natural History Society*, Vol.92, No.2, 166–173.

Sitaramam, V., Krishnakumar, T., Gore, A.P., Paranjpe S.A. and J.G.Shasatri. (1996). Minimum Needs of Poor and Priorities Attached to Them, *Economic and Political Weekly*, Special Number Sept. pp 2499-2505.

Taha,H.A.(1982) *Operations Research*, Macmillan, New Delhi.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Link, R. F., Pieters R. S. and Rising, G. R. (Eds.) (1972), "*Statistics, A Guide To The Unknown*, Holden-Day, San Francisco.

Thakar, J., Kale, A., Puntambekar, S., Shaikh, I., Vaze, K., Jog, M., Paranjpe, S. A. (2002). Bee-eaters (*Merops orientalis*) Respond to what a predator can see, *Animal Cognition*, DOI 10.1007/s10071-002-0155-6 (e-journal).

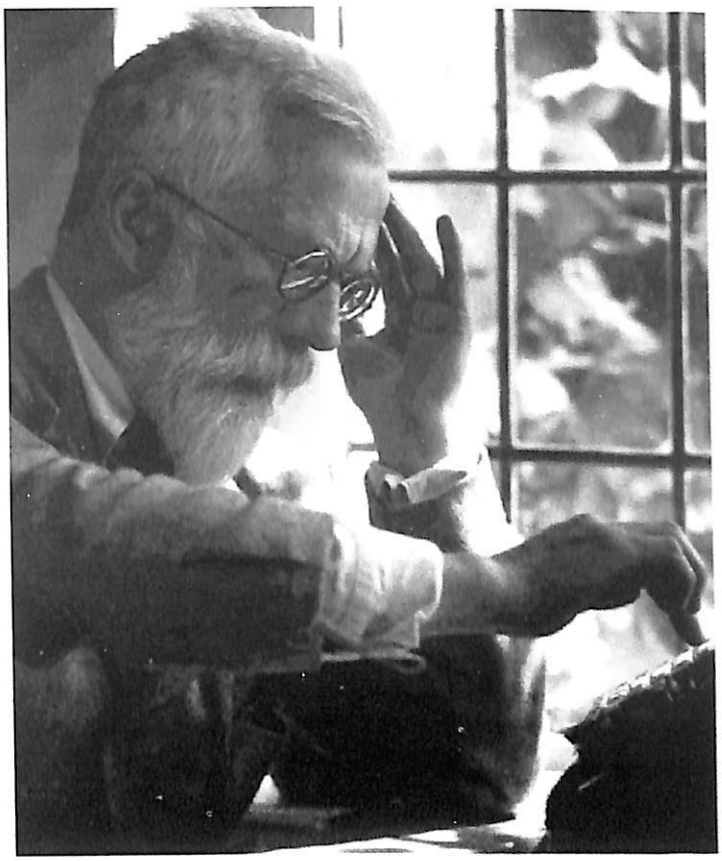
Veena, T. and Lokesha, R. (1993) Association of Drongos with Myna Flocks: Are Drongos Benefitted? *J. Biosciences*, Vol.18, No.1, pp.111–119.

Watve, M. and Sukumar, R. (1995) Parasite abundance and diversity in mammals: Correlates with host ecology, *Proc. Nat. Acad. Sci.* USA, Vol. 92, 8945-8949.

Zeisel, H. (1985) *Say It with Figures*, Harper and Row, NY.

Zeisel, H. and Kaye, D. (1997) *Prove It With Figures*, Springer, NY.

Appendix



R A Fisher

(17 February 1890 – 29 July 1962)

R A Fisher

(17 February 1890 – 29 July 1962)

Some people regard 1900 as the year of birth of Modern Statistics. That year Karl Pearson put forth the chi-square test of goodness of fit. In late 19th century his name was virtually synonymous with statistics. However, if we go by concepts and practice of statistics prevalent today, there is no doubt that the main content is provided by the work of R. A. Fisher. In this sense, Ronald Aylmer Fisher can be regarded as father of modern statistics.

Some of his major contributions to statistics came during his work at the Rothamsted Experimental Station where his job was to examine data on yield of wheat and other crops. One contribution was Analysis of Variance and the other was Design of Experiments. These techniques have been used all over the world in agricultural research. In analysis of variance, a key step is a test for comparison of two independent estimators of variance. He named it 'variance ratio test'. We now use the name F test. This name was given by George Snedecor in honor of Fisher. In proposing new methods for designing experiments, Fisher showed that it is possible to vary many factors at the same time and still get a valid assessment of effects. This was contrary to traditional way of changing one thing at a time. In the area of statistical inference, Fisher gave a firm foundation to statistical methods of estimation by introducing a measure of information in a statistic and the method of maximum likelihood. He enriched the field by introducing concepts of sufficiency and ancilliarity. He introduced the term null hypothesis.

Right from his student days in Cambridge, Fisher was interested in problems of genetics. Around the turn of the 19th century, there were two schools in genetics. Mendelian school focused on inheritance of discrete traits such as hair or eye color. Karl Pearson and others pursued instead, quantitative traits such as plant height, crop yield etc. Fisher showed convincingly that Mendelian theory of particulate inheritance could explain observed patterns of inheritance in quantitative traits as well. This was a landmark result.

Fisher's contributions to the field of evolution are regarded as phenomenal. Richard Dawkins, a famous evolutionary biologist described him as "the greatest of Darwin's successors". Fisher introduced the ideas of gene frequencies, genetic linkage, sexual selection and mimicry. He was the first to use diffusion equations to explain spread of genes. Ecologists and geneticists are known to ask whether it is true that the famous mathematical biologist R.A.Fisher is also a statistician!

Fisher was a great supporter of statistical work in India. He encouraged development of the Indian Statistical Institute and also the Indian Agricultural Statistics Research Institute.

A fine biography of Fisher written by his daughter is-

Box, Joan Fisher (1978) R. A. Fisher: The Life of a Scientist

Fisher used the story of ' lady who preferred tea prepared with milk poured first' to start discussion of Design of Experiments. A book on history of statistics uses that as title.

Salsburg, David (2002) The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century

Prasant Chandra Mahalanobis (PCM)
(29 June 1893–28 June 1972)

Prasant Chandra Mahalanobis (PCM)
(29 June 1893–28 June 1972)

PCM, son of a Professor of physiology in Calcutta University, studied Physics in Cambridge. He came across the journal Biometrika edited by Karl Pearson. He found it interesting and potentially useful to tackle many social problems. During his tenure as a teacher of Physics in the Presidency College Calcutta, PCM gathered around him a group of professors for whom statistics equally interesting. This group founded the Indian Statistical Institute in 1932. In 1959, Government of India declared it an institute of national importance and a deemed university. ISI went on to become one of the most famous places of research in statistics in the world. In 1933 the journal Sankhya was launched. It was modeled on the pattern of Biometrika.

PCM always looked for opportunities to apply statistics to problems of interest to the Indian society. As long as this condition was satisfied, it did not matter which field the problem came from. Anthropologists were interested in comparing different ethnic groups on the basis of body measurements. During the course of these studies PCM developed a formula of comparing and clustering populations using a measure based on all measurements. This formula, D^2, is now named after him as Mahalanobis distance and is used the world over.

He worked as a meteorologist and took up investigation of floods in eastern states. He came up with recommendations for flood control measures which surprised engineers.

PCM became very famous for the sample surveys he organized for consumer expenditure, crop acreage estimation and other purposes. Eventually these activities were taken up by the Government of India under National Sample Survey Organization. Harold Hotelling (a leading statistician from USA) wrote: "No technique of random sample has, so far as I can find, been developed in the United States or elsewhere, which can compare in accuracy with that described by Professor Mahalanobis" and Sir R. A. Fisher commented that "The I.S.I. has taken the lead in the original development of the technique of sample surveys, the most potent fact finding process available to the administration". PCM was regarded highly by the Prime Minister Jawaharlal Nehru. Appointment as member of the Planning Commission gave PCM the opportunity to play a crucial role in the draft of the second five year plan. This was the beginning of rapid industrialization of India.

He received one of the highest civilian awards, the Padma Vibhushan from the Government of India for his contribution to science and services to the country.

A fine biography of PCM is now available.

Rudra, A. (1996), Prasanta Chandra Mahalanobis: A Biography. Oxford University Press

Pandurang Vasudeo Sukhatme (PVS)
(27 July 1911- 28 January, 1997)

Pandurang Vasudeo Sukhatme (PVS)
(27th July 1911 - 28th January, 1997)

PVS was born in village Budhgaon in Sangli District in Maharashtra. He studied mathematics in Fergusson College in Pune and went on to get a Ph D in statistics from the University College London. He was probably the first Indian to get an advanced degree in statistics. He received a D Sc from the same university in 1939.

On return to India, he approached Pandit Madan Mohan Malaviya, the Vice Chancellor of Benaras Hindu University for a job. Panditjee told him that a job would be possible if he could explain how his work would help common man in the country. The young mathematical statistician was embarrassed. He had never thought in such terms. He resolved never again to be stumped by such a question. During the rest of his life he always used his expertise in statistics to solve problems of the people of India and the world.

In 1940 PVS joined the Imperial (Changed to 'Indian' after independence) Council for Agricultural Research and started a small statistics cell. It has now grown into a full fledged research center (Indian Agricultural Statistics Research Institute). He launched many sample surveys to study problems of agriculture in India. In course of this work he adopted a philosophy that any available infrastructure (administrative system, local officials etc) should be used fully. Thus for village surveys local patwari and land records could play a useful role. This was at variance with the approach preferred by Professor P C Mahalanobis. These differences could never be reconciled. When PVS presented lectures on his work in agriculture at the Food and Agricultural Organization of the United Nations, officials were impressed enough to propose that he should extend these methods to all third world countries.

In 1952 PVS joined FAO in Rome as the Head of the Statistics Division. He held this position till 1970. Under his leadership many new initiatives were started in countries in all continents. There was a shortage of skilled personnel. Members of the team in Delhi who had worked under PVS took turns serving in different countries on deputation. This may be the first wave of techies from India doing assignments outside.

After retirement from FAO PVS returned to India and settled in Pune. He continued to work on socially relevant problems. One issue he took up was protein malnutrition. It was argued by some that Indian diets are deficient in protein and this can affect growth and development of children. The suggested remedy was protein supplement. PVS dismissed all such suggestions as invalid. Using data from records of National Institute of Nutrition, Hyderabad, he showed that Indian diets have adequate protein content. Instead the problem is that of poverty. If there is enough earning and people can eat their usual diet to their fill, then they will not be protein deficient. This perspective offers a different solution to the problem. The solution is to put enough money in the pockets of the poor (through employment schemes). He also worked on measurement of poverty.

In 1973 he received the Padma Bhushan award of the Government of India.

Calympudi Radhakrishna Rao
(September 10, 1920)

Calympudi Radhakrishna Rao (born September 10, 1920)

The first place to offer advanced education in statistics in India (or anywhere in the world except UK) was the Calcutta University and one of the first students to get a master's degree in the field was C R Rao. That was in 1943. He worked with R A Fisher and received a Ph D from Cambridge University in 1948 for his thesis "Statistical Problems in Biological Classification" and D. Sc. From the same university in 1965. Thus began one of the most distinguished careers in mathematical statistics.
C R Rao's name is well embedded in text books of mathematical statistics. Two most common references to his name are Crame'r-Rao lower bound and Rao-Blackwell thorem. He is also well known for his score test, orthogonal arrays, generalized inverses. He is a household name in the family of statisticians and now electrical engineers too.

As the head of the Research and Training school of the ISI during 1949-1963, he built an internationally renowned team of young statisticians, economists and mathematicians. The members of this team are now holding prominent positions in several universities in India and abroad.

C R Rao has held several distinguished positions. He occupied the Directorship and the Jawaharlal Nehru Chair at the Indian Statistical Institute, Distinguished Professorship at University of Pittsburg and the Eberly Professorship at Pennsylvania State University.

He has been honored by institutions all over the world. He received honorary doctorates from 27 Universities in 16 countries. In UK he received Fellowship of the Royal Society. In USA he was a recipient of the National Medal of Science. He has been called a living legend in science.

About the Authors

Dr Anil P Gore

Presently Vice-president, Vislation Inc Pune. Professor and Head, Department of Statistics, Pune University (Retired). Recipient of (i) "Distinguished Statistical Ecologist" award of International Association of Ecologists in Italy in 1998, (ii) first "Bose-Nandi Award" of the 'Calcutta Statistical Association' (1999) for the best paper in Applied Statistics. Taught in several universities in India, USA, Malaysia and Canada. Consultant to research institutes and commercial organizations. Author of many newspaper articles in Marathi and Marathi books - "Naracha Narayan" and "Makadcheshta". Co-author of "Statistical Analysis of Non-normal data". Published about 80 research papers in international journals of repute. Member- National Tiger Conservation Authority, Govt. of India.

Dr. Mrs Sharayu A Paranjpe

Professor, Department of Statistics, University of Pune. Awarded "Fellowship of Association of Common-Wealth Universities" (1997). Co-author of two books (1) The Human Face of Clergy (1991) and (ii) "A Course in Mathematical and Statistical Ecology". Published 36 research papers. Currently working as a Principal Statistician in Vislation Inc. Pune. Consultant to research institutes and commercial organizations.

Madhav B Kulkarni

Presently Head, Department of Mathematics and Statistics, B. Y. K. College of Commerce, Nashik. M. Phil from Department of Statistics, University of Pune. Formerly, lecturer Department of Statistics, Visiting Lecturer in the Interdisciplinary School of Health Sciences, University of Pune. Conducts training programs in "Advanced Statistical Computing". Co-author of three books – Common Statistical Tests, Introduction to Discrete Probability and Probability Distributions, Introduction to Statistical Ecology. Author of "Quantitative Techniques" a book written for Yahswantrao Chavan Maharashtra Open University, Nasik. Guided one M. Phil student and has published seven research papers. Written a series of articles "Sankhyashastra Sarvansathi" in local newspapers.

Inside the book....

Statistics is a subject that confuses many, scares some but excites few. This situation is unfortunate and unnecessary. In fact, there is no reason why we the teachers of statistics cannot help laymen recognize the importance of our discipline.

In today's world of exploding information, it is not enough to know the three R's (reading, (w)riting and (a)rithmetic). It is also necessary to learn techniques to consolidate and interpret the continuously bombarding information. Statistics is the science of identification and art of interpreting patterns in numbers.....

"Every individual has his own characteristics and yet the group as a whole follows a pattern (law). Statistics talks about these group properties or population laws."

If every one shares the same suffering, there is no relative deprivation and poverty hurts less in such case. Inequality makes the impact of poverty more severe

Taguchi has become synonymous with consideration of quality and productivity. This Japanese engineer introduced a novel approach to problems of quality in manufacturing

Applied statistics is a venture in which interactive work is a very crucial part of each project. ... If the attitude is right, all this can be a lot of fun.

McCain and Obama have different thinking styles. Whereas McCain tends to be more categorical in his thinking, Obama is more fluid or contextual in the ways he approaches problems..."

One can only hope that the community of statisticians in India will shed its isolation and sloth, rise to the challenge of the new era and in the process bring success and prosperity to their parent institutions and students. Posterity will judge them harshly otherwise

Extract from the Review

"---What sets this book apart from the plethora of texts on statistics is that common statistical concepts, tools, and experimental techniques used by the scientists are introduced in a perceptive and jargon-free manner..."

"--- The book is an easy read and authors do an excellent job of giving an overview of different statistical methods allowing the reader to get comfortable with the theory and eventually think of real-life applications ..."

"... The book fills a void in the literature which was long ignored: extending the knowledge of basic quantitative techniques and principles of data analysis to the common man.."

Current Science, Vol 101, No 12, 25 Dec 2011.