SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Editors: P. C. MAHALANOBIS, C. R. RAO

SERIES B, VOL. 31 DECEMBER 1969 PARTS 3 & 4

A MULTIDISCIPLINARY APPROACH FOR TEACHING STATISTICS AND PROBABILITY*

By C. RADHAKRISHNA RAO

Indian Statistical Institute

1. STATISTICS: A NEW TECHNOLOGY

A few years ago, the Indian Statistical Institute introduced a 4-year course leading to B.Stat. (Bachelor of Statistics) degree, for students with high school education. The object of the course is 'to offer a comprehensive instruction in the theory and practice of statistics and provide at the same time a general education together with the necessary background knowledge in the basic natural and social sciences expected of a professional statistician'. The broad guide lines for the course were formulated by the late R. A. Fisher, the late J. B. S. Haldane and Professor P. C. Mahalanobis all of whom shared some common views on the scope of statistics. Speaking on 'Statistics as a Key Technology' at the 125th Anniversary of the American Statistical Association, Mahalanobis (1965) described the broad scope of statistics and the role of a statistician as follows :

> 'The time has come to introduce educational programs appropriate for statistics as a fully developed technology which calls for the utilization of a wide range of scientific knowledge to help in solving scientific or practical problems. As Fisher has pointed out, *a professional statistician, as a technologist, must talk the language of both theoreticians and practitioners.* The education of a statistician, like that of other technologists, must have a broad base.'

Thus the object of the new course is not to teach statistics as a separate discipline with a well-defined area of study like physics, chemistry or biology, but introduce it as a body of techniques for application in research problems of various disciplines. Or, in other words, teaching has to be problem oriented with emphasis on

*Paper submitted to the International Conference on the teaching of Probability and Statistics at the pre-college level, March 18-27 1969, U.S.A.

collection of live data, their analysis and interpretation rather than drilling the boys in the use of known statistical techniques on what Fisher calls 'mock up data for the use of students only.' The training of a statistician should be on the same lines as that of an engineer or a medical practitioner with emphasis on professional aspects.

In the course that is being currently given, mathematics, statistics and economics constitute the main subjects of study. In addition, courses are given in a number of science subjects, but here the emphasis is not so much on the content of knowledge as on methods, with more of practical exercises involving experimentation and collection of data. Besides physics, chemistry and biology, which are taught during the first two years of the course, lectures are given on selected topics in sociology and psychology during the third year, with emphasis on quantitative approach.

The mathematical content of the high school curriculum in India is not very high. Students admitted to the B. Stat. course would not know calculus or would not have studied it in a rigorous way. The first task was to work out courses in modern mathematics spread over the first three years of study. There was difficulty in the beginning in formulating the syllabii for other subjects such as economics, physics, chemistry, psychology and sociology as only selected topics in these areas were intended to be covered. But during the period of last eight years the Institute has gathered considerable experience, and an integrated syllabus was evolved for mathematics and different branches of social and natural sciences with statistics as a central discipline providing a common bond between subjects. Along with the syllabus, certain methods were developed for teaching statistics and to instill in the students a spirit of true intellectual inquiry. In my lecture today, I shall try to cover both these aspects : the syllabus for an introductory (first year after high school) course in statistics and probability and some methods of teaching statistical techniques.

2. Population projection : a first project

The first batch of students for the B. Stat. course was admitted to the Institute in September 1960. That was the time when hectic preparations were going on in the country for conducting the decinnial census in January 1961. Population or demography seemed to be an appropriate and timely topic to introduce to the students and discuss about it. It is also a subject with a long history and of great current interest. Further, demography has contributed in a significant way to the early development and study of statistics. Thus an early discussion of population problems would also provide the students with an opportunity to glean into the history of statistics.

The first problem set to the class was how to anticipate the Registrar General and predict the population in 1961 in advance of the census. This was a crucial and at the same time a dangerous exercise. The predicted figure would be subject to test a few months later when the census count would be available. If the agreement was good it would inspire some confidence in the students in learning techniques of 'counting chickens before they are hatched,' which are novel to their way of thinking.* On the contrary, the effect could be depressing both for the students as well as the teacher. However, the investigation was worth undertaking. I shall describe how the problem was approached and what the students had learnt, more by experience than by direct teaching, during the process.

Contact with official statistics. The problem gave the students their first introduction to what are called official statistical publications. Population Census volumes had to be consulted to obtain the previous figures which might give some idea of the trend of growth. They had to hunt up for current publications giving annual figures of births, deaths, migration and immigration during the decade 1950-60, which, if available, would reduce the problem of prediction to a simple exercise in arithmetic.

What is a population census ? How is it conducted and what are the difficulties in making a complete enumeration at a national level. What are the likely inaccuracies in the published census figures ? These are some of the questions which had to be discussed to make the students familiar with the data they had to deal with.

How are the data on current births and deaths recorded and published ? How accurate can they be ? If both are under-registered, to what extent can the difference, births-deaths, which is the net addition to population (apart from emigration—migration) be correct ? These constitute another series of questions which have led the students into the realm of reality and exercised their thought. Is there any method of checking the accuracy of published figures and making suitable corrections ? The students were already placed in the thick of the problem and they had to think their way out.

Preparation of schedules. The class had an assignment to prepare the census schedule (questionnaire or data sheet). They had to examine the previous schedules and discuss *what fresh questions* could be added and for what purpose. The designing of a schedule is itself an art and each student produced his own version, which gave an opportunity for discussing some general principles to be followed in such cases.

Stratification. Any method chosen for prediction based on previous census figures could be directly applied on population figures for the country as a whole, or applied separately on different regions and the regional projections added up to get the national figure. It is easy to demonstrate on empirical data that the two procedures lead to numerically different results unless the rate of growth is the same for all the regions, and more accurate results are obtained by the latter procedure. This discussion led to construction of indices of growth for different regions and methods of comparing them.

*It is unfortunate that at an young age we are exposed to stories to drive home the moral, DO NOT COUNT THE CHICKENS BEFORE THEY ARE HATCHED. Nothing can be more devastating and demoralising than this on the young mind which realises the need for foreseeing the future in order to make any progress.

Performance test. When different methods are available, there arises the problem of choosing a method appropriate for a given situation. One way of doing this is to try out different methods in cases where the accuracy of the predicted value can be ascertained by comparison with known figure and to choose the one with the best performance. In the present problem, the different techniques suggested were applied on census figures prior to 1951 to predict the figure for 1951, which was known. By comparing the predicted with the known figure, it was found that two methods gave equally good results compared to a third method. The two methods were then used to predict the 1961 population. The need for testing the performance of a suggested method, before applying in a new situation, was stressed.

Growth model and fitting of constants. The most difficult part of the project was a discussion of the methods employed for prediction. What is a model as distinct from a function which gives the relationship between population and time ? This is a difficult concept, and some time had to be spent in stressing the importance of an equation such as

$$\text{observed value at time } (t) = M(t) + \text{error}$$

where $M(t)$ is a function of t, representing the model. The error term was introduced as a deviation of the observed from the model value and its nature discussed through appropriate examples.

For population projection, a polynomial of a suitable degree or a logistic function may be suggested as a model and simple methods of fitting such functions discussed. By applying these methods, the population for 1961 was estimated to be 420 millions. There was good agreement with the census figure of 439 millions which was available a few months later, the error being less than 5 per cent considering the simplicity of the method used. The exercise was worth undertaking in a number of other ways.

Descriptive statistics. Population figures provide good material for teaching descriptive statistics. For instance the age pyramid which is the histogram of age of individuals in a population, gives an interesting characterization of a population in terms of the relative frequencies of the young and old. Comparisons of age pyramids between sexes, between states or countries and over time (for instance before and after a great war) and writing reports on such studies are excellent exercises for a beginner in statistics. Summarisation of data in terms of centres of location such as mean, median and mode and their comparison, and measures of variation such as inter-quartile range and standard deviation can also be illustrated with age frequency distributions.

The concepts, construction and uses of birth, death and net reproduction rates and life tables could also be discussed in connection with population projection.

The problem of estimating annual population between the census years leads to a discussion of interpolation formulae, which should be a part of the first year course in statistics.

3. Genetics of sex determination : a second project

While throwing of dice, drawing of cards and sampling of beads of different colours from a bag are useful devices in demonstrating and understanding the results of simple chance mechanisms, the knowledge so experienced will remain in abstract if its application to the study of *natural events* is not emphasised. In fact, it would be more interesting to begin with observed sequences of natural events and then examine whether they could be *mimicked* by mechanical chance devices. The problem chosen for investigation was the process of sex determination of children.

The students were sent to a maternity hospital to obtain information on the sex of successive children born, during certain periods of the day over a number of months. A record of over thousand observations is shown in Table 1, where M stands for a male and F for a female child. The students had no knowledge of probability. They were aware that the sex of an unborn child could not be predicted and that roughly half the children born were male. With such background knowledge there was no way of interpreting the observed sequence of random events. But before the discussion on the observed sequence started, the students were asked to carry out a mechanical experiment of drawing beads from a bag containing black and white beads in equal numbers and noting down the colour of the bead drawn each time after thorough mixing. The observed sequence of black and white beads in 1000 draws is given in Table 2.

How does one recognize from an observed series of events such as the occurrence of white and black beads (Table 2) the nature of the chance mechanism that has produced it (such as drawing at random from a bag containing equal numbers of white and black beads) ? How does one compare two series of events and infer that underlying chance mechanisms are similar or not ? Answers to these questions may enable us to infer on the mechanism producing a sequence of events genuinely occurring in nature such as the sex of children of successive births by comparison with a sequence of artificially produced events such as the occurrence of white and black beads in successive draws. We shall refer to the data on children as the real series and the data on the beads as the artificial series.

We shall first carry out the same type of analysis on both the series (of Tables 1 and 2) in an attempt to find differences in their behaviour. An inference of the type that the underlying chance mechanisms for both the series are the same will be a consequence of our inability to distinguish between them by appropriate analysis. In such a situation, since the mechanism for the artificial series is known, the same chance mechanism can be postulated for the real series.

Bernoulli distribution. If we consider a set of three events each of which has two alternatives such as M and F or W and B, then there are 8 possible sets. The frequency distributions of these 8 possible sets in the first 166 sets, in the second 166 sets and for the total of 332 sets are obtained for the artificial and the real series.

TABLE 1. DATA ON SEX OF SUCCESSIVE CHILDREN DELIVERED IN AN INDIAN HOSPITAL OBSERVED DURING CERTAIN PERIODS IN SOME MONTHS IN 1956

		January		
F M M F F	M M M M F	M F M F M	M M F F M	F F M F F
F M F M M	M M M M F	M M M M M	F F F F M	M F M M M
M M M M M	M M F M F	M M F F F	M M F M M	F.F F M F
F M F M M	M F M M M	F F M M F	M F F M M	F M F M M
F F M F M	M F M F F	F M M F F	M F M F F	F M M M F
F F M F M	F M M M M	M F M F F	M F M F M	M F M M F
F F F F F	F F F M M	F M M M F	M M M M F	F M F F F
F M F M M	M M F F F	F M F F F	M M M M M	
		February		
				F F M F F
F F M M M	F F F F M	F F F M F	F M F F M	F F M F F
M M M F M	M F M F M	F F M F M	M F M F M	M M F M M
F M M F F	F M M M F	F F F F M	M M F F F	M M F F M
M F M F M	F M M M M	F F M M F	F M M F M	F M M F M
F F				
		March		
M F F	F M M M M	M M M F M	F F F F F	M M M F M
M F M F F	M F M F F	F F F M M	F M F F M	F M M F M
M F F F F	F M M F M	F M M F F	M M M M M	M M F F M
M M F F M	M M M F M	F F M F M		
		April		
			F M F F M	F F M M M
F M F M M	M M M F M	M M M M M	F F M M M	F M F M F
M M F M M	M M F F M	F M M M M	M M M F F	F M M F M
F M F F M	M F M F F	M M F M F	M F M F M	F F M F M
F F F F M	F M M M F	F M F F F	M M F F F	M M M F F
F F M F F	F M M M F	F M F M F	M F M F M	M M F M F
M F M M F	F M M F F	F M M F M	M M M M M	F M M F F
F F M F M	M F F F M	F M M F F	M F F F M	M F F M F
		July		
F M M M M	F M M M M	F F M F F	F F M M F	F M F M M
F F F M M	F M F F F	F M F M M	F M F M M	M M M M M
M F M F F	M M M M M	F M F M M	M F M M F	F M F M F
M F M M F	F F M M M	M M M F M	M M F F M	M M M F F
F M F F M	M F M F F	F F F F F	M M M M F	F F F M M
F F M M M	M M M M F	M M M M F	F M M F F	F F F M M
F				
		October		
M M M F	F F F M F	F M M F M	M F M M F	M M M M M
M F M F M	F F F F M	F M F F F	F M F M M	M F F F M
M F M M F	M M F F F	F F M F F	F M M M M	M F M M F
F M M F F	M F M M F	F M M F F	M M F F M	F F F M F
F M M F F	M M F M M	M M M M F	F M F F M	M F F M F
F F M M F	F F F M F	F F M F F	F F M F M	F F M F F
M M F M M	F F F M F	M F F M F	M M M F F	F F F F F
M F M F M	M M F F F	M F F M F	M M F M F	M M M F M
M F M M F	M M F F F	F F M F M	F F F M M	M F M M M
M F F F M	M F M F F	M F F F M	M M F F M	M F M M M
M F M M F	F M M M F	F F M M F	F F F F F	F F F M F
M M F M M	M F M F F			

TABLE 2. DATA ON COLOUR OF SUCCESSIVE BEADS DRAWN FROM A BAG CONTAINING EQUAL NUMBERS OF WHITE AND BLACK BEADS

B W W B W	B W W B B	B B B W B	B B W W B	W W W B B
B W B B B	B B W W B	W B W W W	B B W W W	W W W W B
W W B W W	W B B W B	W W W B B	B B B W W	B W B W W
B W W W W	B B W B B	W W B B W	B W W B B	W B B W B
W B W B W	B W B B W	B B B B W	B B B B B	B B W B W
W B W B B	W B W B B	W B W B W	B W B B B	W W B B B
B W W B B	B W W B W	B W B B W	B W B B B	W B W B W
B B B W W	W W W B W	W B W W W	W W W B B	B B W W B
B B B W W	B W W W B	B B W W W	W W B B W	B B B W W
W W B B W	W W B W B	B B W B W	B W W W W	W B W B W
B W B B B	W W W B W	B W B B B	W B B W W	W B W B B
W B W B W	W W B W B	W W B W W	B W W W B	B B B W B
W W W W B	B B W W W	W W W W W	B B B B W	W W B B W
B W B W B	B B B W W	B W W W W	B W B B W	W B B B B
B B W B B	B B W W W	B W B W W	B W B W W	B B B W B
W W W B W	B W W W W	W W W W B	B B W B W	W W W B B
W W B W B	W W W B B	B B B W W	B W B W W	W W W B W
B B B W B	B W W W B	B W W B B	B B W B W	B B B B B
W W B W B	W B W W W	W B B B W	B B W B B	W B W W B
B W B W B	B B W B B	B B B B B	B B W B W	W W W W B
B W W W B	W W B W B	W B W W B	B B B W B	B W W W B
B W B W B	W W B B B	B B W W B	B W B W B	W W B B B
W W W B W	W B B B B	W W W W B	B W W W B	B B B B B
W B B W W	B B B W B	W W B B B	W W B W W	W W B B B
B B B B W	W B W B B	W W B W W	B B B W W	B W B W W
W W B W B	W B W B W	W B W W B	W B W B W	B B B W W
B W B W B	W W W W W	B W W W B	B B W B W	B W B W W
B B B B W	W B W W B	W W B B W	B W W W W	B B B W B
W B W B B	W B W W W	W W B W B	W W W B B	B B B W W
W B W B B	B B B W W	W B B W W	W B W B W	B W W B B
W B W W W	B B B B W	W B B B W	B W W W W	W B B W B
W B W B B	W B B W W	W W W W W	W B B W B	B B W W B
W B B W W	B B B B B	B W W B B	B W W W B	W B B W W
W W B B W	W W W B B	W W W B W	B B W B W	B W B B W
W B W B W	W B W B W	W B B B B	W B W W W	B W B B W
B W W B B	W B B B B	W W W W B	B W W W W	B W B W W
B W B W B	B W B B W	W B W B W	B W W W W	W B B W B
B B W B W	W B W B B	W W W B B	B W B B B	W B W B W
B B W W W	B W W B W	W W W B B	B B B W W	B W B W W
W W W B W	B B W B B	B W B B W	B W W W W	W W W W W

The results are as shown in Table 3. Both the distributions are compared with what may be called 'theoretical expectation according to Bernoullian hypothesis', under which all possible sets are equally likely.

TABLE 3. FREQUENCY DISTRIBUTIONS OF DIFFERENT SETS OF EVENTS

artificial series				expected value	real series			
set type	first 166 sets	second 166 sets	total		total	second 166 sets	first 166 sets	set type
B B B	17	20	37	41.5	46	23	23	M M M
B B W	20	26	46	41.5	36	20	16	M M F
B W B	23	25	48	41.5	49	20	29	M F M
B W W	27	24	51	41.5	50	27	23	M F F
W B B	18	15	33	41.5	37	16	21	F M M
W B W	27	18	45	41.5	37	16	21	F M F
W W B	15	17	32	41.5	43	23	20	F F M
W W W	19	21	40	41.5	34	21	13	F F F
total	166	166	332	332	332	166	166	total

It is interesting to note the following :

(i) In either series, the frequencies for the first and second 166 sets are similar showing that parallel sets of data from the *same* mechanism conform to a certain pattern although they differ in individual frequencies.

(ii) The pattern is as indicated by the Bernoullian hypothesis of equal frequency.

(iii) There is considerable similarity between the artificial and real series in the behaviour of frequencies.

It would indeed be difficult to say as to which analysis relates to artificial data and which to real data, when the lables attached to Tables 1 and 2 are unknown (or withheld).

Binomial distribution. We have made one kind of analysis of the sequences of binary data, which may not always be possible to carry out. For instance we may not have the actual series of events but we may know the number of black beads or the number of males in a set of three events. In such a case we can obtain the frequency distributions of the number of events of one kind. The results are given in Table 4.

TABLE 4. FREQUENCY DISTRIBUTION OF THE NUMBER OF EVENTS OF ONE KIND IN SETS OF THREE

artificial series		expected value	real series	
no. of black beads	frequency		frequency	no. of males
0	40	41.5	34	0
1	128	124.5	130	1
2	127	124.5	122	2
3	37	41.4	46	3
total	332	332	332	total

The expected values of Table 4 are obtained from those of Table 3 by adding over the combinations containing the same number of events of one kind. Thus the expected frequency for 2 black beads is the sum of the expected values for the combinations B B W, B W B and W B B. Again the similarity between the artificial and real data is brought to light by the analysis given in Table 4.

At this stage the derivation of the binomial distribution under the Bernoullian hypothesis could be demonstrated, first in the case of equal probabilities for the two kinds of events.

Let us consider sets of n events instead of 3 chosen in the example. There are 2^n possible sets all of which are equally likely. Thus in N sets, the theoretical frequency of each kind is $N/2^n$. To find the frequency of r events of one kind, we have to add up the frequencies over sets containing r events of one kind. There are precisely $\binom{n}{r}$ such sets, which is the number of combinations of r positions out of n. Hence the required frequency is $N\binom{n}{r}/2^n$, giving the relative frequencies for $r = 0, 1, \ldots, n$, as

$$\binom{n}{0}\frac{1}{2^n}, \binom{n}{1}\frac{1}{2^n}, \ldots, \binom{n}{n}\frac{1}{2^n}$$

which is called a binomial distribution with probability 1/2 for one kind of event.

We shall now examine the frequency distribution of the number of events of one kind in sets of $n = 5$ using the derived theoretical formulae for the expected values. The results are given in Table 5.

TABLE 5. FREQUENCY DISTRIBUTIONS IN SETS OF 5

	artificial data					expected value	real data					
no. of B's	1st 50	2nd 50	3rd 50	4th 50	all sets		all sets	4th 50	3rd 50	2nd 50	1st 50	no. of M's
0	1	2	1	1	5	6.25	5	2	1	1	1	0
1	6	10	6	5	27	31.25	27	8	7	3	9	1
2	19	11	17	17	64	62.50	64	20	12	19	13	2
3	17	16	19	16	68	62.50	64	12	20	17	15	3
4	7	10	6	9	32	31.25	31	8	6	8	9	4
5	0	1	1	2	4	6.25	9	0	4	2	3	5

Histograms are drawn for 4 different sets of 50, 2 different sets of 100 obtained by pooling the first two and the last two sets of 50 each, and for all the 200 sets both for the real and artificial data (see Figures 1 and 2). The following comments can be made.

(i) The histograms for $n = 50$ and 100 of parallel sets of observations although produced by the same mechanism are not the same but are similar in shape.

(ii) The variation in shape between parallel sets is small when the number of observations is increased.

(iii) The situation is the same for the real as well as the artificial data.

(iv) In all cases the shape of the observed histogram conforms closely to the theoretical or expected one and the agreement gets closer as the number of observations increases.

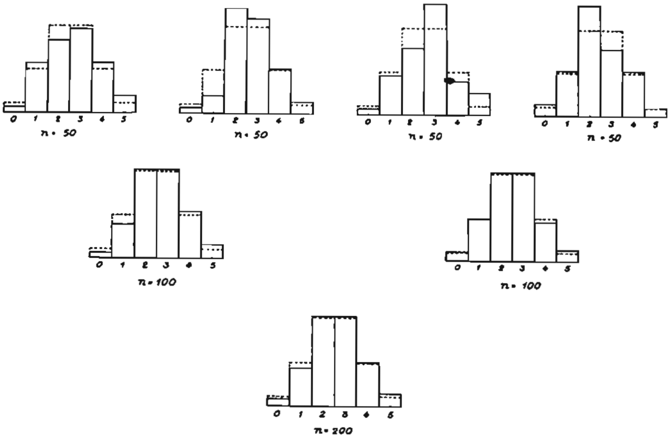


FIG. 1. HISTOGRAMS FOR NUMBER OF MALE CHILDREN IN SETS OF 5 BIRTHS
(n DENOTES THE NUMBER OF SETS)

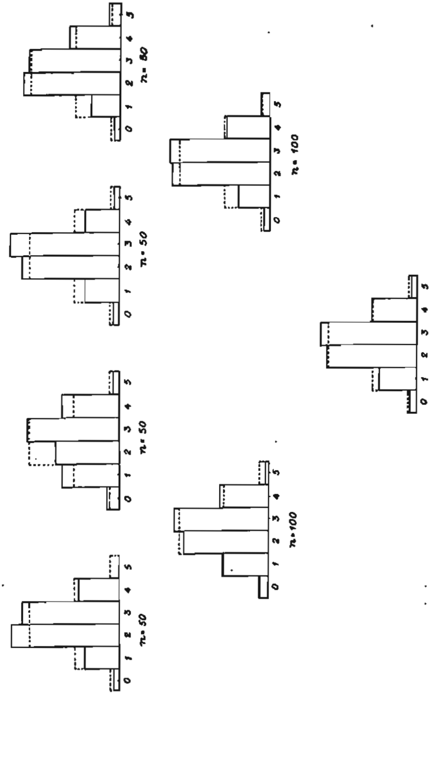


FIG. 2. HISTOGRAMS FOR NUMBER OF WHITE BEADS IN SETS OF 5 DRAWS
(n DENOTES THE NUMBER OF SETS)

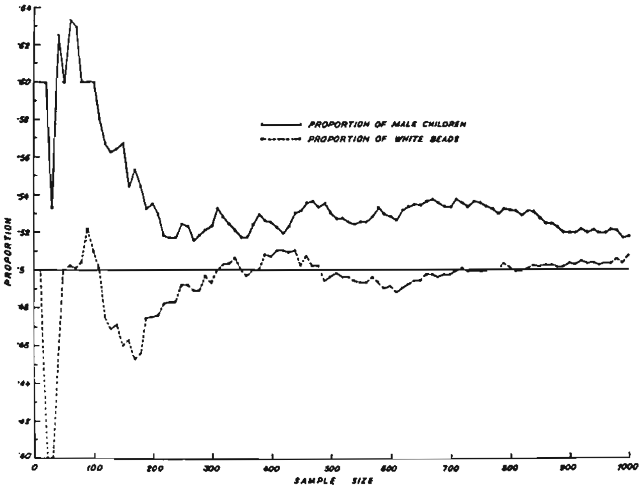


CHART 1. THE RELATIVE FREQUENCIES OF WHITE BEADS AND MALE CHILDREN FOR DIFFERENT SAMPLE SIZES

The limiting frequency. We shall analyse the two series in a slightly different way. From the first n terms of the series, the ratio $r_n = m_n/n$, where m_n is the total number of events of one kind, is computed for $n = 10,20, \ldots$. The graphs of r_n against n for the natural events (ratio of males) and for the artificial events (ratio of white beads) are shown in Chart 1. It is seen that in either case, the graph is characterised by large variations for small values of n, moderate variations for medium values of n and tendency to be constant with minor variations for large values of n. Another feature of interest is the tendency of the graph for artificial data to approach 1/2, suggesting a relationship between the limiting value and the proportion of white beads in the bag. Such a phenomenon could be demonstrated by altering the proportion of white beads in the bag and repeating the experiment. The limiting relative frequency in such case is expected to be the chosen proportion of white beads in the bag.

The graph for real data, while exhibiting the same general features as that for artificial data, tends to a limiting proportion slightly over half demonstrating the possibility of an excess of male over female children at birth.

Distinguishability between chance mechanisms. What kind of sequences would result if the proportion of black beads in the bag had been 1/4 instead of 1/2. A sequence of 500 observations was generated from such a mechanism and frequency distributions were obtained for the number of black beads in sets of 5 events separately for the first 50 sets and the second 50 sets. The results were as shown in Table 6.

TABLE 6. FREQUENCY DISTRIBUTIONS OF NUMBER OF BLACK BEADS

no. of black beads	frequency			
	1st 50 sets	2nd 50 sets	all 100 sets	expected value
0	9	13	22	23.730
1	23	20	43	39.551
2	11	9	20	26.367
3	6	8	14	8.789
4	1	0	1	1.465
5	0	0	0	0.098

Histograms were drawn in a similar way for parallel sets of 50 each and for all the 100 sets. They are compared with the theoretical histograms which are shown by dotted line in Figure 3. It is interesting to observe that histograms from parallel samples behave in a similar way and the general pattern is according to the theoretical expressions for a binomial distribution with probabilities 1/4 and 3/4 for the two kinds of events deduced below. The histograms in Figure 3 are clearly different in shape from those in Figure 2, which are based on equal proportions of black and white beads.

Thus we could distinguish between chance mechanisms by differences in shapes of histograms based on observed data and also identify a chance mechanism by comparing the observed histogram with the expected shape.

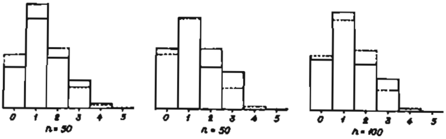


FIG. 3. HISTOGRAMS OF NUMBER OF WHITE BEADS IN SETS OF 5 DRAWS WHEN THE PROPORTION OF WHITE BEADS IS ONE-FOURTH (n = NUMBER OF SETS OF 5 DRAWS)

How do we define the theoretical histogram in the last experiment when the proportion of black beads is 1/4 ? This could be easily done by imagining that the bag contains beads of four colours in equal numbers and one colour is designated as black and rest as white. It is reasonable to make the hypothesis that any given sequence of colours in a set of 5 is equally likely to occur. Then a simple computation shows that the hypothetical frequencies are

$$\binom{5}{r}(1/4)^r(3/4)^{5-r}, \quad r = 0, 1, \ldots, 5$$

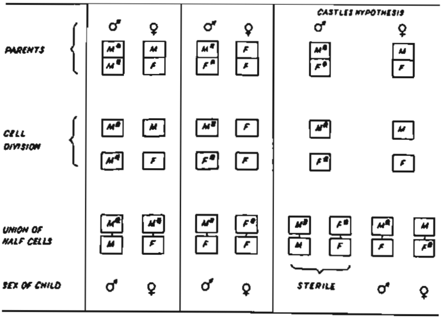
which is the binomial distribution when the proportion of black beads is 1/4.

Inference on the mechanism of sex determination. The analyses of real and artificial data (of Tables 1 and 2) indicate a common underlying structure, and since the mechanism of the latter is known it is reasonable to postulate a *similar* mechanism for the former, i.e., sex of a child is determined by a chance mechanism with equal probabilities for male and female.

It may also be noted that when the chance mechanism is known it is possible to work out the theoretical expectations to which observed events conform. So in practice, if an inference has to be drawn on an unknown mechanism of observed events, it is enough to examine to which theoretical expectations they show close agreement, instead of producing an artificial series for comparison.

Speculative theory of sex determination. At this stage, one could speculate on the physical model for sex determination, which produces a sequence of male and female children, analogous to a sequence of white and black beads in independent draws from a bag. Assuming that a child develops from a cell which is the union of half cells contributed by father and mother, the existence of two sexes shows that the cell of at least one parent is composed of two dissimilar half cells (like black and white beads). The three possible types of parental cells are shown in Table 7. Our inference on the mechanism of sex determination implies that the four possible combinations of half cells one from each parent are equally frequent. The three possible types of parental cells are all consistent with the inferred chance mechanism. Therefore, there is a need for collecting further evidence to choose one among them.

TABLE 7. THREE POSSIBLE HYPOTHESES ABOUT PARENTAL REPRODUCTIVE CELLS



● INDICATES THAT THE CELLS BELONG TO FATHER

Data on pedigrees whose members suffer from a rare malady like haemophilia throw some light on this problem. These data show that a woman may carry this disease without showing any symptoms and when married to a normal male produces male children half of whom on the average inherit the disease while the other half are completely free and female children half of whom on the average are carriers like herself while the other half are completely free. Or in other words a woman carrying but not exhibiting the disease produces sons of two different types and also daughters of two different types in equal numbers on the average. We shall examine which of the three possibilities in Table 7 support the observations.

Let us suppose that a woman is a carrier when the M half of her cell is defective under the models 1 and 3 of Table 7. Labelling this half as M_d we find that all sons receive M_d while no daughter receives M_d under model 1 and vice-versa under model 3. This contradicts observed facts that two types of sons and two types of daughters are born. Similar contradiction arises if the F half is assumed to be defective. The hypothesis that both M and F halves are defective is clearly untenable. There is only one possibility left, viz. that of model 2 in Table 6. It is easily verified by assuming one F to be defective in the femalo, that four types of children are possible with

$$\begin{array}{cccc} M^{\bullet} & M^{\bullet} & F^{\bullet} & F^{\bullet} \\ F & F_d & F & F_d \end{array}$$

equal frequency on the average. The first two types have the male sex and last two female sex. Then half the number of daughters are like the mother while the other half are defect-free. With the further hypothesis that a male is affected when the F half is defective we have two types of sons, one type exhibiting the disease and another completely free. Thus model 2 is consistent with observed data. We have demonstrated how by a purely statistical approach it is possible to probe into the constitution of male and female cells and establish an essential difference.

4. Scrutiny and editing of primary record : cross examination of data

In investigations where observations are recorded in the field and investigators have no chance of repeating the measurements, the problem of 'scrutinising and editing' the primary records assumes paramount importance. The magnitude and the proportion of recording errors are sometimes extremely large in field investigations and the analysis based on such data may result in wrong conclusions. The importance of scrutiny or what Fisher calls, cross examination of data, must be emphasized right at the start of statistical training.

The Indian Statistical Institute undertakes the processing of huge volume of socio-economic data collected from all over India by the National Sample Survey. In addition, the scientists at the Institute produce large mass of data from laboratory investigations, which are subject to statistical analysis. Thus, there is enough live material for the students to worry their heads.

Scrutiny is indeed a difficult job. No definite rules can be laid down in deciding whether a recorded figure is correct or not without having the opportunity to check it from the original source. The students have to learn mostly from experience. Reading of some published papers on the subject will help the student in understanding the problem and in providing broad guide lines.

There is an excellent paper by Mahalanobis (1933), who used very ingenious methods in rectifying the errors in Risley's published record of anthropometric measurements, which was considered to be full of inconsistencies and discrepancies by anthropologists. Other publications by Mahalanobis, Majumdar and Rao (1949), Mukherji, Rao and Trevor (1955) and Majumdar and Rao (1958) contain extensive accounts on scrutiny of data.

Attempts have been made, in a recent investigation at the Indian Statistical Institute, to use the computer in the scrutiny of data. Programs for detecting outliers (possibly recording errors), investigator bias, incomplete information etc., written by the students were used for the purpose. Students find such projects extremely interesting and stimulating. Often there is competition among students in detecting errors in primary records.

A statistician should also acquire skill in examining whether data quoted by others are genuine or faked. To bring out the difference between these two types of data, each student in a class of 20 was asked to write down a series of B's and W's imagining the occurrence of events in 50 independent drawings from a bag containing black and white beads in equal numbers. There were 1000 observations providing an imaginary series (which we shall label as 1) from which a frequency distribution of the number of black beads in sets of 5 was obtained. The frequency distributions for the imaginary data (1) and artificial data of Table 2 are compared in Table 8. The imaginary data (1) contained more sets with 2 or 3 black beads than expected under a chance mechanism so that students were biased towards sets in which the difference between black and white beads is small. At this stage one of the students was shown the expected values and asked to write a series of 1000 observations to conform to a binomial distribution. The frequency distribution obtained from imaginary data (2) is also given in Table 8. The tendency, when the expected values were known, was to produce a frequency distribution which agreed more closely with the expected values than was normally possible under a chance mechanism. Generally, faked data are detected by systematic bias towards certain kinds of event and/or by too close an agreement with an assumed hypothesis. An article by Fisher (1936) illustrates such a case from published data.

TABLE 8. FREQUENCY DISTRIBUTION OF ARTIFICIAL AND IMAGINARY DATA

no. of black beads	frequency			
	artificial data of Table 2	imaginary data (1)	expected value	imaginary data (2)
0	5	2	6.25	6
1	27	20	31.25	32
2	64	78	62.50	63
3	68	80	62.50	61
4	32	17	32.25	32
5	4	3	6.25	6

5. Errors of measurements: concept of true value

Study of natural sciences offer an excellent scope for students to get acquainted with errors of measurements, errors due to limitations of measuring instruments and due to investigator bias. Some possible exercises in Physics and Chemistry are given below.

Rods differing slightly in length would be supplied and students would be required to measure the lengths of these rods. The object of the experiment is to compare the results obtained by the same student on different days, or by different students, or by different methods and also to find what is the smallest difference in the lengths of two rods which can be distinguished by such measurements. The experiment has to be carefully designed. Such excercises would supply a basis for the 'operational' aspect of the concept of length.

The data provided by such an experiment would be ideal for illustrating simple statistical techniques—comparison of means and variabilities between students, testing for differences between rods, comparing the efficiencies of different methods of measurement. Hagen's hypothesis about the symmetry of error distribution can be examined through the computation of the third moment, or Pearson's β_1 coefficient. The data will be useful for more sophisticated treatment by analysis of variance later in the course.

Experiments with simple or Krater's pendulum, calorimeter, spectrometer etc, will provide data of similar but slightly more complicated nature for statistical analysis.

Calibration of burette and pipettes, use of balance, preparation of standard solutions and testing of Avogadro's law are some exercises which might involve the students in understanding principles of design of experiments and testing of hypotheses.

6. An investigation in biology: emphasis on quantitative approach

It is not generally known that a coconut tree can be classified as left handed or right handed depending on the direction of its foliar spiral. Some years ago an investigation into the study of this aspect of coconut trees was undertaken by T. A. Davis, Professor of biology at the Indian Statistical Institute. The results are briefly summarised, as the investigation is a good example of statistical approach to a biological problem, worthy of discussion in a statistical course. The questions raised and the evidence provided by observations are as follows. This provides ideal material for introducing the χ^2 and t tests and their limitations.

Is left and right foliar spirality genetically inherited ? The question can be answered by considering parent plants of different combinations of foliar spirality and scoring the progency for the same characteristic. The data collected for this purpose are given in Table 9.

TABLE 9

pollen parent	seed parent	progeny		proportion of left to total
		left	right	
right	right	28	35	0.44
right	left	32	36	0.47
left	right	20	24	0.45
left	left	14	16	0.47

It is seen from the table that the proportion of plants with left foliar spiral does not depend on the type of parents and cousequently there is no genetic basis for the determination of left or right spirality.

However, the deviation of the overall proportion of left to total, is somewhat less than half. This could not be explained until data from various parts of the world could be collected. Table 10 gives the numbers of plants with left and right spirality from 22 countries in the Northern hemisphere and all countries in the Southern hemisphere.

TABLE 10

hemisphere	left	right	proportion of left to total
north	18968	17843	0.515
south	4090	4540	0.473

It is seen that the proportion of left-plants is more than half in the Northern hemisphere and less than half in the Southern hemisphere, which may be the influence of one-way rotation of the earth, as in the explanation of the phenomenon of the bath tub vortex which under well-controlled conditions is shown to be counter-clockwise in the northern hemisphere and clockwise in the southern hemisphere (Shapiro, 1962).

A more exciting part of the investigation is the difference in yield rates of the left and right trees. The figures of annual yields of nuts are given in Table 11.

TABLE 11

category		no. of trees	mean yield	
			12 years (1949–60)	6 years (1955–60)
healthy	L	58	57.69	65.60
	R	70	49.82	54.28
early diseased	L	60	32.95	36.54
	R	66	30.55	33.10
late diseased	L	56	22.05	23.63
	R	64	20.04	20.33

It appears that a left palm tree gives about 10 per cent more yield than the right palm tree, a conclusion of great economic importance though unexplainable at the present stage of investigation.

7. Some practical exercises : multidisciplinary approach

As I have mentioned earlier, the late Professor J. B. S. Haldane took an active interest in discussions on the nature and content of the B. Stat. course. He had suggested a set of practical exercises of an interdisciplinary nature for students in statistics, which are given below.

(1) After previous training in surveying and determinations of elasticity, a tree will be strained by a rope stretched to a neighbouring building, and its deformation observed with a theodolite. The observations will be repeated in a high wind. This may be regarded as an exercise in (a) surveying, or applied trigonometry, (b) statistics, (c) quantitative biology, (d) meteorology.

(2) Before chemical balance is systematically used, its theory will be studied. Its period of oscillation at different loads will then be measured, and the theory verified.

(3) Before the compound microscope is used, its theory will similarly be studied, and the performance of the instrument used will be calculated from optical principles. The theory will then be verified.

(4) Volumetric gas analysis will be taught. The results will be used (a) to verify Avogadro's law (b) to measure human respiration at rest and at work, (c) to analyse coal mine air, determining methane, carbon dioxide, and oxygen. This introduces the notion of chemical controls for the prevention of industrial accidents.

(5) The use of a flame spectrometer will be taught. It should be possible to use the same instrument for (a) physical measurements, (b) soil analysis (in conjunction with experiments on the growth of plants in different soils), (c) ore analysis, (d) human blood plasma analysis.

While many of the biological exercises will not lead themselves to cooperation with the Physicists and Chemists, they will yield data suited for statistical analysis. This is most obvious in the case of the genetical course. But a few other examples are given.

(1) The class will carry out simple anthropometric measurements on one another. These will be used for the calculation of means, variances and correlations.

(2) They will study life cycles in a frog and a silk worm moth. Daily counts of survivors will enable them to construct life tables.

(3) Observations on a simple and rapid piece of animal behaviour, e.g., the successive ascents of a koi fish for air, will be made in such a way as to furnish data for the estimation of time trends and serial correlations. Five hours' continuous observation will give a series of about 80 intervals, which can be treated by the methods used for rainfall records over 80 years.

Throughout the biological and chemical courses at least, a psychologist will attend whenever quantitative data are obtained, which make it possible to compare the performances of different students, e.g., the accuracy with which they make up standard solutions. This will allow them to assess their own and each others' aptitudes, and may obviate the need for practical examinations.

References

Davis, T. A. (1962) : The non-inheritance of asymmetry in *Cocus nucifera*. *J. Genetics*, 58, 42-50.

——— (1963) : The dependence of yield on asymmetry in coconut palms. *J. Genetics*, 58, 186-215.

——— (1968) : Biology in the tropics. *Haldane and Modern Biology*, 327-33, Johns Hopkins Press, Baltimore.

Fisher, R. A. (1936) : Has Mendel's work been rediscovered ? *Annals of Science*, 1, 115-137.

Mahalanobis, P. C. (1933) : A revision of Risley's anthropometric data relating to the tribes and castes of Bengal. *Sankhyā*, 1, 76-105.

——— (1965) : Statistics as a key technology. *American Statistician*.

Mahalanobis, P. C., Majumdar, D. N. and Rao, C. R. (1949) : An anthropometric survey of the United Provinces, 1941—A statistical study. *Sankhyā*, 9, 90-324.

Majumdar, D. N. and Rao, C. R. (1958) : Bengal anthropometric survey, 1945—A statistical study. *Sankhyā*, 19, 201-408.

Mukherji, R. K., Rao, C. R. and Trevor, J. (1955) : *The Ancient Inhabitants of Jebel Moya*, Cambridge University Press.

Shapiro, A. H. (1962) : Bath tub vortex. *Nature*, 196, 1080-1081.

Paper received : July, 1969.