Agriculture and Applied Statistics – I

Book · March 2007

CITATIONS
0

READS
1,459

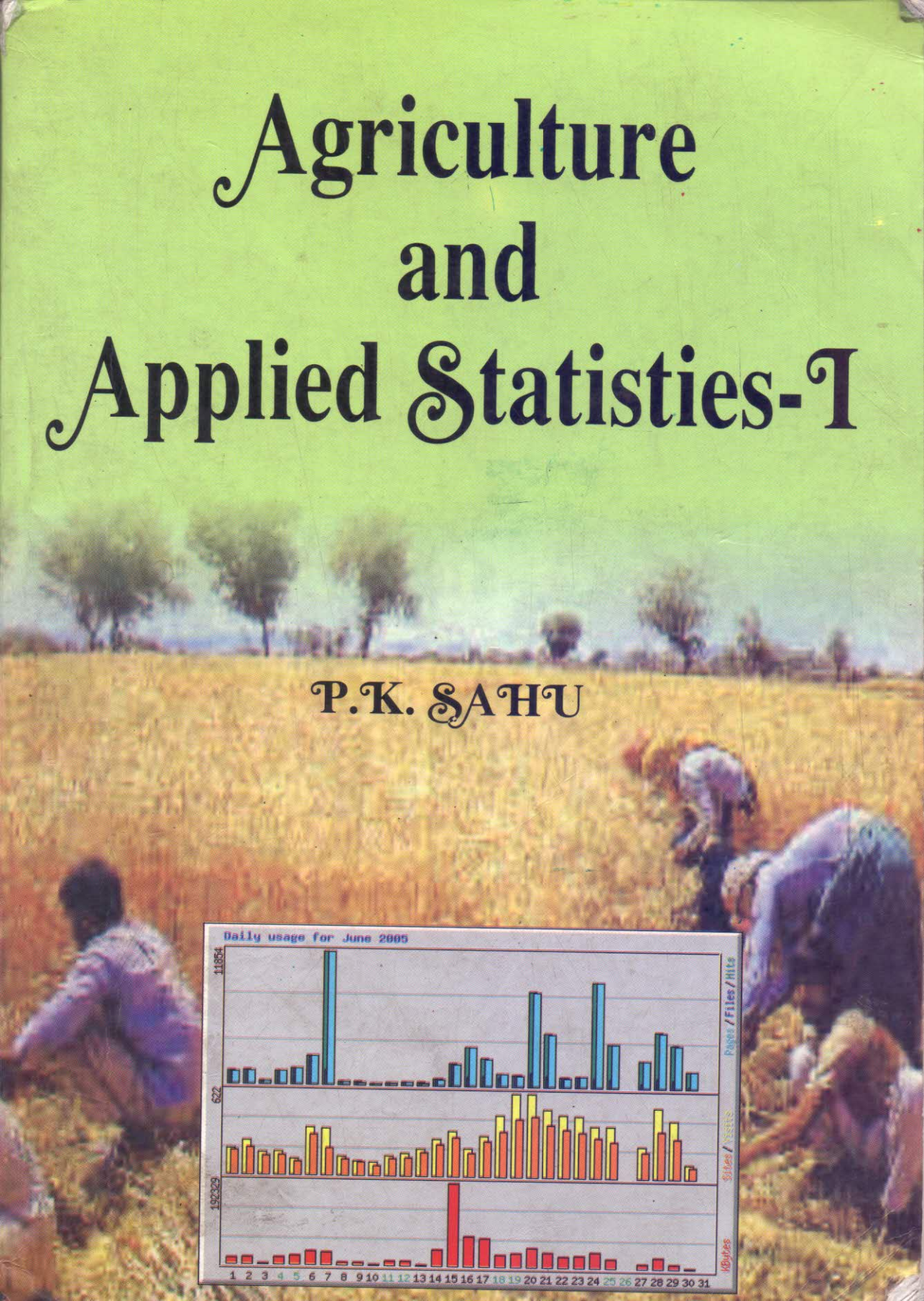
1 author:

Pradip KUMAR Sahu
Bidhan Chandra Krishi Viswavidyalaya
111 PUBLICATIONS **375** CITATIONS

SEE PROFILE

Agriculture and Applied Statisties-I

P.K. SAHU



CONTENTS

Ch. No.	Chapter	Pages
1.	Introduction to Statistics and Its Application	1-7
2.	Information, Processing of Information and its Presentation	8-27
3.	Measures of Central Tendency and Locations	28-51
4.	Measures of Dispersion, Skewness and Kurtosis	52-76
5.	Probability Theory and Its Applications	77-110
6.	Theoretical Probability Distributions and Their Applications	111-154
7.	Sampling Techniques and Their Applications	155-184
8.	Statistical Inference and Its Application	185-270
9.	Correlation Analysis and Its Application	271-302
10.	Regression Analysis and Its Application	303-379
	Questions	380-396
	Reference	397-401

1 INTRODUCTION TO STATISTICS AND ITS APPLICATION

1.1. INTRODUCTION

Long since the term statistics is being used for various purposes in various ways. It is believed that the term statistics was mostly used in the ancient days to understand the political state. In India from Kautilya's Arthashastra usage of statistics is recorded. Most probably the German scholar Gottfried Achenwall coined the word statistics. In any case the word statistics is being used knowingly or unknowingly since time immemorial.

The word statistics is being used in two different senses. The first one implies the presentations of some facts and figures or information-usually known as 'data'. Runs scored in different test matches by Sachin Tendulkar in his test career or the budget presented by the Finance Minister at the Parliament are the examples of statistics used in the sense of data. On the other hand, statistics can be looked as the body of science, which deals with principles, techniques, collections, scrutiny, analysis and drawing inference on a subject of interest. Thus, in brief statistics in singular is a branch of science while statistics in plural means data.

1.2. USE AND SCOPE OF STATISTICS

Starting from the ancient age to the modern times not a single area can be found where statistics is not playing a vital role. Starting from agriculture, biology, education, economics, business, management, medical, engineering, psychology, environment, and space even in the management of war statistics is playing vital roles. Hardly there exists any human activity where statistics is not indispensable.

1.3. SUBJECT MATTER OF STATISTICS

The subject matter of statistics is the study of population rather than the individual unit of the population. Statistics deals with aggregated information on a particular subject in which

2 INFORMATION, PROCESSING OF INFORMATION AND ITS PRESENTATION

2.1. INFORMATION

2.1.1. As has been mentioned one of the ingredients of statistics is the information/data. Information can be of two types:

(*i*) Primary information (data)

(*ii*) Secondary information (data).

Primary information or data refers to the information collected through a pre-designed experiment and/or using fixed sampling technique with a specific objective. For example, if we want to know the intensity of infestation of different pests in a varietal trial of paddy one has to think for setting up of an appropriate field experimentation following a particular design of experiment at appropriate time and then to record the information from the experimental field. Thus, primary data are collected by the experimenter (user) for specific purposes following definite procedure. On the other hand, **secondary data** are those data collected by some agency or user for specific purposes but later on are being used by some other persons or some other agencies for some other purposes. For example, the census carried out after every 10 years gives an in-depth view of the Indian population. But same information is not available for any period between two consecutive censuses. Any one using these census data may be termed as secondary data to the user. The user himself or herself is not involved in planning or collection of data. In many situations data generated by different national and international agencies like Central Statistical Organisation (CSO), National Sample Survey Organization (NSSO), State Planning Board (SPB), Food and Agriculture Organisation (FAO), World Health Organisation (WHO) etc. are used by various researchers or users; to the users these data are secondary data. Secondary information required to be verified about the nature of the data *i.e.* their coverage, the reliability of the data before it is being used.

2.1.2. Information or data are collected on some characters.

Characters are of two types :

(*a*) Qualitative characters and **(*b*) Quantitative characters.**

3 MEASURES OF CENTRAL TENDENCY AND LOCATIONS

In the previous chapter we have seen how to extract and present first hand information from the raw data. With the help of the example of stem borer infestation we have made the frequency distribution table and calculated different characteristics (frequency, cumulative frequency, relative frequency, frequency density etc.). In the same chapter we have also mentioned about the infestation of gallmidge paste in paddy field. Now the question is how to know which pest has infested more or which pest has got lesser infestation. In other words we need to have certain measures by which we can compare the two infestations and conclude which pest is more intense. Moreover, human instinct is to find out certain value(s), which can represent the set of information given in a big table or otherwise. Thus we are always in search of such a measure, which can describe the inherent characteristics of a given set of information.

For example, average height of Indian adult male is160 cm (this dose not mean that all the Indian males are of the height of 160 cm.), which indicates that given any Indian male, his height will tend to lie around the value of 160 cm. Thus given a set of data, we are in search of a typical value below and above which the observations tend to cluster around. Thus, tendency of the observations to cluster around a central value is known as central tendency.

Now, the question is how to measure the central tendency. There are different measures of central tendency *viz*. the mean, median and mode. These are sometimes known as averages.

Before discussing the different types of averages let us try to examine the desirable properties of a good average.

According to Yule, a good average should have the following characteristics:

(*i*) Should be defined rigidly without any ambiguity.

(*ii*) Should be based on all the observations.

(*iii*) Should be easy to calculate.

(*iv*) Should be easy to understand.

(*v*) Should be readily acceptable to mathematical treatments.

(*vi*) Should be least influenced by sampling fluctuations.

In defining a measure of central tendency there should not be any ambiguity. A measure should be clearly and convincingly defined. It is better that it takes care of all the observations which provides a single value to represent a sample or population. A measure, which is difficult to calculate or understand, is least used. As has been mentioned earlier measures are generally used to represent a whole set of observation, so it should be least affected by sampling fluctuation.

3.1. MEAN

Means are of three different types, Arithmetic mean, Geometric mean and Harmonic mean.

(*i*) Arithmetic mean

Arithmetic mean of a set of observations is simply their sum divided by the number of observations and is denoted by $\bar{x}$ or μ, *i.e.* $\bar{x} = \mu = \frac{\sum_{i=1}^{n} x_i}{n}$

where $x_1, x_2, \ldots\ldots x_i \ldots\ldots x_n$ are the values of first, second, third……up to n^{th} observation.

Thus, $\bar{x}$, for the data of stem borer infestation from table 3.1 is $\frac{6+5+\ldots\ldots\ldots+\ldots\ldots+8}{100} = 11.76$ insects/plant.

Table 3.1 : Infestation of stem borer in 100 varieties of paddy.

Variety	*Stem-borer*	*Variety*	*Stem-borer*	*Variety*	*Stem-borer*	*Variety*	*Stem borer*
1	6	26	17	51	18	76	16
2	5	27	22	52	20	77	26
3	10	28	6	53	23	78	20
4	11	29	8	54	9	79	21
5	12	30	16	55	13	80	19
6	13	31	12	56	14	81	7
7	14	32	13	57	8	82	9
8	7	33	14	58	7	83	14
9	8	34	7	59	9	84	22
10	9	35	8	60	13	85	23
11	10	36	11	61	7	86	14
12	11	37	17	62	8	87	26
13	14	38	9	63	11	88	18
14	16	39	10	64	14	89	20

4 MEASURES OF DISPERSION, SKEWNESS AND KURTOSIS

4.1. DISPERSION AND ITS MEASURES

It is our common experience that there are certain varieties of a particular crop, which are very responsive to doses of inputs, and others are not. If these input responsive varieties are provided with high dose of nutrient they come up with very good yields, on the other hand if these varieties are put under input stressed condition then their performance will be very poor. Let us take the following example:

Example 4.1. *Two varieties of rice were put under field trial in ten different farmers plots of same soil conditions. Farmers put input according to their ability and the yields obtained in farmers fields for the two varieties are as follows :*

	Yield (q/ha) obtained in ten farmers' field									
Variety A	*15.00*	*16.50*	*16.00*	*18.00*	*17.00*	*18.50*	*19.00*	*18.50*	*17.50*	*14.00*
Variety B	*10.50*	*21.50*	*22.50*	*9.00*	*24.50*	*11.50*	*23.00*	*20.00*	*17.00*	*10.50*

From the above, one can have the arithmetic means, $\overline{X}_A$ = (15 + 16.5 + ... +14)/10 =17q/ha and $\overline{X}_B$ = (10.50 + 21.50 + ... + 10.50)/10 = 17 q/ha. Thus central tendency, measured in terms of arithmetic mean for the above two varieties are same. But a critical examination of the data reveals that in variety A the yield remains in between 14 q/ha to 19 q/ha on the other hand in variety B it remains between 9 q/ha to 24.5 q/ha. Thus the yield observations are more dispersed in variety B than in variety A, in spite of having same average performance for both the varieties. Thus, from the above results two conclusions can be drawn: (*a*) that maximum yield potentiality and yield variability of variety A is less than the variety B, (*b*) given a better situation more harvest can be obtained from variety B by utilizing its full potentiality.

Example 4.2. ***Runs scored by two cricket players in ten different matches are given as follows:***

	Match1	*Match2*	*Match3*	*Match4*	*Match5*	*Match6*	*Match7*	*Match8*	*Match9*	*Match10*
Player A	*78*	*46*	*64*	*70*	*64*	*67*	*50*	*66*	*50*	*45*
Player B	*115*	*4*	*14*	*25*	*85*	*78*	*99*	*88*	*40*	*52*

It is found that both the batsmen have scored 60 runs on an average (arithmetic mean). But the variability in scoring is more for player B (4-115) than for player A (45-78).

Therefore, from the above two examples it is clear that the measure of central tendency always may not provide an in-depth picture of the data. Variability / scatteredness of the observations of a given set of data is also equally important to extract the inherent characteristics of a variable. **Tendencies of the observations of any variable to remain scattered / dispersed from a central value is known as dispersion of the variable**. So, to some extent, central tendency and dispersion are the two opposite phenomena (characteristics) of a variable.

4.1.1. Measures of Dispersion

Like the measures of central tendency, there is a need to have measures for dispersion also. In fact different measures of dispersions are available in the theory of statistics. Before going details in to the discussion of different measures of dispersion, let us try to examine the characteristics of a good measure of dispersion.

By and large a good measure of dispersion should have the following characteristics :

(*a*) A good measure of dispersion should be defined clearly.

(*b*) A good measure of dispersion should be convincing and easy to understand.

(*c*) A good measure of dispersion should be easy for calculation and mathematical treatment.

(*d*) A good measure of dispersion should be based on all observations

(*e*) A good measure of dispersion should be least affected by the sampling fluctuations.

(*f*) A good measure of dispersion should not be affected badly by the extreme values.

There should not be any ambiguity in defining a measure; it should be clear and rigid in definition. Unless a measure is convincing *i.e.* easily understood and applicable by the user it is of least importance. For further application of a measure, it should be put easily under mathematical treatments. In order to reflect the true nature of the data a good measure should try to take care of all the observations and it should lay equal importance to each and every observation without being affected by the extreme values.

Measures of dispersion can broadly be categorized in to two groups *viz.* (*i*) the **absolute** measures and (*ii*) the **relative** measures of dispersion. Absolute measures of dispersion have the units according to those of variables but relative measures are pure number; as such are unit free measures. Thus unit free measures can be used to compare distributions of different variables measured in different units. Among the different types of absolute measures, **range,**

5 PROBABILITY THEORY AND ITS APPLICATION

Statistics deals with variability and variability refers to a variable. In 2nd chapter we have discussed that a variable associated with chance factor is called variate. Thus, daily rainfall amount is variable but when we say the chances of getting 30mm daily average rainfall in a particular month then it becomes a variate. Infact in statistics we deal with variate. In previous sections we have discussed the variable part of the variates, in this section we shall deal the chance/probability part of the variates.

It is our common experience that the farmers often enquire about the chances of rain at a particular day in which day he is supposed to sow seeds or spray plant protection chemicals etc.; because occurrence of rain has tremendous impact on the operations of crop cultivation and ultimately the success of crop. There exists certain degree of uncertainty. Farmers calculate the chances of 'rain' or 'no rain' with his or her experience/belief etc. The section of statistics which deals with such uncertainty/chances come under the theory of probability. To know the theory of probability, it is useful to have some knowledge of set theory. As such we shall discuss the set theory first and then proceed for discussion of probability.

5.1. SET THEORY

5.1.1. Set

A set is a collection or aggregation of well defined objects/entities. The objects should have common and specific properties according to the rule of definition. For example, when we say a set of odd integers between 1 and 100, definition of odd integers is clear cut and having definite properties like 'not divisible by 2' and all the odd numbers 1, 3, 5, ..., 99 are to be included in the set. Similarly, when we say a set of 'insecticides', then definition and characteristics of insecticides are given and any chemical having the above defined properties should be included in the set.

The individual member of set is called its **element/member/point** of the set. Thus, the integers '1'/ '3'/ '5' etc are the members/points/elements of the set of odd integers between 1 and 100. Generally we use capital letters A, B, C.....X, Y, Z to denote sets and small letters *a, b, c, ...x, y, z* to denote elements. If x is a member of a set 'A' then we express this fact symbolically as $x \in A$ where, $\in$ stands for 'belong to' or '*is an element of*', x is a member or element of the set 'A'. Thus $x \in A$, where,

$A = \{x : 1 \leq x$ (odd integers) $\leq 100\}$ then $3 \in A$ but $4 \notin A$ *i.e.* 4 does not belong to the set A.

A set may be either of the two different types viz. finite set or infinite set. A set containing no element or a finite number of elements is called a **finite set** *e.g.* the set of all statistics book available in the university library. An **infinite set** is a set in which the number of elements is infinite *e.g.* the set of stars in the galaxy, set of all insects, set of all points between 0 and 1.

An infinite set may be **countably infinite** if the elements of the infinite set can be written in some order *i.e.* the infinite set can be put in one-to-one correspondence with the set of natural numbers (1, 2, 3, ...) *e.g.* the set of all odd integers *i.e.* $A = \{x$: all odd integers $\} = \{1, 3, 5, 7$ .....$\}$

A set is called **uncountably infinite** if it is neither finite nor countably infinite *e.g.* $A = \{ x \leq 4 = x \leq 5\}$, we have uncountably infinite numbers between 4 and 5.

Equal set : Two sets 'A' and 'B' are said to be equal if every element of the set 'A' is also an element of the set 'B' and vice-versa. Thus if $A = \{1, 2, 5\}$ and $B = \{2, 5, 1\}$, we write $A = B$.

Null set : A set having no element is called an empty set or null set and is denoted by ϕ.The set $\{0\}$ is not a null set as it contains the element zero. The set of negative integers between 2 and 3, the real roots of $x^2 + 54 = 0$, $A = \{x : x$ is a perfect square of an integer, $27 < x < 35\}$, etc. are the examples of null sets.

Sub set : If every element of a set 'A' is an element of another set 'B' then the set 'A' is called the subset of the set 'B' and it is written as $A \subseteq B$.

Let $A = \{x : 0 < x$ (integer) $\leq 2\} = \{1, 2\}$ and

$B = \{ x : 0 < x$ (integer) $\leq 2\} = \{0,1,2\}$ then $A \subseteq B$.

The subsets of B are ϕ, $\{0\}$, $\{1\}$, $\{2\}$, $\{0, 1\}$, $\{0, 2\}$, $\{1, 2\}$ and $\{0, 1, 2\}$. It is to be noted that the null set is a subset of every set and if a set has n elements then it has 2^n subsets.

We know that pesticides include insecticides, herbicides, fungicides, nematocides etc. If we have two sets A and B such that $A = \{ x$: all herbicides and fungicides$\}$ and

$B = \{x$: all pesticides$\}$ then $A \subseteq B$ (A is a subset of B).

Proper subset: A set 'A' is said to be the proper subset of another set 'B' if all the elements of the set A must belong to the set B but there is at least one element in B which does not belong to A.

Symbolically $A \subset B$ but $A \neq B$.

Let $A = \{1, 2, 3, 4, 5\}$ and

$B = \{0, 1, 2, 3, 4, 5, 6, 7\}$

6 THEORETICAL PROBABILITY DISTRIBUTIONS AND THEIR APPLICATIONS

The objective of statistics is to study the population characteristics because time, accessibility, monetary, and other feasibility constraints it is not always possible to study the population characteristics as such. We take representative part of the population called 'Sample' - study the sample characteristics and infer about the nature of the population from the nature of the sample- 'statistical inference'. The statistical measures like averages SD coefficient of Skewness, Kurtosis etc. what we get from the sample frequency distribution helps us in getting some idea about the nature/ characteristics of the population based on sample observations. If we take into account that, probability is some what limiting form of the relative frequency in frequency distributions. It is natural to represent different values taken by the random variable in terms of probability distribution. Thus we are in search of a functional form (function), which can give us the probabilities at different values of the random variable. That means we are in search of a mathematical form of probability law, which gives us distribution of population variable- known as theoretical probability distribution. These are called 'theoretical' because these are ideal distribution and are hardly expected to reflect in toto the true nature of population distribution. These are meant for a very close approximation of the actual distribution of the population variable. Theoretical distributions are **Univariate, Bivariate** and **Multivariate** depending upon one, two or more variables for which probabilities are considered in distribution. Again, depending upon the nature of the variable a probability distribution may be **discrete** or **continuous probability distribution.** In the following sections we shall discuss some univariate discrete and continuous probability distribution in brief.

6.1. DISCRETE PROBABILITY DISTRIBUTION

6.1.1. Binomial Distribution

Let a random experiment be performed 'm' independent Bernoulli trials each having 'success' or 'failure' with respective probabilities 'p' and 'q' respectively. Then the random

variable X, number of successes out of 'm' trials, is said to follow binomial distribution and its p.m.f. is given by

$$P(X = x) = P(x) = \left\{\binom{m}{x} p^x q^{m-x}\right\}, x = 0, 1, 2, \ldots, m; q = 1 - p.$$

$$= 0 \text{ otherwise.}$$

'm' and 'p' are known as the parameters of the distribution and the distribution is denoted as $X \sim b(m, p)$ *i.e.* the random variable follows binomial distribution with parameters 'm' and 'p'.

Obviously $P(x) \geq 0$

and $$\sum_{x=0}^{m} P(x) = q^m + \binom{m}{1} q^{m-1} p + \binom{m}{2} q^{m-2} p^2 + \ldots + p^m = (q + p)^m = 1$$

Moments of Binomial distribution

As we know that r^{th} order raw moment about the origin is given by

$$v_r = E(X^r) = \sum_{x=0}^{m} x^r \binom{m}{x} p^x q^{m-x}$$

putting $r = 1, 2, 3, \ldots\ldots$ we shall get different moments about origin.

Thus, $$v_1 = \sum_{x=0}^{n} x \binom{m}{n} p^x q^{m-x}$$

$$= \sum_{x=0}^{m} x \frac{m!}{x!(m-x)!} p^x q^{m-x}$$

$$= mp \sum_{x=0}^{m} \frac{(m-1)!}{(x-1)!\{(m-1)-(x-1)\}!} p^{x-1} . q^{\{(m-1)-(x-1)\}}$$

$$= mp \sum_{x=1}^{m} \binom{m-1}{x-1} p^{x-1} q^{(m-1)-(x-1)} = mp \sum_{y=0}^{m-1} \binom{m-1}{y} p^y q^{(m-1-y)}$$

[where, $y = x-1$]

$$= mp\ (q + p)^{m-1} = mp$$

Similarly,

$$v_2 = E(X^2) = E[X\ (X - 1) + X] = E[X\ (X - 1)] + E\ [X]$$

$$= \sum_{x=0}^{m} (x(x-1)) \frac{m(m-1)(m-2)!}{x(x-1)(x-2)!(m-x)!} p^x q^{m-x} + E[X]$$

$$= \sum_{x=2}^{m} \frac{m(m-1)(m-2)!}{(x-2)!(m-x)!} . p^x q^{m-x} + mp$$

7 SAMPLING TECHNIQUES AND THEIR APPLICATIONS

7.1.

In our daily life we are quite familiar with the word 'sample'. If we go to market to buy wheat we ask the retailer to show the sample. The retailer shows a handful of wheat from a stock of huge amount of wheat (say 100kg). We check the sample for its quality assuming that the quality of wheat from which we are supposed to buy (the population) a certain amount of wheat (say 10kg) will be more or less same as that of the sample. If the sample shown to the buyer is not proper representative part of the population then it may lead to wrong decision with regard to buying of the commodity. To avoid this problem we may think for checking the whole population *i.e.* checking the quality of each and every grain of wheat! Simply this is not possible; mainly because of time, labour, and cost involvement. Sometimes, it is not possible also to identify each and every member of the population (infinite population). Same thing is also true for any statistical investigation. Thus we need to have a proper sample following a statistical technique so as to obtain valid inferences about population characteristics based on sample observations and avoid taking any wrong decision.

With the above pretext the whole sampling theory has been developed. Sampling theory can be visualized as consisted of mainly three major components: (*a*) Selecting random sample, (*b*) collection of information from the samples and (*c*) analysis of information to obtain inferences about the population as a whole. Before discussing the above three components, let us have a look on the definition and characteristics of both the 'population' and 'sample'.

7.1.1. Population

A population is a collection or totality of well defined objects (entities) with which a statistician is interested. The observations or entities could refer to anything like persons, plants, animals, objects (like nut bolts, books, pens, pencils, medicines engines etc). Individual member of the population is known as **element or unit** of the population. **Size** (N) of the population is generally referred to the number of observations in the population.

Depending upon the size of the population, a population may be finite or infinite. **A finite population** is a population if it contains finite number of observations *e.g.* germplasms of wheat, a particular type of machine in an industrial survey, a commodity at a particular point of time in different markets. Similarly, **an infinite population** is a population if it contains infinite number of observations /units. For example number of fishes in a particular river, number of stars in a galaxy etc.

A character y is defined on the population and Y_i is the value of the character for i^{th} unit ($i = 1, 2, 3, \ldots N$) of the population. For example the unit may be farm and the character may be the area under a particular crop or the unit may be germplasms of wheat and the characteristic may be the plant height. Now we define a **parameter** as a real valued function of the population values only. For example the population mean $= \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, Population variance $= \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \bar{Y})^2$;

Population coefficient of variation $= C_Y = \frac{\sigma_Y}{\bar{Y}}$ etc.

7.1.2 Sample

A sample is a representative part of population. Thus a sample is a subset of population. **Size (n)** of the sample is number of elements / units with which the sample is constituted of. Definitely, the sample size $n < N$, the population size. There is no hard and fast rule, but generally a sample is recognized as **large sample** if the sample size $n \geq 30$, otherwise **small sample.** If the sample fails to represent the population adequately then there is every chance of drawing wrong inference about the population based on such sample because of the fact that it will overestimate or under estimate some population characteristics. Let us suppose that we want to know the average height of the students of a college. If the college is coeducation college and one draws (*i*) a sample of either boys or girls only, or (*ii*) from a particular class, then the average height obtained from the sample may fail to infer about the true average height of the students of the college (the population). This type of sample is called **biased sample.** On the other hand, an **unbiased sample** is statistically almost similar to its parent population and thus inference about population based on this type of sample is more reliable and acceptable than from biased sample.

Suppose a sample of n units is selected from a population of N units according to some probability scheme and let the sample observations be denoted by $y_1, y_2, y_3, \ldots y_n$. Now we define a **statistic** as a real valued function of the sample values only. For example the sample mean $= \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, sample variance $= s'^2_y = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$; sample coefficient of variation $= c_y = \frac{s'_y}{\bar{y}}$ etc.

8 STATISTICAL INFERENCE AND ITS APPLICATION

8.1.

The main objective of statistics is to study the population behaviour or to draw the inferences about the population from the sample observations. As samples are the part of the population there are tendencies of difference in sample behaviour with that of population behaviour. So the question of how accurately, efficiently the population behaviour can be obtained from sample is always there. Thus the process of knowing the unknown population behaviour from the statistical analysis of the sample behaviour is known as **Statistical Inference.**

In statistical inference mainly there are two type of problems :

(*a*) Some or all information about the population may be unknown and one may be interested to guess or estimate those from the observations. This is known as the problem of **estimation.**

(*b*) Some information or some hypothetical values about the population parameter may be known or available but it is required to be tested how far these information or hypothetical values are acceptable or not acceptable in the light of the information obtained from the sample supposed to have been drawn from the same population. This is known as **testing of hypothesis.**

8.1.1. Problem of estimation

Let (x_1, x_2, x_3 .....x_n) be a random sample drawn from a population having density $f(x/\theta)$ where θ is an unknown parameter. The problem of estimation arises when we attempt to estimate the value θ with the help of the sample observations. There are mainly two types of estimation *viz.* **Point estimation** and **Interval estimation.**

8.1.1.A. Point Estimation

Suppose x_1, x_2, x_3 .....x_n is a random sample from a density $f(x/\theta)$, where θ is unknown parametric value which can assume any value in one dimensional real parameter space Ω.

Let 't' be a function of $x_1, x_2, x_3 \ldots x_n$ so that t is a statistic and hence a random variable. If t is used to estimate θ then t is called a point estimator of θ. If the realized value of t from a sample is used for θ, then t is called a point estimate of θ.

Many estimators based on sample observations can be proposed to estimate the same parameter. Thus to estimate population mean one can propose sample mean, median, mode etc. Now the question is : which estimator is to be used and why? That means there should be some criteria for a good estimator. A good estimator 'u' for an unknown parameter 'θ' is one which minimizes the difference $|\theta - u|$. R. A.

Fisher has given the following criteria for a good estimator :

(*a*) Unbiased estimator

(*b*) Consistent estimator

(*c*) Efficient estimator

(*d*) Sufficient estimator

(*a*) Unbiased estimator : Let t_n be a statistic calculated from a sample $(x_1, x_2, x_3 \ldots x_n)$ of size n from density $f(x/\theta)$. If for all n and θ, $E(t_n) = \theta$ then t_n is called an unbiased estimator. In case t_n be a biased estimator the difference $E(t_n) - \theta$ is the amount of bias.

If x be an $N(\mu, \sigma^2)$ variate, the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ *i.e.* $E(\bar{x}) = \mu$ and the sample mean square $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of the population variance σ^2 *i.e.* $E(s^2) = \sigma^2$

(*b*) Consistent estimator: An estimator t_n is called a consistent estimator of the parameter θ if the probabilistic value of the estimator t_n approaches towards θ as the sample size increases. Thus, a consistent estimator t_n should have the property

$$\underset{n \to \infty}{\text{Lt}}\ P\{|t_n - \theta| > \varepsilon\} = 0 \text{ where } \varepsilon > 0 \text{ is as small as we please.}$$

Thus consistency is large sample property. It is not defined for small sample. And there might be more than one consistent estimator for estimating the same parameter. For example if we have sample from normal population $N(\mu, \sigma^2)$ then sample mean $(\bar{x})$ as well as the sample median (m_d) both are the consistent estimator for population mean. Similarly, the sample variance S_x^2 and the sample mean square s_x^2 are both consistent estimators for σ^2 but s_x^2 is an unbiased estimator for σ^2 and S_x^2 is a biased estimator for σ^2. It is to be noted that if $\underset{n \to \infty}{\text{Lt}}\ E(t_n) = \theta$ and if $\underset{n \to \infty}{\text{Lt}}\ V(t_n) = 0$ then t_n is consistent estimator of θ.

(*c*) Efficient estimator: Let t_1 and t_2 be the two estimators for the same parameter θ. Then 't_1' is called efficient estimator of the parameter θ over the other estimator 't_2' if $V(t_1) < V(t_2)$. Thus a consistent estimator having minimum variance is the efficient estimator. The efficiency of any estimator t_1 is defined as the ratio of the variance of the most efficient estimator 't' to that of the variance of the given estimator t_1 of the same parameter. So, the efficiency of the estimator t_1 is $\frac{V(t)}{V(t_1)}$.

9 CORRELATION ANALYSIS AND ITS APPLICATION

In our daily life we are to deal with a number of variables at a time, instead of a single variable. And it is our common experience that all these variables may not be independent of each other ; rather they tend to vary side by side. Most of the growth and economic variables are found to follow the above characteristics. For example while dealing with yield component analysis for any crop it is found that yield is an ultimate variable contributed / influenced/effected by number of other factors. If we consider the yield of paddy then one can find that the factors like number of hills per square meter, number of tillers per hill, number of effective/panicle bearing tillers per hill, length of the panicle, number of grains per panicle, test (1000 grain) weight of grains etc. are influencing the yield. Variation in one or more of the above mentioned factors resulted in variations of the yield. Thus, yield may vary because of variation in number of hills per square meter or variations in number of tillers per hill or so on. Again, yield may vary because of variation in number of hill per square meter and, or number of tiller per hill and other factors. When we consider variations in one variable due to variation in any other variable then it becomes a bi-variate case. On the other hand when the variations of more than two variables are considered at a time it becomes a multi variate case.

The variations in one variable due to change in other variable(s) may follow a linear $y_i = mx_i + c$; $i = 1, 2, \ldots\ldots, k$ or non-linear (curvilinear) relationship (*e.g.*, $y_i = a_i x_i^{\alpha}$; $i = 1, 2, \ldots\ldots, k$) or any other form.

The problem lies in measuring the change in one variable due to the change in other variables and vice-versa.

In this chapter we will consider only the linear relationship among the variables.

9.1 BI-VARIATE ANALYSIS

9.1.1.

As has been mentioned in the above paragraph that in bi-variate analysis we consider only two characteristics for each individual at a time ; for example, age and height or height

and weight, age and weight, etc. To measure the degree of linear associationship we shall introduce Karl Pearson's correlation coefficients. **Correlation coefficient** measures the **degree** of closeness of the **linear** associationship between any two variables and is given as

$$r_{xy} = \frac{\text{Cov}(x, y)}{S_x \cdot S_y}$$

where (x_1, y_1), (x_2, y_2), (x_3, y_3),..... (x_n, y_n) are n pairs of observations and

(*i*) $Cov\ (x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}$

(*ii*) $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

(*iii*) $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$

Thus,

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2\right)}$$

It may be noted that we have considered two variables x and y irrespective of their dependency.

9.1.2 Properties

(*i*) The correlation coefficient between any two variables is independent of change of origin and scale in value but not in sign.

Let us consider (x_1, y_1), (x_2, y_2), (x_3, y_3),..... (x_n, y_n), n pairs of observations for two characters x and y having means $\bar{x}$ and $\bar{y}$ and variances S_x^2 and S_y^2. We take another two variables such that $u_i = \frac{x_i - a}{b}$ and $v_i = \frac{y_i - c}{d}$; $i = 1, 2, 3 \ldots\ldots n$

and a, b, c and d are constants and a, c are changes in origins and b, d are changes in scales. So, $x_i = a + bu_i$ and $y_i = c + dv_i \Rightarrow \bar{x} = a + b\bar{u}$ and $\bar{y} = c + d\bar{v}$, and $S_x^2 = b^2 S_u^2$ and $S_y^2 = d^2 S_v^2$.

Again,

$$Cov\ (x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{(a + bu_i - a - b\bar{u})\}\{(c + dv_i - c - d\bar{v})\}$$

10 REGRESSION ANALYSIS AND ITS APPLICATION

In the previous chapter we have seen that correlation coefficient measures the degree of linear associationship between any two given variables. Once after getting a good degree of associationship our objective is to find out the actual relationship between or among the variables. The technique by which we can analyze the relationship among the correlated variable is known as 'regression analysis' in the theory of statistics. Francis Galton was to coin the term 'Regression' in his famous paper "Family likeness in stature" in the proceedings of Royal Society, London in 1886. Regression analysis is the study of dependence of one variable (the dependent variable) on one or more independent (explanatory variables) variables.

In agricultural and other experiments mainly three types of variables are recorded : (*a*) the treatments or factors such as variety, insecticide, doses or type of fertilizers, different chemical treatments, different management practices etc., (*b*) Environmental parameters like rainfall, temperature, humidity, sunshine hours, wind speed etc. and (*c*) various responses in the form of different growth and yield parameters, qualitative changes etc. Now the task of a statistician is to work out the actual relationship between or among the variables under study. And this is being accomplished through regression analysis.

In different socio-economic studies different demographic, social, economical, educational etc. parameters are studied to find out the dependence of the ultimate variables, say adoption index, awareness, empowerment status etc. on these parameters. Regression analysis is a technique by virtue of which one can study the relationship.

10.1. OBJECTIVE

Thus the main objective of regression analysis is to estimate and/or predict the average value of the dependent variable given the values for independent /explanatory variables.

Thus in regression analysis the dependent variable is the function of one or more independent variables, and can be represented in the form of

$$Y = f(X_i)$$

This type of regression analysis is known as function/deterministic dependency. But in statistics one deals with random/stochastic variables *i.e.*, the variables having probability distributions. Note that a stochastic variable in simplest form, is a variable which can take a set of values in accordance with given probabilities.

Point may be noted that in correlation analysis we do not consider the dependency of variables on other variables. Given any set of observations for a pair of variables one can work out the degree of linear associationship between them. But in regression analysis there are two group of variables, one variable is treated as dependent variable and the other variables on which the dependent variable depends are the independent variables or the explanatory variables.

Variations in crop yield can be explained very well by the factors like, rainfall, temperature, sunshine, fertilizer, extent of pest and diseases etc.; but certainly will not enable an agronomist to predict exact yield from these factors because all the variables have certain probability distributions, (they are stochastic variables) as a result of which errors will be there in measuring these variables even if we keep other factors constant. Some intrinsic or random variability in dependent variable will always be there, no matter how many explanatory variables are considered and how accurately these are measured. So the above phenomenon is statistical in nature. So in statistical regression the dependent variable is the function of one or more independent variables and the error term, and can be represented in the form of $Y = f(X_i, u_i)$.

10.2. CAUSE-EFFECT RELATIONSHIP ?

As such statistical regression analysis does not imply cause and effect relationship between the explanatory variables and the dependent variable. The idea of causation never comes from statistical theories, it comes from outside the area of statistics. There is no statistical reason to assume that rainfall does not depend upon crop yield but our common (outside statistical) sense suggest that yield can be varied with the change in rainfall, temperature etc. but not the reverse. To know more about causation one should consider Granger test of causality.

10.3. PREDICTION EQUATION

The regression equation can also be used as prediction equation, under the assumption that the trend of change in Y (the dependent variables) corresponding to change in X (or the X_i's) (the independent variables) remains the same. Once the constants are estimated from a given set of observations, the value of the dependent variable corresponding to any value of X (or set of values of X_i's) with in the range of X (or X_i's) can be worked out. To some extent the prediction can be made for Y for the value(s) of X (X_i's) beyond the range but not too far beyond the values taken for calculation.

ABOUT THE AUTHOR

Dr. P.K. Sahu, is presently working as reader in Agricultural Satistics, Bidhan Chandra Krishi viswavidyalaya , West Bengal , India. He has specialized in the field of Agricultural Statistics. He has served the Central University, North Easter Hill University, Medziphama Campus; Central Silk Board, Ministry of Textiles, Government of India. By virtue of his working experience in different agriculture and allied fields he has enriched himself on various problems. He has also served as member of the Under Graduate and Post Graduate Council for faculty of Agriculture, Bidhan Chandra Krishi Viswavidyalaya. The author has published fifty-six research papers in different diverse agriculturol and allied fields with special emphasis on application of statistics in these fields.

ABOUT THE BOOK

This book is an attempt to explain various theories of statistics which can be used in solving numerous related problems. An attempt has been made, in two volumes, to present the theory of statistics in such a way that the students and researchers from biological, agricultural and allied field find it easy to handle and use in addressing many real life problems of their respective fields.

This volume contains ten chapters. The first chapter is to address and explain the subject statistics, its usefulness and application with particular reference to agriculture and allied field.

In each chapter definition or theories are followed by examples from applied fields which will help the reader of this book to understand the theory and application of statistical tool. Attempts have been made to familiarize the problems with example on each topic in a lucid manner. Each chapter is followed by a number of problems which will help the students in gaining confidence on solving those problems. Due care has been taken on varied problems in the field of agriculture and other subjects and the examination need of the students. At the end of the chapters a number of question on each chapter are given.

ISBN 978-81-272-3992-3 **Rs. 225.00**

KALYANI PUBLISHERS