SEMESTER IV MATHEMATICS GENERAL

LECTURE MATERIAL

**STATISTICS**

TARUN KUMAR BANDYOPADHYAY, DEPARTMENT OF MATHEMATICS



**TEXT**: (1) Fundamentals of Mathematical Statistics—Gupta, Kapoor

(2) Fundamentals of Statistics(Vol. II)—Goon, Gupta, Dasgupta



In case you face difficulty in understanding the following material , you may e-mail to me at **tbanerjee1960@gmail.com** stating your Name and Roll No.

**DOWNLOAD SITE: WWW.SXCCAL.EDU**

**Chapter I**

**Frequency Distributions and their Comparison:**

 **Central Tendency**

A Frequency Distribution corresponding to a variable specifies the values the variable takes and the frequencies or the number of times each variate value is taken.

Following are the marks obtained by 60 students in an examination:

22,47,9,42,31,17,13,15,18,13,2,21,27,38,15,0,33,10,34,29,26,16,25,33,36,10,24,2,26,19,14,36,18,25,21,33,35,25,18,28,25,17,38,10,3,31,24,3,12,16,33,18,26,29,27,29,28,35,26,27.

Here the variable is the ‘number of marks’. The data in the above form is called **raw** or **ungrouped data**. This representation of the data does not furnish any useful information and is rather confusing to mind. To make the data more compact and understandable, we arrange the data from the array in ascending or descending order of magnitude to obtain a **Frequency Table.** Take each mark from the data and place a bar( | ) or tally mark against the number when it occurs. Tally marks are recorded in batches of five, the fifth occurrence is shown by putting a cross tally(/) on the first four bars ||||/. We get the following frequency table of marks:

**Frequency Table of Marks in an Examination**

**Marks**: 0 2 3 9 10 12 13 14 15 16 17 18

**Tally marks**:| | || | ||| | || | || || || ||||

**Frequency**: 1 1 2 1 3 1 2 1 2 2 2 4

**Marks**: 19 21 22 24 25 26 27 28 29 31 33 34

**Tally marks**:| || || || |||| |||| ||| || ||| || |||| |

**Frequency**: 1 2 2 2 4 4 3 2 3 2 4 1

**Marks**: 35 36 38 42 47

**Tally marks**: || || || | |

**Frequency**: 2 2 2 1 1

***If the identity of the individuals about whom a particular information is taken is not relevant , nor the order in which the observation arise***, then the first real step of condensation of data is achieved by arranging the data into groups:

**Frequency Table of Marks in an Examination**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Marks (Class)** | **Tally Marks** | **No.of students****(frequency)** | **Cumulative frequency(less than)** | **Cumulative frequency(greater than)** |
| 0-5 | /// | 4 | 4 | 60 |
| 6-10 | / | 1 | 5 | 59 |
| 11-15 | ////| // | 7 | 12 | 58 |
| 16-20 | ////| ////| / | 11 | 23 | 52 |
| 21-25 | ////| / | 6 | 29 | 45 |
| 26-30 | ////| ////| ////| / | 16 | 45 | 29 |
| 31-35 | ////| // | 7 | 52 | 23 |
| 36-40 | ////| / | 6 | 58 | 12 |
| 41-45 | / | 1 | 59 | 5 |
| 46-50 | / | 1 | 60 | 4 |

This type of representation of frequencies is called a **grouped frequency distribution.** The groups 0-5, 6-10,… are called **classes;** 0 and 5 are called the **lower limit** and **upper limit** of the class 0-5 respectively. The difference 5-0=5 between the upper and lower limits of a class is called the **width** of the class. The value =2.5 which lies midway between the lower and the upper limits is called the **mid-value** or **central value** of the class. The **less-than cumulative frequency** (**greater-than cumulative frequency** resp.) corresponding to a class is the total number of observations less than or equal to the upper limit (greater than or equal to the lower limit) of the class. Following points need be kept in mind while classifying given data:

* **Classes should be clearly defined and should not lead to any ambiguity**
* **Classes should be exhaustive**( each of given value should be included in one of the class**es)**
* **Classes should be mutually exclusive and non-overlapping**
* **Classes should be of equal width**
* **Number of classes should neither be too large nor too small; preferably it should lie between 5 and 15.**

**NOTE** A variable, which can take any numerical value within certain range , is called a **continuous variable**. Consider frequency distribution of the continuous variable of ages in years of students in a college. We cannot arrange the data in age groups 16-20,21-25 etc. since there can be students having ages between 20 and 21 years. If the original ‘inclusive’ class intervals( of the form [a,b]) are ,say, 16-20,21-25,…, we calculate the adjustment (lower limit of succeeding class-upper limit of a class)=(21-20)=.5 and change the class intervals to ‘exclusive’ type([a,b)): 15.5-20.5, 20.5-25.5,…. It is understood that age of students whose age is 15.5 and < 20.5 are included in the class interval 15.5-20.5.

**Comparison of Frequency Distributions**

It is frequently necessary to compare two frequency distributions. If they are of different types , a precise comparison is difficult and is usually not required. If they are of same type, a comparison can be made in terms of values of the following four types of measures:

* **Measure of location** or **central tendency** gives a single value around which largest number of values of the variate tend to cluster.
* **The scale parameter** or **measure of dispersion** gives the degree of scatter about the central value. It measures variability or lack of homogeneity of data.
* **Measure of skewness** measuring degree of departure from symmetry
* **Measure of Kurtosis** measuring degree of ‘flatness’ of the ‘top’ as compared with the ‘normal’ curve.

**Characteristics of a good measure of Central Tendency**

* It should be based on all observations
* It should not be affected much by extreme values
* It should be rigidly defined
* It should be easily understandable and easy to calculate
* It should be amenable to algebraic treatment
* It should be least affected by fluctuation of sampling: if a number of samples of same size are drawn from a population, the measure of central tendency having minimum variation among the different calculated values should be preferred.

**Different Measures of Central Tendency**

Various types of measures of location in common use are:

* Arithmetic Mean
* Geometric Mean
* Harmonic Mean
* Median and Quartiles
* Mode

**Arithmetic Mean**

If a variate X takes values x1,…,xn, then the A.M. of the set of observations x1,…,xn, is defined by . If the variate-values are not of equal ‘importance’, we may attach to them ‘weights’ w1,…,wn as measures of their importance; the corresponding weighted mean is defined by

.

In particular, if the variate-value x1 occurs f1 times, x2 occurs f2 times,…, then , where N=f1+…+fn is the total frequency.

**Note** the A.M. of a grouped or continuous frequency distribution is computed by above formula where x’s denote the mid-values of the corresponding class intervals.

**Example 1.1** Find the A.M. of following frequency distribution:

x: 1 2 3 4 5 6 7

f: 5 9 12 17 14 10 6

* **Computation of Mean**

|  |  |  |
| --- | --- | --- |
| **x** | **f** | **fx** |
| 1 | 5 | 5 |
| 2 | 9 | 18 |
| 3 | 12 | 36 |
| 4 | 17 | 68 |
| 5 | 14 | 70 |
|  6 | 10 | 60 |
|  7 | 6 | 42 |
| Total | 73 | 299 |

Thus .

**Example 1.2** Find the A.M. of following frequency distribution:

Marks: 0-10 10-20 20-30 30-40 40-50 50-60

No. of students: 12 18 27 20 17 6

* **Computation of Mean**

|  |  |  |  |
| --- | --- | --- | --- |
| Marks | No. of students(f) | Mid-point(x) | fx |
| 0-10 | 12 | 5 | 60 |
| 10-20 | 18 | 15 | 270 |
| 20-30 | 27 | 25 | 675 |
| 30-40 | 20 | 35 | 700 |
| 40-50 | 17 | 45 | 765 |
| 50-60 | 6 | 55 | 330 |
| **Total** | **100** |  | **2800** |

A.M.=.

**Change of origin and scale**

Let x and u be two variates related by u=. Then . Thus .

Thus mean is dependent on both change of origin a and scale h.

**Example 1.3** Find the A.M. of following frequency distribution:

Marks: 0-10 10-20 20-30 30-40 40-50 50-60

No. of students: 12 18 27 20 17 6

* Let us take the origin a=300 and scale h=50 so that u=.

**Properties of A.M.**

* Algebraic sum of deviations of a set of variate values from their arithmetic mean is zero.
* )==N-N =0, where N=.
* (Mean of the combined distribution) If ,…, be the A.M.s of k distributions with respective frequencies n1,…,nk, then the mean of the combined distribution of frequency N= is given by: =.

**Example 1.4** The average salary of male employees in a firm was Rs. 5200 and that of females was Rs. 4200. The mean salary of all the employees was Rs. 5000. Find the percentage of male and female employees.

* Let n1 and n2 denote respectively the number of male and female employees and and denote their average salary (in Rs.). Then 5000(n1+n2)=5200 n1+4200 n2 implying n1:n2::4:1. Thus percentage of male and female employees in the firm is 80% and 20% respectively.

**Geometic Mean**

If n positive values x1,…,xn occur f1,…,fn times respectively, then geometric mean(G.M.) G of the set of observations is defined by G=, where N=.

**Harmonic Mean**

The harmonic mean H of n non-zero variate values xi with frequencies fi is given by H=.

**Relation between A.M., G.M. and H.M.**

If A,G,H stand for the A.M., G.M. and H.M. respectively of a finite series of positive values **of a variate**, then it can be proved that A.

**MEDIAN**

**Mean can not be calculated whenever there is frequency distribution with open end classes. Also the mean is affected to a great extent by presence of extreme value in the set of observations.** For instance, if salary of 8 persons be Rs. 150,225,240,260,275,290,300 and 1500, the mean salary is Rs. 405, which is not a good measure of central tendency because out of the 8 people, seven get Rs. 300 or less.

Median of a finite set of variate values is the value of the variate which divides it into two equal parts. It is the value which exceeds and is exceeded by same number of observations. Median is thus a positional average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is taken to be the arithmetic mean of the middle terms. Thus , median of the values 25,20,15,35,18, that is, of 15,18,20,25,35 is 20 and the median of 8,20,50,25,15,30, that is, of 8,15,20,25,30,50 is (20+25)/2=22.5.

In case of discrete frequency distribution, median is obtained as follows:

* Find N/2, where N=.
* Find the cumulative frequency (less than type) just greater than N/2
* The corresponding value of the variate is the median.

**Example 1.5** Obtain the median for the following frequency distribution:

x: 1 2 3 4 5 6 7 8 9

f: 8 10 11 16 20 25 15 9 6

* **Calculation of Median**

|  |  |  |
| --- | --- | --- |
| x | f | c.f. |
| 1 | 8 | 8 |
| 2 | 10 | 18 |
| 3 | 11 | 29 |
| 4 | 16 | 45 |
| 5 | 20 | 65 |
| 6 | 25 | 90 |
| 7 | 15 | 105 |
| 8 | 9 | 114 |
| 9 | 6 | 120=N |

N/2=60. The c.f. just greater than N/2 is 65 and the value of x corresponding to 65 is 5. Thus , median is 5.

In case of grouped frequency distribution, median is obtained as follows:

Let us consider the grouped frequency distribution:

Class intervals frequency cumulative frequency

x1 - x2 f1 F1

x2 – x3 f2 F2

… …. ….

xp – xp+1 fp Fp

…. ….. …

xn – xn+1 fn Fn

where Fk=. Let the smallest c.f. greater then N/2 is Fp. Then the **median class** is xp­ – xp+1. We assume that frequency of a class is uniformly distributed over the class interval. Let the c.f. for the class just above the median class be c . Thus (N/2-c) is the frequency of the interval between the median and the lower limit of the median class . the length of the interval corresponding to the frequency (N/2-c) is I, where f is frequency of the median class, I is the length of the class interval of the median class . Hence the median is L0+ I, where L0 is lower limit of the median class.

**Properties of Median**

* Median is a positional average and hence is not influenced by extreme values
* Median can be calculated even in the case of open end intervals
* Median can be located even if the data is incomplete
* It is not a good representative of data if the number of observations is small
* It is not amenable to algebraic treatment
* It is susceptible to sampling fluctuations

**Quartiles** are thsose variate values which divide the total frequency into four equal parts; **deciles** and **percentiles** divide into ten and hundred equal parts respectively. Suppose the values of the variate have been arranged in ascending order of magnitude, then the value of the quartile having the position between the lower extreme and the median , is the first quartile Q1 and that between the median and the upper extreme is the third quartile Q3. The median is the second quartile Q2, is the fifth decile D5 and the fiftieth percentile P50. For a grouped frequency distribution, the quartiles, deciles and percentiles are given by

Qi=l+h, i=1,2,3

Dj= l+h, j=1,…,9

Pk= l+h, k=1,…,99

where l is the lower limit of the class in which the particular quartile/decile/percentile lies, f is the frequency of the class , h is the width of this class, C is the cumulative frequency upto and including the class preceding the class in which the particular quartile/decile/percentile lies and N is the total frequency.

**Example 1.6** Calculate the three quartiles for the following frequency distribution of the number of marks obtained by 49 students in a class:

**Marks No. of students Marks No. of students**

5-10 5 25-30 5

10-15 6 30-35 4

15-20 15 35-40 2

* 1. 10 40-45 2

**Cumulative Frequency Table**

**Class Frequency Cumulative Frequency(less than)**

5-10 5 5

10-15 6 11

15-20 15 26

20-25 10 36

25-30 5 41

30-35 4 45

35-40 2 47

40-45 2 49=N

The cumulative frequency immediately greater than N/4=49/4 is 26; hence to find Q1,

L=15, h=15-10=5, C=11, f=15. Thus Q1= 15+ 5= 15.47 marks.

For median, N/2= 24.5 . Thus the median class is 15-20. Median = 15+ 5=19.5 marks.

To find Q3, we have 3N/4=147/4 . Hence Q3 lies in the class 25-30. L=25, C=36,f=5,h=5. Hence Q3=25+5 =25.75.

**Example 1.7** In a factory employing 3000 persons, in a day 5% work less than 3 hours, 580 work from 3.01 to 4.50 hours, 30% work from 4.51 to 6.00 hours, 500 work from 6.01 to 7.50 hours, 20% work from 7.51 to 9.00 hours and the rest work 9.01 or more hours. What is the median hours of work?

* **Calculation for Median Wages**

|  |  |  |  |
| --- | --- | --- | --- |
| Work Hours | No. of employees(f) | Less than c.f. | Class Boundaries |
| Less than 3 | 5/100 x 3500=150 | 150 | Below 3.005 |
| 3.01-4.50 | 580 | 730 | 3.005-4.505 |
| 4.51-6.00 | 30/100 x 3000=900 | 1630 | 4.505-6.005 |
| 6.01-7.50 | 500 | 2130 | 6.005-7.505 |
| 7.51-9.00 | 20/100 x 3000=600 | 2730 | 7.505-9.005 |
| 9.01 and above | 3000-2730=270 | 3000 | 9.005 and above |

N=3000. The c.f. just greater than N/2=1500 is 1630.the corresponding class 4.51-6.00, whose class boundaries are 4.505-6.005, is the median class. Hence median=l+=4.505+=4.505+1.283=5.79(approx.).

**Example 1.8** An incomplete frequency distribution is given as follows:

**Variable Frequency Variable Frequency**

10-20 12 50-60 ?

20-30 30 60-70 25

30-40 ? 70-80 18

40-50 65 Total(N) **229**

Given that the median value is 46, determine the missing frequencies.

* Let the frequency of the class 30-40 be f1 and that of 50-60 be f2. Then f1+f2=229 -(12+30+65+25+18)= 79.

Since median is given to be 46, 40-50 is the median class. Using formula for median , we get

46=40+ x 10. Hence f1= 33.5= 34(approx.). Hence f2=45.

**Mode**

Let us consider the following statements: The average height of an Indian is 5’6”; the average size of shoes sold in a shop is 7; an average student in a hostel spends Rs. 750 per month. In all the above statements,the average referred to is mode. Mode is the value of the variate which occurs most frequently in a set of observations and around which the other members of the set cluster densely. In other words, mode is the value of the variable which is predominant in the given set of values. In case of discrete frequency distribution, mode is the value of the variable corresponding to maximum frequency. In the following distribution:

x: 1 2 3 4 5 6 7 8

f: 4 9 16 25 22 15 7 3

value of x corresponding to maximum frequency viz. 25 is 4. Hence mode is 4.

In any one of the following cases, mode is determined by the method of grouping:

* If the maximum frequency is repeated
* If the maximum frequency occurs in the very beginning or at the end of the distribution
* If there are irregularities in the distribution

**Example 1.9** Find the mode of the following frequency distribution:

Size(x): 1 2 3 4 5 6 7 8 9 10 11 12

Frequency: 3 8 15 23 35 40 32 28 20 45 14 6

* The distribution is not regular since the frequencies are increasing steadily upto 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. We cannot say that since the maximum frequency is 45, mode is 10. Here we locate mode by the method of grouping as explained below:

**Size Frequency**

**(x) (i) (ii) (iii) (iv) (v) (vi)**

1 3

 11

2 8 26

 23

3 15 46

 38

4 23

 58

5 35 98 73

 75

6 40 107

 72

7 32

 60

8 28 80 100

 48

9 20

 65 93

10 45 79

 59

11 14 65

 20

12 6

The frequencies in column (i) are the original frequencies. Column (ii) is obtained by combining the frequencies two by two.If we leave the first frequency and combine the remaining frequencies two by two, we get column (iii).Combining the frequencies two by two after leaving the first two frequencies results in a repetition of column (ii). Hence, we proceed to combine the frequencies three by three , thus getting column (iv). The combination of frequencies three by three after leaving the first frequency results in column (v) and after leaving the first two frequencies results in column (vi).

The maximum frequency in each column is given in red type. To find mode we form the following table:

**Analysis Table**

|  |  |  |
| --- | --- | --- |
| Column No. | Maximum Frequency(1) | Value or combination of values of x giving max. frequency in (1) (2) |
| (i) | 45 | 10 |
| (ii) | 75 | 5,6 |
| (iii) | 72 | 6,7 |
| (iv) | 98 | 4,5,6 |
| (v) | 107 | 5,6,7 |
| (vi) | 100 | 6,7,8 |

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

**Mode for Continuous frequency distribution**

**Mod**e= l+ h

Where l is the lower limit of the modal class(class having maximum frequency), fm is the maximum frequency, f1 and f2 are the frequencies of the classes preceding and following modal class.

**Example 1.10** The median and mode of the following wage distribution are known to be Rs. 3350 and Rs. 3400 respectively.Find the values of f3,f4,f5:

**Wages (in Rs.) No. of employees Wages (in Rs.) No. of employees**

0-1000 4 4000-5000 f5

1000-2000 16 5000-6000 6

2000-3000 f3 6000-7000 4

3000-4000 f4 **Total** 230

* **Calculation for median and mode**

**Wages (in Rs.) frequency(f) less than c.f.**

0-1000 4 4

1000-2000 16 20

2000-3000 f3 20+f3

3000-4000 f4 20+f3+f4

4000-5000 f5 20+f3+f4+f5

5000-6000 6 26+f3+f4+f5

6000-7000 4 30+f3+f4+f5=N

N=30+f3+f4+f5=230 . Thus f3+f4+f5=200.

Since median is 3350, which lies in 3000-4000, 3000-4000 is the median class. Using median formula,

3350=3000+[115-(20+f3)]. Thus f3=95-0.35f4.

Mode being 3400, modal class is 3000-4000. Using formula for mode,

3400=3000+ ; hence =. Thus f4=100. Hence f3=95-0.35x100=60, f5=40.

**Note** For a symmetrical distribution, mean, median and mode coincide. If the distribution is **moderately asymmetrical**, they obey the following empirical relationship: **mode = 3 median – 2 mean**

**Chapter II**

**Frequency Distributions and their Comparison:**

 **Measures of Dispersion**

A measure of central tendency alone is not enough to have a clear idea about the data unless all observations are almost the same. Moreover two sets of observations may have the same central tendencies whereas variability of data within the sets may vary widely . Consider

Set A: 30 30 30 30 30

Set B: 28 29 30 31 32

Set C: 3 5 30 37 75

All the three sets have same mean and mode; but the amount of variation differs widely amongst the sets.

**Characteristics of an ideal measure of dispersion**

* It should be rigidly defined
* It should be easily understandable and easy to calculate
* It should be based on all observations
* It should be amenable to further mathematical treatment
* It should be least affected by fluctuation of sampling

**Different measres of dispersion**

**Range**: Range is the difference between the maximum and the minimum values of the variate. It is easily understood and easy to calculate but depends only on the two extreme values which themselves are subject to sampling fluctuation; hence range is not a reliable measure of dispersion.

**Quartile Deviation:** quartile deviation or semi-interquartile range is given by (Q3-Q1), where Q1 and Q3 are the first and the third quartiles of the frequency distribution. Quartile deviation is definitely a better measure than the range as it makes use of 50% of data. But since ignores the other 50% of data, it cannot be regarded as a reliable measure.

**Mean Deviation** If xi|fi , i=1,…,n be the frequency distribution, then mean deviation from A (usually mean, median or mode) is defined by s= , =N.

Since mean deviation is based on all the observations, it is a better measure of dispersion as compared to range and quartile deviation. But use of absolute value renders it useless for further mathematical treatment.

**Example 2.1** Calculate Q.D. and M.D. from mean, for the following data:

**Marks**: 0-10 10-20 20-30 30-40 40-50 50-60 60-70

**No. of students**: 6 5 8 15 7 6 3

* **Calculation for Q.D. and M.D. from mean**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Marks | Mid-value | f | d=(x-35)/10 | fd |  | f | c.f. (less) |
| 0-10 | 5 | 6 | -3 | -18 | 28.4 | 170.4 | 6 |
| 10-20 | 15 | 5 | -2 | -10 | 18.4 | 92.0 | 11 |
| 20-30 | 25 | 8 | -1 | -8 | 8.4 | 67.2 | 19 |
| 30-40 | 35 | 15 | 0 | 0 | 1.6 | 24.0 | 34 |
| 40-50 | 45 | 7 | 1 | 7 | 11.6 | 81.2 | 41 |
| 50-60 | 55 | 6 | 2 | 12 | 21.6 | 129.6 | 47 |
| 60-70 | 65 | 3 | 3 | 9 | 31.6 | 94.8 | 50 |
| **Total** |  |  |  | -8 |  | 659.2 |  |

1. Here N=50, N/4=12.75, 3N/4=37.25

The c.f.(less than) just greater than 12.75 is 19. Hence the corresponding class 20-30 contains Q1.

Q1=20+(12.75-11)=22.19

The c.f.(less than) just greater than 37.25 is 41. Hence the corresponding class 40-50 contains Q3.

Q3=40+(37.25-34)=44.64.

Hence Q.D. = (Q3-Q1)=(44.64-22.19)=11.23.

1. = A+=35+ = 33.4 marks. Thus M.D. (from mean) ===13.184.

**Standard Deviation, Variance**

For the frequency distribution xi|fi , i=1,…,n , S.D. is defined by: =, where is the A.M. of the distribution (non-negative value of the square root is considered). 2 is called the variance.

**Note** s2===+(-A)2+2(-A)(xi-)}

=+(-A)2+2(-A)=2+(-A)2, where d=-A.

Thus s2 is least when d=0.that is, when =A. Thus M.D. is least when A= and S.D. is least value of M.D.

**Note (1)** 2==) = +- 2=+-2=. This expression is often used for calculation of 2.

**(2)**If n1 and n2 are the sizes, and be the means and 1 and 2 be the S.D. s of two series, then the S.D. of the combined series of n­1+n2 observations is given by: 2=)+)], where d1=-, d2=- and = is the mean of the combined series.

**Example 2.2** For a group of 200 candidates , the mean and S.D. of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores 43 and 35 were misread as 34 and 53 respectively. Find the corrected mean and S.D. corresponding to the corrected figures.

* Let x be the given variable. Given n=200, =40 and =15. Now 40= = gives =8000.

2=-2 gives =200(225+1600)=365000.

Corrected =8000-34-53+43+35=7991, corrected mean ==39.995

Corrected = 365000-342-532+432+352=364109

Corrected 2=(39.995)2=224.14. Thus corrected ==14.97.

**Example 2.3**  The first of two samples has 100 items with mean 15 and s.d. 3.if the whole group has 250 items with mean 15.6 and s.d. . find s.d. of the second group.

* Here n1=100,=15,1=3, n=n1+n2=250, =15.6, =

= gives =16. Hence d1=-=15-15.6=-0.6, d2=-=16-15.6=0.4

From 2=)+)], 2=4.

**Moments**

The r th (raw) moment of a variable x about any point A , denoted by , is given by =.

The r th (central) moment of a variable x about mean , denoted by , is given by =.

In particular, ===1, ==0, ==2.

**Relation between raw and central moments**

===, where di=xi-A.

=A+=A+. Hence ==+-…+(-1)r)

=-+-…+(-1)r.

In particular, on putting r=2,3,4 and simplifying, we get =/-, =/-3//+2, =+6-3.

**Effect of change of origin and scale on Moments**

Let u=. Then . Thus x-=h(u-). Thus =hr.

**Symmetrical and Skew Distributions**

A distribution is symmetrical when the frequencies are symmetrically distributed about the mean , that is, when the values of the variate equidistant from mean have equal frequencies. For example, the following distribution is symmetrical about its mean 5:

x: 1 2 3 4 5 6 7 8 9

f: 3 4 6 9 10 9 6 4 3

It can be seen that if n is odd, =0 since all the terms cancel in pairs, n being odd and f1=fn, f2=fn-1,…. Thus =0 , for n odd. Hence for a symmetrical distribution, = 0. Thus is a measure of departure from symmetry.

Also for a symmetrical distribution, the mean, median and mode coincide. Further, in the case of such distribution, median lies halfway between the two quartiles.

Skewness means lack in symmetry. It indicates whether the frequency curve is inclined more to one side than the other, that is , whether the frequency curve has a longer tail on one side. **Skewness is positive** if the curve is more elongated to the right side, that is, if the mean of the distribution is greater than the mode; in the reverse case, it is negative. Skewness gives an idea about the direction in which also the extent to which the distribution is distorted from the symmetrical distribution.

For distribution of moderate skewness, an empirical relation holds: mean-mode= 3(mean-median).

Karl Pearson’s coefficient of skewness is given by : coefficient of skewness==.

It is a pure number since the numerator and denominator have the same dimension. It has value zero for a symmetrical distribution.

Bowley’s measure of skewness is .

**Example 2.4** Find out a coefficient of dispersion based on quartile deviation and a measure of skewness from the following table giving wages of 230 persons:

**Wages(in Rs) f c.f. Wages(in Rs) f c.f.**

140-160 12 12 220-240 50 157

160-180 18 30 240-260 45 202

180-200 35 65 260-280 20 222

200-220 42 107 280-300 8 230

* Here N/2=115 and the 115 person has a wage in the class 220-240. Hence median=Q2= 220+x 20 = Rs. 223.20. Similarly, Q1=180+x 20=Rs. 195.71, Q3=240+x 20=Rs. 246.88. It can be shown that mean=Rs. 220.87, S.D.= Rs. 34.52.

Coefficient of dispersion based on quartile deviation= ==0.1156.

Measure of skewness = == - 0.3514

Second measure of skewness==- 0.07446.

**Pearson’s and - Coeficients**

 , , , =-3.

The values of the two coefficients , enable us to know whether the given distribution is symmetrical and whether it is relatively more or less flat than the normal curve. gives a measure of departure from symmetry. Kurtosis measures whether the given frequency curve is relatively more or less flat –topped compared to the normal curve (to be studied later). For a normal distribution , =3. Curves with values of less than 3 are called Platykurtic whereas those with values of greater than 3 are called Lettokurtic. Curves with value equal to 3 are called Mesokurtic.

**Chapter III**

**Theory of Probability**

Basic Terminology

**Random Experiment** : If in each trial(repetition) of an experiment conducted under identical conditions, the outcome is not unique , but may be any one of possible outcomes, then such an experiment is called a random experiment. Examples of random experiments are: tossing a coin, throwing a die, selecting a card from a pack of playing cards etc. in all these cases, there are a number of possible outcomes which can occur but there is an uncertainty as to which of them will actually occur.

**A piece of Information:** A pack of cards consists of four suits called Spades, Hearts and Clubs. Each suit consists of 13 cards, of which nine cards are numbered from 2 to 10, an ace, a king, a queen and a jack(or knave).Spades and clubs are black-faced cards while hearts and diamonds are red-faced cards.

**Outcome:** result of a random experiment is called an outcome.

**Sample Space, Events:** The collection S of all possible outcomes of a random experimepernt is called sample space of the random experiment; any subset of S is an event; a singleton subset of S is an elementary(simple) event. For example, in an experiment which consists of throwing a six-faced die, possible outcomes are 1,2,3,4,5,6. Thus sample space of this experiment is {1,2,3,4,5,6}, {1} is an elementary event; getting an even number, {2,4,6} is an event of this experiment.

**Exhaustive Events:** events E1,…,Ek of a random experiment with sample space S are called exhaustive iff S = E1Ek. in the case of throwing of a die, getting even points( that is, 2,4, or 6) and getting odd points (that is, 1,3, or 5) are exhaustive events.

**Equally Likely Events:** Events are equally likely if there is no reason to expect any one of them compared to others. In the trial of drawing a card from a well-shuffled pack of cards, any of the 52 cards may appear, so that the 52 elementary events are equally likely.

**Exclusive Events:** Events are exclusive if the occurrence of any one of them precludes the occurrence of all others. On the contrary, events are compatible if it is possible for them to happen simultaneously. For instance, in the rolling of two dice, the cases of the face marked 5 appearing on both dice are compatible.

**Favourable Events:** The trials which entail the happening of an event are favourable to the event. For example, in the tossing of a dice, the number of favourable events to the appearance of a multiple of 3 are two viz. getting 3 and 6.

**Classical (a priori) definition of probability**

If a random trial results in n exhaustive, mutually exclusive and equally likely outcomes, out of which m are favourable to the occurrence of an event E, then the probability of occurrence of E , denoted by P(E), is given by P(E)=.

It is clear from definition that 0p1. Since the number of cases in which event A will not happen is n-m, the probability q that the event A will not happen is given by P==1-=1-P(A).

An event A is **certain** to happen if all the trials are favourable to it and then the probability of its happening is unity; for an event which is certain not to happen, the probability is zero.

**Example 3.1** Find the chance that if a card is drawn at random from an ordinary pack, it is one of the court cards.

* Court cards are kings, queens, jacks and their number in a pack of 52 cards is 12, so that the number of favourable events is 12. Hence the probability is 12/52=3/13.

**Example 3.2** What is the chance that a leap year selected at random will contain 53 Sundays?

* A leap year which contains 366 days has 52 Sundays corresponding to 52 weeks and 2 more days.There are following seven possibilities: (1) Sunday, Monday, (2) Monday,Tuesday,(3) Tuesday, Wednesday, (4) Wednesday, Thursday, (5) Thursday, Friday, (6) Friday , Saturday, (7) Saturday, Sunday. Out of these seven possibilities, there are two favourable outcomes, namely (1) and (8). Thus the required probability is 2/7.

**Example 3.3** An urn contains 9 balls, two of which are red, three blue and four black.Three balls are drawn from the urn at random. What is the chance that (1) three balls are of different colours, (2) two balls are of the same colour and third is different, (3) the balls are of the same colour?

* (1) Three balls can be drawn from 9 balls in =84 ways and these are equally likely, exhaustive and mutually likely cases. A red ball can be drawn in 2 ways, a blue in 3 and a black in 4 ways, so that three differently coloured balls can be drawn in 2 x 3x 4=24 ways. Hence the probability is 24/84=2/7.

(2)two blue balls can be drawn in ways and then a red or black ball in 6 ways so that the two blue balls and a different coloured ball can be drawn in 6x=18 ways. Two black balls and a different coloured ball can be drawn in 5x=30 ways. Similarly the number of ways in which two red balls and a different coloured ball can be drawn in 7x=7 ways. Thus the number of ways two balls of same colour and a ball of different colour can be drawn is 18+30+7=55. Thus required probability is 55/84.

(3)Three blue balls can be drawn in 1 way and 3 black balls in or 4 ways so that the corresponding probability is 5/84.

**Limitation of Classical Definition**

This definition breaks down in the following cases:

* If the various outcomes of the random experiment are not equally likely
* If the number of exhaustive outcomes of the random experiment is infinite or unknown

**Von Mises’s statistical (or empirical) definition of probability**

If trials be repeated a great number of times under essentially same conditions, then the limit of the ratio of the number of times that an event happens to the total number of trials, as the number of trials increases indefinitely, is the probability of the happening of the event,provided the ratio approaches a finite and a unique limit.

**Axiomatic definition of Probability**

To an event A(that is, a subset of sample space S) is assigned a real number P(A), called probability of A, satisfying the following properties:

* (Axiom of non-negativity)P(A)0
* (Axiom of certainty) P(S)=1
* (Axiom of Additivity) If {An} is any finite or infinite sequence of **disjoint** events, then P.

**Note** P, **the probability function,**  is otherwise unspecified except it is to satisfy above three axioms.

**Notation** for two events A,B of a sample space S,AB={xS: xA or xB}, AB={xS: xA and xB}, = { xS: x∉A}, A-B={ xS: x and x∉B}, AB can be denoted by A+B, if A and B are disjoint; AB=(B)+(A).

**Example 3.4** Let A,B,C are three arbitrary events. Find expression for the following events:

(1)Only A occurs, (2) Both A,B but not C , occur, (3) All three events occur, (4) At least one occurs, (5) At least two occur, (6) one and no more occurs, (7) two and no more occur, (8) none occurs.

* (1) A , (2) A, (3) A, (4)A, (5) (A) A), (6) () () (), (7)))), (8) .

**Some Theorems on Probability**

**Theorem3.1** Probability of impossible event is zero: P()=0.

* S=S and S, are disjoint events. Using axiom of additivity, P(S)=P(S)=P(S)+P(); hence P()=0.

**: P(A)=0 does not necessarily mean A is impossible event.** In case of continuous random variable X, the probability at a point is always zero: P(X=c)=0.

**Theorem3.2** P(=1-P(A).

* 1=P(S)=P(A)=P(A)+P( (since A=).

**Corollary** 0P(=1-P(A); hence 0P(A)1.

**Lemma** For two events A,B, P(

* and are disjoint events and B=B=B(A)=(); hence by axiom of additivity, P(.

**Corollary** (1)IfA⊆B, then P(P(B)-P(A), (2) P(A)P(B).

**Theorem3.3 (Addition Theorem of Probability)** If A,B are any two events, P(AB)=P(A)+P(B)-P().

* AB=A() and A, are disjoint. Hence P(AB)=P(A)+ P(= P(A)+P(B)-P().

Generalising , for three events A,B,C, we have

P(.

**Example 3.5** If p1=P(A), p2=P(B), p3=P(), express the following in terms of p1,p2,p3: (1) P(), (2) P(, (3)P(B),(4) P(

* (1) P()=1-P(A=1-[P(A)+P(B)-P(AB)]=1-p1-p2-p3.(2) P()=P()=1-P(B)=1-p3. (3) P(B)=P(B)- P(B)=p2-p3.(4) P()=P()+P(B)- P(B)=1-p1+p2-( p2-p3)=1-p1+p3.

**Example 3.6** It two dice are thrown, what is the probability that the sum is (1) greater than 8, (2)neither 7 nor 11?

* (1)If X denotes the sum on the two dice, then we want P(X>8). The required event can happen in the following mutually exclusive cases: X=9, X=10,X=11,X=12. Hence by addition theorem on probability, P(X>8)=P(X=9)+P(X=10)+P(X=11)+P(X=12). In a throw of two dice, the sample space contains 62=36 points. The number of favourable cases can be enumerated as:

X=9: (3,6),(6,3),(4,5),(5,4)

X=10: (4,6),(6,4),(5,5)

X=11: (5,6),(6,5)

X=12:(6,6).

Thus P(X>8)==.

(2)Let A denote the event of getting the sum of 7 and B denote the event of getting the sum of 11 with a pair of dice.

X=7: (1,6),(6,1),(2,5),(5,2),(3,4),(4,3)

X=11: (5,6),(6,5)

Required probability= P()=1-P(AB)=1-[P(A)+P(B)] (since A and B are disjoint events)=1-=.

**Example 3.7** Two dice are tossed. Find the probability of getting an even number on the first die or a total of 8.

* Let A be the event of getting an even number on the first dice and B be the event that the sum of points obtained on the two dice is 8. The events are represented by the following subsets of the sample space S:

A={2,4,6} X {1,2,3,4,5,6}, B={(2,6),(6,2),(3,5),(5,3),(4,4)}. Here AB={(2,6),(6,2),4,4)}.

Required probability is P(AB)=P(A)+P(B)-P(AB)==.

**Example 3.8** An integer is chosen at random from first two hundred natural numbers. What is the probability that the integer is divisible by 6 or 8?

* Sample space of the random experiment is {1,2,…,200}. The event A ‘ integer chosen is divisible by 6’ is given by {6,12,…,198}; the event B‘ integer chosen is divisible by 8’ is given by {8,16,…,200}. LCM of 6 and 8 is 24. Hence a number is divisible by 6 and 8 iff it is divisible by 24. Thus AB={24,48,…,192}. Hence required probability is P(AB)=P(A)+P(B)-P(AB)==.

**Example 3.9** The probability that a student passes Physics test is 2/3 and the probability that he passes both Physics test and English test is 14/45.The probability that he passes at least one test is 4/5. What is the probability that he passes English test?

* Let A be the event that the student passes the Physics test and B be the event that he passes English test. Given P(A)=, P(AB)=, P(AB)=. We want P(B). From P(AB)=P(A)+P(B)-P(AB), we get P(B)=.

**Example 3.10** An investment consultant predicts that the odds against the price of a certain stock will go up during the next week are 2:1 and the odds in favour of the price remaining the same are 1:3.What is the probability that the price of the stock will go down during the next week?

* Let A denote the event ‘ stock price will go up’ and B be the event ‘stock price will remain same’. Then P(A)=, P(B)=. Thus P(AB)=P(A)+P(B)=. Hence the probability that the stock price will go down is given by P()=1- P(AB)=1-=.

**Example 3.11** An MBA applies for a job in two firms X and Y. The probability of his being selected in firm X is 0.7 and being rejected at Y is 0.5. The probability of at least one of his applications being rejected is 0.6. What is the probability that he will be selected in one of his firms?

* Let A and B denote the events that the person is selected in firms X and Y respectively. Then P(A)=0.7, P=0.5. Thus P=1-0.7=0.3, P(B)=1-0.5=0.5 and 0.6= P()= P+ P P(). The probability that the person will be selected in one of the twofirms X or Y is given by: P(AB)=1- P()=1-[ P P P(]=1-(0.3+0.5-0.6)=0.8.

**Example 3.12** Three newspapers A,B and C are published in a certain city. It is estimated from a survey that 20% read A, 16% read B, 14% read C, 8% read both A and B, 5% read both A and C, 4% read both B and C, 2% read all three. Find what percentage read at least one of the papers?

* Let E,F,G denote the events that a person reads newspapers A,B and C respectively. Then we are given: P(E)=, P(F)=, P(G)=, P(EF)=, P(EG)=, P(GF)=, P)= .

The required probability that a person reads at least one of the newspapers is given by

P)=P(E)+P(F)+P(G)- P(EF)- P(EG)- P(GF)+ P)==0.35.

**Example 3.13** A card is drawn from a pack of 52 cards. Find the probability of getting a king or a heart or a red card.

* Let A,B and C denote the event ‘card drawn is a king’, ‘card drawn is a heart’ and ‘card drawn is a red card’ respectively. Then A,B,C are not mutually exclusive.

AB: card drawn is king of hearts ; n(AB)=1

B=B( since B⊆C): card drawn is a heart ;n(B)=13

A: card drawn is a red king; n(A)=2

A= A: card drawn is the king of hearts; n(A)=1.

Thus P(A)=, P(B)=, P(C)=, P(AB)=, P(B)=, P(A)=, P(A)=. Thus required probability is P()=P(A)+P(B)+P(C)- P(AB)- P(B)- P(A)+ P(A)=.

**Conditional Probability**

In many situations we have the information about the occurance of an event A and are required to find out the probability of the occurrence of another event B. This is denoted by P(B/A). For example, if we know that a card drawn from a pack is black, we may need to calculate the probability that it is the ace of spade.

Let us take the problem of throwing a fair die twice. Suppose same number of spots do not appear in both the throws and we are required to find the probability that the sum of number of spots in the two throws is six.

A patient comes to a doctor with his family history that his elders suffered from high blood pressure. He wants to know the probability of the event that he will also suffer from high blood pressure.

**Definition** Let A and B be two events. The **conditional probability** of event B supposing event A has occurred, is defined by P(B/A)=, if P(A)>0.

**Note** Let B1,…,Bk be mutually exclusive events. The conditional probability of given that event A has occurred is given by P=.

**Example 3.14** Ifa card drawn from a packis black, represented by event A, find the probability of the event B that the card drawn is ace of spade .

* Number of black cards in a pack of 52 cards is 26. P(A)=26/52=1/2. Out of 26 black cards, only one is ace of spade. The event AB contains only one point; thus . Hence P(B/A)===.

**Example 3.15** An experiment is conducted by throwing a fair dice twice. Let A be the event that same number of spots do not turn up in two throws and B be the event that sum of the spots is 6. Find P(B/A).

* A includes all 36 points of the sample space except (1,1),(2,2),(3,3),(4,4),(5,5) and (6,6). Thus P(A)=30/36=5/6. Points favourable to event B are (1,5),(2,4),(3,3),(4,2),(5,1). Points common to A and B , that is AB , are (1,5),(2,4),(4,2),(5,1). Thus P(AB)=4/36. Thus P(B/A)===0.133.

**Example 3.16**  10% of patients feel they suffer and are really suffering from TB, 30% feel they suffer but actually do not suffer, 25% do not feel they are suffering but are suffering and remaining 35% neither feel nor suffering from TB. Find the probility of events E1,E2,E3,E4, where E1: person who suffers from TB and feels he suffering from TB, E2: person has TB and does not feel, E3:person feels he has TB and does not suffer from TB, E4: person feels and has TB.

* Let us define events: A: person feels he has TB, B: person suffers from TB. P(AB)=0.1, P(A=.3, P(B)=.25, P()=.35. Thus P(A)= P(AB)+ P(A=0.1+0.3=0.4, P(B)= P(AB)+ P(B)=.10+.25=.35, P()= P(B)+P()=0.25+0.35=0.6,P()= P(A)+ P()=0.3+0.35=0.65. Hence

P(E1)=P(B/A)=, P(E2)=P(B/)==0.417,P(E3)= P(A/)=0.462, P(E4)=P(A/B) = 0.286.

**Example 3.17** There are two lots of manufactured item. Let one contain 40 pieces whereas lot two contains 50 pieces.it is known that former lot contains 25% defective pieces and the later one 10%. We flip a coin and select a piece from a lot one if it turns with head up; otherwise we select a piece from lot 2. Find the probability that a selected piece will be defective.

* Let A,B stand for the event ‘piece is selected from lot 1’ and ‘piece is selected from lot 2’. Since the probability of turning head up=1/2, we have P(A)=1/2 and P()=1/2. Lot 1 has 10 defective and 30 non-defective pieces; lot 2 has 5 defective and 45 non-defective pieces. Given P(B/A)=1/4, P(B/=1/10. Thus P(AB)=P(B/A)P(A)=1/8, P(B)= P(B/ P()=1/20. Hence the probability that the selected item is defective is P(AB)+ P(B)=0.175.

**Independent Events**

If we draw two cards from a pack of cards in succession, then the results of the two draws are independent if the cards are drawn **with replacement** and are not independent if the cards are drawn **without replacement**.

**Definition** An event A is independent of another event B iff P(A/B)=P(A). This definition is meaningful when P(A/B) is defined, that is, when P(B)0.

**Theorem3.4** If two events A and B are such that P(A)0, P(B)0 and A is independent of B, then B is independent of A.

* P(A/B)=P(A)⇒ ⇒=P(A)P(B) ⇒P(B/A)===P(B).

**Theorem3.5**  If A,B are events with positive probilities, then A and B are independent iff =P(A)P(B).

**Theorem3.6** If A and B are independent, then (1) A and , (2) and B , (3) , are independent.

* Since A and B are independent, =P(A)P(B). P(A)=P(A)-=P(A)-P(A)P(B)=P(A)P(). P()=P()=1-P(AB)=1-[P(A)+P(B)-)]=1-P(A)-P(B)+P(A)P(B)=[1-P(A)][1-P(B)]=P()P().

**Example 3.18** If , then show that P(A)P(

* A=()=)= ⇒A⊆⇒ P(A)P(.

**Example 3.19** Let A and B be two events such that P(A)=3/4, P(B)=5/8. Show that (a) P(AB), (b) .

* (a) A⊆A.

(b) ⊆B⇒P(P(B)=. Also, B)) ⇒

**BAYES’ THEOREM**

**Theorem3.7** If E1,E2,…,En are mutually disjoint events with P(Ei)0(i=1,…,n), then for any arbitrary event A which is a subset of such that P(A)>0,

P(Ei/A)=

**Example 3.20**  Suppose that a product is produced in three factories X,Y, and Z. It is known that factory X produces thrice as many items as factory Y and that factory Y and Z produces same number of items. 3% of the items produced by each of the factories X and Z are defective and 5% of those manufactured in Y are defective. All the items in the three factories are stocked and an item of the product is selected at random. (1) What is the probability that the item selected is defective? (2) if an item selected at random is found to be defective, what is the probability that it was produced in factory X,Y,Z respectively?

* Let the number of items produced by factories X,Y, and Z be 3n,n,n respectively. Let E1,E2,E3 be the events that the items are produced by factory X,Y and Z respectively and let A be the event that the item being defective. Then P(E1)==0.6, P(E2)=0.2, P(E3)=0.2. Also, P(A/E1)= P(A/E3)=0.03, P(A/E2)=0.05(given).
1. The probability that an item selected at random from the stock is defective is given by P(A)==.6 x .03+.2x.05+.2x.03=.034.
2. By Bayes’ rule, the required probabilities are given by :

P(E1/A)==, P(E2/A)==, P(E3/A)=.

**Example 3.21**  In 2002 there will be three candidates for the position of the principal –Mr. x, Mr. y and Mr.z—whose chances of getting the appointment are in the ratio 4:2:3 respectively. The probability that Mr. x if selected would introduce co-education in the college is 0.3. The corresponding probabilities for Mr. y and Mr.z are 0.5 and 0.8. (1) What is the probability that there will be co-education in 2003? (2) if there is co-education in 2003, what is the probability that Mr. z is the principal?

* Let us define the events

A: introduction of co-education, E1: Mr. x is selected as principal

 E2: Mr. y is selected as principal, E3: Mr. z is selected as principal.

 P(E1)=4/9, P(E2)=2/9, P(E3)=3/9, P(A/E1)=3/10, P(A/E2)=5/10, P(A/E3)=8/10.

1. The required probability that there will be coeducation in the college in 2003 is

P(A)=P[(AE1)]= P(AE1)+ P(AE2)+ P(AE3)

=P(E1)P(A/E1)+ P(E2)P(A/E2)+ P(E3)P(A/E3)==.

1. The required probability is given by Bayes’ rule:

P(E3/A)===.

**Chapter IV**

**Random Variable and Distribution Function**

In many random experiments, we are interested not in knowing which of the outcomes has occurred but in the numbers associated with them. For example, when n coins are tossed, one may be interested in knowing the number of heads obtained. When a pair of dice are tossed, one may seek information about the sum of points. Thus, we associate a real number with each outcome of a random experiment. In other words, we are considering a function whose domain is the set of all possible outcomes and whose range is a subset of the set of reals.

**Definition** Let S be the sample space associated with a given random experiment. A real-valued function X: S) is called a one-dimensional random variable(**r.v**.).

**Notation** If x is a real number, the set of all wS such that X(w)=x is denoted by X=x. Thus P(X=x)=P{w: X(w)=x}. Similarly P(Xa)= P{w}, P(a<Xb)=P{w:X(w)(a,b]}

**Example 4.1** Consider the random experiment of tossing a coin. Then S={w1,w2}, w1=H,w2=T. Define X:{w1,w2}{0,1} by X(w1)=1, X(w2)=0. X is a r.v.

A function X:S→R2 is a two-dimensional random variable.

**Example 4.2**: If a dart is thrown at a circular target, the sample space S is the set of all points w on the target.By imagining a co-ordinate system placed on the target with the origin at the centre, we can consider a two-dimensional random variable X which assigns to every point w of the circular region , its rectangular co-ordinates (x,y)

**Example 4.3** If a pair of dice is tossed , then S={1,2,3,4,5,6}X{1,2,3,4,5,6} .Let X be the random variable defined by X(i,j)=max{i,j}. Then

P(X=1)=P{(i,j):X(i,j)=1}=P{(1,1)}=1/36, P(X=2)=P{(1,2),(2,2),(2,1)}=3/36.

**Distribution Function**

**Definition** Let X be a random variable(r.v.). The function F : (-∞,∞)→[0,1] defined by F(x)=P{t: X(t)≤x} is the **distribution function** (d.f.) of the r.v. X.

**Note:** To emphasize the r.v. X, we sometimes denote F(x) by FX(x).

**Properties of Distribution Function**

1. If F is the d.f. of the r.v. X and if a<b, then P(a<X≤b)=F(b)-F(a).
* The events a<X≤b’ and X≤a are disjoint and their union is the event X≤b. Hence, by addition theorem of probability,

P(a<X≤b)+P(X≤a)=P(X≤b). Hence the result.

 **Corollary:** P(a≤X≤b)=P{(X=a) (a<X≤b)}=P(X=a)+ P(a<X≤b) =

 P(X=a)+ [F(b)-F(a)]. When P(X=a)=0, the events

 a≤X≤b and a<X≤b have same probability F(b)-F(a)

1. 0≤F(x) ≤1. If x<y, then F(x) ≤F(y).

**Discrete Random Variable**

A r.v. which can assume only at most countable number of real values is a discrete random variable. Example of discrete random variable are marks obtained in a test, number of accidents per month etc.

**Probability Mass Function**

If X is a one-dimensional discrete r.v. taking at most a countable number of values x1,x2,…, then the probabilistic behaviour of X at each xi is described by its probability mass function.

**Definition** If X is a discrete r.v. having distinct values x1,x2,…, then the function pX(x), or simply p(x), defined by p(x)=P(X=xi)=pi, if x=xi and =0, if x≠xi, i=1,2,… is called probability mass function(p.m.f.) of r.v. X.

**Note** (1)The set {(x1,p1),(x2,p2),…} specifies the probability distribution of the r.v. X.

(2)P(xi)≥0 , for all i and

**Example 4.4** Let S={H,T} be the sample space corresponding to the random experiment of tossing of a ‘fair’ coin. Let X be the r.v. defined by X(H)=1, X(T)=0. X has only two distinct values, namely, 0 and 1. The corresponding p.m.f. is given by: p(1)=P(X=1)=P(H)=1/2, p(0)=1/2.

**Example 4.5** A r.v. X hasthe following p.m.f.:

xi: 0 1 2 3 4 5 6 7

pi: 0 k 2k 2k 3k k2 2k2 7k2+k

(1)Find k, (2)Evaluate P(X<6),P(X≥6) and P(0<X<5), (3) If P(X≤a)>1/2, find the minimum value of a, (4) determine the p.d.f. of X.

* (1) Since , k+2k+2k+3k+k2+2k2+7k2+k=1 giving 10k2+9k-1=0, which gives k=1/10 or -1. Since p2=k cannot be negative, -1 is rejected and k=1/10.

 (2)P(X<6)=P(X=0)+P(X=1)+…+P(X=5)=1/10+2/10+2/10+3/10+1/100=81/100. Now P(X≥6)=1-P(X<6)=1-81/100=19/100.

1. P(X≤a)>1/2. By trial, we get a=4.
2. The p.d.f. of X is given by:

 X: 0 1 2 3 F(x)=P(X≤x): 0 k=1/10 3k=3/10 5k=5/10

X: 4 5 6 7

F(x) 8k=4/5 8k+k2 8k+3k2 9k+10k2

**Example 4.6** If p(x)=x/15, x=1,2,3,4,5; =0, elsewhere be the p.m.f. of a r.v.X. Find (1) P{X=1 or 2}, (2) P{.

* (1) P{X=1 or 2}=P(X=1)+P(X=2)=1/15+2/15=1/5.
1. P{.

**Example 4.7** An experiment consists of three independent tosses of a fair coin. Let X=the number of heads, Y=the number of head runs,Z=the length of head runs, a head run being defined as consecutive occurrence of at least two heads, its length then being the number of heads occurring together in three tosses of the coin. Find the probability function of (1)X, (2) Y, (3)Z,(4)X+Y and (5)XY.

* **Elementary Event Random Variables**

 **X Y Z X+Y XY**

HHH 3 1 3 4 3

HHT 2 1 2 3 2

HTH 2 0 0 2 0

HTT 1 0 0 1 0

THH 2 1 2 3 2

THT 1 0 0 1 0

TTH 1 0 0 1 0

TTT 0 0 0 0 0

(1)Obviously X is a r.v. which can take the values 0,1,2, and 3. p(3)=P(HHH)==, p(2)=P(HHT)=P(HHT)=P(HTH)+P(THH)=1/8+1/8+1/8=3/8. Similarly p(1)=3/8, p(0)=1/8.

(2) probability distribution of Y: p(0)=5/8, p(1)=3/8.

(3)probability distribution of Z: p(0)=5/8, p(1)=0,p(2)=2/8,p(3)=1/8.

(4) probability distribution of U=X+Y:p(0)=1/8, p(1)=3/8,p(2)=1/8,p(3)=2/8,p(4)=1/8.

(5) probability distribution of V=XY: p(0)=5/8,p91)=0,p(2)=2/8,p(3)=1/8.

**Continuous Random Variable**

**Definition** A r.v. X is continuous iff X takes all values between two unequal real numbers.

**Probability Density Function**

Consider a small interval (x, x+dx) of length dx about x. Let f(x) be any continuous function of x so that f(x)dx represents the probability that X falls in the infinitesimal interval (x, x+dx). Symbolically, P(x=fX(x)dx. fX is called probability density function (p.d.f.) of the r.v. X.

The probability for a variate value to lie in the interval [a,b] is P(aXb)=.

The p.d.f. f(x) of a r.v. X has the properties:

f(x)0, =1 (since gives total probability), P(X=c)==0(where c is any value of the variate X)

**Various measures of central tendency, dispersion, skewness and kurtosis for continuous probability distribution**

The formulae for these measures in case of discrete frequency distribution can be easily extended to the case of continuous probability distribution by simply replacing pi=fi/N by f(x)dx, xi by x and summation over ‘i’ by integration over the specified range of the variable X.

Let f(x) be the p.d.f. of a r.v. X, [a,b] being the range of X. Then

A.M. =, (central)=, (about x=A)=

Median is the point which divides the total area into two equal parts: if M be the median, then

Mode is the value of x for whixh f(x) is maximum. Mode is the solution of f/(x)=0 and f//(x)<0, provided it lies in [a,b].

**Example 4.8** The diameter of an electric cable , say X, is assumed to be a continuous random variable with p.d.f. f(x)=6x(1-x), 0x1. (1) Check that f(x) is a p.d.f., (2) determine the median b of the distribution.

* (1) (by direct calculation); hence f(x) is p.d.f. of r.v. X. (2) P(X<b)=P(X>b)⇒⇒ b=1/2, lying in [0,1].

**Example 4.9** Suppose that the life in hours of a certain part of radio tube is a continuous random variable X with p.d.f. given by f(x)=100/x2, when x100; =0, elsewhere. (1) What is the probability that all of three such tubes in a given radio set will have to be replaced during the first 150 hours of operation?(2)What is the probability that none of the original tubes will have to be replaced during the first 150 hours of operation?(3)what is the probability that a tube will last less than 200 hours if it is known that the tube is still functioning after 150 hours of service?

* (1)p=P(X150)=. By compound probability theorem, the probability that all three of original tubes will have to be replaced during the first 150 hours =p3=1/27.

(2)The probability that a tube is not replaced during the first 150 hours of operation is P(X>150)=1-P(X150)=1-p=2/3. By compound probability theorem, the probability that none of the three tubes will have to be replaced during the first 150 hours =q3=8/27.

(3)Probability that a tube will last less than 200 hours given that the tube is still functioning after 150 hours is P(X<200|X>150)===0.25.

**Example 4.10** The amount of bread (in hundreds of pounds) x that a certain bakery is able to sell in a day is found to be a numerical valued random phenomenon with a probability function specified by the p.d.f. f(x) given by f(x)=kx, 0x<5; =k(10-x), 5x<10; =0, otherwise. (1) find the value of k such that f(x) is a p.d.f., (2)what is the probability that the number of pounds of bread that will be sold tomorrow is (a) more than 500 pounds, (b) less than 500 pounds and (c) between 250 and 750 pounds? (3) Denoting by A,B,C the events that the pounds of bread sold are as in (2)(a),(2)(b) and (2)(c) respectively, find P(A|B),P(A|C). Are (1) A,B independent, (2) A,C independent?

* (1)=1 gives k=1/25.

(2)(a)P(510)==0.5

 (b)P(00.5

 (c)P(2.5=3/4

(3)From (2)(a),(b),(c), P(A)=0.5, P(B)=0.5, P(C)=3/4. The events AB and A are given by: AB=, A:5<X<7.5. Thus P(AB)=0,P(AC)==3/8.

P(A)P(C)=1/2 X ¾=3/8=P(A), P(A)P(B)=1/4 P(AB). Thus A,C are independent and A,B are not independent.

**Expectation of a r.v.**

Let the r.v. X take values x1,…,xn with probabilities p1,…,pn. Let X take value xi , fi number of times; let N=f1+…+fn. Mean of X is given by . Let N. Using the statistical definition of probability, limiting value of mean of X , .

**Definition** Expectation of X, E(X)=.

Thus E(X) may be regarded as the limiting value of the average value of X realized in N random experiments as N. Generalising, if f(X) is a function of X, f(X) will take values f(x1),…,f(xn) with frequencies f1,…,fn and the average value of f(X) in N experiments is f(x1)+…+f(xn) and as N, this approaches to E(f(X))=p1f(x1)+…+pnf(xn). In particular, =E[(X-a)r],r],2]=.

**Example 4.11** What is the expectation of the number of failures preceding the first success in an infinite series of independent trials with constant probability p of success in a trial?

* If X denotes the number of failures preceding the first success , we find that X takes the values 0,1,2,3,… with probabilities p,qp,q2p,q3p,…, where q=1-p. Thus probability density function is f(x)=qrp, r=0,1,2,…. Hence E(X)=0.p+1.qp+2.q2p+3.q3p+…=pq(1+2q+3q2+…)=pq(1-q)-2 (since q<1)=q/p = 1/p-1.

**Properties of Expectation**

1. Addition Theorem of Expectation: If X,Y are r.v., then E(X+Y)=E(X)+E(Y).
2. Multiplication Theorem of Expectation: If X,Y are independent r.v., E(XY)=E(X)E(Y).
3. If X is a r.v. and a,b are constants, then E(aX+b)=aE(X)+b, provided all the expectations exist.
4. If X0,then E(X)0.
5. If X,Y are r.v. and X(t)Y(t), forall t, then E(X)E(Y), provided all expectations exist.

**Example 4.12** Let X be a r.v. with the following probability distribution:

x: -3 6 9

P(X=x): 1/6 ½ 1/3

Find E(X) and E(X2) and using laws of expectation, evaluate E(2X+1)2.

* E(X)==(-3)=, E(X2)=. Then

E(2X+1)2=4E(X2)+4E(X)+1=209.

**Example 4.13**  Two unbiased dice are thrown. Find the expected values of the sum of numbers of points on them.

* The probability function of X(sum of number of heads on two dice) is

x: 2 3 4 … 12

P(X=x): 1/36 2/36 3/36 … 1/36

E(X)==7.

**Chapter V**

**Special Probability Distributions**

**Binomial Distribution**

Let a series of n trials be performed in which occurance of an event is called a ‘success’ and its a non-occurrence is called a ‘failure’.Let p be the probability of a success and q=1-p is the probability of a failure. We shall assume that trials are independent and probability p of success is same in every trial. The number of successes in n trials may be 0,1,2,…,n and is a randam variate. The probability of x succeses and n-x failures in a series of n independent trials in a specified order(say) SSFSFFF… FSF (S represents success and F represents failure) is given by compound probability theorem: P(SSFSFFF… FSF)=P(S)…P(F)=p..p(x factors)q…q(n-x factors)=pxqn-x. But x successes in n trials can occur in ways and the probability for each one of these ways are same, viz. pxqn-x. Hence the probability of x successes in n trials *in any order* is given by the addition theorem of probability by the expression pxqn-x. The probability distribution of the number of successes so obtained is called  **Binomial probability distribution**, for the obvious reason that the probabilities of 0,1,…,n successes viz. qn,p1qn-1,p2qn-2,…,pn are the successive terms of the binomial expansion of (q+p)n.

**Definition** A random variable X is said to follow binomial distribution with parameters n and p, written as X if it assumes only non-negative values and its p.m.f. is given by P(X=x)=p(x)=pxqn-x, x=0,1,…,n, q=1-p; =0,otherwise. A random variable which follows binomial distribution is called a **binomial variate**.

For a binomial distribution, following conditions must hold:

* Number of trials n is finite
* Trials are independent
* Probability of success p is constant for each trial
* Each trial results in one of two mutually exclusive and exhaustive outcomes, termed ‘success’ and ‘failure’

**Example 5.1**  Ten coins are tossed. Find the probability of getting at least seven heads.

* P=q=1/2, n=10. The probability of getting x heads in a random throw of 10 coins is p(x)==, x=0,1,…,10. Hence probability of getting at least 7 heads is P(X7)= p(7)+p(8)+p(9)+p(10)=.

**Example 5.2**  A and B play a game in which their chances of winning are in the ratio 3:2. Find A’s chance of winning at least three games out of the five games played.

* Let p be the probability that A wins the game. N=5,p=3/5, q=2/5. The probability that out of 5 games played, A wins ‘x’ games is given by:P(X=x)=, x=0,1,…,5.

The required probability that A wins at least three games is given by P(X3)== 0.68.

**Example 5.3**  A multiple-choice test consists of 8 questions with 3 answers to each question(of which only one is correct). A student answers each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4 and the third answer if he gets 5 or 6. To get a distinction, the syudent must secure at least 75% correct answers. If there is no negative marking, what is the probability that the student secures a distinction?

* Since there are three answers to each question, out of which only one is correct, the probability of getting a correct answer to a question is p=1/3,so that q=2/3. The probability of getting r correct answers in a8-question set is P(X=x)=p(x)= , x=0,1,…,8.

Hence the required probability of securing a distinction (that is, to get correct answers to at least 6 out of 8 questions) is given by: p96)+p(7)+p(8)==0.0197.

**Example 5.4**  The probability of a man hitting a target is ¼. (1) if he fires 7 times , what is the probability of his hitting the target at least twice? (2) How many times must he fire so that the probability of his hitting the target at least once is greater than 2/3?

* p=probability of the man hitting the target =1/4, q=1-p=3/4.

p(x)=probability of getting x hits in 7 shots=, x=0,1,…,7.

1. Probability of at least 2 hits=1-[p(0)+p(1)]=1-[]=.
2. Probabilty of at least one hit in n shots=1-p(0)=1-. It is required to find n such that 1- > , that is, >. Taking logarithms on both sides and simplifying, n>=3.8. Thus required number of shots is 4.

**Example 5.5**  In a Binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 successes are 0.4096 and 0.2048 respectively. Find the parameter ‘p’ of the distribution.

* Let X~B(n,p) where n=5, p(1)=0.4096, p(2)=0.2048. P(X=x)=px(1-p)5-x, x=0,1,..,5. Given p(1)=p1(1-p)5-1 = 0.4096,p(2)=p2(1-p)5-2=0.2048. Dividing, we get 2, p=0.2.

**Moments of Binomal Distribution**

Thus mean of Binomial distribution is np. It can be verified that =n(n-1)p2+np, (central)=npq,=npq(q-p).

**Note** If X~B(n,p), then mean=np, variance=npq. Hence variance<mean for a Binomial variate.

**Example 5.6**  The mean and the variance of a binomial distribution are 4 and 4/3 respectively. Find P(X1).

* Let X~B(n,p). Then np=4, npq=4/3. q=1/3.p=1-q=2/3. Hence n=4/p=6. Thus P(X1)=1-P(X=0)=1-qn=1-=0.99863.

**Poisson Distribution**

Poisson Distribution is a limiting case of Binomial Distribution under the following conditions:

* n, the number of trials , is indefinitely large, that is, n
* p, the constant probability of success for each trial is indefinitely small, that is, p0
* np=λ (say) is finite.

 A r.v. X is said to follow a Poisson distribution if it assumes only non-negative values and its p.m.f. is given by p(x,)=P(X=x)=, x=0,1,…,λ>0; =0,otherwise.

λ is known as the parameter of the distribution; we write X~P(λ) to denote X is a Poisson variate with parameter λ.

Following are some examples of Poisson variate:

* the number of typographical errors per page in typed material
* the number of defective screws per box of 100 screws
* the number of bacterial colonies in a given culture per unit area of microscope slab
* the number of deaths in a district in one year by a rare disease

**Moments of Poisson Distribution**

 λ, E(X2)= λ2+ λ,== λ, = λ.

**Example 5.7**  Find the probability that at most 5 defective fuses will be found in a box of 200 fuses if experience shows that 2 percent of such fuses are defective.

* n=200, p=probability of defective fuses=2%=.02. Since p is small, we may use oisson distribution. λ =mean number of defective pins=np=200(.02)=4. Thus required probability =P(X5)==e-4=.785.

**Example 5.8**  Six coins are tossed 6400 times. Using Poisson distribution, find the approximate probability of getting six heads r times.

* The probability of getting six heads in one throw of six coins (a single trial) is p=, assuming head and tail are equally probable. λ =np=6400=100. Thus required probability of getting 6 heads r number of times is P(X=r)=, r=0,1,2,…

**Example 5.9**  In a book of 520 pages, 390 typographical errors occur. Assuming Poisson law for the number of errors per page, find the probability that a random sample of 5 pages will contain no error.

* The average number of typographical error per page in the book is λ=390/520=0.75.

Using Poisson probability law,the probability of x errors per page is given by P(X=x)==, x=0,1,2,…. The required probability that a random sample of 5 pages will contain no error is given by [P(X=0)]5=(e-0.75)5=e-3.75.

**Normal Distribution**

**Definition** A r.v. X is said to have a normal distribution **with parameters**  (called ‘mean’) and 2(called ‘variance’) if its p.d.f. is given by the probability law: f(x;.

**Note** (1) When a r.v.is normally distributed with mean and standard deviation , it is customary to write X~N(2). If X~N(2), then Z=~N(0,1); Z is called corresponding standard normal variate. The p.d.f. of standard normal variate Z is given by . The corresponding distribution function, denoted by Φ(z)=P(Z.

**Few** **Properties of distribution function of standard normal variate**

* Φ(-z)=1- Φ(z)
* P(a Φ Φ, where X~ N(2).

**Chief Characteristics of the Normal Distribution and** **Normal Probability curve**

The normal probability curve with mean and s.d. is given by f(x)=,-. The curve has the following properties:

* The curve is bell-shaped and symmetrical about the line x=
* Mean,median and mode of the distribution coincide
* As x increases numerically, f(x) decreases rapidly, the maximum probability [p(x)]max= occurring at x=.
* Since f(x) (being probability)0, for all x, no portion of the curve lies below the x-axis
* x-axis is an asymptote to the curve f(x)
* =0, r=0,1,2,…
* mean deviation about mean=4/5 (approx.), quartile deviation=2(approx.)
* Area property: P( P(.

**Example 5.10**  For a certain normaldistribution, the first moment about 10 is 40 and the fourth moment about 50 is 48. What is the arithmetic mean and s.d. of the distribution?

* (about X=10)=40. Thus mean=10+=50. Also (about X=50)=48, that is, 4=48( since mean =50). But for a normal distribution with s.d. , 4=34=48 giving =2.

**Example 5.11**  X N(12,4). (a) Find the probability of (1) X20, (2)X20, (3)0X. (b) Find x such that P(X>x)=0.24.

* (a) For X=20, Z==2. Thus P(X20)=P(Z2)=0.5-P(0=0.5-0.4772=0.0228.

 P(X20)=1- P(X20)=1-.0228=.9722.

 P(0X=P(-3Z0)=P(0=0.49865.

(b)When X=x, Z==z1(say).GivenP(X>x)=P(Z>z1)=0.24;thus P(0<Z<z1)=0.26. From normal table, z1=0.71(approx..) Hence =0.71 giving x=14.84.

**Example 5.12**  The mean yield for one-acre plot is 662 kilos with s.d. 32 kilos. Assuming normal distribution, how many one acre plots in a batch of 1000 plots would you expect to have yield (1) over 700 kilos,(2)below 650 kilos and (3) what is the lowest yield of the best 100plots?

* If the r.v. X denotes the yield (in kilos) for one-acre plot, then X~ N(2), with , =32.
1. The probability that a plot has a yield over 700 kilos is given by P(X>700)=P(Z>1.19) [z= = 0.5-P(0Z1.19)=0.5-0.3830=0.1170. Hence in a batch of 1000 plots, the expected number of plots with yield over 700 kilos is 1000 x 0.117=117.
2. Required number of plots with yield below 650 kilos isgiven by 1000 x P(X<650)=1000 X P(Z<-0.38)[z=] =1000 x P(Z>0.38)=1000 X [0.5-P(00.38)]=1000[0.5-0.1480]=352.
3. The lowest yield , say, x,of best 100 plots is given by:P(X>x)==0.1. When X=x, Z==z1 (say) such that P(Z>z1)=0.1⇒ P(0Z1)=0.4⇒z1=1.28(approx.) (from normal tables). Thus x=662+32z1=702.96. Hence the best 100 plots have yield over 702.96 kilos.

**Example 5.13**  The marks obtained by a number of students for a certain subject are assumed to be approximately normally distributed with mean value 65 and s.d. 5. If 3 students are taken at random from this list, what is the probability that exactly 2 of them will have marks over 70?

* Let the r.v. X denote the marks obtained by the given set of students in the given subject. Given that X~ N(2), with , =5. The probability that a randomly selected student from the given set gets marks over 70 is given by p=P(X>70)=P(Z>1)=0.5-P(0Z=0.5-0.3413=0.1587. Since this probability is same for each student of the set, the required probability that out of 3 students selected at random from the set, exactly 2 will have marks over 70, is given by the binomial probability law: p2(1-p)=3 x (0.1587)2 x (0.8413)=.06357.

**Chapter VI**

**Correlation and Regression**

Often we come across situations in which our focus is simultaneously on two or more possibly related variables . If change in one variable affects a change in the other variable, the variables are said to be correlated. If increase in values of one variable results in increase in the corresponding values of the other variable, variables are said to be positively correlated; if increase in one variable results in decrement of values of the other variable, variables are negatively correlated. Correlation is said to be perfect if deviation in one variable is followed by a corresponding and proportional deviation in the other variable.

**Scatter Diagram**

It is simplest diagrammatic representation of bivariate data. Thus for the bivariate distribution (xi,yi), i=1,…,n; if the values of the variables X and Y are plotted along the x-axis and y-axis respectively in the x-y plane, the diagram of dots so obtained is known as scatter diagram. From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not: if the points are very close to each other, we should expect high correlation between the variables.

**Karl** **Pearson’s coefficients of correlation**

Let (xi,yi),i=1,…,n be a bivariate distribution of two r.v.s X and Y. Correlation coefficient between X,Y , usually denoted by r(X,Y) or by rXY, is a numerical measure of **linear relationship** between themand is defined as : r(X,Y)=, where Cov(X,Y)=E[{X-E(X)}{Y-E(Y)}]==-+ = -, ,.

**Note** r(X,Y) is independent of units of measurement of X,Y.

Karl Pearson’s correlation coefficient is based on the **assumptions**:

* There is a linear relationship between the r.v.s, that is, if the paired observations of both the variables are plotted on a scatter diagram, the plotted points will approximately be concurrent
* The variations in the two variables follow the normal law.

**Limits for value of correlation coefficient**

r(X,Y)=, where ai= xi-, bi= yi-. Now by Schwartz inequality, . Hence r21. Thus -1.

**Note** if r=+1, correlation is perfect and positive; if r=-1, correlation is perfect and negative.

**Effect of change of origin and scale of reference on correlation coefficient**

If U= , then . Hence if h,k are of the same sign, then ; if h,k are of opposite sign,.

**Note** r is independent of origin a,b.

**Note** two independent variables are uncorrelated; converse may not hold.

* If X,Y are independent, Cov(X,Y)=0; hence r(X,Y)==0. Two uncorrelated variables may not be independent:

X: -3 -2 -1 1 2 3

Y: 9 4 1 1 4 9

XY: -27 -8 -1 1 8 27

r(X,Y)=0 but X,Y are dependent: Y=X2.

**Example 6.1**  Calculate the correlation coefficient for the following heights(in inches) of fathers (X) and their sons(Y):

X: 65 66 67 67 68 69 70 72

Y: 67 68 65 68 72 72 69 71

* **Calculation for correlation coefficient**

**X Y X2 Y2 XY**

65 67 4225 4489 4355

66 68 4356 4624 4488

67 65 4489 4225 4355

67 68 4489 4624 4556

68 72 4624 5184 4896

69 72 4761 5184 4968

70 69 4900 4761 4830

72 71 5184 5041 5112

………………………………………………………………………………………..

544 552 37028 38132 37560

r(X,Y)===0.603.

**Short-cut Method**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | U=X-68 | V=Y-69 | U2 | V2 | UV |
| 65 | 67 | -3 | -2 | 9 | 4 | 6 |
| 66 | 68 | -2 | -1 | 4 | 1 | 2 |
| 67 | 65 | -1 | -4 | 1 | 16 | 4 |
| 67 | 68 | -1 | -1 | 1 | 1 | 1 |
| 68 | 72 | 0 | 3 | 0 | 9 | 0 |
| 69 | 72 | 1 | 3 | 1 | 9 | 3 |
| 70 | 69 | 2 | 0 | 4 | 0 | 0 |
| 72 | 71 | 4 | 2 | 16 | 4 | 8 |
| TOTAL |  | 0 | 0 | 36 | 44 | 24 |

=0,=0, Cov(U,V)===3, =36=4.5, =44=5.5. Thus rUV===0.603.

**Example 6.2**  A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results: n=25, It was however later discovered at the time of checking that he had copied down two pairs as (6,14),(9,6) while the correct values are (8,12),(6,8). Obtain the correct value of correlation coefficient.

* Corrected 125-6-8+8+6=125,corrected 100-14-6+12+8=100

Corrected =650-62-82+82+62=650, corrected =460-142-62+122+82=436,

Corrected =508-6 X 14-8 X 6+8 X 12+6 X 8=520.

Corrected = x 125=5, Corrected = x100=4.

Cov(X,Y)==4/5.=1, =36/25. Hence corrected rXY=0.67.

**Regression Analysis**

Regression Analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of data.

If the variables in a bivariate distribution are related, the corresponding points in the scatter diagram will cluster round some curve called’ curve of regression’. If the curve is a straight line, it is called ‘line of regression’ and there is said to be linear regression between the variables.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus ‘line of regression’ is the line of best fit and is obtained by principle of least squares.

Let us suppose that in the bivariate distribution (xi,yi), i=1,…,n, X is independent and Y is dependent variable. Let the line of regression of Y on X be Y=a+bX **(1)**. (1) represents a family of straight lines for different values of a and b. The problem is to find a and b corresponding to the line of ‘best fit’.

According to the principle of least squares, we have to find a,b so that E= is minimum. From the principle of maxima and minima, 0= and 0= which gives b **(2)**, **(3)**

Equations (2) and (3) are known as **normal equations** for estimating a and b. From (2), on dividing by n, we get **(4)** Thus the line of regression of Y on X passes through (,).

Now A(say)=Cov(X,Y)= ⇒=A+. **(5)**

⇒.**(6)**

From (3),(5) and (6), A+=a+b(). From (4) and (6), A=b giving b=A/.

Since the regression line of Y on X passes through (,) and has slope b=A/, its equation is

Y-(X-), that is, Y-r(X-).

Similarly the equation of line of regression of X on Y is given by X-=r(Y-).

**Note** In case of perfect correlation r=1 and in that case the equations of two regression lines coincide: =.

**Regression Coefficients**

bYX= r and bXY= r are called regression coefficient of Y on X and of X on Y respectively.

**Properties of regression coefficients**

* bYX bXY=r2. Thus r=. Since r=, bYX=, bXY=, it may be noted that sign of correlation coefficient is same as that of regression coefficients, since the sign of each depends on that of A. Thus, if the regression coefficients are positive, r is positive; if the regression coefficients are negative, r is negative. Hence the sign to be taken before the square root is that of the regression coefficients.
* If one of the regression coefficients is >1, then the other must be <1: bYX bXY=r21; if bYX=>1, then bXY<1.
* Regression coefficients are independent of change of origin but not of scale: if U=, V=, then bXY=bUV.

**Example 6.3** Obtain the equations of two lines of regressions for the following data. Also obtain the estimate of X for Y=70.

X: 65 66 67 67 68 69 70 72

Y: 67 68 65 68 72 72 69 71

* Let U=X-68, V=Y-69. Then ,5.5, Cov(U,V)=3,r(U,V)=0.6. Since correlation coefficient is independent of change of origin, r=r(X,Y)=r(U,V)=0.6.

=68+=68,=69+=69,===2.12,==2.35.

Equation of line of regression of Y on X is: Y-=r(X-), or, Y=0.665 X+23.78

Equation of line of regression of X on Y is: X-=r(Y-), or, X=0.54Y+30.74

To estimate X for given Y, we use line of regression of X on Y.If Y=70, estimated value of X is given by =0.54 X 70+30.74=68.54.

**Example 6.4** In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are available: Variance of X=9, Regression equations: 8X-10Y+66=0,40X-18Y=214. What are values of(1), (2)rXY, (3)?

* (1)Since () is the point of intersection of the lines of regression, solving given equations of lines of regression simultaneously, we get =13,=17.

(2) Comparing given equations of regression lines Y=, X=, we get bYX=, bXY=. Hence r2= bYX. bXY=. Hence r=. Since both the regression coefficients are positive, r==0.6.

(3)we have bYX=r. ; hence = , giving =4.

**Example 6.5** Find the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata from the following:

 Kolkata Mumbai

Average price 65 67

Standard Deviation 2.5 3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8.

* Let the prices (in Rs.) in Kolkata and Mumbai be denoted by X and Y respectively. Given =65, =67,=2.5,=3.5, r=0.8. We want Y corresponding to X=70.

Line of regression of Y on X is: Y-= r.(X-) ,or, Y=67+0.8 X (X-65).

When X=70, =67+0.8 X (70-65)=72.6.

Thus most likely price in Mumbai corresponding to the price of Rs .70 in Kolkata is Rs. 72.60.

**Example 6.6** Can Y=5+2.8 X and X=3-0.5Y be the estimated regression equations of Y on X and of X on Y respectively?

* Line of regression of Y on X is: Y=5+2.8X ;thus bYX=2.8

Line of regression of X on Y is:X=3-0.5Y; thus bXY=-0.5

This is not possible, since the regression coefficients bYX, bXY must be of the same sign. Hence given equations can not be taken as lines of regression.

**Curvilinear Regression**

In many situations, variables X and Y may be related non-linearly. Extending the method of finding regression lines using method of least square, we like to fit a parabolic curve Y=a+b1X+b2X2 to the given set (x1,y1),(x2,y2),…,(xn,yn) of n observations.

Using principle of least squares, we have to determine the constants a,b1,b2 so that E= is minimum. Equating to zero the partial derivatives of E w.r.t. a,b1,b2, we obtain the **normal equations** :

0==-2, 0=, 0=.

Simplifying, ,,. Solving these equations simultaneously, we get a,b1,b2 corresponding to the curve of best fit.

**Example 6.7** For 10 randomly selected observations, following data were recorded:

**Overtime hours(X):** 1 1 2 2 3 3 4 5 6 7

**Additional units(Y):** 2 7 7 10 8 12 10 14 11 14

Fit a parabolic curve to above data using method of least squares.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No. | X | Y | X2 | X3 | X4 | XY | X2Y |
| 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 |
| 2 | 1 | 7 | 1 | 1 | 1 | 7 | 7 |
| 3 | 2 | 7 | 4 | 8 | 16 | 14 | 28 |
| 4 | 2 | 10 | 4 | 8 | 16 | 20 | 40 |
| 5 | 3 | 8 | 9 | 27 | 81 | 24 | 72 |
| 6 | 3 | 12 | 9 | 27 | 81 | 36 | 108 |
| 7 | 4 | 10 | 16 | 64 | 256 | 40 | 160 |
| 8 | 5 | 14 | 25 | 125 | 625 | 70 | 350 |
| 9 | 6 | 11 | 36 | 216 | 1296 | 66 | 396 |
| 10 | 7 | 14 | 49 | 343 | 2401 | 98 | 686 |
| Total | 34 | 95 | 154 | 820 | 4774 | 377 | 1849 |

Corresponding normal equations are: 10a+34b1+154b2=95,34a+154b1+820b2=377, 154a+820b1+4774b2=1849. Solving , a=1.80,b1=3.48,b2=-0.27. Thus regression equation of Y on X is: Y=1.80+3.48X-0.27X2.

**Chapter VII**

**Index Numbers**

An index number may be defined as a measure of the average change in a group of related variables over two different situations. The group of variables may be the prices of a specified set of commodities, the volumes of production in different sectors of an industry, the marks obtained by a student in different subjects and so on. The two different ‘situations’ may be either two different times or two different places.

 The most commonly used index number is the index number of prices. Let p0 and p1 denote the prices of a commodity in suitable units in two different situations denoted by ‘0’and ‘1’. Any change in the price of the commodity from ‘0’ to ‘1’ may be expressed either in absolute or relative terms. The absolute change is p1-p0; the relative change is given by p1/p0, which is called a price relative. The problem is to combine these various individual changes in prices and get a measure of the overall change in the prices of the set of commodities. A **price index number** is a sort of average of these individual price relatives, and it measures the price changes of all the commodities collectively.

Although different commodities may have peculiar characteristics in their price fluctuations, it has been empirically found that , taken as a whole, the distribution of price relatives is bell-shaped with a marked central tendency, provided the base period is in the recent past. Hence we are justified in taking an appropriate measure of central tendency in combining the different price relatives.

Let us denote by p0i the price of i th commodity in the base period and by p1i the price of this commodity in the current period (i=1,…,k). If we use the arithmetic mean of price relatives for constructing the index number, then I01= is a simple or unweighted index number.

**Choice of weights**

The commodities included in the index number are not all of equal importance. For instance, in constructing a wholesale price index for India, ‘rice’ should have greater importance than ‘tobacco’. So the problem of weighting different commodities included in the index number according to their importance deserves attention. If we ignore weights, we get an inappropriately weighted index. If wi be the weight attached to the price relative for the i th commodity, then we get the weighted A.M.

I01= . Choosing different weight system, we get different index numbers:

* Choosing wi=q0i (the base period quantities) we get Laspeyres’ index: I01=.
* Choosing wi=q1i (the current period quantities) we get Paasche’s index: I01=
* Choosing wi=(q1i+q0i)/2, we get Edgeworth-Marshall index: I01=
* Fisher’s ‘ideal’ index: I01=

**Example 7.1** Table below gives the wholesale prices (p) and quantities produced (q) of a number of commodities in Delhi. Calculate Laspeyres’, Paasche’s, Edgeworth-Marshall and Fisher’s index numbers for the year 1985 , with the year 1982 as base.

**Commodity 1982 1985**

 **p q p q**

**Rice** 277.92 1.1 366.67 6.2

**Wheat** 176.25 106.0 186.58 116.9

**Jowar** 151.00 4.2 182.57 5.5

**Barley** 121.83 2.4 181.25 1.0

**Bajra** 156.75 13.1 155.75 6.1

**Gram** 273.00 1.0 498.83 0.6

* Let p0i,q0i and p1i,q1i denote the prices and quantities for 1982 and 1985, respectively. Then

=22241.229,=23921.766,=24399.034,=26519.314.

Thus Laspeyres’ Inex==107.56

 Paasche’s Index==108.69

Edgeworth-Marshall Index==108.15

and Fisher’s ‘ideal’ index number==108.12.

**Chapter VIII**

**Analysis of Time Series**

Time series is a series of observations recorded at different points or intervals of time. Maximum temperature of a place for different days of a month, yearly production of coal for last 20 years, monthly sales figure of some product are examples of time series data.

Let yt denotes the value of the variable y at time t (t=1,..,n).In case the figures relate to n successive periods (and not points of time), t is to be taken as the mid-point of the t th period.

**Components of time series**

A graphical representation of a time series shows continual change over time, giving us an overall impression of haphazard movement. A critical study of the series will, however, reveal that the change is not totally haphazard and a part of it, at least, can be accounted for. The systematic part which can be accounted for may be attributed to several broad factors: (1)**secular trend**, (2)**seasonal variation**, (3)**cyclical variation**. Separation of the different components of a time series is of importance, because it may be that we are interested in a particular component of the systematic variation or that we want to study the series after eliminating the effect of a particular component. It may be noted that it is the systematic part of the time series which may be used in forecasting.

**Secular Trend** or trend of a time series is the smooth, regular, long-term movement of the series if observed long enough. Sudden or frequent changes are incompatible with the idea of trend .

**Seasonal variation**

Seasonal variation stands for a periodic movement in a time series where the period is not longer than one year. It is the component which recurs or repeats at regular intervals of time. Example of seasonal fluctuation may be found in the passenger traffic during the 24 hours of a day, sales of a departmental store during the 12months of a year etc. The study and measurement of this component is of prime importance in certain cases. The efficient running of any departmental store , for example, would necessitate a careful study of seasonal variation in the demand of the goods.

**Cyclical Fluctuation**

By cyclical fluctuation we mean the oscillatory movement in a time series, the period of oscillation being more than a year. One complete period is called a cycle. The cyclical fluctuations are not necessarily periodic, since the length of the cycle as also the intensity of fluctuations may change from one cycle to another.

**Irregular Fluctuation**

This component is either wholly unaccountable or are caused by such unforeseen events as wars, floods, strikes etc.

**Estimation of secular trend in a time series by elimination of seasonal and cyclical fluctuation**

In order to measure the trend , we are to eliminate from the time series the other three components. If the period of seasonal fluctuations be a year, then the yearly totals or tearly averages will be free from the seasonal effect. Thus, in determining the trend from monthly data, it is customary to start with the yearly totals or averages, which are free from seasonal effects. The monthly trend values can be obtained from the annual trend values by interpolation.to eliminate the other two components, viz. the cyclical and the irregular, we may consider the following methods:

**Method of moving averages**

The simple moving average of period k of a time series gives a new series of arithmetic means, each of k successive observations of the time series. We start with the first k observations. At the next stage, we leave the first and include the (k+1)th observation. This process is repeated until we arrive at the last k observations. Each of these means is centered against the time which is the mid-point of the time interval included in the calculation of the moving average. Thus when k, the period of moving average, is odd, the moving average values correspond to tabulated time values for which the time series is given. When k is even,the moving average falls midway between two tabulated values. In this case, we calculate a subsequent two-item moving average to make the resulting moving average values correspond to the tabulated time periods.

The interpolation ofsimple moving averages is very simple. A k-point moving average may be interpreted as the estimated value for the middle of the period covered from successive linear curves fitted through the first k points, through the 2nd to the (k+1)th values and so on, and lastly through the last k points.

Consider the first k points y1,…,yk. Let the origin be shifted to the middle of the period so that =0. The normal equations for fitting a curve Y=a+bt through y1,…,yk are

=ka+b,

So that . Hence the estimated value for the middle of the period covered ,that is, for t=0, from the curve Y=+t is , which is the first moving average value. Similarly it can be shown that the estimated value from the fitted linear curve through y2,…,yk+1 would be , the second moving average value and so on.

A moving average with a properly selected period will smooth out cyclical fluctuations from the series and give an estimate of the trend. The central problem in this method is thus the selection of an appropriate method which will eliminate all fluctuations that d raw the series away from the trend.

**Example 8.1** Apply 3-year moving average to the following data on production of cements. Plot the data and the trend values on the same graph.

Year: 1992 1993 1994 1995 1996 1997 1998 1999

Output(in ‘000 tons) 1542 1447 1552 2102 2612 3195 3597 3567

* **Determination of trend values by 3-yearly moving average**

|  |  |  |  |
| --- | --- | --- | --- |
| year | Output(in ‘000 tons) | 3-yearly moving total | 3-yearly moving average(Trend values) |
| 1992 | 1542 | .. | .. |
| 1993 | 1447 | 4541 | 1513.7 |
| 1994 | 1552 | 5101 | 1700.3 |
| 1995 | 2102 | 6266 | 2088.7 |
| 1996 | 2612 | 7909 | 2636.3 |
| 1997 | 3195 | 9344 | 3114.7 |
| 1998 | 3537 | 10299 | 3433.0 |
| 1999 | 3567 | .. | .. |

**Example 8.2** Work out the trend values by 4-yearly moving average from the following data on production of iron ore( in ‘000 tons)

**Year:** 1983 ’84 ’85 ’86 ’87 ’88 ’89 ’90 ’91 ’92 ‘93

**Production:** 110125118 134 121 132 145 155 159 148 162

* Year production 4-yearly 4-yearly 4-yearly(centered)

 moving total moving average moving average

1983 110

1984 125

 487 121.75

1985 118 123.125

 498 124.5

1986 134 125.370

 505 126.25

1987 121 129.620

 532 133.00

1988 132 135.625

 553 138.25

1989 145 143.000

 591 147.75

1990 155 149.750

 607 151.75

1991 159 153.870

 624 156.00

1992 148

1993 162

**Example 8.3** Calculate the trend values by the method of moving averages from the following data on quarterly production(in ‘000 tons):

 **year**

**Quarter 1995 1996 1997**

**I** 15 15 20

**II** 19 22 21

**III** 21 23 25

**IV** 18 20 20

Is it possible to find the trend value for the first quarter of 1998 by the above method? Justify.

Year Quarter Production 4-Quarter 4-Quarter 4-Quarter(centered)

 (in’000 tons) moving total moving average moving average

 I 15 ----

1995 II 19 ----

 73 18.25

 III 21 18.250

 73 18.25

 IV 18 18.625

76

I 15 19.00

 78 19.250

1996 II 22 19.50

 80 19.750

 III 23 20.00

 85 20.625

 IV 20 21.25

 84 21.125

 I 20 21.00

 86 21.250

1997 II 21 21.50

 86 21.500

 III 25 21.50

 ----

 IV 20 -----

It is not possible to find the trend values for the 1st quarter of 1998 by the moving average method since there is no specific mathematical equation which can be used for interpolation or prediction purposes.

**Method of mathematical curves**

The trend values obtained by the method of moving averages , even though fairly smooth, is not representable by a simple mathematical formula.Since there does not exist any mathematically expressed trend equation, the method fails to achieve the main objective of trend analysis, that is, the interpolation and extrapolation of trend values.Therefore, attempt is made to fit the observed time series with a fairly simple mathematical curve. The fitting of mathematical curve has two parts: (1) determination of the appropriate trend curve, (2) determination of unknown parameters involved in the equation. From the graphical representation of the given time series , an investigator may guess the nature of the which fits the data best. The method is subjective in this sense. Determination of unknown constants appearing in the trend equation can be done by method of least squares.

**Example 8.4** Following table gives the number of hospital beds in West Bengal for the years 1979 to 1986. Plot of year versus no. of beds suggest that a linear trend will be appropriate to fit to the given data. The necessary data are done in table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | No.of beds | t=2(year-mid-period) | tyt | t2 | Tt=a0+a1t |
| 1979 | 55477 | -7 | -388339 | 49 | 55938 |
| 1980 | 58045 | -5 | -290225 | 25 | 57365 |
| 1981 | 58448 | -3 | -175344 | 9 | 58792 |
| 1982 | 59876 | -1 | -59876 | 1 | 60219 |
| 1983 | 61894 | 1 | 61894 | 1 | 61646 |
| 1984 | 63734 | 3 | 191202 | 9 | 63073 |
| 1985 | 64667 | 5 | 323335 | 25 | 64500 |
| 1986 | 65319 | 7 | 457233 | 49 | 65927 |
| Total | 487460 | 0 | 119880 | 168 |  |

Since =0, the normal equations are 487460=8a0, 119880=168a1 so that a0=60932.5, a1=713.57. The linear trend equation is , therefore, Tt=60932.5+713.57t.

**Example 8.5** table below shows the data on passenger-kilometer(millions) for Indian Railways during 1983 to 1989. Fit a quadratic trend :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Pass-Kilo | T=year-1986 | tyt | t2yt | t2 | t4 |
| 1983 | 6096 | -3 | -18288 | 54864 | 9 | 81 |
| 1984 | 6379 | -2 | -12758 | 25516 | 4 | 16 |
| 1985 | 6774 | -1 | -6774 | 6774 | 1 | 1 |
| 1986 | 7327 | 0 | 0 | 0 | 0 | 0 |
| 1987 | 7516 | 1 | 7516 | 7516 | 1 | 1 |
| 1988 | 7863 | 2 | 15726 | 31452 | 4 | 16 |
| 1989 | 8427 | 3 | 25281 | 75843 | 9 | 81 |
| Total | 50382 | 0 | 10703 | 201965 | 28 | 196 |

Here =0,=0. Hence the normal equations are

0==7a0+28a1, 10703==a2, 201965=28a0+196a1. Solving, a0=7176.63, a1=5.20, a2=382.25. Thus trend equation is Tt=7176.63+5.20t+382.25t2.

Quadratic Trend fitted to the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | T=year-1986 | a2t | a1t2 | Trend Tt=a0+a2t+a1t2 |
| 1983 | -3 | -1146.75 | 46.80 | 6076.68 |
| 1984 | -2 | -764.5 | 20.80 | 6432.93 |
| 1985 | -1 | -382.25 | 5.20 | 6799.58 |
| 1986 | 0 | 0 | 0 | 7176.63 |
| 1987 | 1 | 382.25 | 5.20 | 7564.08 |
| 1988 | 2 | 764.5 | 20.80 | 7961.93 |
| 1989 | 3 | 1146.75 | 46.80 | 8370.18 |