**Bivariate data**

The ‘bivariate’ is used to describe situations in which two characters are measured on each individual or item, the characters being represented by two variables. Statistical data relating to the simultaneous measurement of two variables are called bivariate data. The observations on each individual are then paired, one for each variable – (, ), (, ), (, ), ..., (, ).

For example, the variables may be (i) the amount of rainfall and yield of a certain crop, (ii) the height and weight of a group of children, (iii) income and expenditure of several families, (iv) ages of husband and wife, (v) marks obtained in mathematics and physics by the students in a class etc.

There are two main problems involved in such studies:

1. **Problem of correlation**: The data may reveal some association between and , and we may be interested to measure numerically the strength of this association between the variables. Such a measure will determine how well a linear or other equation explains the relationship between the variables. This is the *problem of correlation*. *In short, correlation is concerned with the measurement of the strength of association between variables.*
2. **Problem of regression:** There may be one variable of particular interest, and the other variable (regarded as an auxiliary variable) may be studied for its possible aid in throwing some light on the former. In such a case, one is then interested in using a mathematical equation for making estimates or predictions regarding the principal variable. This equation is known as regression equation, and the problem of making predictions on the basis of the equation is called the *problem of regression*. In short, *regression is concerned with the ‘prediction’ of the most likely value of one variable when the value of the other variable is known*.

**Bivariate frequency distribution**

Bivariate data on height (in inches) and weight (in lbs.) of 15 individuals are shown in the following table.

**Table 1:** Bivariate data on height and weight of 15 individuals

|  |  |  |
| --- | --- | --- |
| Person | Height (inches) | Weight (lbs) |
| 1 | 68 | 135 |
| 2 | 64 | 127 |
| 3 | 69 | 149 |
| 4 | 62 | 133 |
| 5 | 67 | 145 |
| 6 | 63 | 114 |
| 7 | 67 | 152 |
| 8 | 68 | 157 |
| 9 | 66 | 131 |
| 10 | 70 | 149 |
| 11 | 70 | 130 |
| 12 | 59 | 118 |
| 13 | 72 | 139 |
| 14 | 71 | 136 |
| 15 | 67 | 136 |

When a large number of pairs of observations are available, it becomes necessary to condense the data in the form of a two-way table, called ***bivariate frequency table or bivariate frequency distribution***. The method of constructing bivariate frequency distribution can be summarized as follows:

* Find out the maximum and minimum values in each of the - and -series, based on which chose the number of class intervals and the class limits.
* If class intervals are chosen for -series and class intervals are chosen for the -series, we construct a table with rows and columns, thus giving rectangular spaces, called cells.
* The class limits of the -series and -series are shown as row-headings and column-headings respectively.
* Each pair of observations is now represented by a tally mark (/) placed side by side, in the appropriate cell. Every fifth tally mark, in any cell, is however placed across the preceding four to form a group of 5 tally marks.
* When all pairs of observations have been shown in the two-way table by tally marks, these marks are counted and the numbers of observations falling in each cell, i.e. the cell frequencies, are counted.
* A two-way table thus formed with cell frequencies and grand total is called bivariate frequency distribution.

The bivariate frequency distribution of bivariate data on height and weight presented in Table 1 is shown below:

**Table 2:** Bivariate frequency distribution of height and weight

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Height (inches) | Weight (lbs.) | | | | | | |
|  | 111-120 | 121-130 | 131-140 | 141-150 | 151-160 | Total |
| 58-61 | 1 |  |  |  |  | 1 |
| 62-65 | 1 | 1 | 1 |  |  | 3 |
| 66-69 |  |  | 3 | 2 | 2 | 7 |
| 70-73 |  | 1 | 3 | 1 |  | 4 |
| Total | 2 | 2 | 6 | 3 | 2 | 15=N |

**Marginal distribution:** A univariate frequency distribution, derived from the bivariate frequency distribution, irrespective of the values assumed by the other variable, is called a Marginal distribution**.**

The totals of cell frequencies shown in the last row and the last column in a bivariate frequency table are called marginal totals or marginal frequencies.

In case of Table 2, the marginal totals in the last column represent class frequencies of the distribution of height, called ‘marginal distribution of height’ whose class intervals are shown in the first column. Similarly, the marginal totals in the last row represent class frequencies of the distribution of weight, called ‘marginal distribution of weight’, whose class intervals are shown in the first row.

**Conditional distribution:** A univariate frequency distribution, derived from the bivariate frequency distribution, for a specified value (or class interval) of the other variable, is called a conditional distribution.

There is only one marginal distribution of and one marginal distribution of . But there is one conditional distribution of corresponding to each class interval of ; and similarly, corresponding to each class interval of , there is one conditional distribution of . For example, the conditional distribution of height (when weight 131 lb -140 lb) is given by the cell values in the third column and the conditional distribution of weight (when height 70-73 inches) is given by the cell values in the fourth row.

The conditional mean value is the arithmetic mean calculated from the conditional distribution.

**Scatter diagram:** When statistical data relating to the simultaneous measurement on two variables are available, each pair of observations can be geometrically represented by a point on the graph paper – the values of one variable being shown along -axes and those of the other variable along -axis. If there are pair of observations, finally the graph paper will contain points. This diagrammatic representation of bivariate data is known as scatter diagram.

**Correlation**

The word correlation is used to denote the degree of association between variables. If two variables and are so related that changes in the magnitude of one variable tend to be accompanied by changes in the magnitude of the other variable, they are said to be correlated.

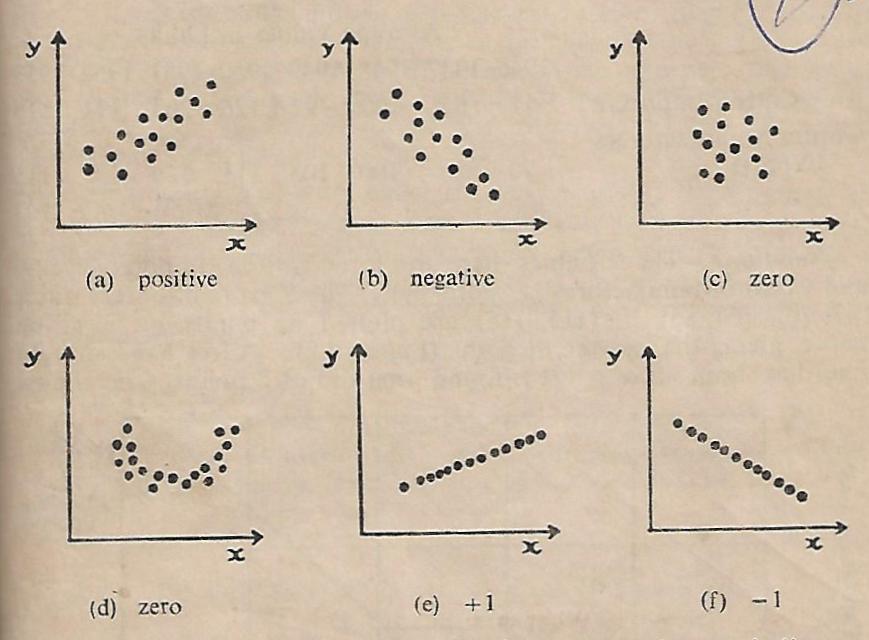
If tends to increase as increases, the variables are said to be ***positively correlated***.

If tends to decrease as increases, the variables are said to be ***negatively correlated***.

If values of are not affected by changes in the values of , the variables are said to be ***uncorrelated***.

Correlation may also be *linear* or *non-linear*. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, the correlation is said to be linear, because the scatter diagram will show a linear path. Here we shall be concerned with *linear correlation or simple correlation* only. This is measured by *correlation coefficient*.

The easiest way to understand the type and degree of correlation between two variables is to examine the scatter diagram. Different types and degrees of correlation exhibited by scatter diagrams are shown below.

****

**Fig. 1**: Scatter diagrams showing different types and degrees of correlation

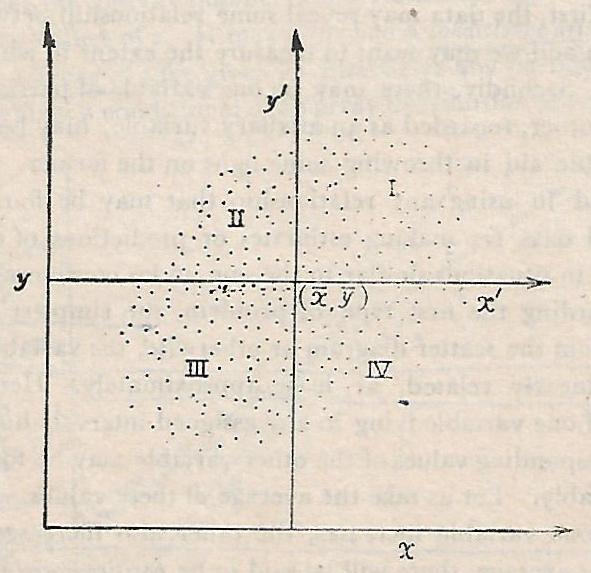
**Correlation coefficient**

In the scatter diagram, let us take two axes of co-ordinates for and . The origin of the new axes must be the point (, ), in terms of the original co-ordinates. The points of the scatter diagram may now be seen as distributed over four quadrants (I-IV) of the (,) plane.

It may be noted that

* in quadrant I, and are both positive
* in quadrant II is negative and is positive
* in quadrant III and are both negative
* in IV is positive and negative

Hence, is positive for all points occurring in quadrants I and III, while it is negative for all points in quadrants II and IV.



**Fig. 2:** A scatter diagram with the four quadrants of the (,) plane

* In case of positive correlation, the general tendency of the points is to lie in quadrants I and III, so that in the sum the positive values outweigh the negative ones; and the sum thus becomes a positive quantity.
* In case of negative correlation, the general tendency of the points is to lie in quadrants II and IV, so that in the sum the negative values now outweigh the positive ones; and the sum thus becomes a negative quantity.
* On the other hand, when there is no correlation, the points are equally distributed over the four quadrants, and the sum becomes zero, as the set of positive values (products) and that of negative values (products) just balance each other.
* Consequently, a natural measure of correlation will seem to be the sum

(1)

But this sum is also dependent on some factors that have nothing to do with the correlation between the variables. For example, (i) it depends on the number of pairs of values of and which are taken into account, i.e. ; (ii) it also depends on the units in which the variables and are measured and also on their variability.

The first defect can be removed by dividing the sum by the number of pairs and the second defect can be eliminated by dividing the sum by the product of the standard deviations, and (both assumed to be ). Thus, the measure of correlation becomes as

(2)

This measure i.e. is called the (product-moment) *correlation coefficient* of the variables and . Often, it is represented as .

The quantity in the numerator of is called the covariance of and [represented by ], in analogy with the term variance that is used in case of a single variable. Since the standard deviations, and are the positive square-roots of the variances of and [represented by and ], one may also write

(3)

Again,

Similarly,

Hence, may be expressed in the alternative form alternative form as

(4)

The last form is found to be the most convenient for computing from raw data.

**Calculation of correlation coefficient from group data**

Suppose the values of and are given in the form of a bivariate frequency table with classes for and classes for . Let us denote by the mid-point of the class of and by the mid-point of class of . These new classes define the ()th cell of the bivariate frequency table. Let denote the frequency in this cell. Then the correlation coefficient is obtained using the following formula:

where , and , and

**Properties of the correlation coefficient**

1. The correlation coefficient of and is a pure number and is independent of the units of measurement of and .
2. The correlation coefficient is independent of the choice of both origin and scale of observations. This means that if and , where , , , ( and positive) are four arbitrarily chosen constants, then .

***Proof:*** Let and be the variables, and (,), (,),..., (,) denote pairs of observations. Let us change the origins of and to and , the units of measurement (i.e. scales) to and respectively, and write

and (1)

where , , , ( and positive). We have to prove that the correlation coefficient between and is the same as that between and , i.e. .

From (1), and

and

Hence,

(since is positive)

Similarly, (since is positive)

Again,

Therefore,

1. If and (i=1,2,...,n) where , , , are arbitrary constants, then , if , are of the same sign, and , if they are of opposite sign.

***Proof:*** , and . These results are true whatever be the signs of and . But since the standard deviation is always positive, we must have and

If and are of the same sign, the product will be positive, exactly equals to , which is always positive. Therefore,

, so that

If and are of opposite signs, the product will be negative but numerically equal to . Therefore,

, so that

1. The correlation coefficient lies between -1 and +1; i.e. cannot exceed 1 numerically. This implies that .

***Proof:*** Let and be the variables, and (,), (,),..., (,) denote pairs of observations with means , and standard deviations and respectively. If we write,

and

Then

Similarly,

Again,

where denotes the correlation coefficient between and .

Now,

or,

or,

or,

or,

or, , i.e.

Similarly,

or,

or,

or,

or,

or, , i.e.

Combining the results and , we get .

**Example:** Computation of correlation coefficient from given numerical data

Calculate the coefficient of correlation from the following data:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2.52 | 2.49 | 2.49 | 2.45 | 2.43 | 2.42 | 2.41 | 2.40 |
|  | 730 | 710 | 770 | 890 | 970 | 1020 | 970 | 1040 |

Since the correlation coefficient of correlation is unaffected by changes of origin and scale, let us write and . Then the correlation coefficient can be computed easily from the values in the following table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sl No. |  |  |  |  |  |  |  |
| 1 | 2.52 | 730 | 7 | -17 | 49 | 289 | -119 |
| 2 | 2.49 | 710 | 4 | -19 | 16 | 361 | -76 |
| 3 | 2.49 | 770 | 4 | -13 | 16 | 169 | -52 |
| 4 | 2.45 | 890 | 0 | -1 | 0 | 1 | 0 |
| 5 | 2.43 | 970 | -2 | 7 | 4 | 49 | -14 |
| 6 | 2.42 | 1020 | -3 | 12 | 9 | 144 | -36 |
| 7 | 2.41 | 970 | -4 | 7 | 16 | 49 | -28 |
| 8 | 2.40 | 1040 | -5 | 14 | 25 | 196 | -70 |
| Total | 19.61 | 7200 | 1 | -10 | 135 | 1258 | -395 |

= =

=

= =

**Limitations of correlation coefficient**

1. In linear correlation, it is assumed that there is a straight line relationship between the variables. A small value of therefore indicates only a poor linear type relationship between the variables. This however, does not rule out the possibility that the association is very close, but the relationship is non-linear.

For example let us consider the following data:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|  | 9 | 4 | 1 | 0 | 1 | 4 | 9 |

Here, , , ,

Therefore,

Hence,

i.e. the correlation coefficient between and is zero. But it may be noticed that and are bound by the relation . So, and are not independent. Thus, the correlation coefficient may be zero, even when the variables are not independent.

Therefore, before using as a measure of the degree of association between the variables, it is advisable to draw a scatter diagram and see whether the pattern of points is linear or not.

2) Also a high value of does not imply there is a direct cause and effect relationship between the variables. For example, there will be high correlation between age of husband and age of wife, although none may be said to be caused by other. The high correlation is really an effect of the prevailing social norms that lead older husbands to have older wives (and vice versa) and younger husbands to have younger wives (and vice versa).

The high value of may be generated solely due to the influence of a third variable affecting both. In this case, the effect of the third variable should be eliminated from the first two and then the partial correlation between them should be found out.

**3)** Sometimes it may happen that two series of observations show a high correlation coefficient even though there is no logical basis for any relationship between them. For example, a renowned statistician observed a high correlation between the stork population in Oslo over a period of several years and the number of babies born there each year. Yet it is hard to develop a theory as why this should be. Such correlation is said to be ***Spurious correlation or Non-sense correlation***. One should apply common-sense in deciding whether the association indicated by the value of is real or spurious.

Another example of a spurious relationship can be seen by examining a city's [ice cream](https://en.wikipedia.org/wiki/Ice_cream) sales. The sales might be highest when the rate of drowning in city [swimming pools](https://en.wikipedia.org/wiki/Swimming_pool) is highest. To claim that ice cream sales cause drowning, or vice versa, would be to imply a spurious relationship between the two. In reality, a [heat wave](https://en.wikipedia.org/wiki/Heat_wave) may have caused both. The heat wave is an example of a hidden or unseen variable, also known as a [confounding variable](https://en.wikipedia.org/wiki/Confounding_variable).

Another example of spurious relation may be the correlation between the cost of higher education and cost of living. Often it is noticed that both the cost of higher education and the cost of living increase. However, this change in both variables does not necessarily mean there is a causal relationship between the two. The increase in the cost of living isn't necessarily the cause of the increase in higher education tuition. The change in both costs may be attributed to the rise in inflation or other macroeconomic factors, resulting in a spurious correlation.

4) If the data are not reasonably homogeneous the coefficient of correlation may give a misleading picture of the extent of association. For example, if the scatter diagram shows the points in separate clusters or groups, the correlation coefficient based on all groups taken together may be very high; yet if separate values of are computed from each group, they may be close to zero. If some reasonable basis can be found for separating the data into groups, it is desirable to compute values of for each group.

For example, consider the scatter diagram of experience level and number of errors in an insurance company.

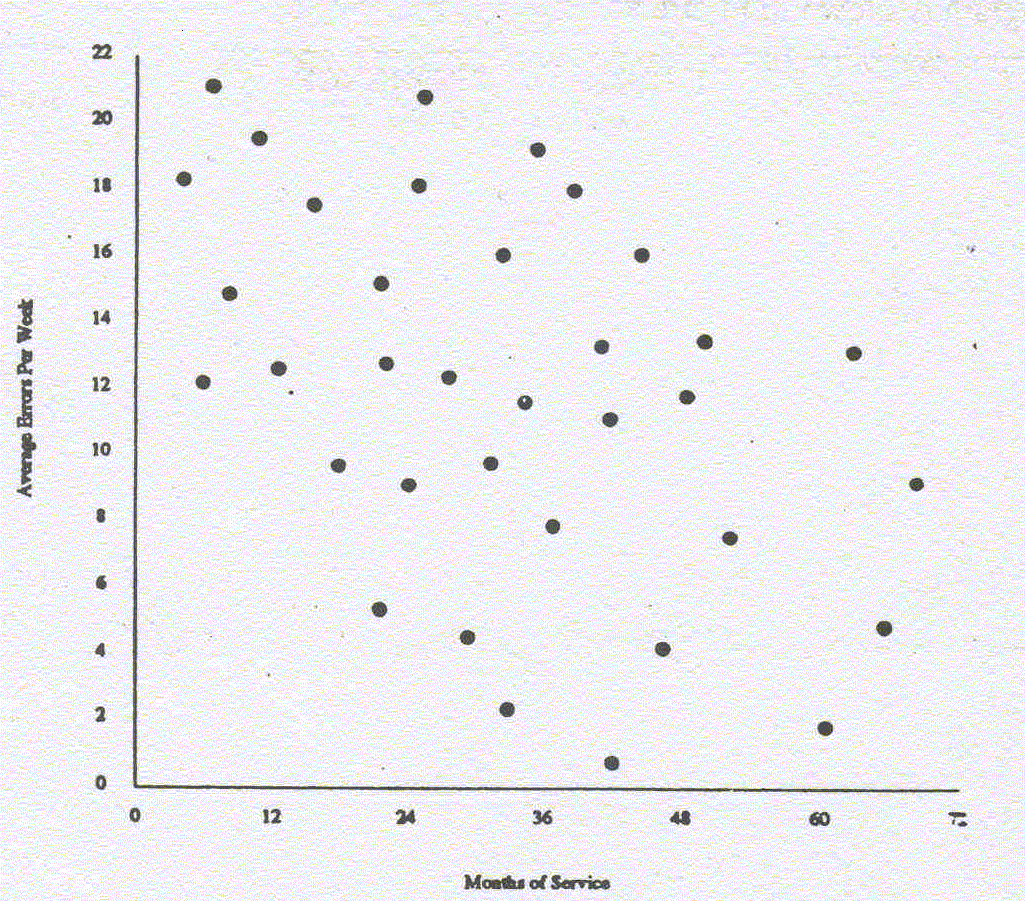


Fig. 2: Scatter diagram of errors in the insurance company

The scatter diagram indicated some correlation between experience level and number of errors.

Now look at the following scatter diagram:

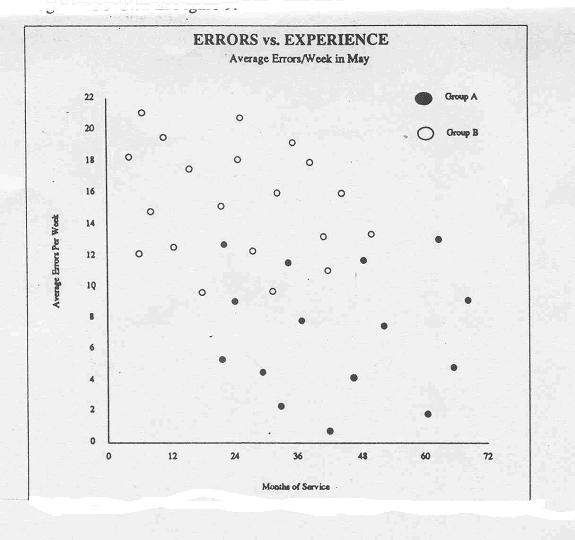


Fig. 3: Stratified scatter diagram of errors in the insurance company

When looked at each group separately, it was found that there was no real correlation between experience and number of errors. The cause of the errors was more related to the lack of support systems in group B than anything inherent to experience. It just happened that the more experienced clerks were in group A.

**Exercises**

**Exercise 1:** While calculating the coefficient of correlation between two variables x and y, the following results were obtained: = 25, , , , , . It was however later discovered at the time of checking that two pairs of observations () were copied (6, 14) and (8, 6) while the correct values were (8, 12) and (6, 8) respectively. Determine the correct value of the coefficient of correlation. [Ans. r=2/3]

**Exercise 2:** Marks of 10 students in Mathematics and Statistics are given below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Math () | 32 | 38 | 48 | 43 | 40 | 22 | 41 | 69 | 35 | 64 |
| Stat () | 39 | 31 | 38 | 43 | 33 | 11 | 27 | 76 | 40 | 59 |

Calculate product-moment correlation coefficient and its standard error [].

**Exercise 3:** The following table gives the index numbers of industrial production in a country and the number of registered unemployed persons in the same country during the eight consecutive years. Calculate the coefficient of correlation and comments on the results.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1954 | 1955 | 1956 | 1957 | 1958 | 1959 | 1960 | 1961 |
| Index of industrial production | 100 | 102 | 103 | 105 | 106 | 104 | 103 | 98 |
| No. of registered unemployed (in K) | 10.5 | 11.4 | 13.0 | 11.5 | 12.0 | 12.5 | 15.6 | 20.8 |

