# 1. Introduction

Welcome to the series of e-learning modules on bivariate data and plotting the bivariate data. In this module we learn about the meaning of bivariate data, purpose of having bivariate data, identifying independent and dependent variables in the bivariate data and plotting them to find the relation between the two variables, constructing contingency table using the scatter plot and the differences between univariate and bivariate data.

By the end of this session, you will be able to:

- Define Bivariate data
- Understand the purpose of having bivariate data
- Identify independent and dependent variables
- Plot the bivariate data
- Identify the relationship between two variables using scatter plots
- Define the types of relationship like positive, negative etc
- Illustrate the use of contingency table
- Construct contingency table using scatter plots
- Differentiate between univariate and bivariate data

If we see the word bivariate data, it has two parts, bivariate and data.
Data means information or knowledge.
Now let us split the term bivariate as ‘bi’ and ‘variate’.
The meaning of ‘bi’ is 2 and ‘variate’ means variables.
Hence if we have any data set with two variables, then it is called bivariate data.

Bivariate data consists of two *quantitative variables* for each individual.
Bivariate data deals with relationships between these two variables.
The purpose of bivariate data is to analyze and explain this relationship.

For example, in large health studies of populations, it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol of each individual
Or say, economic studies may be interested in, among other things, personal income and years of education.
Also, most of the university admissions committees ask for an applicant's high school grade point average and standardized admission test scores.

The relationship between 2 variables is often of interest.
For example, are height and weight related? Are age and heart rate related? Are income and taxes paid related? Is a new drug better than an old drug? Does the weather depend on the previous day’s weather? Exploring and summarizing such relationships is the current goal.
There can be positive relationship, negative or inverse relationships and then there are variables which have no relationship with one another.
A positive relationship would exist if as one variable increases, the other variable increases or, if as one variable decreases, the other variable decreases. And a negative or inverse relationship would exist of as one variable increases, the other variable decreases or if as one variable decreases, the other variable increases.

# 2. Examples of Positive Relationship and Negative Relationship

An example of a positive relationship would be between the two variables temperature and population at a beach. It is commonly known that as the weather gets warmer more people go to the beach. This means that, as the temperature increases the amount of people at the beach will increase.
This also means that as the temperature decreases there are fewer people at the beach. Both of these situations are positive relationships because as one variable either decreases or increases the other variable does the same.

An example for negative relationship would be, say, often as the amount of time you exercise increases, the less time it takes to run a mile. Here the two variables, negatively related. That is as one variable increases, the other decreases. Hence here exercise time and time required to run a mile are negatively correlated.

While examining a relationship between two sets of variables is, it is useful to know whether one of the variables depends on the other.
Consider a case where a study of comparing the heights of the company employees against their salaries.
In this case it is not appropriate to designate one variable as independent and other as dependent.

Here is a case where the age of the company employees are compared with their annual salaries. In this case, the age of the employee is the independent variable where as the salary of the employee is the dependent variable.

It is useful to identify the independent and dependent variables whenever possible since it is the usual practice when displaying the data on the graph to place the independent variable on the horizontal X axis and the dependent variable on the vertical axis Y axis.

We can denote bivariate data either in raw form or in tabulated form.
Here we write two variables and values taken by these two variables.
Let us consider the following example.

The number of hours of study and marks scored in the test of nine students are recorded. Here the example contains two variables. One, the number of hours of study and two, the test scores.

Since here we have only 09 students, we can write in the raw form as given in a table. Observe that if a student studies for 3 hours, he secure 91 marks, if he studies for 1 hour he scores 86 and so on.

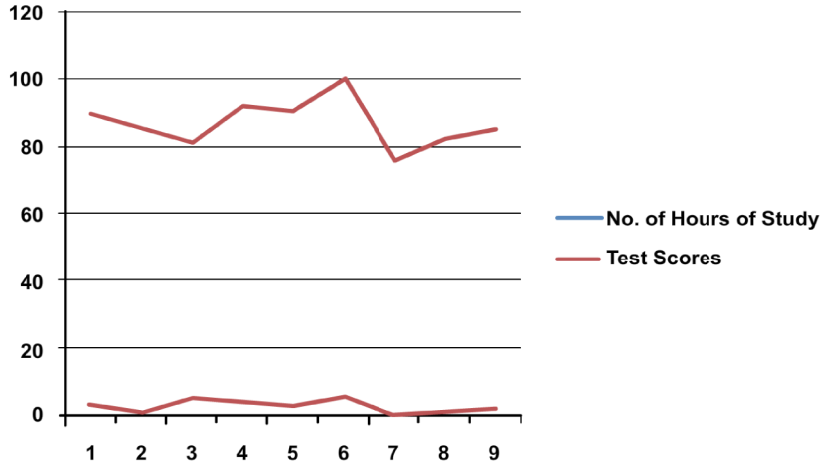
**Figure 1**

| Number of Hours of Study | Test Scores |
|---|---|
| 3 | 90 |
| 1 | 86 |
| 5 | 84 |
| 4 | 92 |
| 3 | 91 |
| 5 | 100 |
| 0 | 76 |
| 1 | 82 |
| 2 | 85 |

# 3. Relationship Between two Variables - Illustrations

To see whether there is any relationship between two variables, we can sketch the variables on a graph and examine.

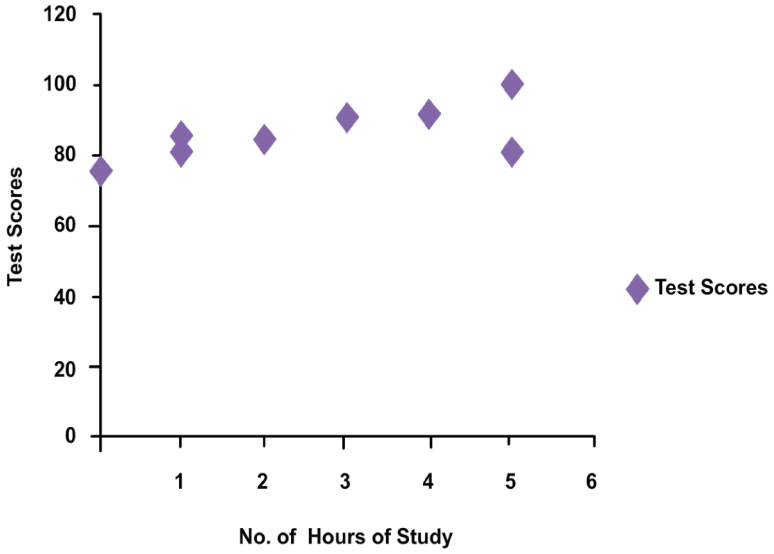
**Figure 2**




From the graph observe that, as number of hours of study increases, which is shown by blue line, the test score also increases and number of hours of study decrease, the test score also decreases.
Hence we can say there is positive association between number of hours of study and test scores.

For the same data we can have a scatter plot also.

**Figure 3**



Here, the marks scored by a student depend on how much time the student spends studying. Hence the time spent by the student is the independent variable which controls the other is taken along X axis, and the marks scored by the student is taken along the Y axis.

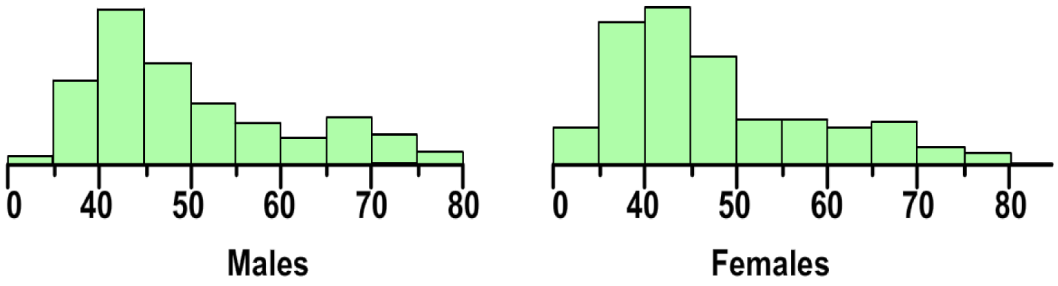
The scatter plot shows that as number of hours of study increases, the test scores also increases. Hence there is a positive relationship between the two variables.
Observe that scatter plot gives a better idea than that of a line graph.

We can draw different graphs to identify the nature of two distributions and can verify whether the two variables are related or not.

Here is another illustration.
Suppose a dataset consisting of 282 pairs of spousal ages are available then it is too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages.

We know that each variable can be summarized by a histogram as shown below.

**Figure 4**




Observe that, each distribution is fairly skewed with a long right tail.

We can learn much more by displaying the bivariate data in a graphical form using scatter plot which maintains the pairing.

**Figure 5**


There are two important characteristics of the data revealed in the scatter graph.

First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. Second, when one variable, here 'Y' increases with the second variable, which is 'X' in this case, we say that X and Y have a positive association.

In the next illustration, we have data indicating the marks scored by 10 students in mathematics and statistics.

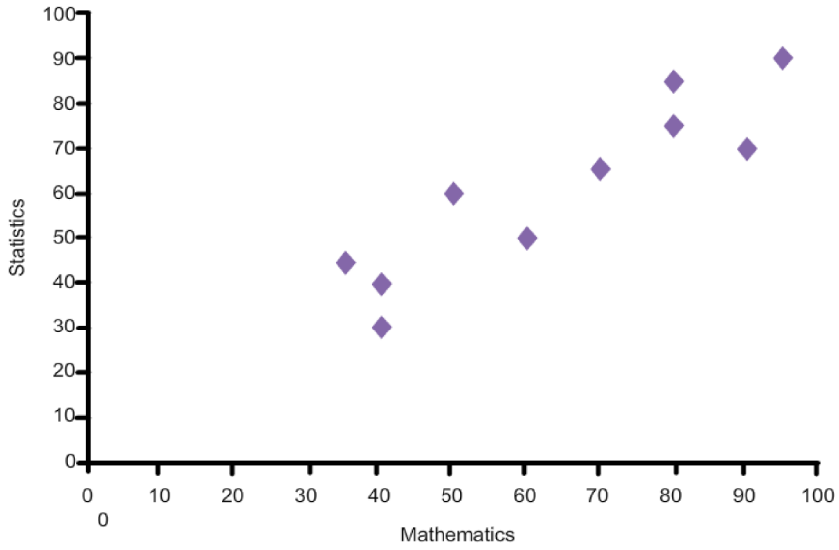
**Figure 6**

| Marks in Statistics | Marks in Mathematics |
|---|---|
| 45 | 35 |
| 70 | 90 |
| 65 | 70 |
| 30 | 40 |
| 90 | 95 |
| 40 | 40 |
| 50 | 60 |
| 75 | 80 |
| 85 | 80 |
| 60 | 50 |

If one is interested to know about a student's mathematical ability, one can compare the marks in statistics and mathematics scored by a student.
Here it is difficult to identify the independent and dependent variable, since we cannot control one variable to have the other. But by intuition, if a student has good mathematical ability, he or she is expected to score well in statistics. Hence we take Marks in mathematics in X axis and the marks in Statistics along the Y axis.

To see whether there is any relationship between these two variables, we can draw the scatter plot as follows.

**Figure 7**



From the scatter diagram, we can see that there is a positive association between the two variables. That is, the student who has scored more in statistics has scored more in mathematics also.

Next, if we are interested to whether the soil temperature affects the germination time of the seed, consider the following bivariate data which gives the soil temperature and the germination time at various places.

**Figure 8**

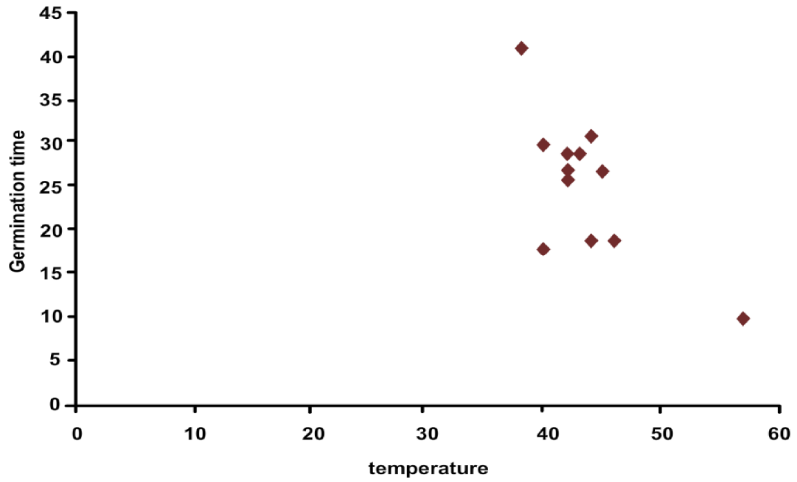
| Temperature | Germination time |
|---|---|
| 57 | 10 |
| 42 | 26 |
| 40 | 30 |
| 38 | 41 |
| 42 | 29 |
| 45 | 27 |
| 42 | 27 |
| 44 | 19 |
| 0 | 18 |
| 46 | 19 |
| 44 | 31 |
| 43 | 29 |

The table in the right shows the soil temperature and the germination time required for the seeds at various places.

# 4. Construction of Scatter Plot

To construct the scatter plot, first we identify the independent and dependent variable. In this case one is dependent on the other. Here the temperature controls the time needed for the germination of seeds. Hence we can consider the temperature as independent variable and germination time as dependent variable. Therefore we take Temperature along the X axis and Germination time along the Y axis.

**Figure 9**




From the above scatter plot we can see that there is slightly negative association between soil temperature and germination time as the line has moved from left top to right bottom.

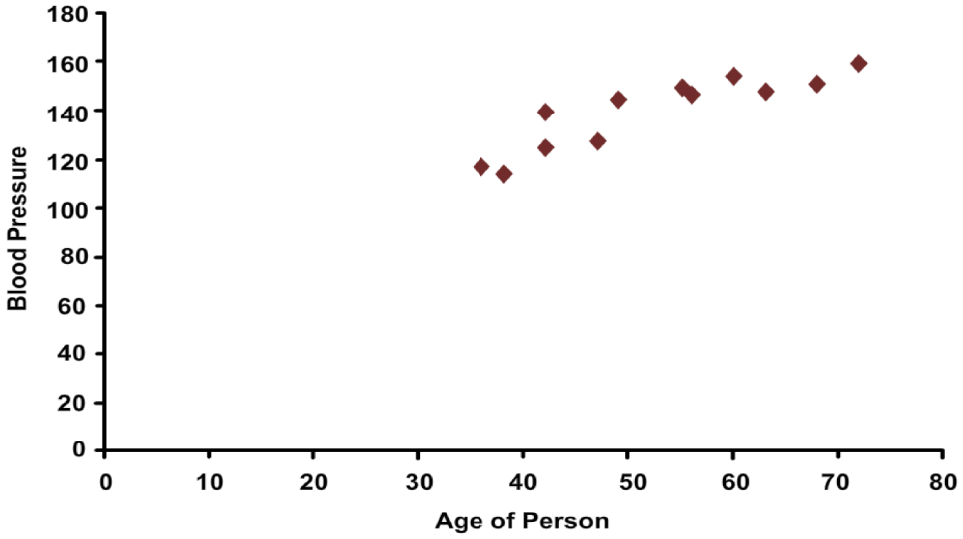
In this example, we are required to verify whether the age of a person is related to blood pressure of that person or not. Let us consider the following sample of twelve men.

**Figure 10**

| Age of a person | Blood Pressure |
|---|---|
| 56 | 147 |
| 42 | 125 |
| 72 | 160 |
| 36 | 118 |
| 63 | 149 |
| 47 | 128 |
| 55 | 150 |
| 49 | 145 |
| 38 | 115 |
| 42 | 140 |
| 68 | 152 |
| 60 | 155 |

Let us create a scatter plot for the above data. Since age of a person influence the Blood Pressure we can consider age as independent variable and Blood Pressure as dependent variable. Hence we take age along the X axis and Blood pressure along the Y axis.

**Figure 11**




From the scatter plot observe that as age increases, the blood pressure also increases. We do not get a straight line but the points are clustered around a straight line showing the positive association between age of a person and blood pressure.

Next, Let us consider the data related to arm span and height of a person given in centimetres. Here we are interested to verify whether there is any relation between these two variables.

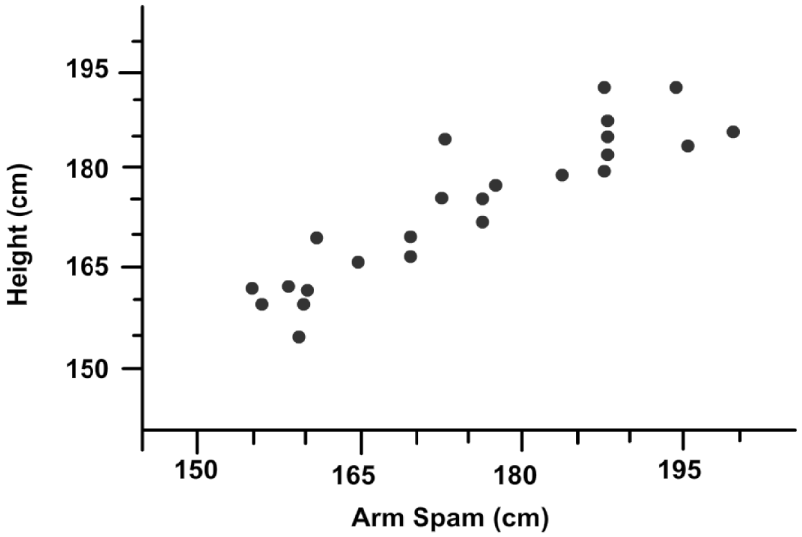
The data given below are sorted by arm span. Here we can take any variable as independent and other as dependent as one does not control the other but we have to remember that they are interdependent.

**Figure 12**

| Arm Span | Height | Arm Span | Height |
|---|---|---|---|
| 156 | 162 | 177 | 173 |
| 157 | 160 | 177 | 176 |
| 159 | 162 | 178 | 178 |
| 160 | 155 | 184 | 180 |
| 161 | 160 | 188 | 188 |
| 161 | 162 | 188 | 187 |
| 162 | 170 | 188 | 182 |
| 165 | 166 | 188 | 181 |
| 170 | 170 | 188 | 192 |
| 170 | 167 | 194 | 193 |
| 173 | 185 | 196 | 184 |
| 173 | 176 | 200 | 186 |

Hence to construct the scatter plot, we take arm span along the X axis and height of the person along Y axis.

**Figure 13**




From a scatter plot observe that as arm span increases, the height also increases. Therefore there is a positive association between the two variables arm span and height of the person.

We shall look at another example.

A group of students wanted to know whether there was a relationship in the height from which a ball was dropped and its rebound height. Using a basketball, they dropped the ball from each of 11 heights three times and measured how high it rebounded. Both the height from which the ball was dropped and the height of the rebound were measured in inches, from the bottom of the ball.

The data are given in the following table.

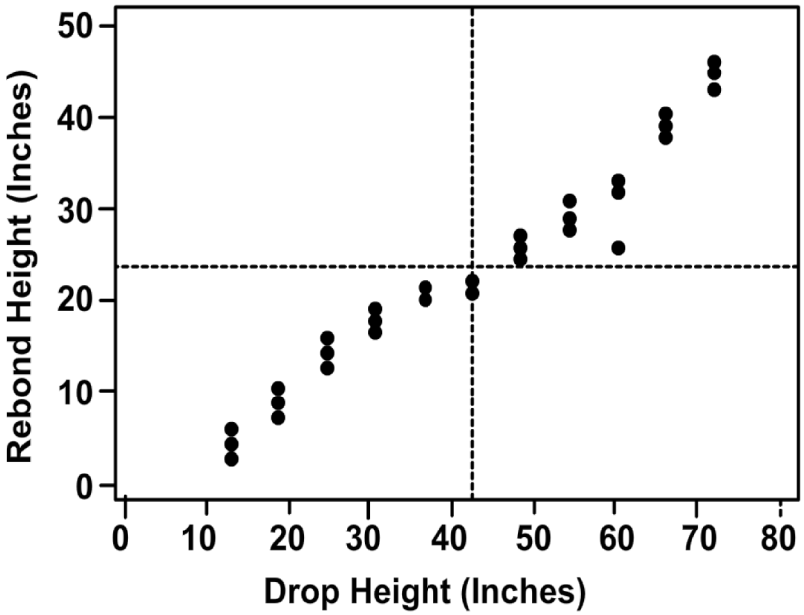
**Figure 14**

| Drop Height | Rebound Measurement 1 | Rebound Measurement 2 | Rebound Measurement 3 |
|---|---|---|---|
| 12 | 3 | 6 | 5 |
| 18 | 7 | 8 | 11 |
| 24 | 13 | 14 | 16 |
| 30 | 19 | 18 | 17 |
| 36 | 20 | 21 | 20.5 |
| 42 | 22 | 21.5 | 21 |
| 48 | 24 | 26 | 25 |
| 54 | 27 | 28 | 30 |
| 60 | 25 | 31 | 32 |
| 60 | 25 | 31 | 32 |
| 66 | 37 | 39 | 38 |
| 72 | 45 | 44 | 42 |

Let us Create a scatter plot of the data and describe the relationship between drop height and

rebound height. Drop height is the explanatory variable because this is the variable that is controlled during the study. Rebound height is the response variable because the rebound height was measured for a given drop height. Thus, drop height is on the X axis and rebound height is on the *Y* axis.

**Figure 15**




Observe that The ball did not always have the same rebound height when it was dropped repeatedly from a specific drop height; there was variability in the rebound heights for a given drop height. The rebound height tends to increase in a linear manner as the drop height increases, though the relationship is certainly not an exact one.

Now, we shall look at what a contingency table is.
If there is a huge data, that is number of observations under each variable is very large, then the data can be written in a tabular form either by giving frequencies that each value take when the number repeats or by giving class intervals and seeing how many belong to each of the class intervals. And this arrangement is known as contingency table.

Hence, whenever we denote the data in tabulated form, that that table is known as contingency table also referred to as cross tabulation or cross tab. In a contingency table the rows correspond to one criterion and the columns to the other. Most often it is used to record and analyse the relationship between two or more categorical variables. It displays the frequency distribution of the variable in a matrix format.

Now, let us suppose that we have two variables - sex (male or female) and handedness (right or left handed). Further suppose that 100 individuals are randomly sampled from a very large population as part of a study of sex differences in handedness. A contingency table can be created to display the numbers of individuals who are male and right-handed, male and left-handed, female and right-handed, and female and left-handed.

Such a contingency table is shown below.

**Figure 16**

| | Male | Female | Total |
|---|---|---|---|
| Right-handed | 65 | 60 | 125 |
| Left-handed | 32 | 43 | 75 |
| Total | 97 | 103 | 200 |

The numbers of the males, females, and right- and left-handed individuals are called marginal totals.

The grand total, i.e., the total number of individuals represented in the contingency table, is the number in the bottom right corner.

The table allows us to see at a glance that the proportion of men who are right-handed is about the same as the proportion of women who are right-handed although the proportions are not identical.

The significance of the difference between the two proportions can be assessed with a variety of statistical tests including Pearson's chi-square test, the G-test, Fisher's exact test, and Barnard's test, provided the entries in the table represent individuals randomly sampled from the population about which we want to draw a conclusion.

If the proportions of individuals in the different columns vary significantly between rows or vice versa, we say that there is a *contingency* between the two variables.
In other words, the two variables are *not independent*.
If there is no contingency, we say that the two variables are *independent.*

# 5. Illustrations

In a survey of 150 persons, two variables were collected.
Gender - Male or Female, and
Marital status – Single/ Married /Widowed/ Divorced.
The following contingency table was constructed to summarise the information contained in the sample.

**Figure 17**

| Marital Status / Gender | Single | Married | Widowed | Divorced | Total |
|---|---|---|---|---|---|
| Female | 38 | 36 | 1 | 7 | **82** |
| Male | 40 | 14 | 4 | 10 | **68** |
| Total | 78 | 50 | 5 | 17 | **150** |

From the contingency table, we can easily say that

- Number of single males is 40
- The number of males is 63
- The number of widows is 5
- Total number of divorcees are 17 etc

And we can apply any statistical test to see whether the variables are independent or not.

In the next example, we consider the data regarding the arm span and height of a person given as follows.

**Figure 18**

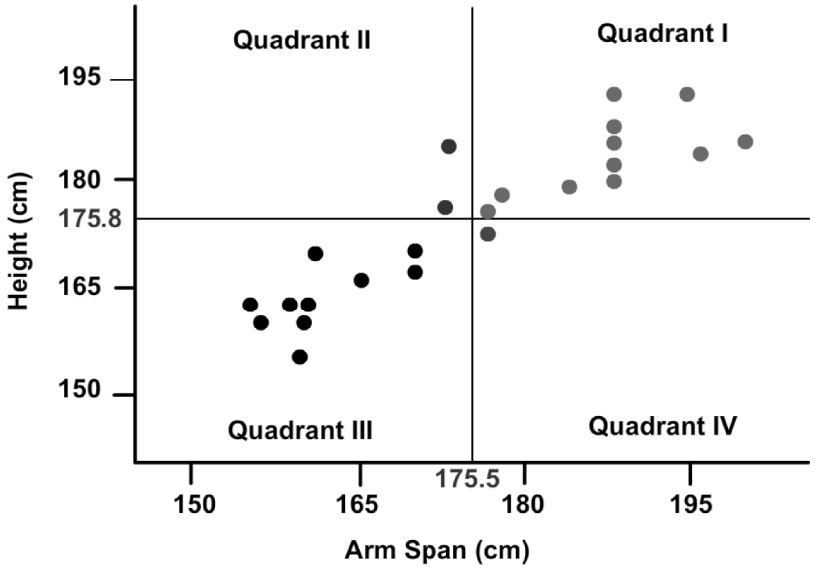
| Arm Span | Height | Arm Span | Height |
|---|---|---|---|
| 156 | 162 | 177 | 173 |
| 157 | 160 | 177 | 176 |
| 159 | 162 | 178 | 178 |
| 160 | 155 | 184 | 180 |
| 161 | 160 | 188 | 188 |
| 161 | 162 | 188 | 187 |
| 162 | 170 | 188 | 182 |
| 165 | 166 | 188 | 181 |
| 170 | 170 | 188 | 192 |
| 170 | 167 | 194 | 193 |
| 173 | 185 | 196 | 184 |
| 173 | 176 | 200 | 186 |

Since data is sorted by arm span, while constructing the scatter plot, we take arm span along

X axis and Height of a person along the Y axis.
Let's take a look at the scatter plot with the quadrants indicated.
We find the average height and average arm span using the given data which is equal to 174.8 and 175.5 respectively. Hence we draw the lines parallel to X axis and Y axis at the average points and divide the whole graph into four quadrants as shown below.

Let's take a look at the scatter plot with the quadrants indicated.

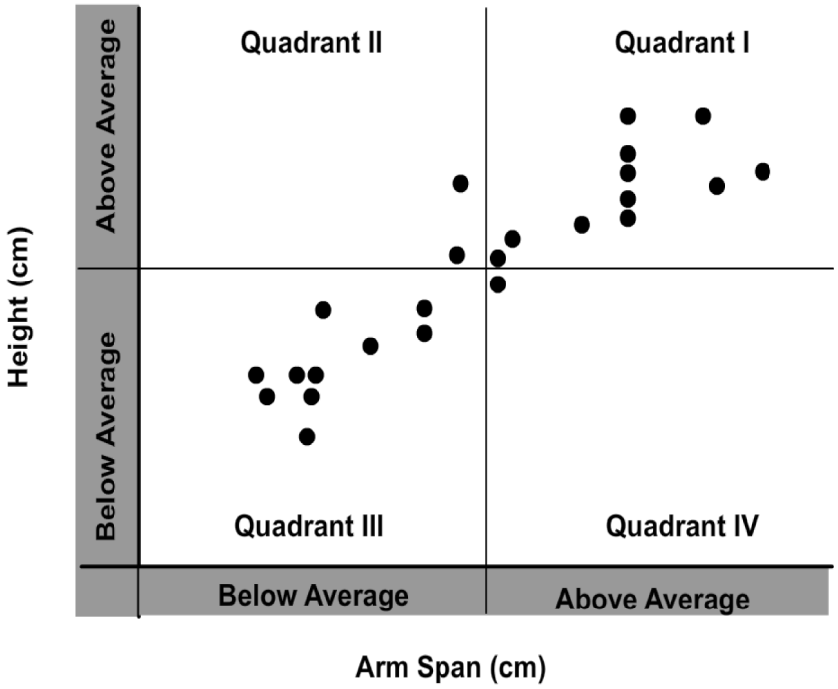
**Figure 19**




Observe that,
Quadrant I has points that correspond to people with above-average arm spans and heights.
Quadrant II has points that correspond to people with below-average arm spans and above-average heights.
Quadrant III has points that correspond to people with below-average arm spans and heights.
Quadrant IV has points that correspond to people with above-average arm spans and below-average heights.

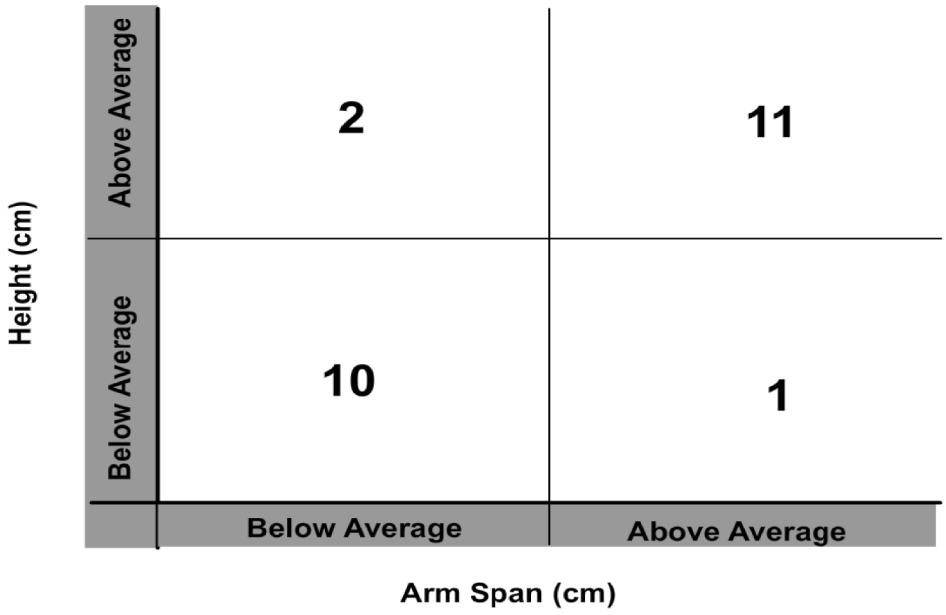
The diagram here summarises the said observations.

**Figure 20**




If you count the number of points in each quadrant on the scatter plot, you get the following summary, which is called a contingency table:

**Figure 21**




Let us now look at how to differentiate univariate data from bivariate data.
Univariate data involves single variable whereas, bivariate data involves two variables.
Univariate data does not deal with causes or relationship but bivariate data deals with causes or relationships.
The major purpose of the univariate analysis is to describe whereas the major purpose of the

bivariate analysis is to explain.
Also, using bivariate data we can identify independent and dependent variables, while using univariate data we cannot.

Using univariate data we can find the measure of:

- Central tendency namely, mean, mode, median.
- Dispersion namely, range variance, maximum, minimum, quartiles, standard deviation etc.
- We can find frequency distributions
- We can draw bar graph, histogram, pie chart, line graph, box and Whisker plot.

Whereas, using bivariate data we can do the analysis of two variables simultaneously.
We can find the correlation between two variables.
We can compare and find the relationship between them, find the causes of relationship and give explanations about the same.
We can write the tables where one variable is contingent on the values of the other variable.

Here's a summary of our learning in this session:

- Meaning of Bivariate data.
- Purpose of having bivariate data.
- Identification of independent and dependent variables.
- Plotting the bivariate data.
- Identification relationship between two variables using scatter plots.
- Types of relationship like positive, negative etc.
- Use of contingency table.
- Constructing contingency table using scatter plots.
- Difference between univariate and bivariate data