Bivariate Relationships

Contents

Describing a Bivariate Relationship

We have so far studied how do we describe and study Univariate Data, that is data having only one variable.

Now we shall study to describe a Bivariate data, that is a data having two variables. The Bivariate data can either have :

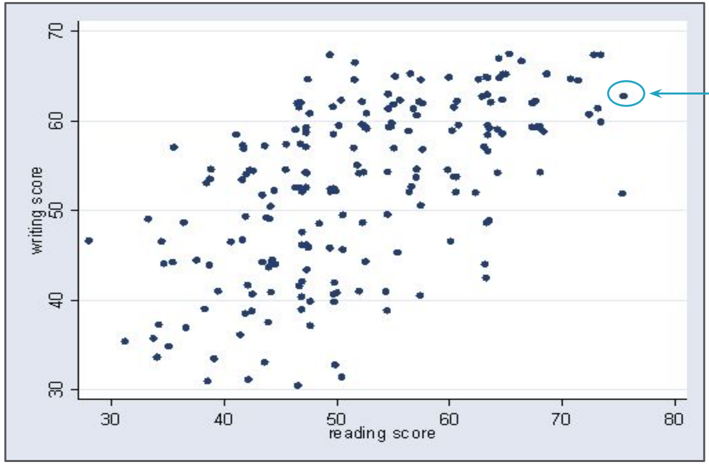
- Two Numeric Variables
- Two Categorical Variables
- One Numeric and One Categorical Variable

The relationship between two numeric continuous variables can be described using :

- Scatter Plot : Scatter plot provides nature of relationship graphically
- Co-Relation Coefficient : Correlation coefficient measures degree of linear relationship
- Simple Linear Regression : Simple Linear Regression gives equation of the type
  Y=a +bX where in you can also predict the value Y for any given value of X.

What is a Scatterplot?

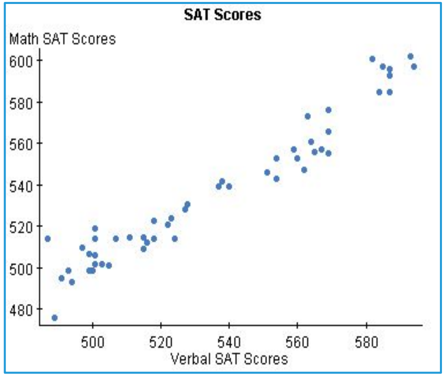
- A scatter plot consists of a X axis (the horizontal axis), a Y axis (the vertical axis), and a series of dots.
- The X-axis and Y-axis represent the values of one variable each.
- Each dot on the scatterplot is one observation from a data set representing the corresponding variable value on X and Y axis respectively
- This plot can be used only for two numeric continuous variables



This dot represents a person having reading score of approx. 74 and writing score of 62.

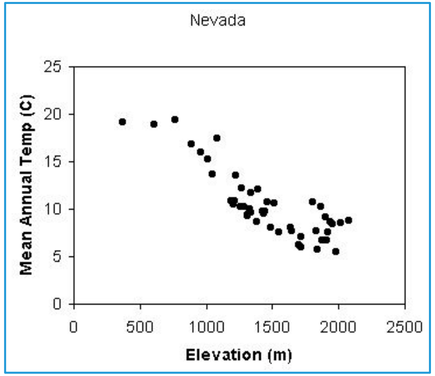
Interpreting a Scatterplot

Positive Correlation



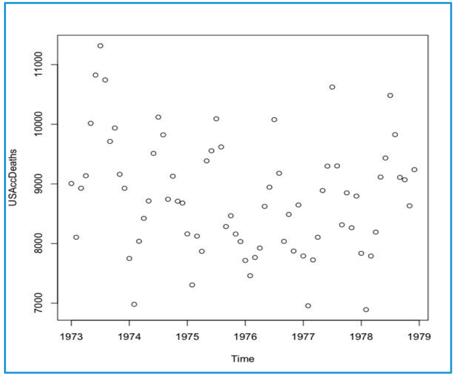
This is a positive sloping (upward) graph.
As the value of one variable increases, the value of other variable also increases.

Negative Correlation



This is a negative sloping (downward) graph.
As the value of one variable increases, the value of other variable tends to decrease.

No Correlation



This is a graph with random pattern.
There is no connection between the two variables. If value of one variable increases, other might increase/decrease.

Pearson's Coefficient of Correlation

The Pearson's correlation coefficient numerically measures the strength of a linear relation between two variables

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} = \frac{cov\,(X,Y)}{sd(x)sd(y)}$$

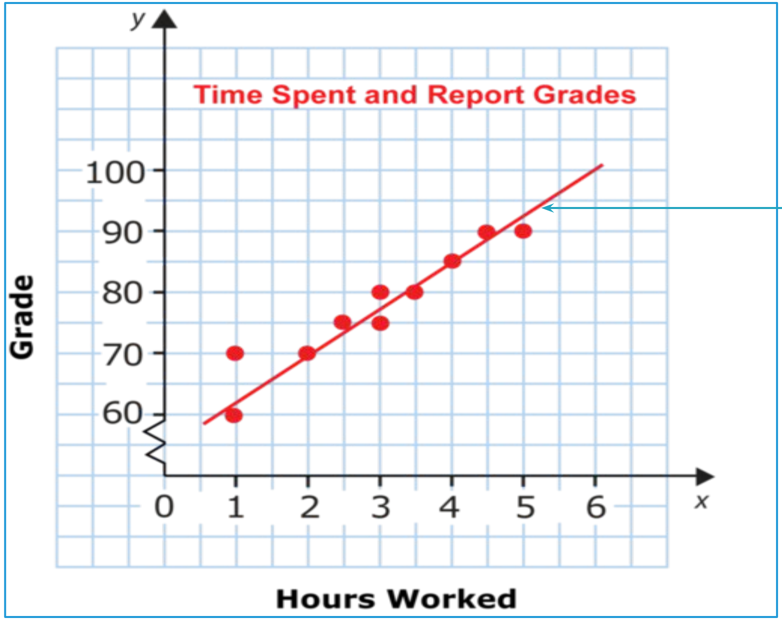
RANGE -1 <= r < =1	
Positive Correlation	r > 0
Negative Correlation	r < 0
No Correlation	r = 0

- The two variables can be measured in entirely different units.
- Example, you could correlate a person's age with their blood sugar levels. Here, the units are completely different.
- It is not affected by change of Origin and Scale

* Both Covariance and Pearson's correlation coefficient can be used only for continuous Numeric variables

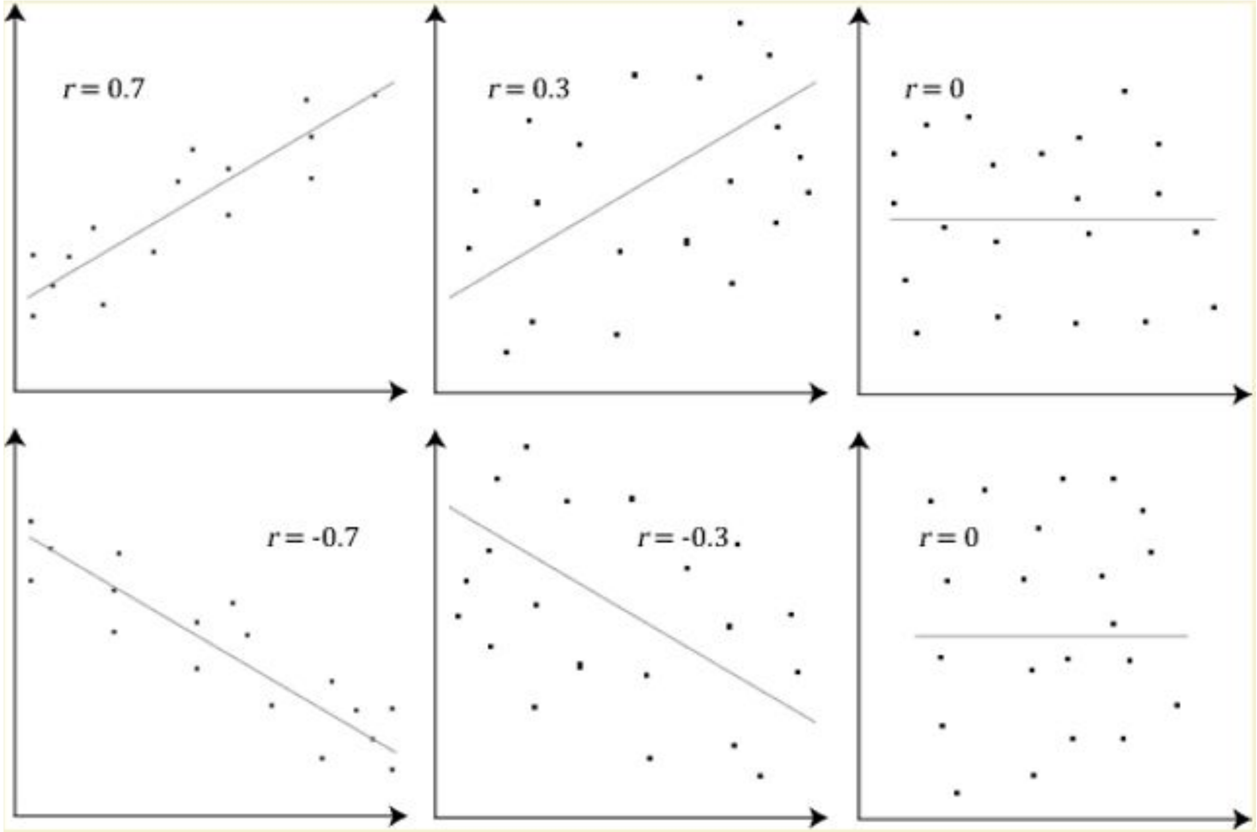
Line of Best Fit : Regression Line

A line of best fit (or "trend" line) is a straight line that best represents the data on a scatter plot.



Line of Best Fit : This line may pass through some of the points, none of the points, or all of the points.

Relationships and r



Simple Linear Regression

The equation of line of best fit is used to describe relationship between two variables

Mathematical form of simple linear regression : $\boldsymbol{Y} = \boldsymbol{aX} + \boldsymbol{b} + \mathrm{e}$

Where,

a : Intercept (The value at which the fitted line crosses the y-axis i.e. X=0)
b : Slope of the Line
e : error which is assumed to be a random variable

NOTE : a and b are population parameters which are estimated using sample

Here, variable Y is known as a 'Dependent' variable, that 'depends on' X which is known as the 'Independent' variable.

Application Areas

Scatter Plot	It is useful in visualising the relationship between any two variables as an initial step. - Life expectancy and the number of cigarettes smoked per day - Literacy rate and life expectancy in a particular region
Correlation Coefficient	It gives the exact numeric measure of the extent of bivariate relationship. - Distance between home & office and the time taken to get there - Size of car engine and cost of car insurance
Simple Linear Regression	It is very useful in predicting the value of one variable given the value of another in a bivariate scenario. - Number of bedrooms and cost of home insurance - Scores in the final exam given the scores in mock test

Case Study - 1

Background

- A company conducts different written tests before recruiting employees. The company wishes to see if the scores of these tests have any relation with post-recruitment performance of those employees.

Objective

- To study the correlation between Aptitude and Job Proficiency.
- Predict the Job proficiency for a given Aptitude score.

Available Information

- Sample size is 33
- Independent Variables: Scores of tests conducted before recruitment on the basis of four criteria – Aptitude, Test of English, Technical Knowledge, General Knowledge
- Dependent Variable job_prof: Job Performance Index calculated after an employee finishes probationary period (6 months)

Data Snapshot

Job_Proficiency

Variables

Observations

empno	aptitude	testofen	tech_	g_k_	job_prof
1	86	110	100	87	88
2	62	62	99	100	80
3	110	107	103	103	96
4	101	117	93	95	76
5	100	101	95	88	80
6	78	85	95	84	73
7	120	77	80	74	58
8	105	122	116	102	116

Columns	Description	Type	Measurement	Possible values
Empno	Employee Number	numeric	-	positive values
aptitude	Aptitude Score of the Employee	numeric	-	positive values
Testofen	Test of English	numeric	-	positive values
tech_	Technical Score	numeric	-	positive values
g_k	General Knowledge Score	numeric	-	positive values
Job_prof	Job Proficiency Score	numeric	-	positive values

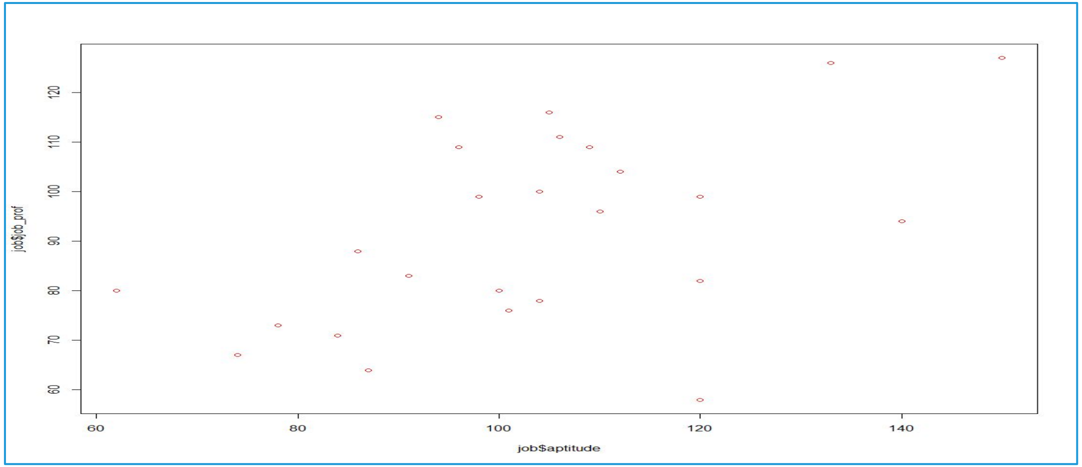
Scatter Plot in R

```
# Importing Data
job<-read.csv("Job_Proficiency.csv",header=T)
```

```
# Scatterplot
plot(job$aptitude,job$job_prof,col="red")
```

- **plot()** gives a scatterplot of the two variables mentioned.
- **col=** provides color to the points.

```
# Output
```



Pearson Correlation Coefficient in R

```
# Correlation
cor(job$aptitude,job$job_prof)
[1] 0.5144107
```

cor() calculates Pearson Correlation Coefficient for the two variables mentioned.

Pearson Correlation Coefficient	0.5144

There is positive relation between aptitude and job proficiency but the relation is of moderate degree.

Simple Linear Regression in R

```
# Simple Linear Regression
model1<-lm(job_prof~aptitude, data=job)
model1
```

lm() gives the linear regression model

```
# Output
Call:
lm(formula = job_prof ~ aptitude, data = job)

Coefficients:
(Intercept)     aptitude
    41.3216       0.4922
```

Inferences : Simple Linear Regression

Dependent Variable : Job Proficiency

Independent Variable : Aptitude

Intercept	Aptitude
41. 3216	0.4922

Equation : Job Proficiency = 41. 3216 + 0.4922 * Aptitude

Here Job Proficiency changes by 0.4992 units with a unit change in aptitude.

Case Study - 2

To learn more Descriptive Statistics in R, we shall consider the below case as an example.

Background

Data of 100 retailers in platinum segment of the FMCG company.

Objective

To describe bivariate relationships in the data

Sample Size

Sample size: 100
Variables: Retailer, Zone, Retailer_Age, Perindex, Growth, NPS_Category

Data Snapshot

Retail Data

Variables

Retailer	Zone	Retailer_Age	Perindex	Growth	NPS_Category
1	North	<=2	81.84	3.04	Promoter

Observations

Columns	Description	Type	Measurement	Possible values
Retailer	Retailer ID	numeric	-	-
Zone	Location of the retailer	character	East, West, North, South	4
Retailer_Age	Number of years doing business with the company	character	<=2, 2 to 5, >5	3
Perindex	Index of performance based on sales, buying frequency and buying recency	numeric	-	positive values
Growth	Annual sales growth	numeric	-	positive values
NPS_Category	Category indicating loyalty with the company	character	Detractor, Passive, Promoter	3

Summarizing Two Categorical Variables

Using Frequency/Cross Tables describing the counts, percentages, etc. is a very basic and most useful way in summarizing two categorical variables.

```
#Importing Data
retail_data <-read.csv("Retail_Data.csv", header=TRUE)
```

```
# Frequency Tables
freq <- table(retail_data$Zone,retail_data$NPS_Category)
freq
```

	Detractor	Passive	Promoter
East	5	9	1
North	5	13	7
South	7	9	16
West	6	10	12

table() in R, gives the frequency of counts of the two variables mentioned.

```
# Percentage Frequency Tables
prop.table(freq)
```

	Detractor	Passive	Promoter
East	0.05	0.09	0.01
North	0.05	0.13	0.07
South	0.07	0.09	0.16
West	0.06	0.10	0.12

prop.table() in R, gives the frequency expressed as percentage of total count.

Summarizing Two Categorical Variables

```
prop.table(freq,1)
```

	Detractor	Passive	Promoter
East	0.33333333	0.60000000	0.06666667
North	0.20000000	0.52000000	0.28000000
South	0.21875000	0.28125000	0.50000000
West	0.21428571	0.35714286	0.42857143

- **prop.table()** in R, gives the frequency expressed as percentage of total count.
- (,1) expresses the frequency as percentage of row count whereas (,2) would express it as percentage of column count.

Summarizing Two Categorical Variables

```
# Installing package - "gmodels"
install.packages("gmodels")
library(gmodels)
```

```
# Frequency Table using "gmodels" package
CrossTable(retail_data$Zone,retail_data$NPS_Category)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  100


                 | retail_data$NPS_Category
retail_data$Zone | Detractor |   Passive |  Promoter | Row Total |
-----------------|-----------|-----------|-----------|-----------|
            East |         5 |         9 |         1 |        15 |
                 |     0.696 |     1.321 |     3.585 |           |
                 |     0.333 |     0.600 |     0.067 |     0.150 |
                 |     0.217 |     0.220 |     0.028 |           |
                 |     0.050 |     0.090 |     0.010 |           |
-----------------|-----------|-----------|-----------|-----------|
           North |         5 |        13 |         7 |        25 |
                 |     0.098 |     0.738 |     0.444 |           |
                 |     0.200 |     0.520 |     0.280 |     0.250 |
                 |     0.217 |     0.317 |     0.194 |           |
                 |     0.050 |     0.130 |     0.070 |           |
-----------------|-----------|-----------|-----------|-----------|
           South |         7 |         9 |        16 |        32 |
                 |     0.018 |     1.294 |     1.742 |           |
                 |     0.219 |     0.281 |     0.500 |     0.320 |
                 |     0.304 |     0.220 |     0.444 |           |
                 |     0.070 |     0.090 |     0.160 |           |
-----------------|-----------|-----------|-----------|-----------|
            West |         6 |        10 |        12 |        28 |
                 |     0.030 |     0.191 |     0.366 |           |
                 |     0.214 |     0.357 |     0.429 |     0.280 |
                 |     0.261 |     0.244 |     0.333 |           |
                 |     0.060 |     0.100 |     0.120 |           |
-----------------|-----------|-----------|-----------|-----------|
    Column Total |        23 |        41 |        36 |       100 |
                 |     0.230 |     0.410 |     0.360 |           |
-----------------|-----------|-----------|-----------|-----------|
```

CrossTable() in R, gives the frequency of counts of the two variables mentioned

Summarizing Two Categorical Variables

```
# Frequency Table using 'gmodels' package
CrossTable(retail_data$Zone,retail_data$NPS_Category,prop.r = FALSE,
prop.c = FALSE)
```

prop.r= removes the row proportion from the output and **prop.c=** removes the column proportion

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Table Total |
|-------------------------|


Total Observations in Table:  100


                 | retail_data$NPS_Category
retail_data$Zone | Detractor |   Passive |  Promoter | Row Total |
-----------------|-----------|-----------|-----------|-----------|
            East |         5 |         9 |         1 |        15 |
                 |     0.696 |     1.321 |     3.585 |           |
                 |     0.050 |     0.090 |     0.010 |           |
-----------------|-----------|-----------|-----------|-----------|
           North |         5 |        13 |         7 |        25 |
                 |     0.098 |     0.738 |     0.444 |           |
                 |     0.050 |     0.130 |     0.070 |           |
-----------------|-----------|-----------|-----------|-----------|
           South |         7 |         9 |        16 |        32 |
                 |     0.018 |     1.294 |     1.742 |           |
                 |     0.070 |     0.090 |     0.160 |           |
-----------------|-----------|-----------|-----------|-----------|
            West |         6 |        10 |        12 |        28 |
                 |     0.030 |     0.191 |     0.366 |           |
                 |     0.060 |     0.100 |     0.120 |           |
-----------------|-----------|-----------|-----------|-----------|
    Column Total |        23 |        41 |        36 |       100 |
-----------------|-----------|-----------|-----------|-----------|
```

Summarizing Three Categorical Variables

```
# Three Way Frequency Table
```

```
table1 <-
table(retail_data$Zone,retail_data$NPS_Category,retail_data$Retailer_Age)
ftable(table1)
```

```
                   <=2 >5 2 to 5

East  Detractor      2  1      2
      Passive        3  3      3
      Promoter       0  1      0
North Detractor      2  1      2
      Passive        1  6      6
      Promoter       1  6      0
South Detractor      1  4      2
      Passive        2  3      4
      Promoter       3 10      3
West  Detractor      1  2      3
      Passive        1  8      1
      Promoter       0 11      1
```

ftable() in R, gives the frequency of counts of the three variables in one table itself.

Quick Recap

In this session, we learnt the basics of Bivariate Relationships

Bivariate Data	• Bivariate data can either have :
	  • Two Numeric Variables
	  • Two Categorical Variables
	  • One Numeric and One Categorical Variable

Scatter Plot	• Each dot on the scatterplot is one observation from a data set representing the corresponding variable value on X and Y axis respectively. Here X & Y are continuous variables.
Pearson's Correlation Coefficient	• Numerically measures the strength of a linear relation between two variables
Simple Linear Regression	• The equation of the line of best fit used to describe relationship between two variables
Cross Tables	• Tables for summarizing categorical variables.