DESCRIPTIVE STATISTICS

Dr. Hategekimana Fidele

Adventist University of Central Africa (AUCA)

May 03, 2023

PART II. BIVARIATE DISTRIBUTION

1. Introduction

Many engineering and scientific problems are concerned with determining a relationship between a set of variables.

For instance, in teaching, the teacher might be interested in the relationship between the rate of attendance (variable X) and the performance of the students in an exam (variable Y).

Knowledge of such a relationship would enable us to predict the marks expected in an exam for various possible rates of attendance.

In many situations, there is a single response variable Y, also called the **dependent variable**, which depends on the value of a set of input, also called **independent, or explanatory variables** X.

We assume that the response variable Y assumes values $y_1, y_2, \ldots, y_n$ respectively for n values $x_1, x_2, \ldots, x_n$ of the response variable X by a simple linear function

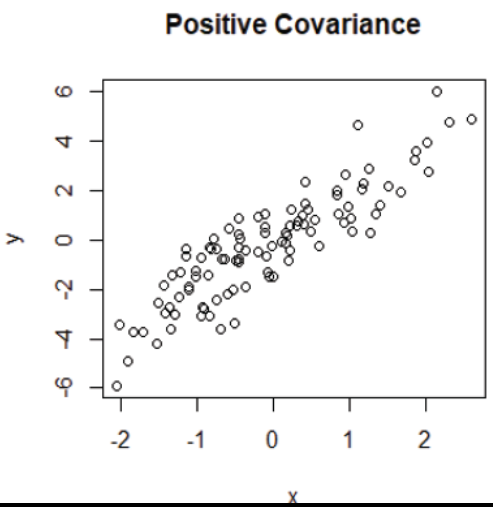
$$y_i = \alpha + \beta x_i$$

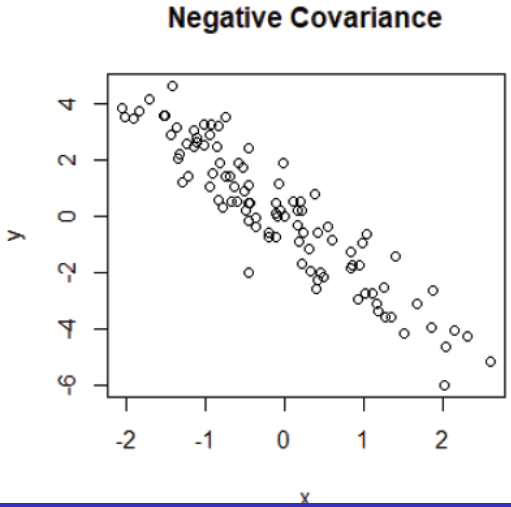
When we want to measure the relationship between two variables, **covariance and correlation coefficient** are used.

Covariance measures how two variables are related:

- a positive covariance indicates that a variable tends to increase when the other variable increases;
- a negative covariance indicates that a variable tends to decrease when the other variable increases;
- Zero covariance indicates that the change of a variable is random whenever the other variable increases or decreases.

The following scatter diagrams illustrate the relationship between two variables and the covariance.





2. Covariance

The covariance of the bivariate distribution of the variable X and Y, whose the values of sample are here indicated,

x_i	x_1	x_2	x_3	x_4	x_5	...	x_n
y_i	y_1	y_2	y_3	y_4	y_5	...	y_n

is given by the formula;

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

EXAMPLE

The following table shows the initial speed (miles per hour) and stopping distance (feet) of a car.

Initial speed	11	22	32	41	51
Stopping distance	8.2	32.8	82.0	144.4	236.2

Find the covariance between the initial speed and the stopping distance.

Here, $\overline{x} = 31.4$ and $\overline{y} = 100.72$. The covariance can be obtained as

i	x_i	$\overline{y_i}$	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
1	11	8.2	−20.4	−92.52	1887.41
2	22	32.8	−9.4	−67.92	638.45
3	32	82.0	0.6	−18.72	−11.23
4	41	144.4	9.6	43.68	419.33
5	51	236.2	19.6	135.48	2655.41
Total					5589.36

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{5589.36}{5-1} = 1397.34$$

By developing the formula of Covariante of the variable X, Y, we obtain its equivalent formula here below;

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

Using the data of the previous example, we obtain:

$$Cov(x, y) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{21402.4 - 5 * 31.4 * 100.72}{5-1} = 1397.34$$

the same value.

Covariance describes how two variables are related. It indicates whether the variables are positively or negatively related. However, the covariance cannot be an efficient measure of the relationship between two variables, because it is also affected by the magnitude of the variables.

3. Coefficient of Correlation

To obtain a measure relationship with a standard unit of measurement, we use the correlation coefficient r. The correlation coefficient is a scaled version of covariance. The correlation coefficient ranges from -1 to 1. The correlation coefficient is equal to the quotient of the covariance and the product of the standard deviations of the two variables

$$r = \frac{Cov(x, y)}{s_x s_y}$$

Where,

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

and

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

To facilitate the calculation, we can use the following substitution formulae: $X_i = x_i - \bar{x}$ and $Y_i = y_i - \bar{y}$ thus, the covariance $Cov(x, y)$ and coefficient of variation r are given by;

$$Cov(x, y) = \frac{\sum_{i=1}^{n} X_i Y_i}{n - 1}$$

,

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2$$

and

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} Y_i^2$$

Using these 3 formulae to find the covariance and the coefficient of variation of the data of the example, we obtain:

Solution:

Consider the table generated by the following R codes letting $Xi = x_i - \bar{x}$, $Yi = y_i - \bar{y}$, $Xi2 = Xi^2$ and $Yi2 = Yi^2$;

```
Console  Jobs
~/
> x<-c(11,22,32,41,51)
> y<-c(8.2,32.8,82.0,144.4,236.2)
> d<-data.frame(x,y)
> library(plyr)
> d1<-mutate(d,x,y,Xi=x-mean(x),Yi=y-mean(y),
+            Xi2=Xi^2,Yi2=round(Yi^2,2),XiYi=round(Xi*Yi,2))
> d1
   x     y    Xi     Yi    Xi2      Yi2    XiYi
1 11   8.2 -20.4 -92.52 416.16  8559.95 1887.41
2 22  32.8  -9.4 -67.92  88.36  4613.13  638.45
3 32  82.0   0.6 -18.72   0.36   350.44  -11.23
4 41 144.4   9.6  43.68  92.16  1907.94  419.33
5 51 236.2  19.6 135.48 384.16 18354.83 2655.41
> summarize(d1,sum(Xi2),sum(Yi2),sum(XiYi))
  sum(Xi2) sum(Yi2) sum(XiYi)
1    981.2 33786.29   5589.37
>
```

The covariance:

$$Cov(x,y) = \frac{\sum_{i=1}^{n} X_i Y_i}{n-1} = \frac{5589.37}{5-1} = 1397.342$$

For the coefficient of variation,

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 = \frac{981.2}{5-1} = 245.3$$

and

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} Y_i^2 = \frac{33786.29}{5-1} = 8446.573$$

Thus,

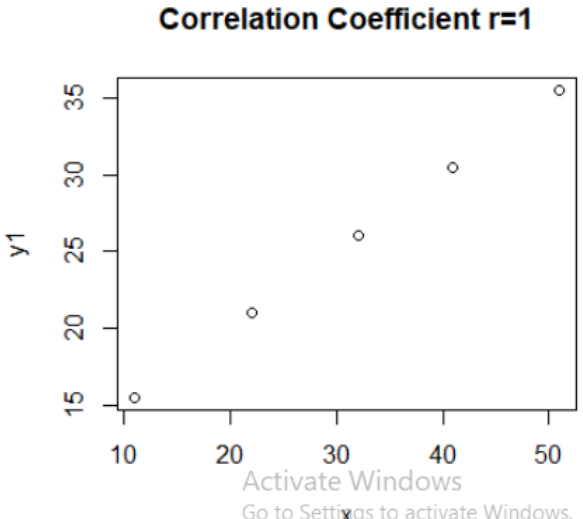
$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{1397.342}{\sqrt{245.3 * 8446.573}} = 0.970764$$

4. Scales of the coefficient of correlation

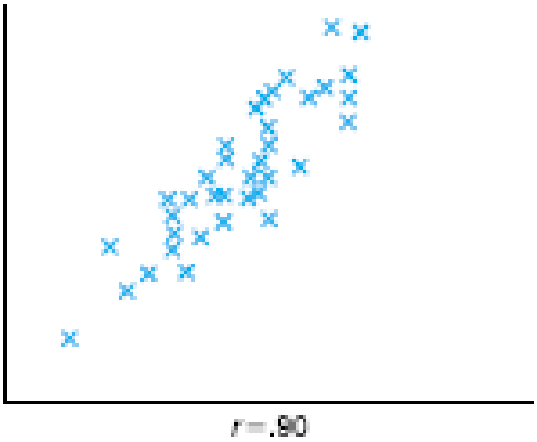
The coefficient of correlation has the following ranking or scales:

- $|r| = 1$: pure correlation coefficient. This happens when all the (x_i, y_i) are points of the unique straight line
- $0.9 \leq |r| < 1$: Very high strong coefficient of correlation: many points (x_i, y_i) are on the same straight line except few of them.
- $0.7 \leq |r| < 0.9$: High correlation; a good number of points (x_i, y_i) will be on the same straight line.
- $0.5 \leq |r| < 0.7$: Moderate correlation; many points (x_i, y_i) will deviate from the straight line.

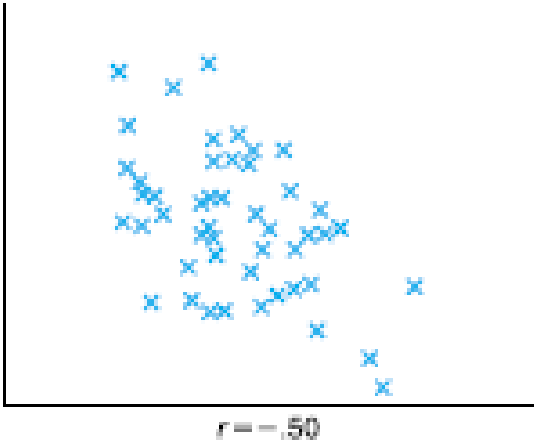
For the given example, $r = 1$ is a pure correlation coefficient. All the points (x_i, y_i) are on the same straight line



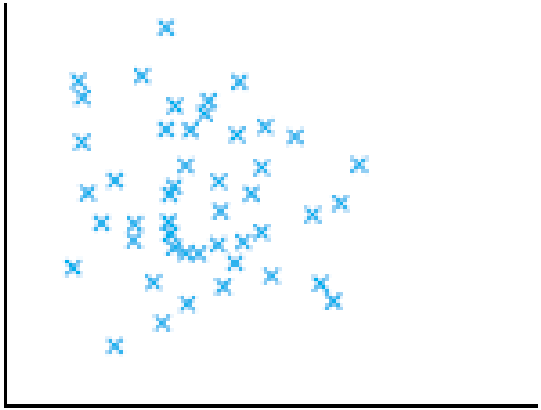
For the given example, $r = 0.9$ is a very strong correlation coefficient. All the points (x_i, y_i) are not necessarily on the same straight line, but they are very concentrated near and all along an imaginary straight line



For the given example, $r = -0.5$ is a moderate correlation coefficient. All the points (x_i, y_i) are not necessarily on the same straight line, but they are widely spread far from an imaginary straight line



For the given example, $r = 0$ is a moderate correlation coefficient. All the points (x_i, y_i) are not necessarily on the same straight line, but they are widely spread far from an imaginary straight line



6. Linear regression and Prediction

The objective of the study of the relationship between two variables is to establish a linear equation between these two variables. This linear equation is called **the regression**. Depending on the value to predict, we classify the regression equations in two categories:

- The regression equation of y on x: express y in terms of x and it is used to predict the value of y at the specified value of x;

$$y = bx + (\bar{y} - b\bar{x})$$

where

$$b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

- The regression equation of x on y express x in terms of y and it is used to predict the value of x at the specified value of y;

$$x = by + (\bar{x} - b\bar{y})$$

where

$$b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} Y_i^2}$$

Example:

A study has been conducted to find the relationship between the shelf spaces (in square feet), variable x, to predict the monthly sales (in thousand dollars) of milk, variable y. A random sample of 11 grocery stores is selected and the results are given below.

x	5	7	8	9	10	12	13	15	16	18	20
y	3.2	4.4	2.8	3.8	4.7	5.2	4.6	5.4	5.6	5.4	6.1

i. Represent the bivariate by a scatter diagram. Does the scatter diagram glimpses a linear relationship between shelf spaces dimensions and monthly sales?

ii. Find the covariance;

iii. Find the correlation coefficient;

iv. Find the monthly sales in dollars of grocery for the shelf spaces of 14 square feet;

v. What should be the dimension of the shelf spaces (in square feet) needed for making a monthly sales of 6.2 thousand dollars?