Correlation

Definition, types of correlation, scatter diagram, Karl Pearson's coefficient of correlation (linear correlation)

Deependra Dhakal

Assistant Professor

Agriculture and Forestry University

https://rookie.rbind.io

Meaning and definition

- Suppose we have a sample of n pairs for which each pair represents the measurement of two variables, X and Y. If a scatterplot of Y versus X shows a general linear trend, then it is natural to try to describe the strength of the linear association.
- The systematic interrelationship between the two continuous related variables say, X and Y is termed as correlation. When only two variables are involved, the correlation is called simple correlation. If more than two variables are involved, the correlation is said to be multiple correlation.
- When the variables move in the same direction, i.e., increase in one variable causes and increase in other variable and *vice versa*, such type of correlation is called positive/direct correlation. In general, grain yield of wheat and the number of grains per spike are positively correlated.
- By analogy, negative correlation is said to occur when increase in one variable is followed by decrease in other. For example, grain yield of wheat and severity of disease in the field are negatively correlated.

Anscombe's Quartet

Table 1: Anscombes quartet is a set of 4 (x, y) data sets that were published by Francis Anscombe in a 1973 paper Graphs in statistical analysis.

SN	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Rothamsted Oats

Table 2: A dataset from RCB experiment, carried out at Rothamsted facility, of oats taking measurements on straw and grain with 9 fertilizer treatments.

SN	grain	straw
1	61.375	83
2	68.75	130
3	64.25	100
4	65.5	96
5	79.625	130.5
6	79.25	122
7	. . .	. . .
8	. . .	. . .
9	. . .	. . .
10	82.125	175.5
11	83.75	140.5
12	84.75	122
13	83.875	192.5
14	89	188
15	93.25	162

- We take the two datasets (above) and show (graphically) following types of correlation in scatterplot:
  - Positive correlation
  - Negative correlation
  - Dinosaur correlation

Scatterplot diagram

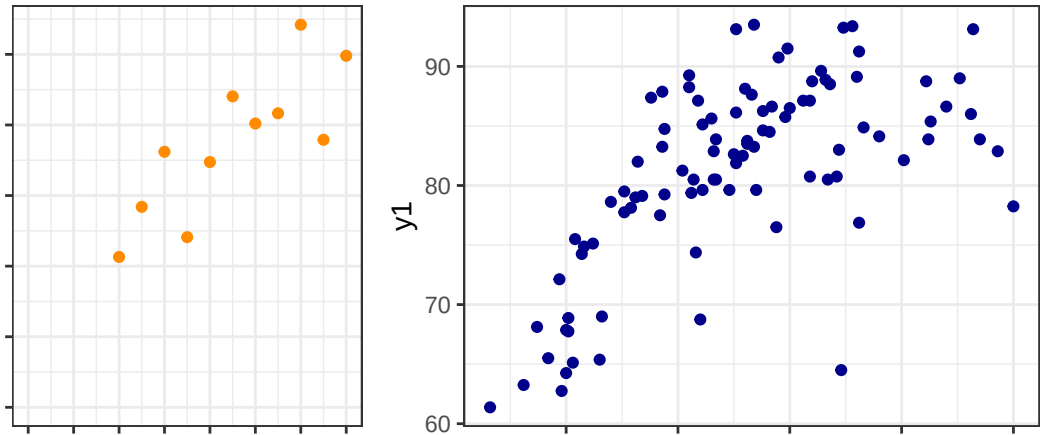

Figure 1: Association between variables in two datasets (Anscombe's Quartet: Left; Rothamsted experiment: Right)

```
## [1] 9.00000000000000 9.00000000000000 9.00000000000000 9.00000000000000
## [5] 7.50090909090909 7.50090909090909 7.50000000000000 7.50090909090909

## [1] 11.00000000000000 11.00000000000000 11.00000000000000 11.0000000000
## [5]  4.12726909090909  4.12762909090909  4.12262000000000  4.1232490909

## [1] 0.816420516344840 0.816236506000243 0.816286739489598 0.8165214368
```

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	num
numeric	x1	0	1	9.00000000000000	3.31662479
numeric	x2	0	1	9.00000000000000	3.31662479
numeric	x3	0	1	9.00000000000000	3.31662479
numeric	x4	0	1	9.00000000000000	3.31662479
numeric	y1	0	1	7.50090909090909	2.03156813
numeric	y2	0	1	7.50090909090909	2.03165673
numeric	y3	0	1	7.50000000000000	2.03042360
numeric	y4	0	1	7.50090909090909	2.03057851

Mathematical representation of correlation

Karl pearson's coefficient of correlation (linear correlation)

Correlation coefficient for bivariate frequency distribution