Correlation Analysis

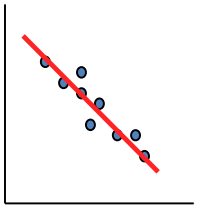
- Analysis of the degree of association between two random variables.
  - Example:
    - Association between shear strength and weld diameter.
    - Association between weight of a person and his blood pressure.
    - Association between hours of training imparted on safety and number of accidents in plant per year.

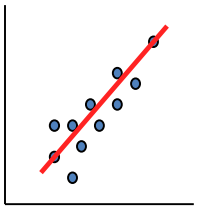
- ❑ **Form (linear or non-linear)**
- ❑ **Direction (positive or negative)**
- ❑ **Strength (none, weak, strong, perfect)**

Caution: It is an empirical study, so logical/feasible association to be identified prior to analysis.

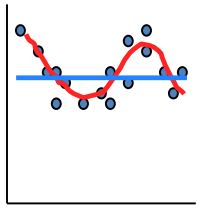
Forms of correlation

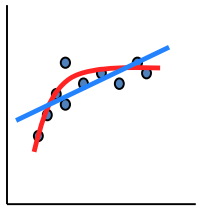
Linear



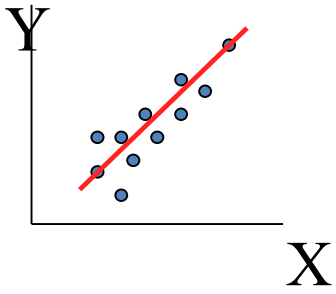


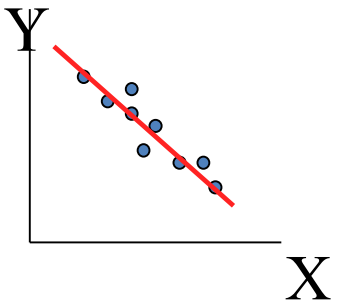
Non-linear





Directions of correlation





Positive

- X & Y vary in the same direction
- As X goes up, Y goes up

Negative

- X & Y vary in opposite directions
- As X goes up, Y goes down

Strength or degree of correlation

- measured by correlation coefficient (r) which range from -1 to +1
- Zero means “no relationship”
- The farther the r is from zero, the stronger the relationship

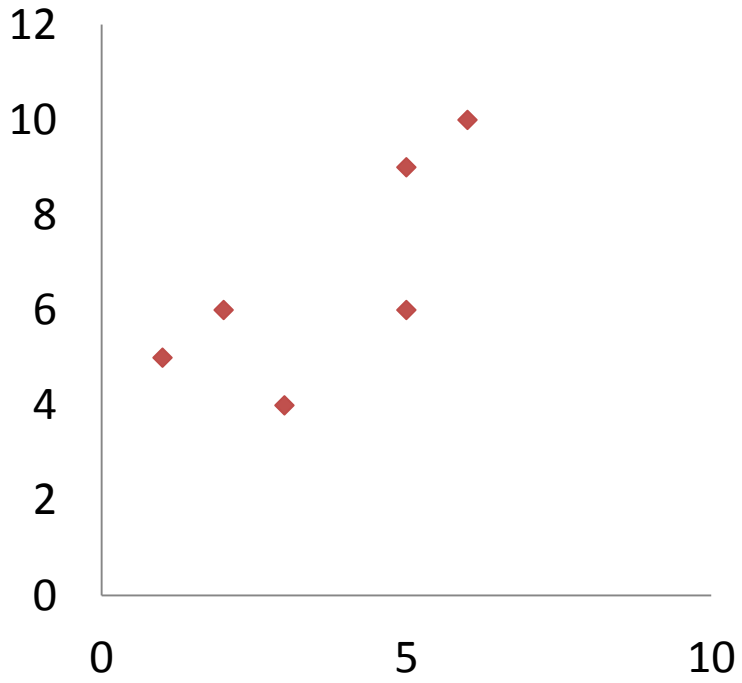
Methods of determining correlation

1. Scatter Plot
2. Spearman's Rank-correlation coefficient
3. Phi-Correlation Coefficient
4. Cramer's V Coefficient
5. Karl Pearson's coefficient of correlation(r)

1. Scatter Plot

The values of the two variables are plotted on a graph paper. Pictorial examination of the relationship between two quantitative variables.

study hour(X)	grade(Y)
5	9
2	6
1	5
6	10
3	4
5	6



2. Spearman's Rank-correlation coefficient

- Two variables measured on ordinal scale (ranks).

$$r = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Example

Ten salesmen were deputed for a training program, and at the end of the program they have to undergo a test conducted by the program director. They were ranked based on the scores obtained in the test. After six months they were again ranked based on the sales performance. The ranks are given in the following table. Find whether there is any existence of relationship between a salesman undergoing the training program and his performance.

Sales man	Rank on test	Rank on sales performance
1	4	5
2	6	8
3	1	3
4	3	1
5	9	7
6	7	6
7	10	9
8	2	2
9	8	10
10	5	4

Sales man	Rank on test	Rank on sales performance	d_i	d_i^2
1	4	5	-1	1
2	6	8	-2	4
3	1	3	-2	4
4	3	1	2	4
5	9	7	2	4
6	7	6	1	1
7	10	9	1	1
8	2	2	0	0
9	8	10	-2	4
10	5	4	1	1

$$\sum d_i^2 = 24$$

$$r_s = 1 - \frac{6.24}{10(10^2 - 1)} = 0.855$$

- Example: Six products A, B, C, D, E & F, manufactured by a particular company are sold in both rural and urban areas. The table below shows the ranks of preferences of customers from rural as well as urban areas. Show whether there exists any preference of products among customers of rural and urban areas.

products	RURAL	URBAN
A	2	1
B	5	4
C	1	3
D	3	2
E	6	5
F	4	6

3. Phi-Correlation Coefficient

- Two variable measured on a dichotomous scale
- Example: correlation between eye colour (Blue, Black) and Gender (Male, Female).
  - A total of 100 respondents comprising of 55 men and 45 women were contacted and their eye colour was identified as either blue of black.

Observed	Male	Female	Total
Black	45	7	52
Blue	10	38	48
Total	55	45	100

Expected	Male	Female	Total
Black	55*52/100 = 28.6	45*52/100 = 23.4	52
Blue	55*48/100 = 26.4	45*48/100 = 21.6	48
Total	55	45	100

$$\phi = \sqrt{\frac{\sum \frac{(o-e)^2}{e}}{N}}$$

$$\phi = 0.66$$

4. Cramer's V Coefficient

- Two variable measured on a nominal scale (any number of categories)
- Example: correlation between credit card ownership (Yes, No) and Occupational status (Employees, Professionals, Self-employed).

Observed	Employees	Professionals	Self-employed	TOTAL
No	2	10	27	39
Yes	20	23	18	61
Total	22	33	45	100

$$V = \sqrt{\frac{\sum \frac{(o-e)^2}{e}}{N \times \min\{r-1, c-1\}}}$$

Expected	Employees	Professionals	Self-employed	TOTAL
No	8.58	12.87	17.55	39
Yes	13.42	20.13	27.45	61
Total	22	33	45	100

$$V = \sqrt{\frac{17.5}{100 \times \min\{2-1, 3-1\}}}$$

$$= \sqrt{\frac{17.5}{100}} = 0.418$$

5. Pearson's Correlation Coefficient

- Two variable measured on a ratio/interval scale
- Example: correlation between shear strength and weld diameter.

$$r = \frac{S_{XY}}{[S_{XX} S_{YY}]^{1/2}}$$

$$S_{XY} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{XX} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{YY} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Computation Table for *r*

Observations	X	Y	X^2	Y^2	XY
1 2 . . . . n					
	Σx	Σy	Σx^2	Σy^2	Σxy

Example

Find correlation coefficient between monthly income and net savings.

Employee no.	1	2	3	4	5	6	7	8	9
Monthly income (hundreds Rs.)	780	360	980	250	750	820	900	620	650
Net savings (hundreds Rs.)	84	51	91	60	68	62	86	58	53

Employee no.	Monthly income (hundreds Rs.)	Net savings	X^2	Y^2	XY
1	780	84	608400	7056	65520
2	360	51	129600	2601	18360
3	980	91	960400	8281	89180
4	250	60	62500	3600	15000
5	750	68	562500	4624	51000
6	820	62	672400	3844	50840
7	900	86	810000	7396	77400
8	620	58	384400	3364	35960
9	650	53	422500	2809	34450
	6110	613	4612700	43575	437710

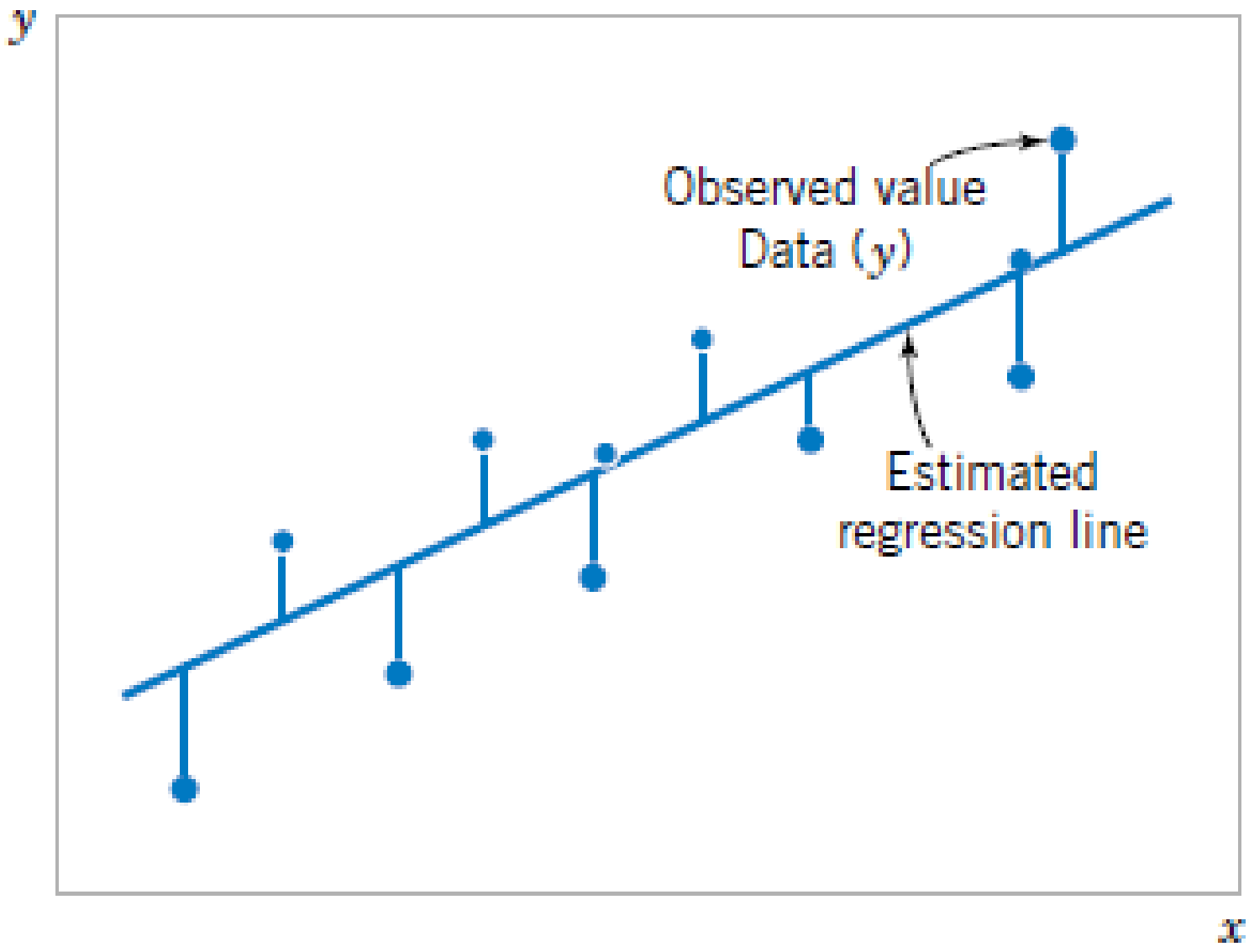
Sxy	21551.11
Sxx	464688.9
Syy	1822.889

r = 0.740472

Regression Analysis

- Existence & degree of association : Correlation
- Extent of causal relationship: Regression

- Simple Linear regression model:
  - Estimated y is $\hat{y} = a + b\,x$


y
Observed value
Data (y)
Estimated
regression line
x

Least square method

- If $y_i = a + b\,x_i + e_i$

  i.e., actual y = ŷ + error value
- Then minimize the squared sum of e_i

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [\hat{y}_i - (a + bx_i)]^2$$

- Solving the following two normal equations for a and b

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

- Alternatively

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}$$

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

Coefficient of Determination

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{\left[\sum xy - \frac{\sum x \sum y}{n}\right]^2}{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}$$

- R^2 : Proportion of variation of values of y explained by the regression model.
- $0 \leq R^2 \leq 1$
- $R^2 = 1$, indicates the regression line is a perfect estimation of linear relationship between x & y.
- $R^2 = 0$, indicates no relationship

Example: Sales manager intends to see the relationship between the constituents of a food product and the consumer's preference. He identified a potential costumer and got his preferences on a 1-9 scale on 10 different alternative products with varying protein contents.

Consumer's rating attempts	Preferences (Y)	Protein (X)
1	3	4
2	7	9
3	2	3
4	1	1
5	6	3
6	2	4
7	8	7
8	3	3
9	9	8
10	2	1

Protein (x)	Preferences (y)	xy	x^2	y^2
4	3	12	16	9
9	7	63	81	49
3	2	6	9	4
1	1	1	1	1
3	6	18	9	36
4	2	8	16	4
7	8	56	49	64
3	3	9	9	9
8	9	72	64	81
1	2	2	1	4
43	43	247	255	261

Sxy	62.1
Sxx	70.1
Syy	76.1

b =	0.886
a =	0.491
R^2 =	0.723

- The normal equations: $10a + 43b = 43$

  $43a + 255b = 247$
- so estimated $b = 0.886$ and $a = 0.491$
- Regression line : $\hat{y} = 0.491 + 0.886x$
- The regression coefficient $b = 0.886$ indicates the change in consumer's preference with unit change in protein contents.
- Coefficient of Determination , $R^2 = 0.723$

  It implies that 72.3% of the variation in preference levels is explained by the estimated line and the remaining 27.7% of the variation may be explained either by other variables or errors in measurements or both.