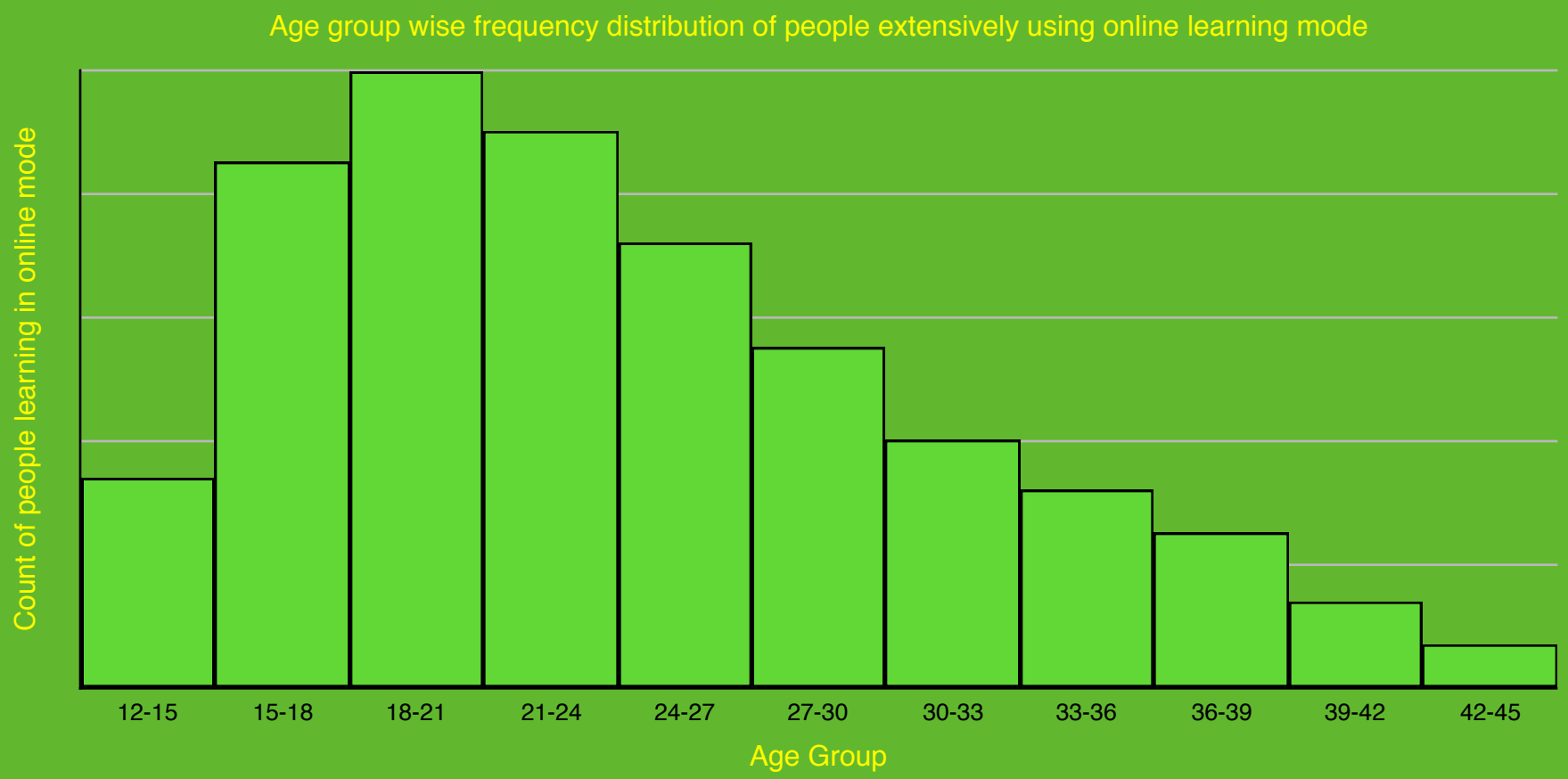
Basics of Statistics – Part 2

Converting data into information



Instructors

Mousum Dutta
Chief Data Scientist, Spotle.ai
IIT Kharagpur

Dr Sourish Das
Assistant Professor,
Chennai Mathematical Institute
Common Wealth Rutherford Fellow,
University of Southampton

Skewness

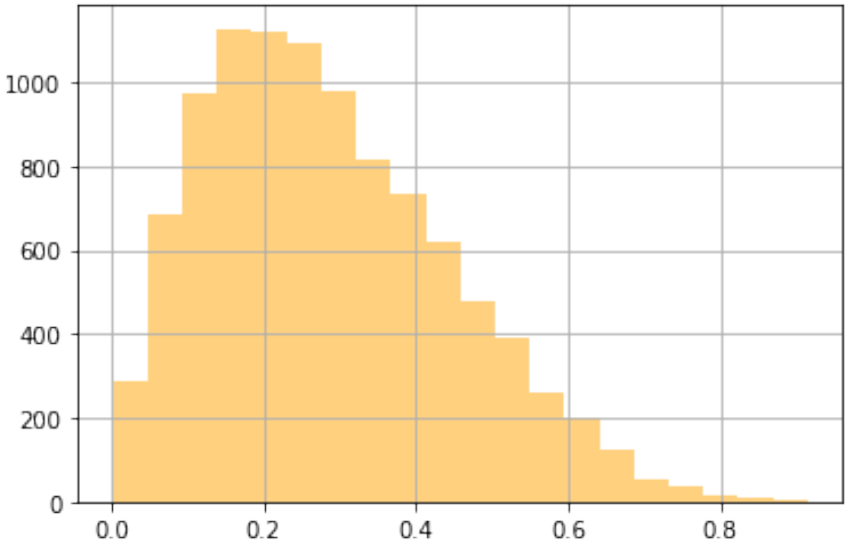
Skewness represents the degree of distortion of a distribution from a symmetrical bell curve or normal distribution from its mean to the left or to the right, in a set of data. Skewness tells us the extent to which a distribution differs from a normal distribution.

Histogram gives us a general impression about skewness.

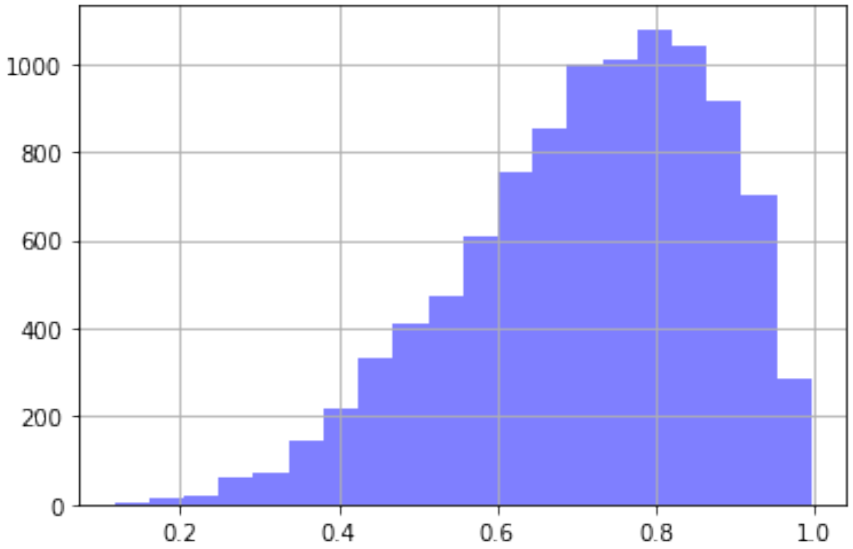
Skewness

- If in a data set, the volume of the data is concentrated at the left, producing a long tail at the right, the distribution is called a right skewed or positively skewed distribution.
- Similarly, if the peak is toward the right creating a long tail at the left the distribution is called a left skewed or negatively skewed distribution.

Positive (0.59)



Negative (-0.60)



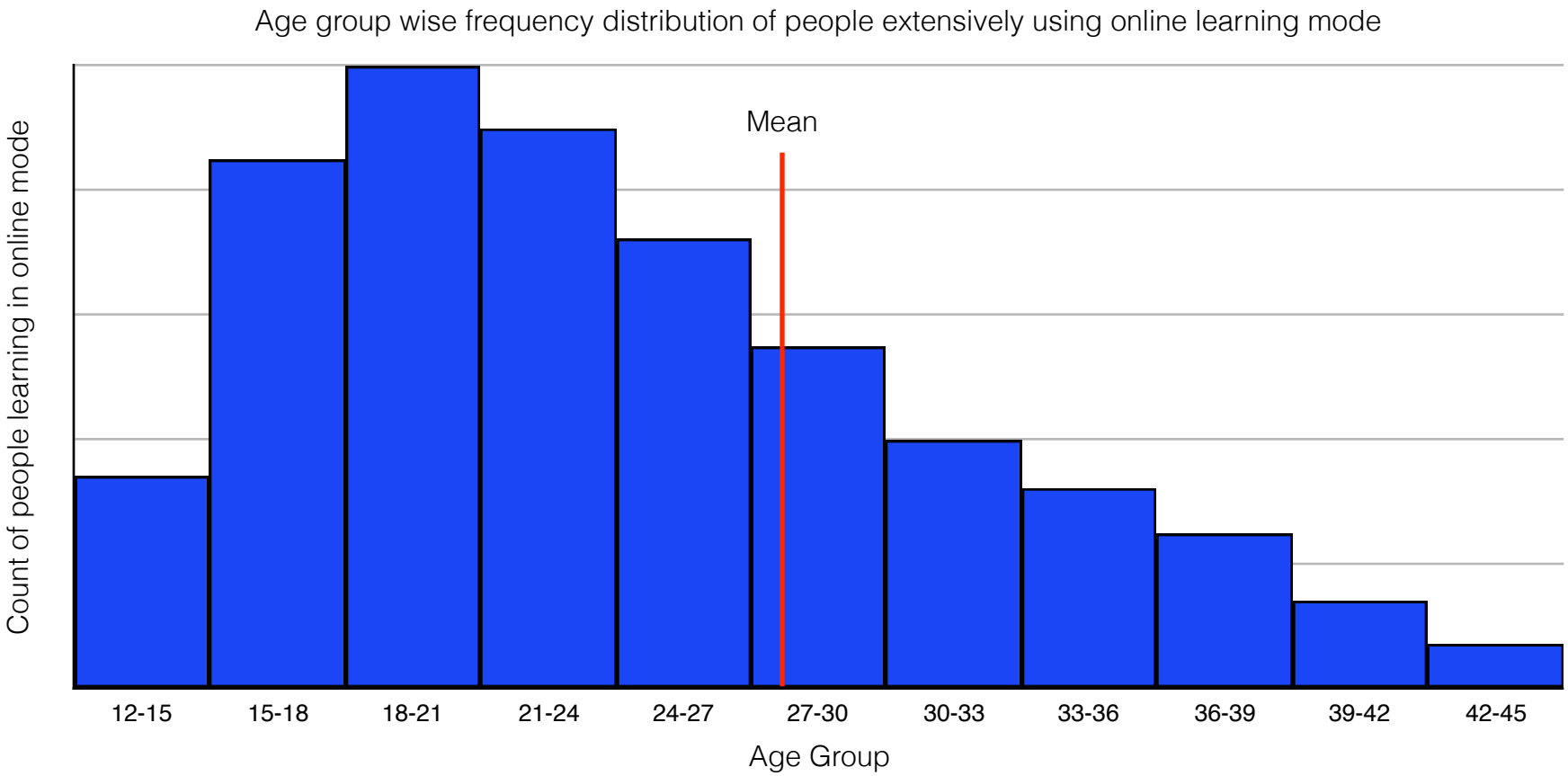
Skewness – Example

The data in the table shows the count, adjusted to a convenient unit, of people learning in online mode classified into age group.

Age group 12-15	34
Age group 15-18	85
Age group 18-21	100
Age group 21-24	90
Age group 24-27	72
Age group 27-30	55
Age group 30-33	40
Age group 33-36	32
Age group 36-39	25
Age group 39-42	14
Age group 42-45	7

Skewness – Example

If we draw a histogram we get a right skewed or positively skewed distribution as below.



Skewness

The formula for computing skewness is

$$Skew = \frac{\frac{1}{n}\sum_1^n (x_i - \bar{x})^3}{\left(\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2\right)^{\frac{3}{2}}}$$

This formula cannot compute skewness without some bias. But this is the most popular formula for computing skewness.

Kurtosis

There is another feature we need to consider while talking about the distributions and that is the shape of the tail of the distribution on the far left or the far right. Kurtosis is the measure of the thickness or heaviness of the tails of a distribution.

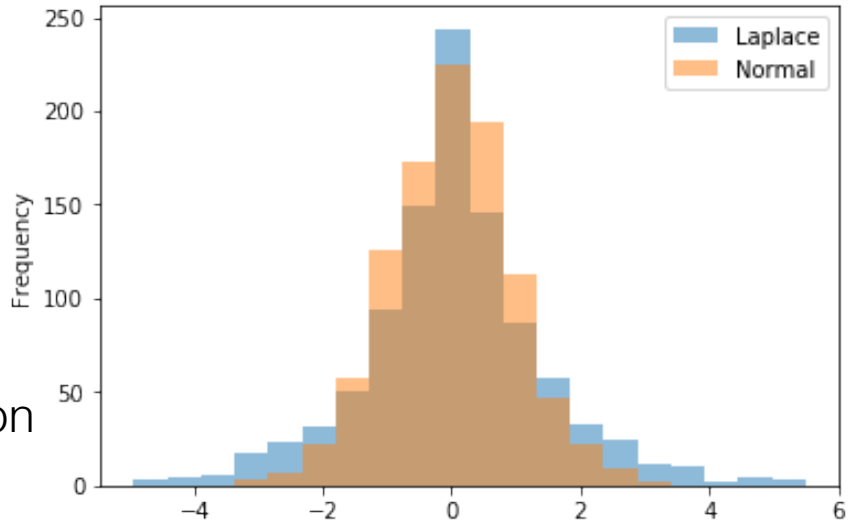
The formula for computing kurtosis is

$$Kurtosis = \frac{\frac{1}{n}\sum_1^n (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_1^n (x_i - \bar{x})^2\right)^2}$$

This formula cannot compute kurtosis without some bias. But this is the most popular formula for computing kurtosis.

Kurtosis

Excess kurtosis means the distribution of event outcomes have lots of instances of outlier results, causing "fat tails" on the distribution curve. This means the event in question is prone to extreme outcomes.



Let us now consider Laplace distribution

- Excess kurtosis for Laplace = 2.78
- Laplace has fat tail
- It is the tails that mostly account for kurtosis, not the central peak.
- Distributions with higher kurtosis are mostly used in extreme event analysis or catastrophic modelling.

Box - Whisker Plot

Five number summary of a variable, written in increasing order, Min, Q1, Q2, Q3, Max provide information on the centre and variation of the variable in a compact manner.

Box - Whisker plot is a diagram based on the five number summary of a data set. Here instead of minimum we take minimum value which is not considered as an outlier and similarly, instead of maximum we take maximum value which is not considered as an outlier.

One way of computing modified Min and Max as
Min = Q1 - 1.3 * (Q3 - Q1)
Max = Q3 + 1.3 * (Q3 - Q1)

Box - Whisker Plot – Example

Let us consider the following Data set:

15, 20, 25, 28, 24, 26, 27, 26, 25, 31, 33

Here we get the following:

Q1 = 24.5
Q2 = 26.0.
Q3 = 27.5
IQR = Q3 - Q1 = 3
Min = Q1 - 1.3*IQR = 24.5 - 1.3 * 3 = 20.6
Max = Q3 + 1.3 * IQR
= 27.5 + 1.3 * 3
= 31.4

Box - Whisker Plot – Example

And if we plot the output on a graph we get the following Box - Whisker Plot.

Here:

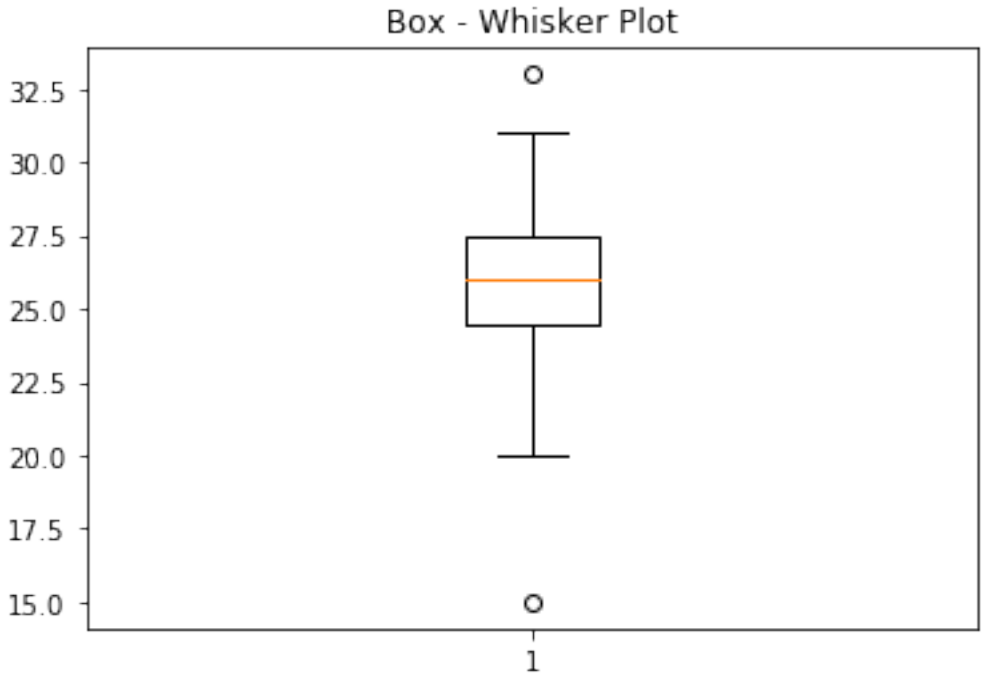
Q1 = 24.5 Q2 = 26.0. Q3 = 27.5 IQR = Q3 - Q1 = 3

Min = Q1 - 1.3*IQR = 24.5 - 1.3 * 3 = 20.6

Max = Q3 + 1.3 * IQR

= 27.5 + 1.3 * 3

= 31.4



Scatter Plot

The most effective way to display the relation between two quantitative variables is a scatterplot.

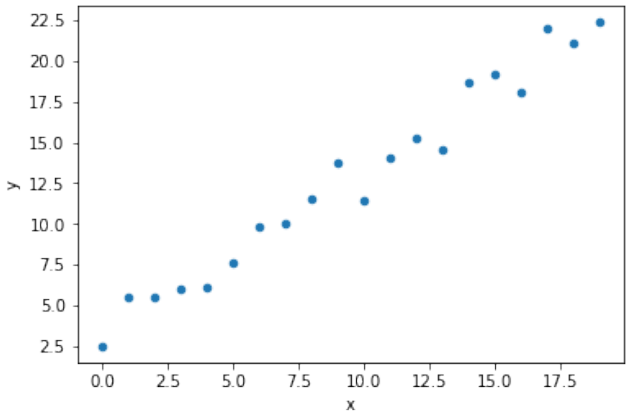
A scatterplot shows the relationship between two quantitative variables measured on the same individuals.

The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

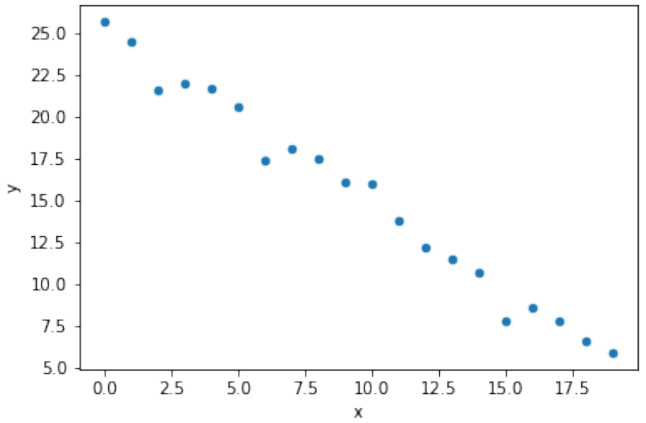
Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Scatter Plot

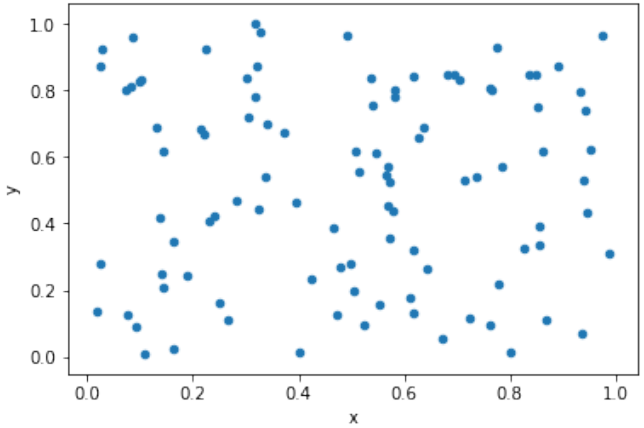
Positive Association



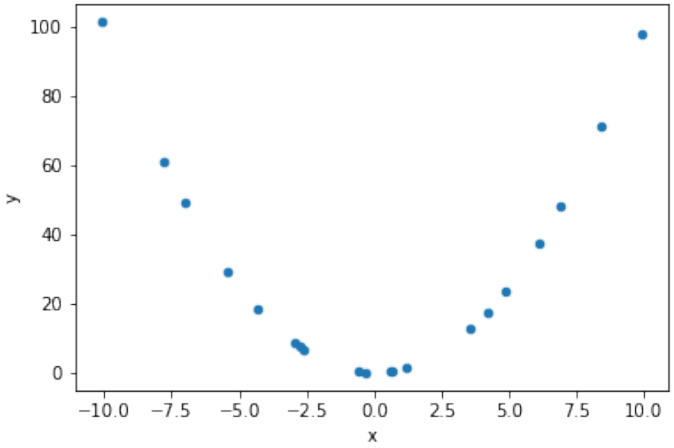
Negative Association



Zero Association



Nonlinear Association



Scatter Plot

To interpret a scatterplot, look first for an overall pattern.
This pattern should reveal the direction, form and strength of the relationship between the two variables.

The important form of the relationships between variables are linear relationships, where the points in the plot show a straight-line pattern. Curved relationships and clusters are other forms.

Two variables are positively associated when most of the values of one variable increase as most of the values of other variable increase.
Two variables are negatively associated when most of the values of one variable decrease as most of the values of other variable increase, and vice versa.

The strength of relationship is determined by how close the points in the scatterplot lie to a simple form such a line.

Correlation Coefficient

The scatterplot provides a visual impression of the nature of relation between two variables in a bivariate data set.

Our visual impression of the closeness of the scatter to a linear relation can be quantified by calculating a numerical measure, called the sample correlation coefficient.

The computational formula for correlation co-efficient is as below:

$$r_{xy} = \frac{\sum_1^n \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_1^n \left(x_i - \bar{x}\right)^2 \quad \sum_1^n \left(y_i - \bar{y}\right)^2}}$$

Correlation Coefficient

Let us talk about some important features of the correlation coefficient.

Positive correlation indicates positive association between the variables, and negative correlation indicates negative association.

The correlation always falls between -1 and 1. Values near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as correlation moves away from 0 toward 1. The extreme values r = 1 occur only in the case of a perfect linear relationship.

Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x, y or both.

Correlation measures the strength of only a linear relationship between two variables. Correlation does not describe curved relationships between variables.

The correlation is strongly affected by few outlying observations.

References

Agresti, A. & Finlay, B., Statistical Methods for the Social Sciences, 3rd Edition. Prentice Hall, 1997.

Anderson, T. W. & Sclove, S. L., Introductory Statistical Analysis. Houghton Mifflin Company, 1974.

Johnson, R.A. & Bhattacharyya, G.K., Statistics: Principles and Methods, 2nd Edition. Wiley, 1992.

Weiss, N.A., Introductory Statistics. Addison Wesley, 1999.