Module 2

Data Analytics with Python – Statistics

Section

Math for Data Science – Basic Statistics

Correlation and Covariance

Covariance

- Covariance is a statistical technique used for determining the relationship between the movement of two random variables. In short, how much two random variables change together.
- Provides a measure of the strength of the correlation between two or more sets of random variates.
- It is the relationship between two variables in a given data set.
- Example: Let E(x) be the expected value of a given variable x, and E(y) be the expected value of variable y, then the covariance between x and y is given by:

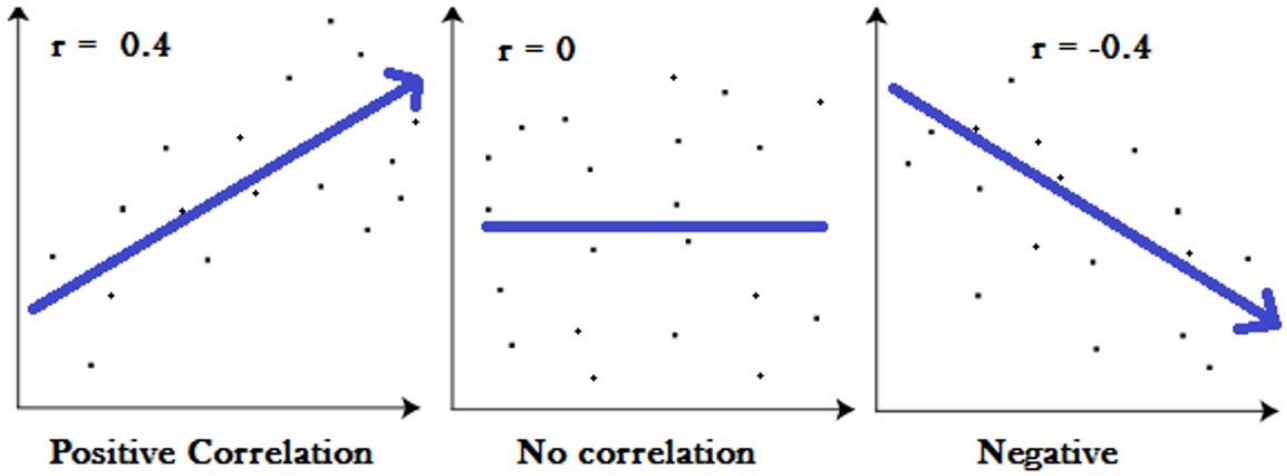
$$cov(x, y) = E[xy] - E[x]\,E[y]$$

Correlation

- It is the measure of the strength of the relation between two variables. How strongly two variables are connected is defined as the correlation.
- It depends upon two variables, change in one variable effect a change in second variable.
- Its value lies in the range of -1 and +1.
- Is related to covariance by the given formula: $cor(x,y) = cov(x,y)/\sigma x\,\sigma y$

Correlation Coefficient

- A correlation coefficient is a way to put a value to the relationship.
- The following graph shows the correlation of -1, 0 and 1.



Basis for comparison	Covariance	Correlation
Definition	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1
Change in scale	Affects covariance	Does not affect the correlation
Unit-free measure	No	Yes

Business logic with correlation Analysis

- As the old saying goes, ''correlation is not causation.'' Even still, correlation can be a useful measure for predicting the future.
- If you look at the analysis of a publicly traded company, you will quickly find yourself sorting through large amounts of data. In fact, the point of having all this data available should be to help us in making informed investment decisions.
- The process of making those decisions is often made easier by using standard statistical correlations. **Correlations** provide us with measurements of the actual relationships between two or more variables.

Measure of shape

Kurtosis

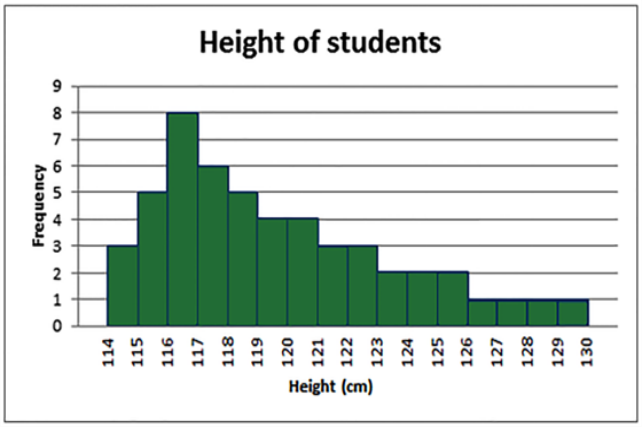
- Measure of how tall and sharp the central peak is, relative to a standard bell curve.
- Also called as the measure of tailed Ness of a data distribution.
- Outliers impact the kurtosis of a data distribution.
- Formula to calculate kurtosis:

$$\frac{\sum_{i=1}^{N} \frac{(X_i - \bar{X})}{N}}{s^4}$$

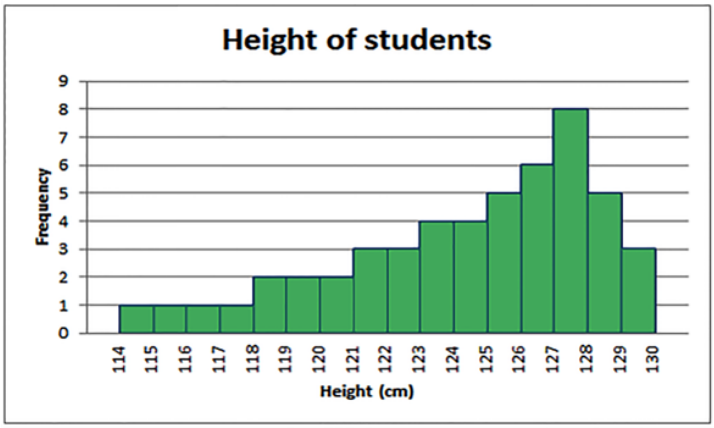
Where Xi is the individual value, X is the mean, N is the total number of values and s is the standard deviation

Skewness

The amount and direction of skew in data distribution, i.e., deviation from the horizontal symmetry.
Two types of skewed distribution:



Positively skewed distribution



Negatively skewed distribution

Linear Algebra

Matrix

A matrix is a two-dimensional array that has a fixed number of rows and columns and contains a number at the intersection of each row and column. A matrix is usually delimited by square brackets.

Example: The following is an example of a matrix having two rows and three columns:

$$A = \begin{bmatrix} 1 & 5 & 0 \\ 2 & 9 & 1 \end{bmatrix}$$

Dimension of a matrix
The number of rows and columns of a matrix constitutes its dimension. If a matrix has K rows and L columns, we say that it is a K X L matrix, or that it has dimension K X L.

Example: Define a matrix

$$A = \begin{bmatrix} 1 & 5 & 0 \\ 2 & 9 & 1 \end{bmatrix}$$

The matrix A has 2 rows and 3 columns. So, we say that is a 2 X 3 matrix.

Elements of a matrix
The numbers contained in a matrix are called elements of the matrix (or entries, or components). If is a matrix, the element at the intersection of row k and column l is usually denoted by A_{kl} and we say that it is the kl^{th} element of A.
Example: Let A be a 2 X3 matrix defined as follows:

$$A = \begin{bmatrix} 1 & 5 & 0 \\ 2 & 9 & 1 \end{bmatrix}$$

The element of at the intersection of the 2^{nd} row and the 1^{st} column, i.e., its A_{21}-th element is 2.

Vectors

If a matrix has only one row or only one column it is called a vector.
A matrix having only one row is called a **row vector**.

Example: The 1 X 3 matrix

$$A = [1\ 2\ 3]$$

is a row vector because it has only one row.

A matrix having only one column is called a **column vector**.

Example The 2X1 matrix

$$A = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

is a column vector because it has only one column.

Inverse Of Matrix

The inverse of matrix is another matrix, which on multiplication with the given matrix gives the multiplicative identity. For a matrix A, its inverse is A^{-1}, and $A.A^{-1} = I$. Let us check for the inverse of matrix, for a matrix of order 2 × 2, the general formula for the inverse of matrix is equal to the adjoint of a matrix divided by the determinant of a matrix.

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$A^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

$$A^{-1} = \frac{1}{|A|}\,\text{Adj}\,A$$

The inverse of matrix exists only if the determinant of the matrix is a non-zero value. The matrix whose determinant is non-zero and for which the inverse matrix can be calculated is called an invertible matrix.

Transpose Of a Matrix

If A is a K X L matrix, its **transpose**, denoted by A^T, is the L X K matrix such that the (l,k)th element of A^T is equal to the (k, l) -th element of A. In other words, the columns of A^T are equal to the rows of A (equivalently, the rows of A^T are equal to the columns of A).

Example Let A be the 2 X 3 matrix defined by

$$A = \begin{bmatrix} 7 & 5 & 0 \\ 5 & \frac{1}{2} & 1 \end{bmatrix}$$

Its transpose A^T is the following 3 X 2 matrix:

$$A^T = \begin{bmatrix} 7 & 5 \\ 5 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$$

Eigenvalues and Eigenvectors

Let A be an n X n matrix (Square Matrix).

1. An ***eigenvector*** of A is a *nonzero* vector v in Rn such that Av=λv, for some scalar λ.
2. An ***eigenvalue*** of A is a scalar λ such that the equation Av=λv has a *nontrivial* solution.

If Av=λv for v not equal to 0, we say that λ is the ***eigenvalue for*** v, and that v is an ***eigenvector for*** λ.

The German prefix “eigen” roughly translates to “self” or “own”. An eigenvector of A is a vector that is taken to a multiple of itself by the matrix transformation T(x)=Ax, which perhaps explains the terminology. On the other hand, “eigen” is often translated as “characteristic”; we may think of an eigenvector as describing an intrinsic, or characteristic, property of A.

Note- Eigenvalues and eigenvectors are only for square matrices.

Summary

- How strongly two variables are connected is defined as the correlation.

- Covariance is a statistical technique used for determining the relationship between the movement of two random variables. In short, how much two random variables change together.

- Kurtosis Measure of how tall and sharp the central peak is, and skewness is the deviation of data from the horizontal symmetry.

- A matrix is a two-dimensional array that has a fixed number of rows and columns and contains a number at the intersection of each row and column.

- If a matrix has only one row or only one column it is called a vector.

- The inverse of matrix is another matrix, which on multiplication with the given matrix gives the multiplicative identity. The inverse of matrix exists only if the determinant of the matrix is a non-zero value

- Transpose of matrix is a transformation of matrix rows in respective columns.