UNIT-2

Random Variables

- Random Variables are basically a function that maps from the set of sample spaces to a set of real numbers.
- Random variables are defined over a sample space of any random experiment.
- Values of random variables correspond to the outcomes of the random experiment.
- There are two basic types of random variables,
  - Discrete Random Variables
  - Continuous Random Variables
- We define random variable a function which maps from sample space of an experiment to the real numbers. Mathematically, Random Variable is expressed as,
    - $X: S \rightarrow R$
  - where,
    - X is Random Variable (It is usually denoted using capital letter)
    - S is Sample Space
    - R is Set of Real Numbers
- Suppose a random variable X takes m different values i.e. sample space X = {x1, x2, x3.....xm} with probabilities
    - $P(X = xi) = pi$
      - where $1 \leq i \leq m$
- The probabilities must satisfy the following conditions :
  - $0 \leq pi \leq 1$; where $1 \leq i \leq m$
  - $p1 + p2 + p3 + \ldots\ldots + pm = 1$ Or we can say $0 \leq pi \leq 1$ and $\sum pi = 1$
- Hence possible values for random variable X are 0, 1, 2.
  - X = {0, 1, 2} where m = 3
  - P(X = 0) = (Probability that number of heads is 0)
    - = P(TT) = 1/2×1/2 = 1/4
  - P(X = 1) = (Probability that number of heads is 1)
    - = P(HT | TH) = 1/2×1/2 + 1/2×1/2 = 1/2
  - P(X = 2) = (Probability that number of heads is 2)
    - = P(HH) = 1/2×1/2 = 1/4
- Here, you can observe that, $(0 \leq p1, p2, p3 \leq 1/2)$
  - p1 + p2 + p3 = 1/4 + 2/4 + 1/4 = 1
- For example,
- Suppose a dice is thrown (X = outcome of the dice). Here, the sample space S = {1, 2, 3, 4, 5, 6}.
  - The output of the function will be:
    - P(X=1) = 1/6
    - P(X=2) = 1/6
    - P(X=3) = 1/6
    - P(X=4) = 1/6
    - P(X=5) = 1/6
    - P(X=6) = 1/6

Discrete Random Variable

- A random variable X is said to be discrete if it takes on a finite number of values. Th probability function associated with it is said to be
- PMF = Probability Mass Function P(xi), if
  - $0 \leq pi \leq 1$
  - $\sum pi = 1$ where the sum is taken over all possible values of x

Discrete Random Variables Example

- Let S = {0, 1, 2}

xi	0	1	2
Pi(X = xi)	P1	0.3	0.5

- Find the value of P (X = 0)

Solution:

- We know that the sum of all probabilities is equal to 1. And P (X = 0) be P1
  - P1 + 0.3 + 0.5 = 1
  - P1 = 0.2
  - Then, P (X = 0) is 0.2

Continuous Random Variable

- A random variable X is said to be continuous if it takes on an infinite number of values.
- The probability function associated with it is said to be PDF (Probability Density Function).
- PDF (Probability Density Function)
  - If X is a continuous random variable. P $(x < X < x + dx) = f(x)dx$ then,
  - $0 \leq f(x) \leq 1$; for all x
  - $\int f(x)\, dx = 1$ over all values of x
  - Then P (X) is said to be PDF of the distribution.

Continuous Random Variables Example

Example: Find the value of P (1 < X < 2)

- Such that,
  - $f(x) = kx3; 0 \leq x \leq 3 = 0$
- Otherwise f(x) is a density function.

Solution:

- If a function f is said to be a density function, then the sum of all probabilities is equal to 1.
- Since it is a continuous random variable Integral value is 1 overall sample space s.
  - $\int f(x)\, dx = 1$
  - $\int kx3\, dx = 1$
  - K[x4]/4 = 1
  - Given interval, $0 \leq x \leq 3 = 0$
  - K[34 – 04]/4 = 1

- K(81/4) = 1
- K = 4/81

- Thus,
  - P (1 < X < 2) = k×[X4]/4
  - P = 4/81×[16-1]/4
  - P = 15/81

Need of statistics in Data Science and Big Data Analytics

Terminologies associated with statistics

- **Population:** It is an entire pool of data from where a statistical sample is extracted. It can be visualized as a complete data set of items that are similar in nature.
- **Sample:** It is a subset of the population, i.e. it is an integral part of the population that has been collected for analysis.
- **Variable:** A value whose characteristics such as quantity can be measured. it can also be addressed as a data point, or a data item.
- **Distribution:** The sample data that is spread over a specific range of values.
- **Parameter:** It is a value that is used to describe the attributes of a complete data set (also known as 'population'). Example: Average, Percentage
- **Quantitative analysis:** It deals with specific characteristics of data- summarizing some part of data, such as its mean, variance, and so on.
- **Qualitative analysis:** This deals with generic information about the type of data, and how clean or structured it is.

Data associated with statistics is of many types. Some of them have been discussed below.

- **Categorical data** represents characteristics of people, such as marital status, gender, food they like, and so on. It is also known as 'qualitative data' or 'yes/no data'. It takes numerical values like '1', '2', where these numbers indicate one or other type of characteristics. These numbers are not mathematically significant, which means it can't be associated with each other.
- **Continuous data** deals with data that is immeasurable, and can't be counted, which basically continual forms of values are. Predictions from a linear regression are continuous in nature. It is a continuous distribution that is also known as probability density function.
- On the other hand, **Discrete values** can be measured, counted, and are discontinuous. Predictions from logistic regression are considered to be discrete in nature. Discrete data is non-continuous, and density concept doesn't come into the picture here. The distribution is known as probability mass function.

Descriptive statistics:

- **Inferential statistics:** It deals with drawing inferences/conclusions on the sample data set which is obtained from the population (entire data set) based on the relationship identified between data points in the data set. It helps in generalizing the relationship to the entire dataset. It is important to remember that the dataset drawn from the population is relevant and represents the population accurately.
- **Regression:** The term 'regression' which is a part of statistics and machine learning, talks about how data can be fit to a line, and how every point from the straight line gives some insights. In terms of machine learning, it can be understood as tasks that can be solved without explicitly being programmed. They discuss how a line can be fit to a given set of data points, and how it can be further extrapolated for the predictions to be done.
- **Maximum likelihood**: It is a method that helps in finding values of parameters for a specific model. The values of the parameters have to be such that the likelihood of the predictions that occur have to be maximum in comparison to the data values that were actually observed. This means the difference between the actual and predicted value has to be less, thereby reducing the error and increasing the accuracy of the predictions.

Measures of Central Tendency:

- When we work with numerical data, it seems apparent that in most set of data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as central tendency.
- Some more commonly used measures, namely arithmetic mean, median and mode.

1. Arithmetic Mean:

- The mean or average is the most popular and well known measure of central tendency.
- It can be used with both discrete and continuous data, although its use is most often with continuous data.
- The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

Example1: Find the Average marks obtained by student 64,69,72,72,75,65.

Solution: for ungrouped data A.M.$=\bar{x}=\sum x/n$

$=417/6$

$=69.5$

The average marks are=69.5.

Example2: Find the A.M. for the following

Solution: Grouped data discrete case

No of days spent	1	2	3	4	5	6	7	8
No of patient	5	6	5	10	8	4	3	2

No of days spent(x)	1	2	3	4	5	6	7	8	Total
No of patient(f)	5	6	5	10	8	4	3	2	43=N=$\sum f$
fx	5	12	15	40	40	24	21	16	Σfx =173

A.M.= $\Sigma fx/\sum F=173/43$

$=4.02$

Ex 3: Find the arithmetic mean for the following

Monthly sales	frequency
100-120	15
120-140	35
140-160	50
160-180	60
180-200	30
200-220	10

Solution:

For grouped data continuous variate case

monthly sales CI(Class Interval)	Frequency(f)	X(mid point of CI)	fx
100-120	15	110	1650
120-140	35	130	4550
140-160	50	150	7500
160-180	60	170	10200
180-200	30	190	5700
200-220	10	210	2100
total	N=200	**Σfx** =31700	

Arithmetic mean= $\Sigma fx/\sum f$, where $\sum f=N=200$

$= 31700/200$

$=158.5$

2. Median

- Median represents the middle value for any group.
- It is the point at which half the data is more and half the data is less.
- Median helps to represent a large number of data points with a single data point.
- The median is the easiest statistical measure to calculate.
- For calculation of median, the data has to be arranged in ascending order, and then the middlemost data point represents the median of the data.

The following steps are helpful while applying the median formula for ungrouped data.

- Step 1: Arrange the data in ascending or descending order.
- Step 2: Secondly, count the total number of observations 'n'.
- Step 3: Check if the number of observations 'n' is even or odd.

Median Formula When n is Odd

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

Median Formula When n is Even

$$\text{Median} = \frac{\frac{n^{th}}{2}\text{obs.} + \left(\frac{n}{2}+1\right)^{th}\text{obs.}}{2}$$