**Normal Distribution**

**Recap**

* Probability distributions
* Discrete vs continuous
* Binomial
* Poisson

**Topics**

* Continuous probability distribution - Normal Distribution

**Introduction**

The normal distribution is a continuous distribution and plays a very important and pivotal role in statistical theory and practice.

It plays an important role in the area of statistical inference. One area of application is called statistical quality control.

Its importance is also due to the fact that in practice, the experimental results very often seem to follow the normal distribution or the bell-shaped curve.

**The Structure**

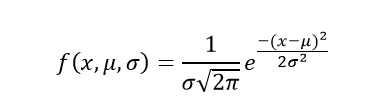
The normal curve is symmetrical and is defined by its mean (mu) and its standard deviation (sigma). The shape of the curve is shown below (draw a normal curve). The mean, mode and median of the distribution have the same value.

**Characteristics of the Normal Curve**

1. All normal curves are symmetrical about the mean. This means that the number of units in the data below the mean is the same as the number of units above the mean. Therefore, the mean and median have the same value.
2. The height of the normal curve is at its maximum at the mean value. Thus, the mean and the mode coincide.
3. Hence, mean, median and mode are the same in a normal curve.

**Normal Curve Function**

The height of the normal curve Y at any value of the random continuous variable x is given by the following equation:



**Common Applications**

1. Statistical quality control
2. Sociological studies
3. Healthcare

**Construction of the Curve**

The bell-shaped curve is constructed from the probability distribution of the data where X-axis denotes the various values of the continuous variable (x) and the Y-axis denotes the probability distribution or the relative frequency of such occurrence of the variable.

As an illustration, let us develop the normal distribution curve by first constructing a histogram of a frequency distribution of grouped data.

The X-Axis and the Y-axis represent respectively the class interval at its mid-point.

The corresponding relative frequency, which is also the probability that the random variable will have a value within the given class interval.

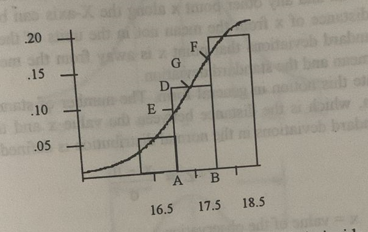
Data here

|  |  |  |  |
| --- | --- | --- | --- |
| Ages (CI) | Mid-point (X) | (f) | Relative Frequency |
| 16 and upto 18 | 16.5 | 4 | 0.04 |
| 17 and upto 18 | 17.5 | 14 | 0.14 |
| 18 and upto 19 | 18.5 | 18 | 0.18 |
| 19 and upto 20 | 19.5 | 28 | 0.28 |
| 20 and upto 21 | 20.5 | 18 | 0.18 |
| 21 and upto 22 | 21.5 | 14 | 0.14 |
| 22 and upto 23 | 22.5 | 4 | 0.04 |
|  | Total | 100 | 1.00 |

The histogram for this data would be given below:

**Interesting Interpretations**

The development of the smooth normal curve from a histogram has some interesting interpretations. It should be noted that since the height of each rectangle of the histogram represents the relative frequency of the corresponding class interval enclosed by the given rectangle, and since the width of each rectangle is the same, it can be concluded that the area of a given rectangle has the same proportion to the total area of all the rectangles as the relative frequency of the class interval represented by that rectangle.



If in the above figure the area of the triangle DEG which is inside the rectangle but outside the curve is equal to the area of the triangle CFG, which is outside the rectangle but inside the curve, then it can be concluded that the area of the rectangle ABCD is equal to the area under the curve AEFB bounded by the interval AB.

This is true for every rectangle.

*What is the meaning of this?*

This means that the area under the curve bounded by the class interval for any given class represents the relative frequency of that class. In other words, the area under the curve lying between any two vertical lines at points A and B along the X-axis represents the probability that the random variable x takes on value in that interval bounded by A and B.

This means that by finding the area under the curve between any two points along the X-axis, we can find the percentage of data occurring within these two points.

It should also be noted that the relationship between the curve and the histogram is true to a large extent except that while the histogram is close-ended the curve is not.

**Area under the Normal Curve**

Since all the normal curves are defined by the values of mu and sigma, with all other characteristics being the same, it is possible to standardize the normally distributed variable into a form so that a single table of areas under the normal curve between different points is applicable, no matter what the units of the original data are.

The area under the curve between the mean and any point x which is 1 standard deviation away from the mean is always the same no matter what the values of mu and sigma may be.

This is also true for the area under the curve between the mean and a point x which is 2 standard deviations away from the mean or ½ standard deviation away or any number of standard deviations away.

This means that the area under the normal curve between this mean any other point x along the X-axis can be identified by simply knowing the distance of x from the mean not in the units of the data but in terms of number of standard deviations that point x is away from the mean irrespective of the values of the mean and the standard deviation.

**General form**

The number of standard deviations Z for an observation, which is the distance between the value x and the mean in terms of number of standard deviations in the normal distribution is defined by;

Z = x - mu / sigma

Then the desired area under the curve can be found by referring to the table of the standardized normal distribution and picking the value of the entry which corresponds to the calculated value of Z from the above formula.

The computed value Z is also known as the Z-score or standardized normal deviation. The value of Z itself follows a normal probability distribution with a mean of 0 and a standard deviation of one.

This probability distribution is known as the standard normal probability distribution. This method allows us to use only one table of areas for all types of normal distribution.