Probability Density Functions and Cumulative Density Functions

skilltohire · Follow

4 min read · Jul 30, 2020

A probability density function for a constant random variable is a function whose value at any given point in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample space.

For a continuous function, the probability density function (PDF) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points as shown;

$$\int_a^b f(x)dx = Pr[a \leq X \leq b]$$

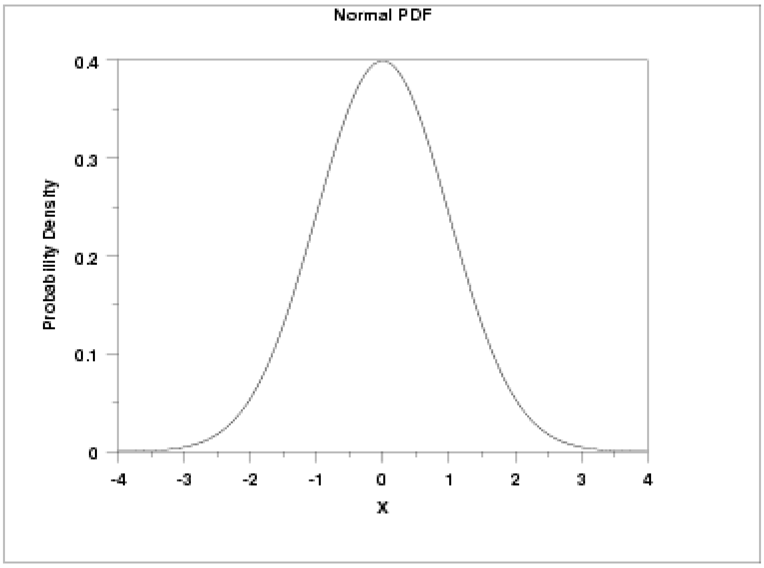
https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

For a discrete distribution, the PDF is the probability that the variate takes the value x as shown;

$$f(x) = Pr[X = x]$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

The PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variable's PDF over that range that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is non negative everywhere, and its integral over the entire space is equal to 1.



The cumulative distribution function (CDF) is the probability that the variable takes a value less than or equal to x. That is;

$$F(x) = Pr[X \leq x] = \alpha$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

For a continuous distribution, this can be expressed mathematically as;

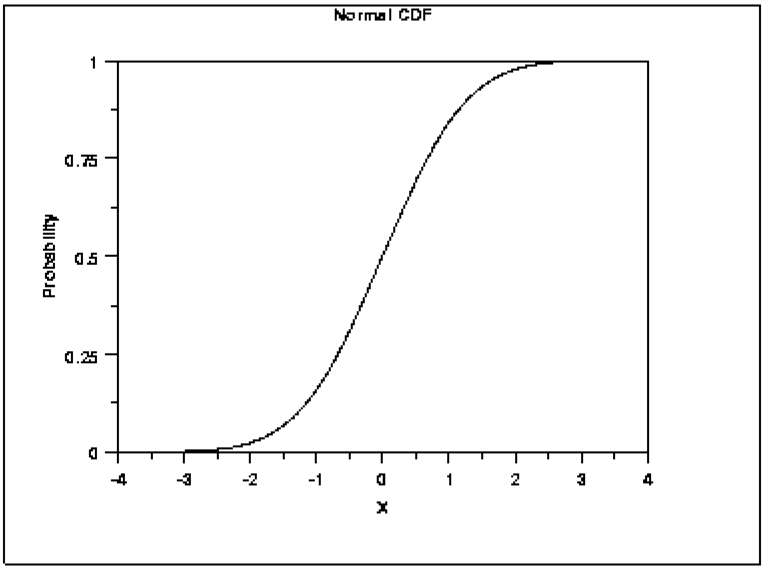
$$F(x) = \int_{-\infty}^{x} f(\mu) d\mu$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

For a discrete distribution, the cdf can be expressed as;

$$F(x) = \sum_{i=0}^{x} f(i)$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm



https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

There are few terms which we usually associate with PDF's and CDF's and they are;

Percentiles

Skewness

Let us understand them in an intuitive manner;

Percentiles tell you how a value compares to other values. The general rule is that if value X is at the K percentile, then X is greater than K% of the values. The percentile rank can be understood easily with an example. It is that amount of fraction of people who scored lower than you (or the same) in an examination. So for instance if i am in the 90th percentile then, i actually did well as or better than 90% of the people who took the examination.

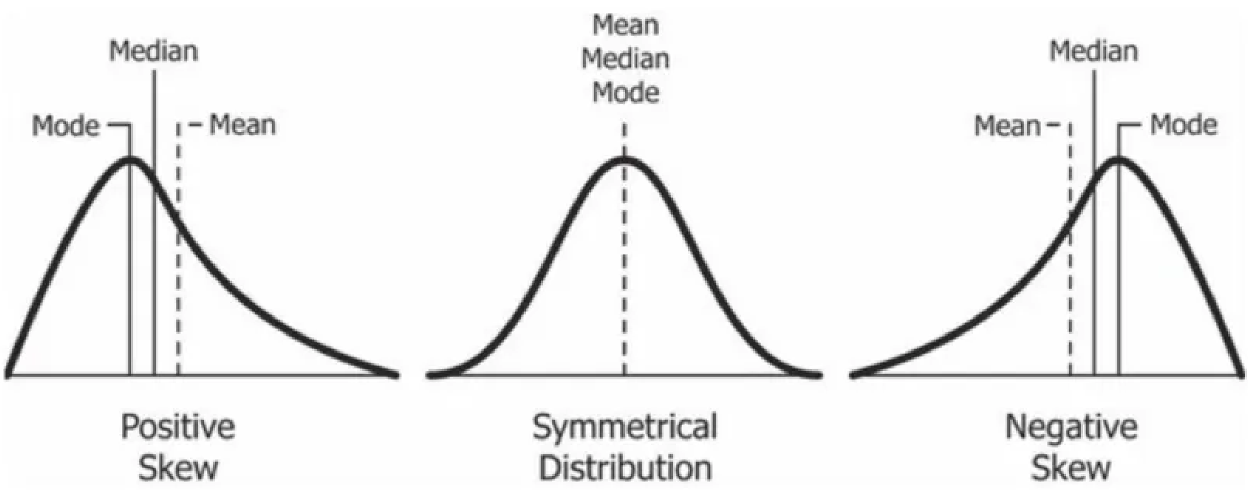
Percentile rank takes a value and computes its percentile rank in a set of values whereas percentile takes a percentile rank and computes the corresponding value. The ordinal rank(n) can be calculated using the following formula, note that P denotes percentile here;

$$n = \left\lceil \frac{P}{100} \times N \right\rceil.$$

https://wikimedia.org/api/rest_v1/media/math/render/svg/269ac418ad7fb85b46cf8847af7037815ba7748a

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

If for our distribution, the mean value is greater than the median i.e the average of all the data points is greater than the middle value in the data set then the data is said to be positively skewed data. Similarly, if median is greater than the mean value for a data set then the data is said to be negatively skewed. When we say 'skewed left', we mean that the left tail is long relative to the right tail. Similarly when we say 'skewed right', we mean that the right tail is long relative to the left tail.



https://upload.wikimedia.org/wikipedia/commons/thumb/c/cc/Relationship_between_mean_and_median_under_different_skewness.png/868px-Relationship_between_mean_and_median_under_different_skewness.png

For univariate data $Y_1, Y_2, \ldots, Y_N$, the formula for skewness is:

$$g_1 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^3 / N}{s^3}$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

The above formula is termed as Fisher-Pearson coefficient of skewness. Here, Y bar denotes mean and s denotes standard deviation. N is the number of data points that we have. At times in software programs, we use adjusted Fisher-Pearson coefficient of skewness which is computed as;

$$G_1 = \frac{\sqrt{N(N-1)}}{N-2} \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3}$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean is equal to median is equal to mode.

Probability Density | Cdf | Statistics | Distribution