**Binomial Distribution**

* **Aim: -** How we can use binomial distribution in R.
* **Theory: - Binomial Distribution**

In probability theory and statistics, the **binomial distribution** is the discrete probability distribution that gives only two possible results in an experiment, either **Success or Failure**.

In binomial probability distribution, the number of ‘Success’ in a sequence of n experiments, where each time a question is asked for yes-no, then the boolean-valued outcome is represented either with success/yes/true/one (probability p) or failure/no/false/zero (probability q = 1 − p). A single success/failure test is also called a [Bernoulli trial](https://byjus.com/maths/bernoulli-trials-binomial-distribution/) or Bernoulli experiment, and a series of outcomes is called a **Bernoulli process**. For n = 1, i.e. a single experiment, the binomial distribution is a **Bernoulli distribution**. The binomial distribution is the base for the famous binomial test of statistical importance.

* **Formula:** The binomial distribution formula is for any [random variable](https://byjus.com/maths/random-variable/) X, given by;

|  |
| --- |
| P (x: n, p) = nCx px (1-p) n-x  Or  P(x:n,p) = nCx px (q)n-x |

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

The binomial distribution formula can also be written in the form of n-Bernoulli trials, where nCx = n!/x!(n-x)!. Hence,

**P(x:n,p) = n!/[x!(n-x)!].px.(q)n-x**

Binomial Distribution Mean and Variance

For a binomial distribution, the mean, variance and standard deviation for the given number of success are represented using the formulas

Mean, μ = np

Variance, σ2= npq

Standard Deviation σ= √(npq)

Where p is the probability of success

q is the probability of failure, where q = 1-p

* **Properties:**

The properties of the binomial distribution are:

* There are two possible outcomes: true or false, success or failure, yes or no.
* There is ‘n’ number of independent trials or a fixed number of n times repeated trials.
* The probability of success or failure remains the same for each trial.
* Only the number of success is calculated out of n independent trials.
* Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.
* **Example**

As we already know, binomial distribution gives the possibility of a different set of outcomes. In real life, the concept is used for:

* Finding the quantity of raw and used materials while making a product.
* Taking a survey of positive and negative reviews from the public for any specific product or place.
* By using the YES/ NO survey, we can check whether the number of persons views the particular channel.
* To find the number of male and female employees in an organisation.
* The number of votes collected by a candidate in an election is counted based on 0 or 1 probability.
* **Binomial Distribution in R**
* R has four in-built functions to generate binomial distribution. They are described below:

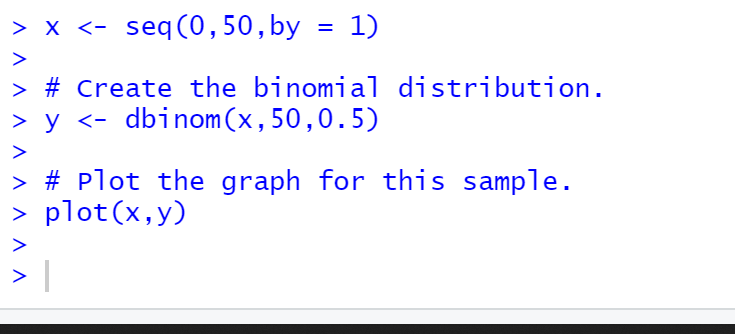
1. dbinom(x, size, prob)
2. pbinom(x, size, prob)
3. qbinom(p, size, prob)
4. rbinom(n, size, prob)

* Following is the description of the parameters used :
* **x** is a vector of numbers.
* **p** is a vector of probabilities.
* **n** is number of observations.
* **size** is the number of trials.
* **prob** is the probability of success of each trial.

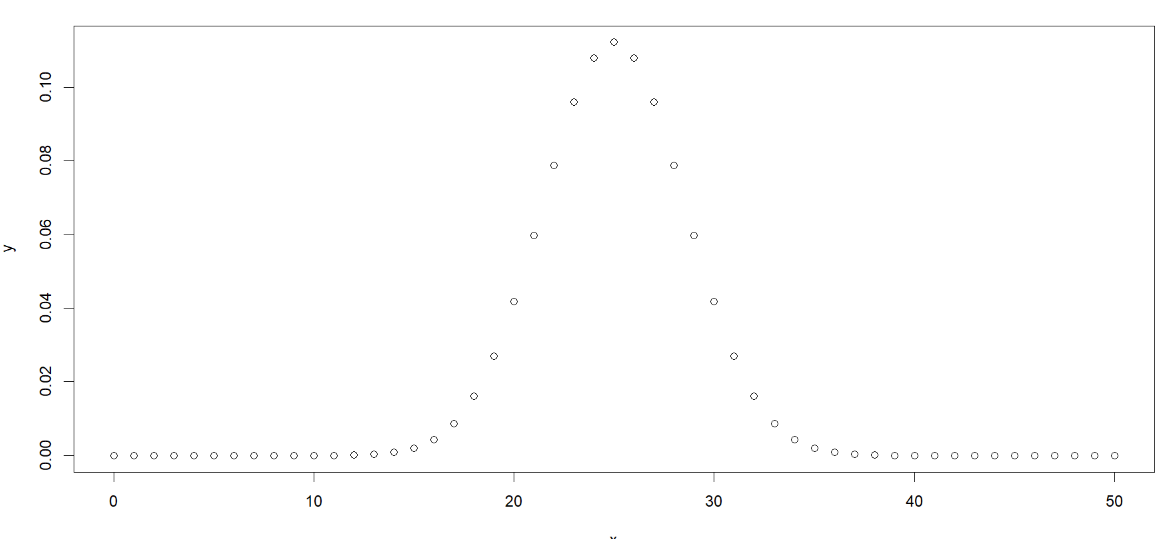
* **dbinom:**

This function gives the probability density distribution at each point.

* **Code:**

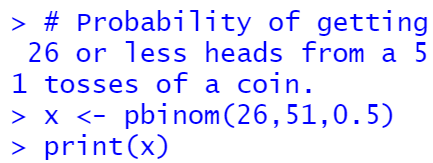
****

* **Output:**



* **pbinom:**

This function gives the cumulative probability of an event. It is a single value representing the probability.

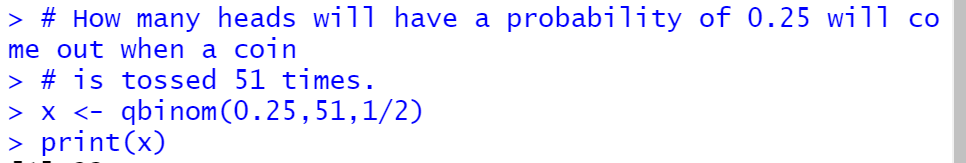
* **Code:**
* ****
* **Output:**

[1] 0.610116

* **qbinom:**

This function takes the probability value and gives a number whose cumulative value matches the probability value.

* **Code:**



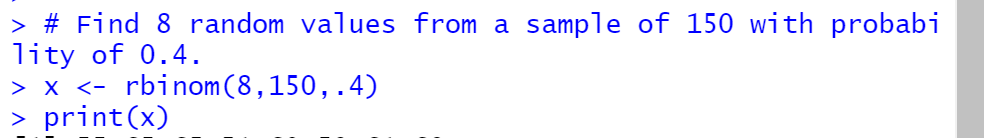
* **Output:**

[1] 23

* **rbinom:**

This function generates required number of random values of given probability from a given sample.

* **Code:**



* **Output:**

[1]  60 67 74 58 55 71 65 61

**Poisson Distribution**

* **Aim:** How we can use poisson distribution in R.
* **Theory:** Poisson Distribution

**Poisson distribution**, in [statistics](https://www.britannica.com/science/statistics), a [distribution function](https://www.britannica.com/science/distribution-function) useful for [characterizing](https://www.britannica.com/dictionary/characterizing) events with very low probabilities of occurrence within some definite time or [space](https://www.britannica.com/science/space-physics-and-metaphysics).

### Assumptions and validity

The Poisson distribution is an appropriate model if the following assumptions are true:[[14]](https://en.wikipedia.org/wiki/Poisson_distribution#cite_note-Koehrsen2019-14)

* *k* is the number of times an event occurs in an interval and *k* can take values 0, 1, 2, ... .
* The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
* The average rate at which events occur is independent of any occurrences. For simplicity, this is usually assumed to be constant, but may in practice vary with time.
* Two events cannot occur at exactly the same instant; instead, at each very small sub-interval, either exactly one event occurs, or no event occurs.
* **Formula:**

The formula for the Poisson distribution function is given by:

**f(x) =(e– λ λx)/x!**

Where,

e is the base of the logarithm

x is a Poisson random variable

λ is an average rate of value

### Poisson Distribution Mean and Variance

For Poisson distribution, which has λ as the average rate, for a fixed interval of time, then the mean of the Poisson distribution and the value of variance will be the same. So for X following Poisson distribution, we can say that λ is the mean as well as the variance of the distribution.

Hence: E(X) = V(X) = λ

where

* E(X) is the expected mean
* V(X) is the variance
* λ > 0

**Properties:**

The Poisson distribution is applicable in events that have a large number of rare and independent possible events. The following are the properties of the Poisson Distribution. In the Poisson distribution,

* The events are independent.
* The average number of successes in the given period of time alone can occur. No two events can occur at the same time.
* The Poisson distribution is limited when the number of trials n is indefinitely large.
* mean = variance = λ
* np = λ is finite, where λ is constant.
* The standard deviation is always equal to the square root of the mean μ.
* The exact probability that the random variable X with mean μ =a is given by P(X= a) = μa/ a! e -μ
* If the mean is large, then the Poisson distribution is approximately a normal distribution

### **Example:**

The Poisson distribution may be useful to model events such as:

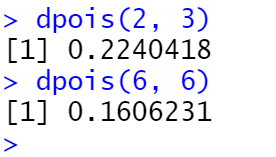
* the number of meteorites greater than 1 meter diameter that strike Earth in a year;
* the number of laser photons hitting a detector in a particular time interval; and
* the number of students achieving a low and high mark in an exam.
* **Poisson Distribution in R**
* There are four Poisson functions available in R:
* dpois()
* ppois()
* qpois()
* rpois()

* **dpois:**

This function is used for illustration of Poisson density in an R plot. The function dpois() calculates the probability of a random variable that is available within a certain range.

**Syntax:** dpois(k,λ,log)  
  
**where,** 

***k:****number of successful events happened in an interval***λ *:*** *mean per interval****log:****If TRUE then the function returns probability in form of log*

* **Code:**
* ****
* **Output:**

[1] 0.2240418

[1] 0.1606231

* **ppois:**

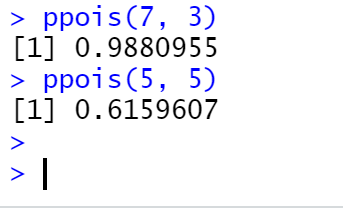
This function is used for the illustration of cumulative probability function in an R plot. The function ppois() calculates the probability of a random variable that will be equal to or less than a number.

**Syntax:** ppois(k, λ, lower.tail, log)  
  
**where,** 

***k:****number of successful events happened in an interval***λ *:*** *mean per interval*

***lower.tail:****If TRUE then left tail is considered otherwise if the FALSE right             tail  is considered****log:****If TRUE then the function returns probability in form of log*

* **Code:**



* **Output:**

[1]0.9880955

[1] 0.6159607

* **rpois:**

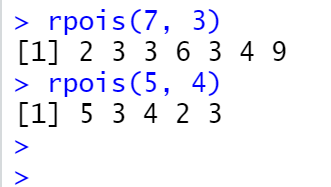
The function rpois() is used for generating random numbers from a given Poisson’s distribution.

**Syntax:** rpois(q, λ)  
  
**where,** 

***q:****number of random numbers needed*

**λ *:*** *mean per interval*

* **Code:**



* **Output:**

[1] 2 3 3 6 3 4 9

[1]  5 3 4 2 3

* **qpois:**

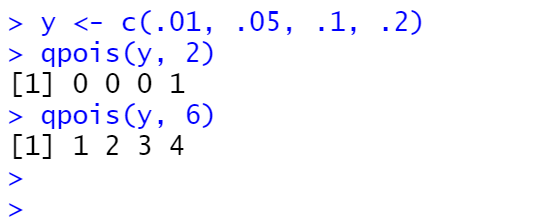
The function qpois() is used for generating quantile of a given Poisson’s distribution.   
In probability, quantiles are marked points that divide the graph of a probability distribution into intervals (continuous ) which have equal probabilities.

**Syntax:** qpois(k, λ, lower.tail, log)  
  
**where,** 

***k:****number of successful events happened in an interval***λ *:*** *mean per interval*

***lower.tail:****If TRUE then left tail is considered otherwise if the FALSE right             tail  is considered****log:****If TRUE then the function returns probability in form of log*

* **Code:**



* **Output:**

[1] 0 0 0 1

[1] 1 2 3 4

**Normal Distribution:**

* **Aim: :** How we can use normal distribution in R.
* **Theory:** Normal Distribution

In probability theory and statistics, the **Normal Distribution**, also called the **Gaussian Distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other [probability distributions](https://byjus.com/maths/probability-distribution/), therefore supporting the usage of the word ‘normal ‘as in about the one, mostly used.

The Normal Distribution is defined by the [probability density function](https://byjus.com/maths/probability-density-function/) for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

f(x) ≥ 0 ∀ x ϵ (−∞,+∞)

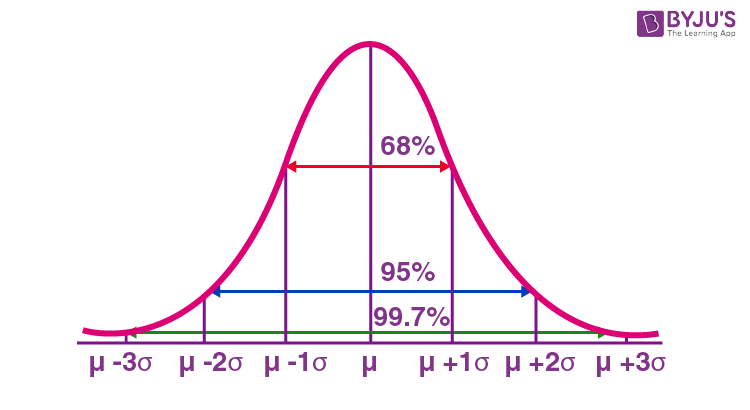
And -∞∫+∞ f(x) = 1

Normal Distribution Standard Deviation

Generally, the normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out. If the standard deviation is smaller, the data are somewhat close to each other and the graph becomes narrower. If the standard deviation is larger, the data are dispersed more, and the graph becomes wider. The standard deviations are used to subdivide the area under the normal curve. Each subdivided section defines the percentage of data, which falls into the specific region of a graph.

Using 1 standard deviation, the Empirical Rule states that,

* Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean- one Standard Deviation and Mean + one standard deviation)
* Approximately 95% of the data falls within two standard deviations of the mean. (i.e., Between Mean- two Standard Deviation and Mean + two standard deviations)
* Approximately 99.7% of the data fall within three standard deviations of the mean. (i.e., Between Mean- three Standard Deviation and Mean + three standard deviations)



Thus, the empirical rule is also called the 68 – 95 – 99.7 rule.

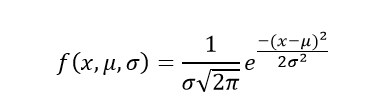
Every normal distribution can be converted to the standard normal distribution by turning the individual values into *z*-scores.

*Z*-scores tell you how many standard deviations away from the mean each value lies.

We convert normal distributions into the standard normal distribution for several reasons:

* To find the probability of observations in a distribution falling above or below a given value.
* To find the probability that a sample mean significantly differs from a known population mean.
* To compare scores on different distributions with different means and standard deviations.
* **Formula:**

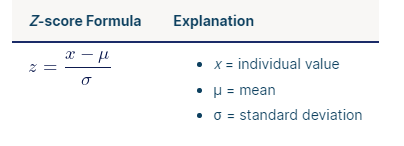
The probability density function of normal or gaussian distribution is given by;



Where,

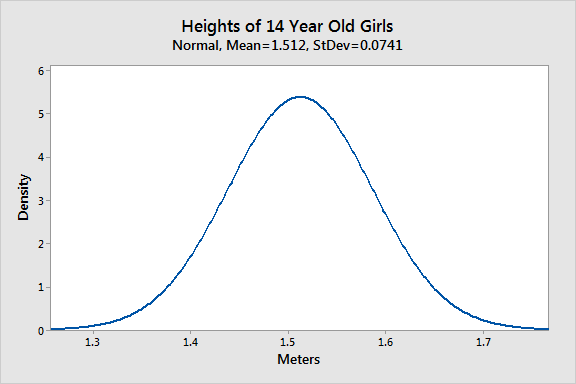
* x is the variable
* μ is the mean
* σ is the standard deviation

to find the z-score of a value.



* **Properties:**
* The [mean](https://www.scribbr.com/statistics/mean/), [median](https://www.scribbr.com/statistics/median/) and [mode](https://www.scribbr.com/statistics/mode/) are exactly the same.
* The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
* The distribution can be described by two values: the mean and the [standard deviation](https://www.scribbr.com/statistics/standard-deviation/).
* **Example:**

Height data are normally distributed. The distribution in this example fits real data that I collected from 14-year-old girls during a study. The graph below displays the probability distribution function for this normal distribution.



As you can see, the distribution of heights follows the typical bell curve pattern for all normal distributions. Most girls are close to the average (1.512 meters). Small differences between an individual’s height and the mean occur more frequently than substantial deviations from the mean. The standard deviation is 0.0741m, which indicates the typical distance that individual girls tend to fall from mean height.

The distribution is symmetric. The number of girls shorter than average equals the number of girls taller than average. In both tails of the distribution, extremely short girls occur as infrequently as extremely tall girls.

* **Normal Distribution in R**

R has four in-built functions to generate binomial distribution. They are described below:

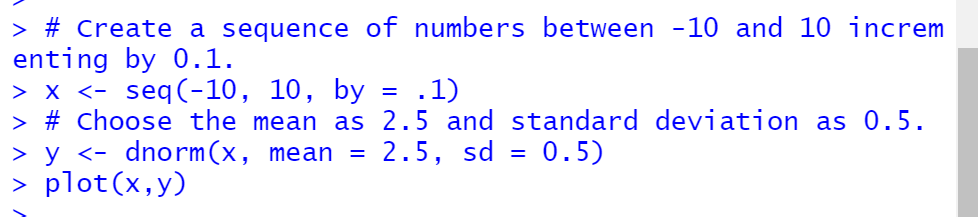
1. dnorm(x, mean, sd)
2. pnorm(x, mean, sd)
3. qnorm(p, mean, sd)
4. rnorm(n, mean, sd)

* Following is the description of the parameters used :
* **x** is a vector of numbers.
* **p** is a vector of probabilities.
* **n** is number of observations.
* **mean** is the mean value of the sample data. It's default value is zero.
* **sd** is the standard deviation. It's default value is 1.

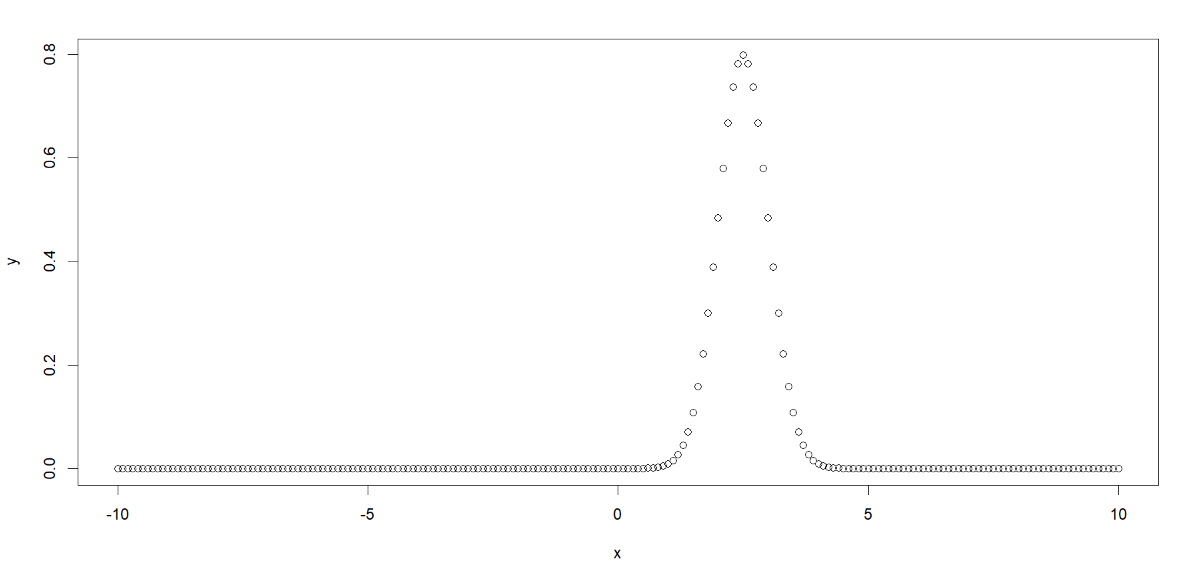
* **dnorm():**

This function gives height of the probability distribution at each point for a given mean and standard deviation.

* **Code:**

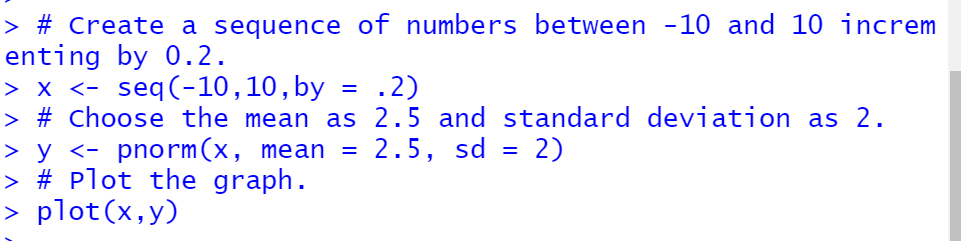


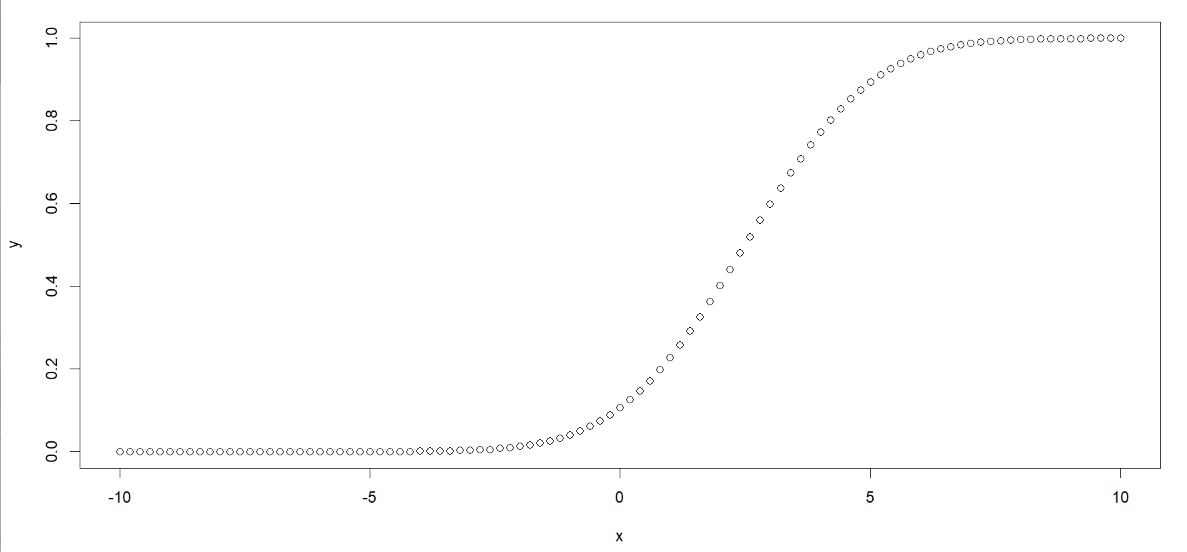
* **Output:**

****

* **pnorm():**

This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".

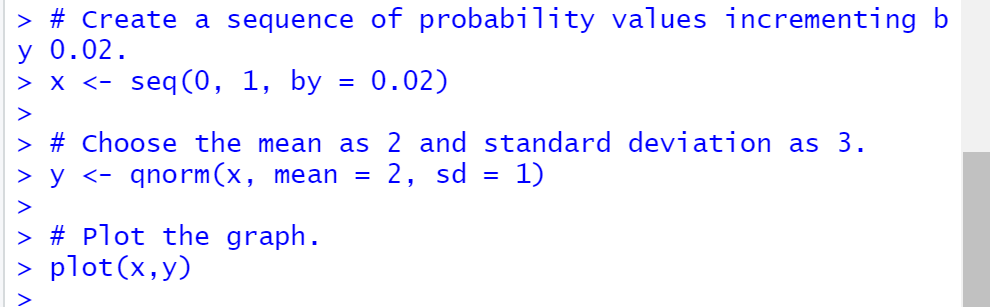
* **Code:**
* ****
* **Output:**



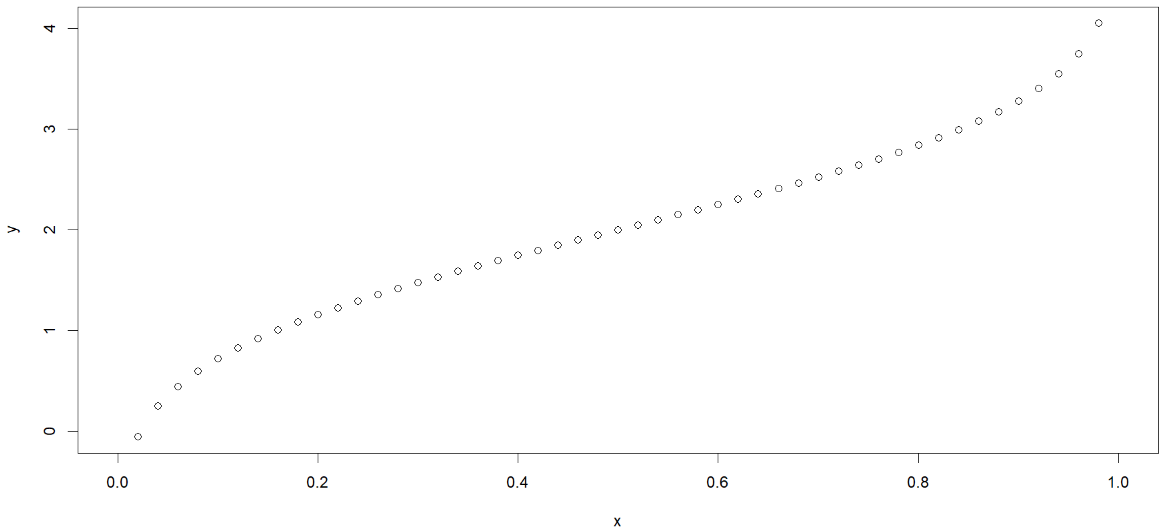
* **qnorm():**

This function takes the probability value and gives a number whose cumulative value matches the probability value.

* **Code:**



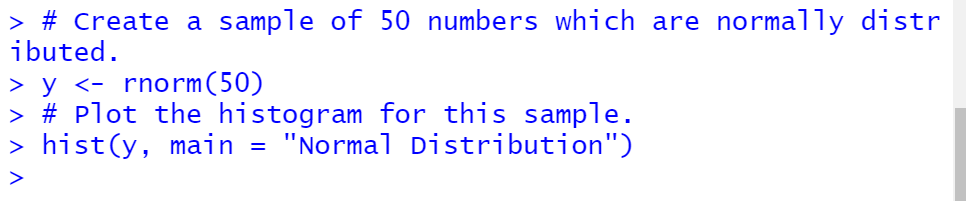
* **Output:**

****

* **rnorm():**

This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.

* **Code:**



* **Output:**

