Random Generation in R

Alireza Sheikh-Zadeh, Ph.D.

Random number generation in R based on different types of probability distributions

We review three types of distributions:

1- Empirical Discrete Distributions

2- Parametric Discrete Distributions

3- Parametric Continuous Distributions

Empirical Discrete Distributions

Example: The random variable x is discrete (possible values are 1, 2, 3 and 4) and the chance of each outcome is as follows:

x	$P(X = x)$
1	0.7
2	0.2
3	0.08
4	0.02
total	1.0

```
# Let's create 1000 random numbers based on the empirical distribution (the t
able above).

nsim = 1000

xRange <- c(1,2,3,4)
p <- c(0.7, 0.2, 0.08, 0.02)

x <- sample(xRange, nsim, p, replace= TRUE)
head(x)

## [1] 1 2 2 3 1 1

table(x) # the table of count or frequency
```

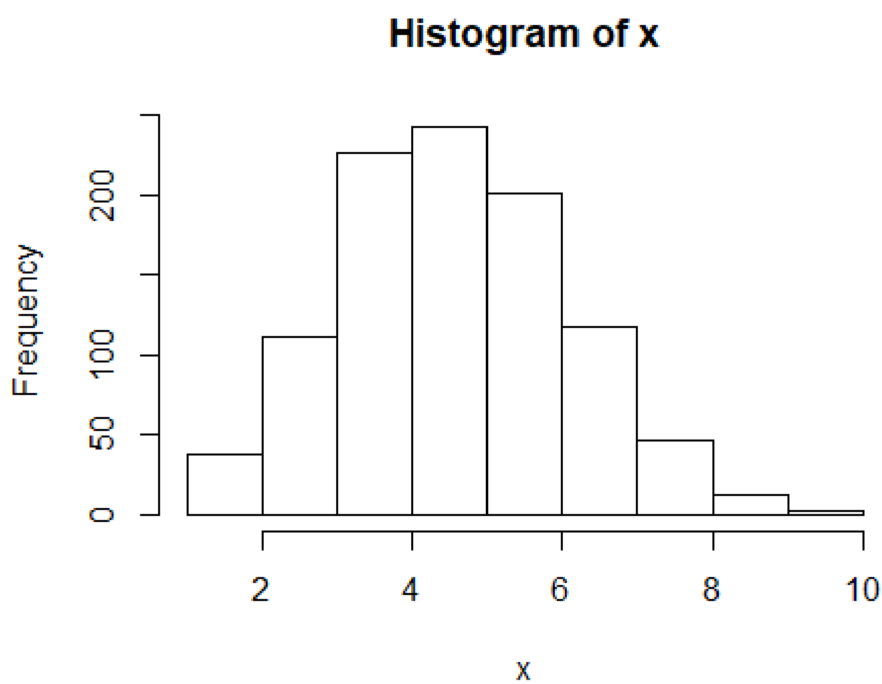
```
## x
##   1   2   3   4
## 701 198  84  17
```

Parametric Discrete Distributions

Binomial distribution

Here, we create pseudo-random numbers for binomial distribution as a known parametric discrete distribution. In probability theory and statistics, the **binomial distribution** with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent experiments.

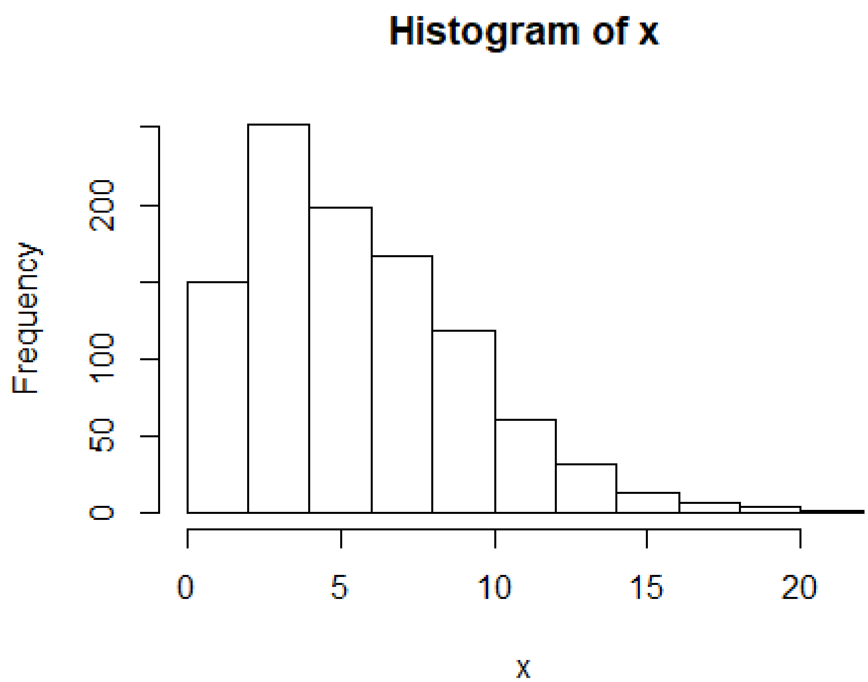
```
nsim = 1000

# Binomial with p = 0.5, n = 10
# p is the probability of success and n is the number of trials
x = rbinom(nsim, 10, 0.5) # x is the number of successes in 10 trials
hist(x)
```



Negative Binomial distribution

The **negative binomial distribution** is a discrete probability distribution that models the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted *r*) occurs.

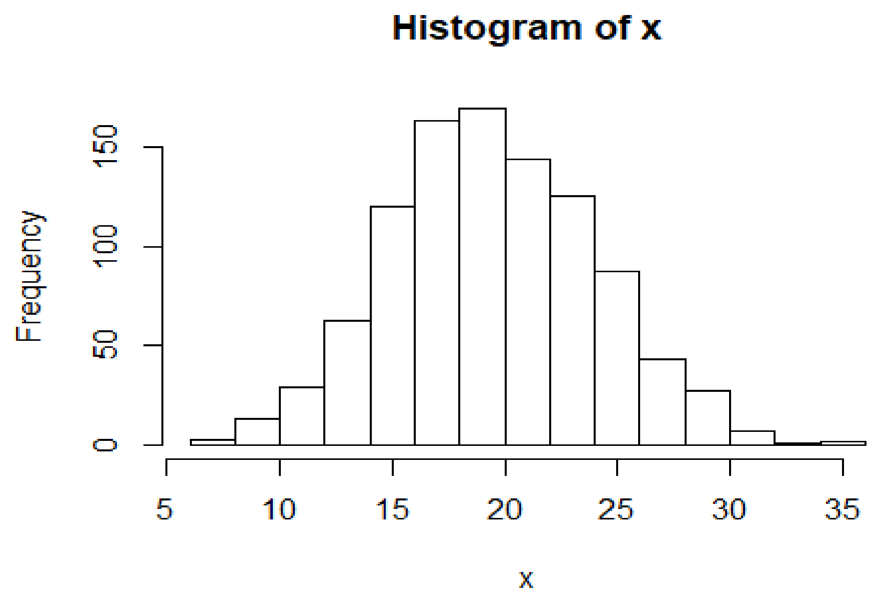
```
# Negative Binomial with p = 0.5, r = 6
# (p is the probability of success and r is the target number of successes)
x = rnbinom(nsim, size = 6, p = 0.5)  # x is the number of failures untill ac
heiving the rth success.
hist(x)
```



Poisson distribution

the **Poisson distribution** is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

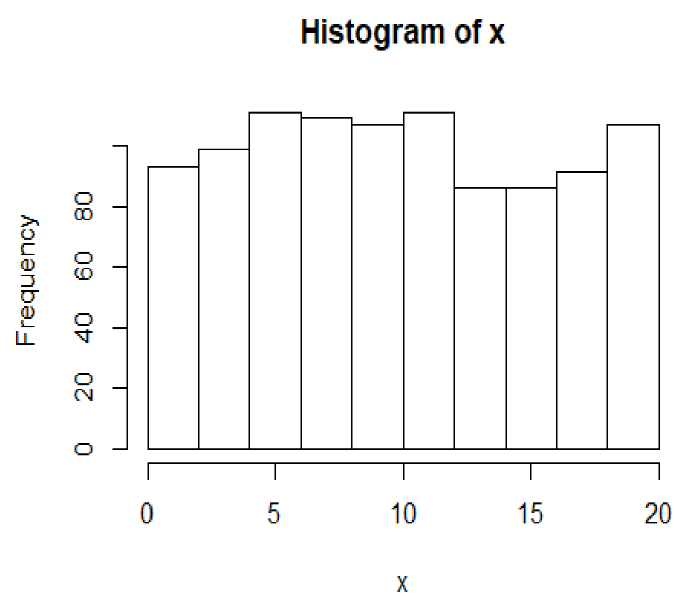
```
# Poisson distribution with lambda = 20 (For example, the arrival rate to a r
estaurant is 20 per hours)
x = rpois(nsim, lambda = 20)
hist(x)
```



Discrete Uniform distribution

The **discrete uniform distribution** is a symmetric probability distribution wherein a finite number of values are equally likely to be observed; every one of n values has equal probability $1/n$.

```
# Discrete Uniform distribution with min = 1, and max = 20.
x = round(runif(nsim, min=0.5, max=20.5),0)
hist(x)
```

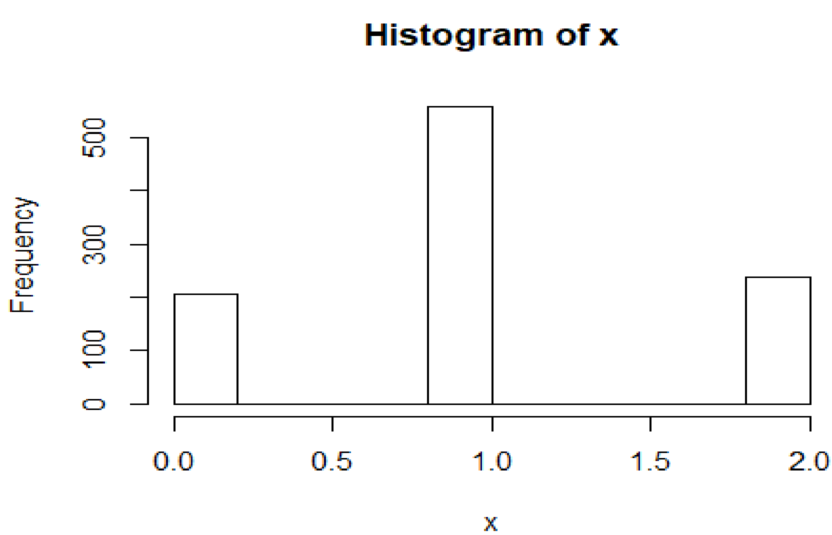


```
table(x)

## x
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
## 45 48 52 47 53 58 49 60 53 54 55 56 48 38 38 48 43 48 56 51

# OR you simulate discrete uniform data by sampling with replacement without
weighting (compare it with the empirical discrete distribution).
x = sample(1:20, nsim, replace = T)

table(x)

## x
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
## 61 40 52 51 43 48 46 51 55 66 54 45 58 55 39 51 48 36 54 47
```

Hypergeometric distribution

The **hypergeometric distribution** is a discrete probability distribution that describes the probability of *x* successes (random draws for which the object is drawn has a specified feature) in *k* draws, *without* replacement, from a finite population of size *n+m* that contains exactly *m* objects with that feature, wherein each draw is either a success or a failure.

```
# Check out the Wikipedia for more information (https://en.wikipedia.org/wiki
/Hypergeometric_distribution) with m = 2, n = 8, k = 5
# m: the number of white balls in the urn.
# n: the number of black balls in the urn.
# k: the number of balls drawn from the urn.
x = rhyper(nsim, m = 2, n = 8, k = 5) # x is the number of white balls

between those 5 balls drawn from the urn (sampling is without replacement)
hist(x)
```

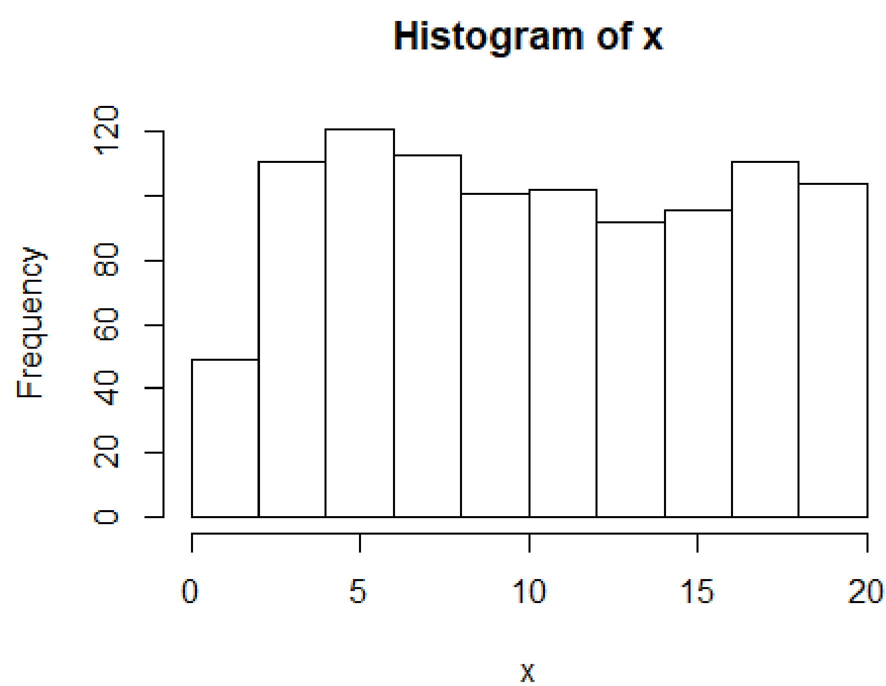


Parametric Continuous Distributions

Continuous uniform distribution

The **continuous uniform distribution** or **rectangular distribution** is a family of symmetric probability distributions. The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.

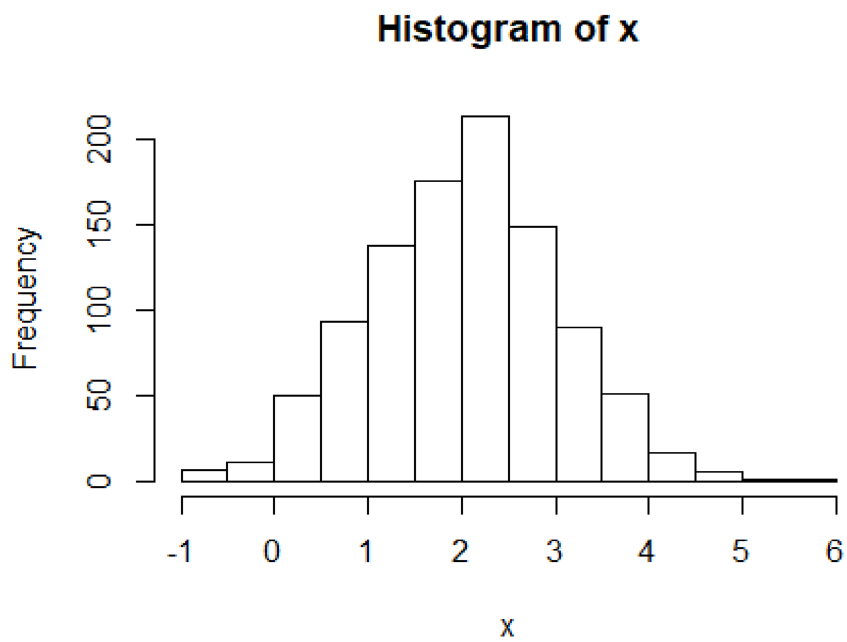
```
nsim = 1000

# Uniform distribution with min = 1, and max = 20.
x = runif(nsim, min=1, max=20)
hist(x)
```



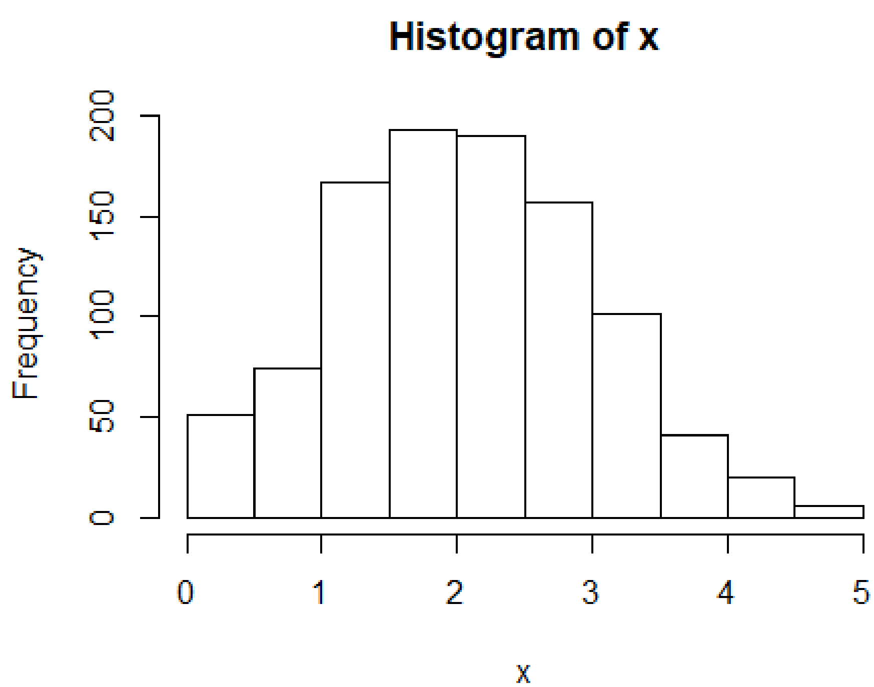
Normal distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

```
# Normal Distribution with mean = 2, and sd = 1
x = rnorm(nsim, mean = 2, sd = 1)
hist(x)
```



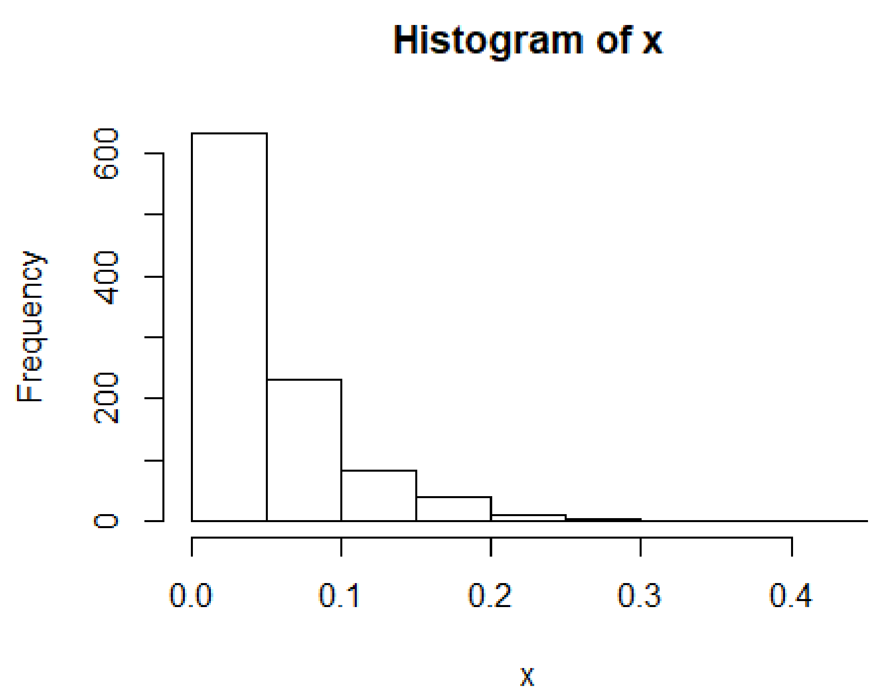

```
# Truncated normal Distribution with a = 0, b = 5, mean = 2, and sd = 1
# Goal is to have a lower and upper bound for the produced numbers.
library(truncnorm) # You may need to install "truncnorm"
x = rtruncnorm(nsim, a = 0, b = 5, mean = 2, sd = 1)
hist(x)
```



Exponential distribution

The **exponential distribution** is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.

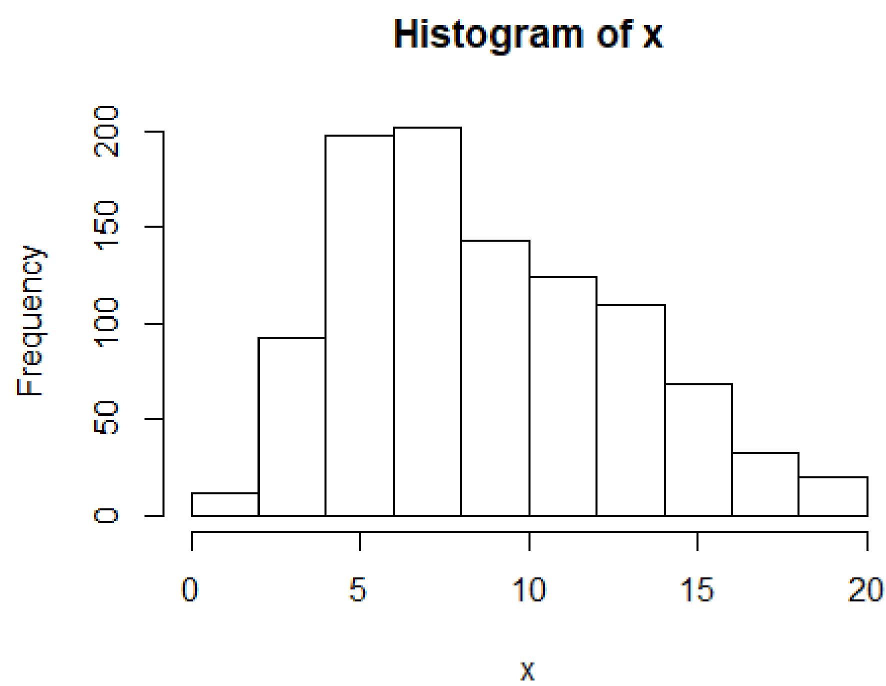
```
# Exponential Distribution with lambda = 20 (number of arrival per unit of time)
x = rexp(nsim, rate = 20) # x is the time between arrivals.
hist(x)
```



Triangular distribution

The **triangular distribution** is a continuous probability distribution with a lower limit *a*, upper limit *b,* and mode *c*, where $a < b$ and $a \le c \le b$.

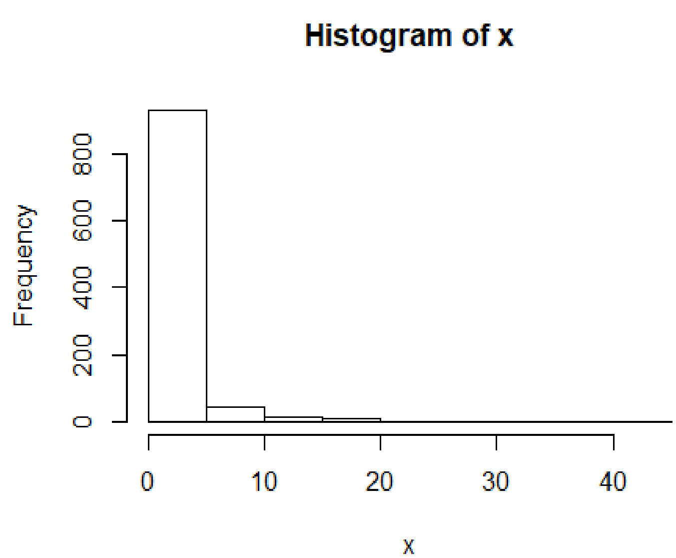
```
# Triangular Distribution with a = 1, b = 20, c = 5
# a: minimum value
# b: maximum value
# c: most likely value
library(triangle)

## Warning: package 'triangle' was built under R version 3.5.3

x = rtriangle(nsim, a=1, b=20, c=5)
hist(x)
```



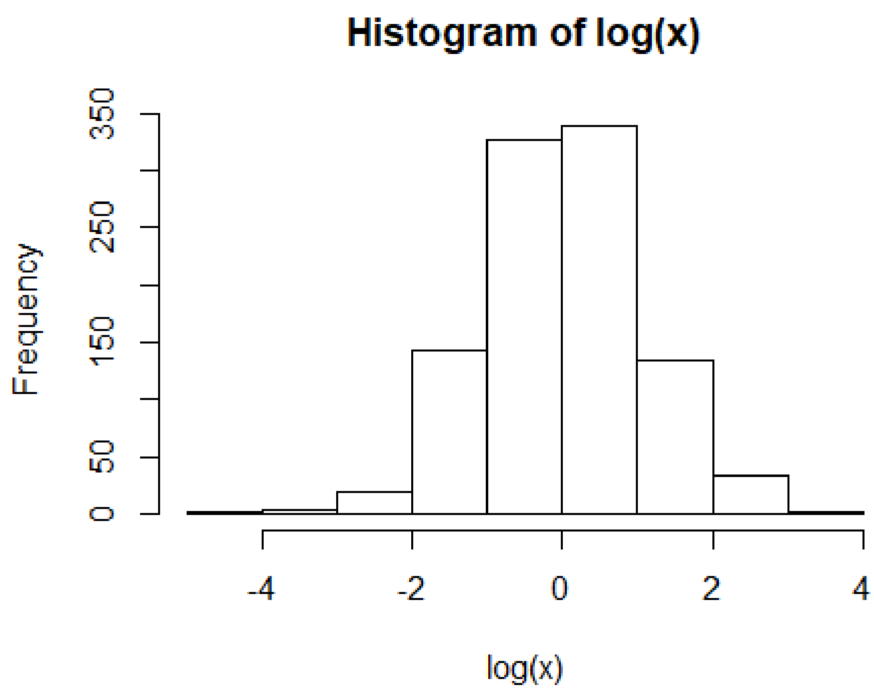
Log-normal distribution

A **log-normal (or lognormal) distribution** is a continuous probability distribution of a random variable whose logarithm is normally distributed.

```
# Log-Normal Distribution with meanlog = 0, sdlog = 1
x = rlnorm(nsim, meanlog = 0, sdlog = 1)
hist(x)
```



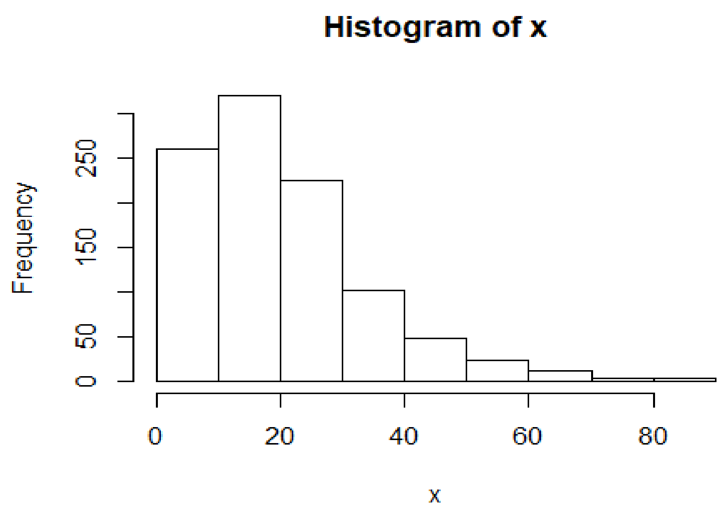
```r
mean(log(x))

## [1] 0.01390393

sd(log(x))

## [1] 1.06073

# log of x is normally distributed.
hist(log(x))
```



Gamma distribution

The **gamma distribution** is a two-parameter family of continuous probability distributions. The exponential distribution, normal, and chi-square distribution are special cases of the gamma distribution.

```r
# Gamma distribution with parameters shape = 2 and scale = 10.
x = rgamma(nsim, shape = 2, scale = 10)
hist(x)
```




```
# When shape = 1, gamma is exactly similar to exponential distribution with r
ate = 1/scale

dgamma(2, shape = 1, scale = 10)

## [1] 0.08187308

dexp(2, rate = 1/10)

## [1] 0.08187308

# for larger shape values, gamma merges to normal distribution.
y = rgamma(nsim, shape = 20, scale = 10)
hist(y)
```