Distributions

October 8, 2017

1 Central Limit Theorem

The central limit theorem states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean μ and a variance $\frac{\sigma^2}{N}$ where N is the sample size for each mean and not the number of samples. The fascinating thing about the central limit theorem is that no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution. Two important properties of N are:

```
1) the more it increases the distribution looks more like a normal distribution
2) the variance decreases.
```

2 Confidence Interval

A confidence interval (CI) is a type of interval estimate (of a population parameter) that is computed from the observed data.

Let assume that we toss a coin 100 times and the result is 42 heads and 58 tails. We can say that $X_i \sim B(p)$

We can ask:

```
1) What is the best estimate of p
2) How confident are we of that estimate?.
```

Let's apply the Central Limit Theorem. By CLT we know that $\sum_{i=1}^{100} X_i \dot{\sim} N(100p, 100p(p-1)$

So , 95% of the time we expect p to be between $100p - 1.96\sqrt{100p(1-p)}$ and $100p + 1.96\sqrt{100p(1-p)}$

In our case we have observed 42 heads, so we estimate p as $\hat{p} = \frac{42}{100} = 0.42$ The CI will be $42 + -1.96\sqrt{42 \times 0.58} = (32.33, 51.67)$ So we can say that we're 95% confident that p is in the interval .3233 to .5167. We're 95% confident that the true probability of getting a head is in this interval. So if we want to ask a question do we think this is a fair coin. That it is reasonable that this is a fair coin because one half is in this interval.

What does it mean when I say we're 95% confident?

Under the frequentist paradigm, what this means is we have to think back to our infinite hypothetical sequence of events. So if we were to repeat this trial an infinite number of times, or an arbitrary large number of times. Each time we create a confidence interval in this way based on the data we observe. Than on average 95% of the intervals we make will contain the true value of p.

3 Cumulative distribution function

The cumulative distribution function (CDF) exists for every distribution. We define it as $F(x) = P(X \leq x)$ for random variable X. If X is discrete-valued, then the CDF is computed with summation $F(x) = \sum_{t=-\inf}^{x} f(t)$ where $f(t) = P(X = x)$ is the probability mass function (PMF) that we have already seen. If X is continuous, the CDF is computed with an integral $F(x) = \int_{t=-\inf}^{x} f(t)dt$ where f(t) is the probability density function (PDF). The examples below use binomial (discrete) distribution and an exponential (continuous) distribution. **Example**: Suppose

$$Y \sim Binomial(5, 0.6)$$

. Then

$$F(1) = P(X \leq 1) = \sum_{-\inf}^{1} f(t) = \sum_{-\inf}^{-1} 0 + \sum_{0}^{1} \binom{5}{t}^{t} (1-0.6)^{5-t}$$

$$= \binom{5}{0}^{0} (1-0.6)^{5-0} + \binom{5}{1}^{1} (1-0.6)^{5-1} = 0.4^5 + 5(0.6)(0.4)^4 \approx 0.087$$

Example Suppose $Y \sim Exp(1)$. Then $F(2) = P(Y \leq 2) = \int_{-\inf} 2e^{-t} I_{\{t \geq 0\}} dt = \int_0 2e^{-t} dt = -e^{-t}\Big|_{t=0}^{t=2} = -(e^{-2} - e^0) = 1 - e^{-2} \approx 0.865$

The CDF is convenient for calculating probabilities of intervals. Let a and b be any real numbers with $a < b$ Then the probability that X falls between a and b is equal to $P(a < Xb) = P(Xb)P(Xa) = F(b)F(a)$.

4 Quantile function

The CDF takes a value for a random variable and returns a probability. Suppose instead that we start with a number between 0 and 1, call it p, and we wish to find the value x so that P(Xx) =p.The value x which satisfies this equation is called the p quantile (or 100 p percentile) of the distribution of X.

Example: In a standardized test, the 97th percentile of scores among all test-takers is 23. Then 23 is the score you must achieve on the test in order to score higher than 97% of all test-takers. We could equivalently call q = 23 the .97 quantile of the distribution of test scores.

Example: The middle 50% of probability mass for a continuous random variable is found between the .25 and .75 quantiles of its distribution. If $Z \sim N(0, 1)$, then the .25 quantile is 0.674 and the .75 quantile is 0.674. Therefore, $P(0.674 < Z < 0.674) = 0.5$.

5 Distributions

5.1 Discrete Distributions

A discrete random variable X is described by a probability mass functions (PMF), which we will also call "distributions," f(x)=P(X =x). The set of x-values for which f (x) > 0 is called the support. Note, if the distribution depends on unknown parameter(s) we can write it as f (x;) (preferred by frequentists) or f(x|) (preferred by Bayesians). The difference between the frequentist and Bayesian paradigm will be covered later.

5.1.1 Bernoulli Distribution

It's used when we have two possible outcomes, or the cases where we have a success or a failure.

Well, to denote this, let's say a random variable x follows a Bernoulli distribution with probability p, where p is probability of success $X \sim B(p)$. The failure or tales, x = 0 has probability $1 - p$.

We can write this as a function for all the different possible outcomes.

And say what's the probability that the random variable x takes a value of little x given a specific value of p?

$f(X = x \mid p) = f(x \mid p) = p^x(1-p)^{1-x} I_{x \in 0,1}(x)$. This is referred to as the probability mass function. It gives the probability of different outcomes of the random variable.

The indicator function takes precedence in the order of operations so we always evaluate it first, this is a way we can avoid doing things such as taking the log or the square root of a negative number.

The expected value of x as we sum over all possible outcomes. Little x, we sum up x times the probability random variable takes up variable x.

$E[X] = \sum_x P(X = x) = (1)p + (0)(1-p) = p$

One possible outcome is one, it takes that with the probability of p. Another possible outcome is 0, it takes that with the probability 1- p. So the expected value for Bernoulli is just the probability p.

Similarly we can talk about the variance which is the square root of the standard deviation.

$Var(X) = p(1-p)$

```
In [42]: #install.packages('mc2d')
         library(mc2d)# load the module that contains the 'dbern' function.
         # find the propability of X = 0 when p = 0.7
         dbern(0, 0.7) # dbern(x, p) -> 0.3
```

0.3

5.1.2 Binomial Distribution

The generalization of the Bernoulli when we have N repeated trials is a binomial. It fits to repeated trials each with a dichotomous outcome such as success-failure, healthy-disease, heads-tails.

Binomial is just the sum of the N independent Bernoullis. We can say X follows a binomial distribution with parameters n and p ($X \sim Bin(n, p)$). In this case, the probability function, probability that X takes some value little x is given p.

$$f(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

for

$$x \in \{1, ..., n\}$$

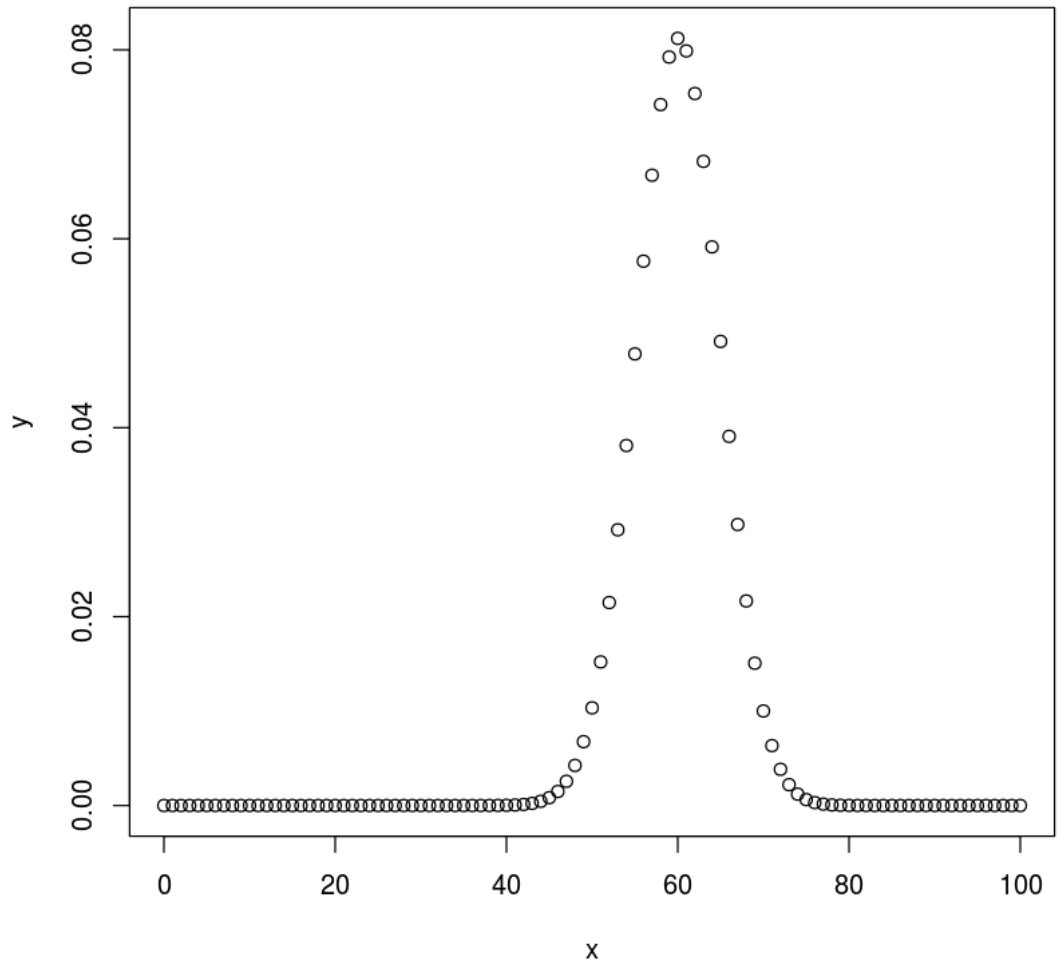
The expected value for binomial is

$$E[X] = np$$

The variance for binomial is

$$Var[X] = np(1-p)$$

```
In [43]: # Let's take a look at it
         x <- seq(0,100,by=1)
         y <- dbinom(x,100,0.6)
         plot(x, y)
```



Example
Let that that a coin $X \sim Bin(100, 0.43)$. Compute the following:

```
In [44]: n <- 100
         p <- 0.43
         # P(X  = 48)
         dbinom(48, n, p) # 4.8%

         # P(X >= 45)
```

```r
sum(dbinom(45:n, n, p)) # 37.9%

# P(X < 40)
1 - sum(dbinom(40:n, n, p)) # 24%

# P(40 <= X <= 50)
sum(dbinom(40:60, n, p)) # 75.9%

#Find the quantiles 0.025
qbinom(0.025, n, p) # 33

#Find the quantiles 0.975
qbinom(0.975, n, p) # 53
```

```
0.0480202964916629
0.379403243670669
0.24062828589424
0.759151640374164
33
53
```

5.1.3 Geometric Distribution

The geometric distribution is the number of trials needed to get the first success, i.e., the number of Bernoulli events until a success is observed, such as the first head when flipping a coin. It takes values on the positive integers starting with one (since at least one trial is needed to observe a success).

$$X \sim Geo(p)$$

$$P(X = x \mid p) = p(1-p)^{x-1}$$

for

$$x = 1, 2...$$

$$E[X] = \frac{1}{p}$$

If the probability of getting a success is p, then the expected number of trials until the first success is $\frac{1}{p}$.

Example: What is the probability that we flip a fair coin four times and don't see any heads? This is the same as asking what is $P(X > 4)$ where $X \sim Geo(1/2)$.

$$P(X > 4) = 1P(X = 1)P(X = 2)P(X = 3)P(X = 4)$$

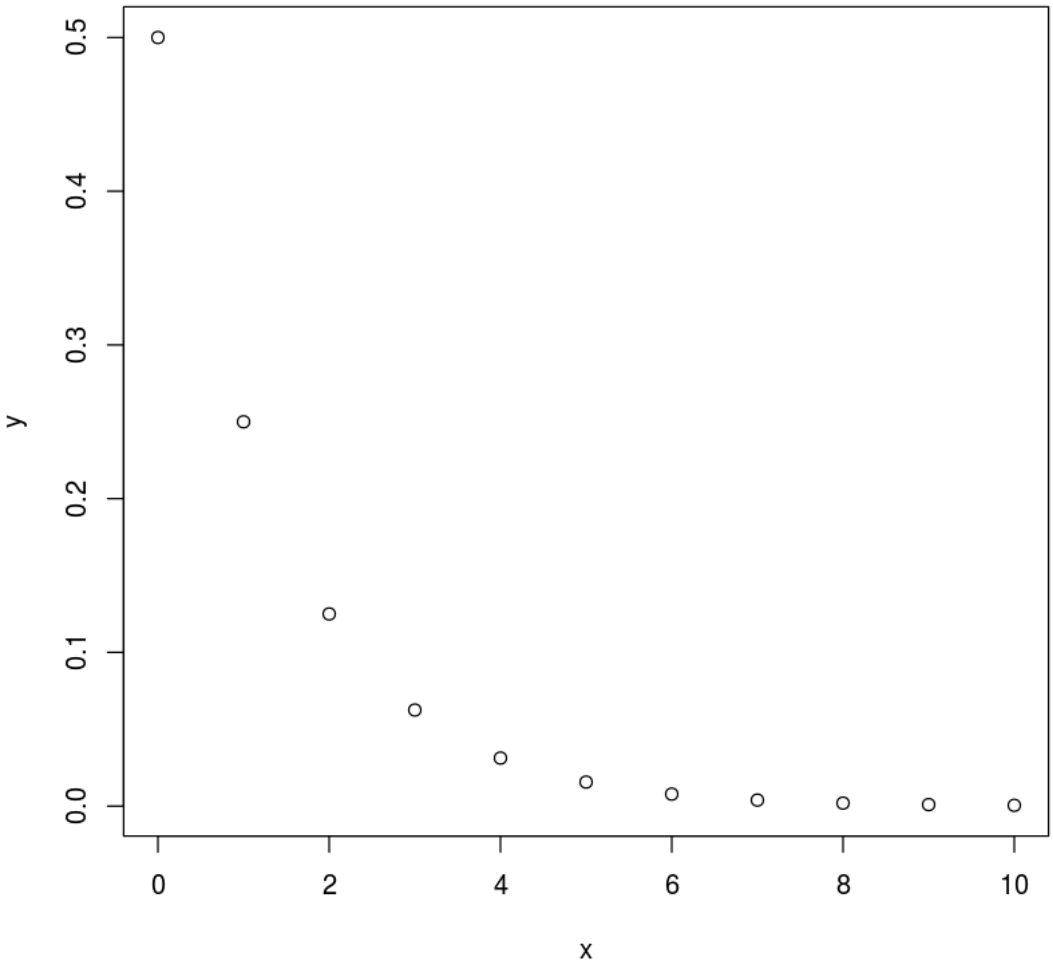
$$= 1(1/2)(1/2)(1/2)(1/2)(1/2)^2(1/2)(1/2)^3$$

$$= 1/16$$

. Of course, we could also have just computed it directly, but here we see an example of using the geometric distribution and we can also see that we got the right answer.

In [45]:
```
p <- 0.5

# P(X  > 4)
1 - sum(dgeom(0:3, p))

#Find the quantiles 0.975
qgeom(0.975, p) # 5 tosses

# Let's take a look at it
x <- seq(0,10,by=1)
y <- dgeom(x, 0.5)
plot(x, y)
```

0.0625
5



5.1.4 Multinomial Distribution

Another generalization of the Bernoulli and the binomial is the multinomial distribution, which is like a binomial when there are more than two possible outcomes. Suppose we have n trials and there are k different possible outcomes which occur with probabilities p_1 , , p_k . For example, we are rolling a six-sided die that might be loaded so that the sides are not equally likely, then n is the total number of rolls, $k = 6$, $p1$ is the probability of rolling a one, and we denote by x_1 , . . . , x_6 a possible outcome for the number of times we observe rolls of each of one through six, where $\sum_{i=1}^{6} x_i = n$ and $\sum_{i=1}^{6} p_i = 1$.

$$p(x_1, ..., x_n \mid p_1, ..., p_n) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

The expected number of observations in category i is

$$E[X_i] = np_i$$

$$Var[X_i] = np_i(1 - p_i)$$

Example

Let's assume that we have observed a player of paper-rock-scissor and have calculated the followng probabilities for each move : $P_{paper} = 0.2$, $P_{rock} = 0.3$, $P_{scissor} = 0.5$.

```
In [46]: p1<- 0.2; p2<- 0.3; p3<- 0.5;

         # Find the probability of the other player playing scissors
         dmultinom(c(0,0,1), prob=c(p1,p2,p3))

         # Find the probability for the other player playing 5 of each
         dmultinom(c(5,5,5), prob=c(p1,p2,p3))

         # Find the probability for the other player playing 9 times scissor
         dmultinom(c(0,0,9), prob=c(p1,p2,p3))
```

0.5
0.0183891708000001
0.001953125

5.1.5 Poisson Distribution

The Poisson distribution is used for counts, and arises in a variety of situations. The parameter $\lambda > 0$ is the rate at which we expect to observe the thing we are counting.

$$X \sim Pois(\lambda)$$

$$P(X = x \mid \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

for for x = 0, 1, 2, . . .

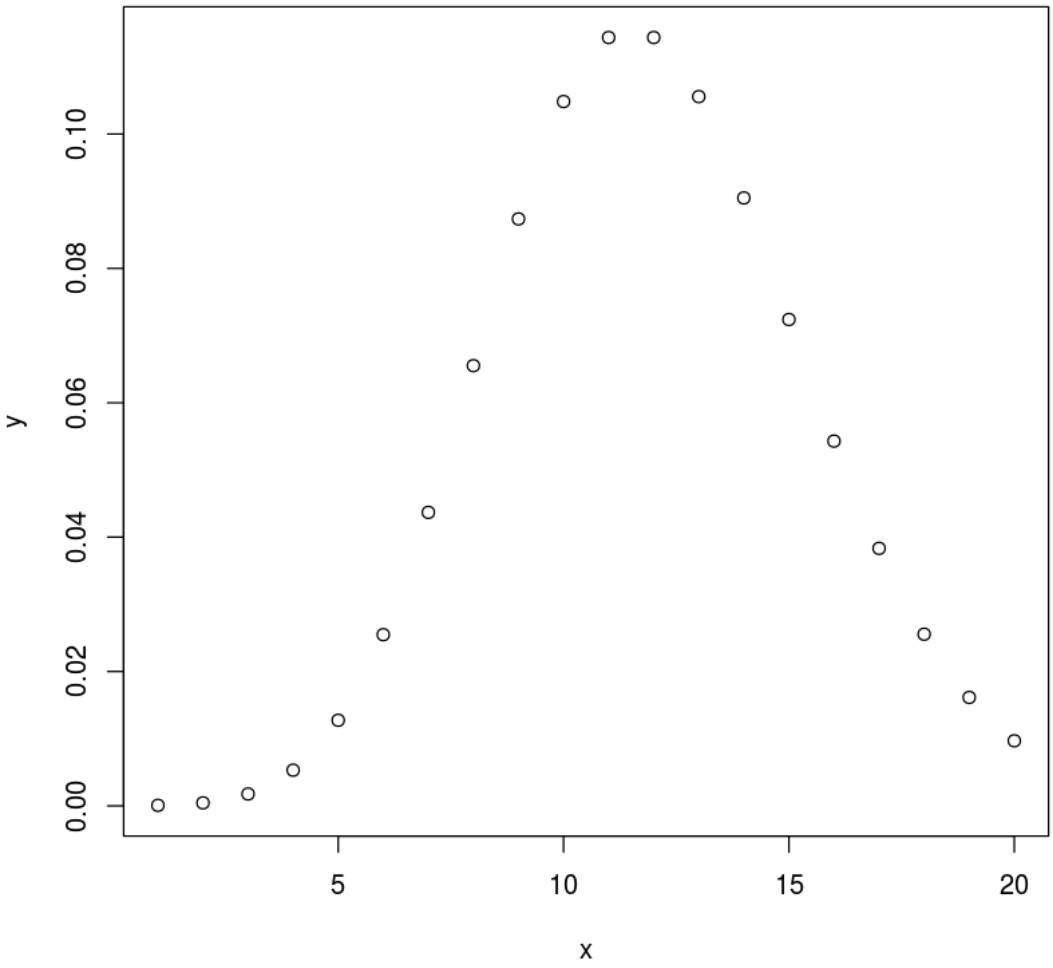
$$E[X] = \lambda$$

$$Var[X] = \lambda$$

** Example** If there are twelve cars crossing a bridge per minute on average. Compute the following:

```
In [47]: lambda <- 12
         # Find the probability of having sixteen or less cars crossing the bridge in a particul
         ppois(16, lambda)

         # Find the probability more than 17 cars crossing the bridge in a minute
         1 - ppois(16, lambda)

         #OR

         ppois(16, lambda, lower = FALSE) # upper tail

         #Let's visualise the following distribution
         x <- 1:20
         y <- dpois(x, lambda)
         plot(x, y)
```

0.898708992560162
0.101291007439838
0.101291007439838



5.2 Continuous Distributions

We can define a continuous random variable based on its probability density function or PDF for short. The PDF is sort of proportional to the probability that the random variable will take a particular value. In differential calculus sense because it can take an infinite number of possible values. The key idea is that if you integrate the PDF over an interval, it gives you the probability that the random variable would be in that interval.

Let :

$$X \sim U[0,1]$$

$$f(x) = \begin{cases} 1, & \text{if } 0 \leq x \leq 1. \\ 0, & \text{otherwise.} \end{cases} = I_{\{0 \leq x \leq 1\}} \tag{1}$$

What's the probability that X will be between 0 and one-half?

$$P(0 \leq x \leq \tfrac{1}{2}) = \int_0^{\frac{1}{2}} f(x)dx = \int_0^{\frac{1}{2}} dx = \frac{1}{2}$$

What's the probability that X will be one-half?

$$P(x = \frac{1}{2}) = 0$$

we just going to get a 0 because there are an infinite number of possible outcomes. The probability of taking any particular x is going to be 0

5.2.1 Exponential Distribution

The exponential distribution is often used to model the waiting time between random events. Indeed, if the waiting times between successive events are independent from an $Exp(\lambda)$ distribution, then for any fixed time window of length t, the number of events occurring in that window will follow a Poisson distribution with mean $t\lambda$.

$$X \sim Exp(\lambda)$$
$$f(x \mid \lambda) = \lambda e \lambda x I_{x \geq 0}(x)$$
$$E[X] = \frac{1}{\lambda}$$
$$Var[X] = \frac{1}{\lambda^2}$$

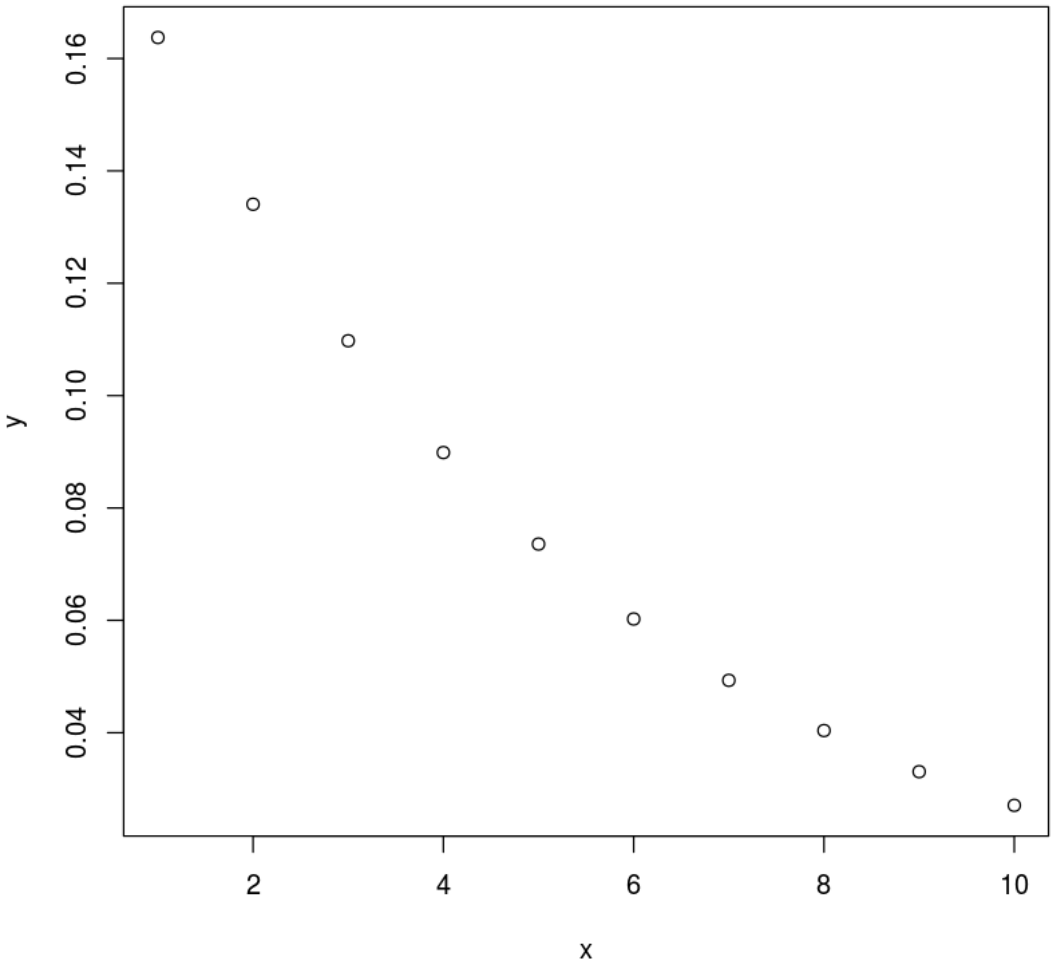
Similar to the Poisson distribution, the parameter is interpreted as the rate at which the events occur.

Example

The busses in a town have no schedule: they just wander around on their routes. Suppose that the distribution of interarrival times is exponential with $\lambda = 5$.

```
In [48]: lambda <- 5
         # Find the probability for the next bus to come in less than 2 minutes.
         pexp(2, rate=1/lambda)

         #Let's visualise it
         x <- 1:10
         y <- dexp(x, rate = 1/lambda)
         plot(x, y)
```

0.329679953964361



5.2.2 Gamma distribution

If $X_1, X_2, ..., X_n$ are independent (and identically distributed $Exp(\lambda)$) waiting times between P successive events, then the total waiting time for all n events to occur $Y = \sum_{i=1}^{n} X_i$ will follow a gamma distribution with shape parameter $\alpha = n$ and rate parameter $\beta = \lambda$.

$$X \sim Gamma(\alpha, \beta)$$

$$f(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y) I_{\{y \geq 0\}}(y)$$

$$E[Y] = \frac{\alpha}{\beta}$$

$$Var[Y] = \frac{\alpha}{\beta^2}$$

where $\Gamma()$ is the gamma function, a generalization of the factorial function which can accept non-integer arguments. If n is a positive integer, then $\Gamma(n) = (n1)!$. Note also that $\alpha > 0$ and $\beta > 0$.

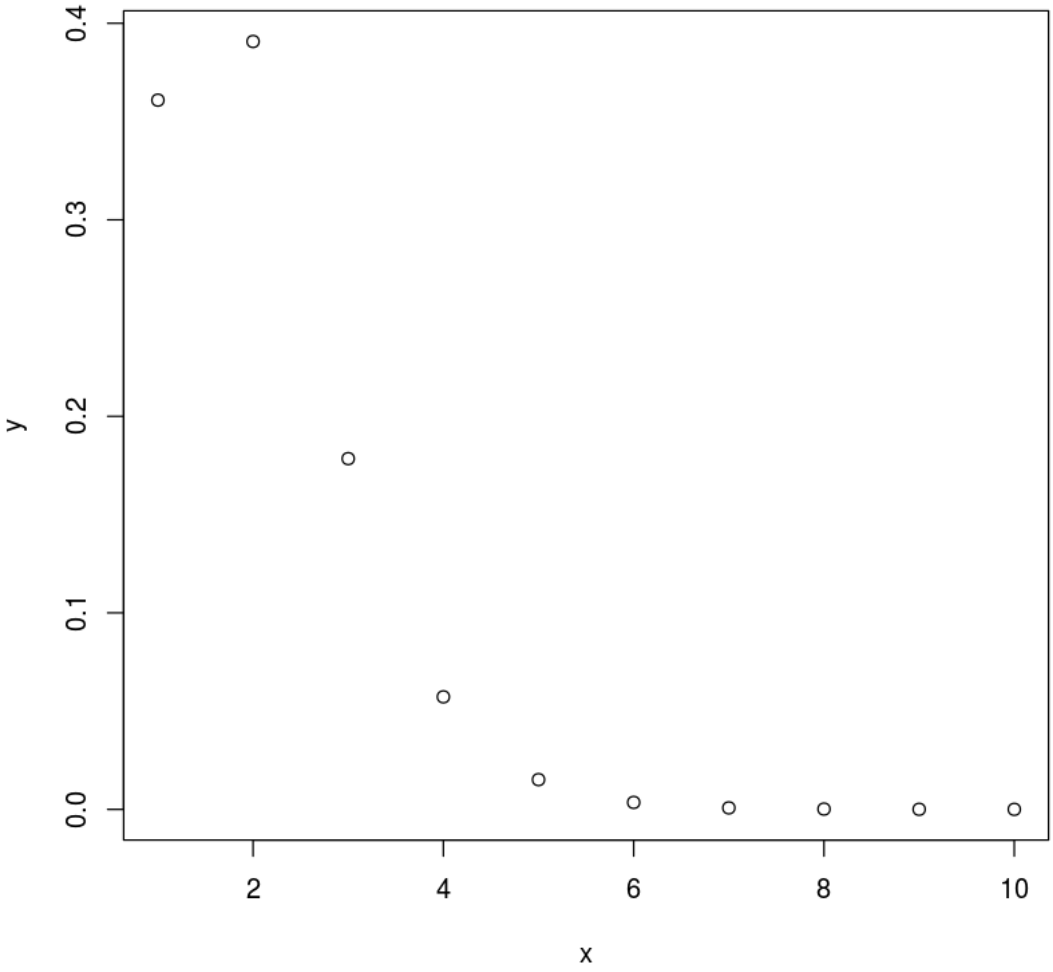
The exponential distribution is a special case of the gamma distribution with = 1. The gamma distribution commonly appears in statistical problems. It is used to model positive-valued, continuous quantities whose distribution is right-skewed. As α increases, the gamma distribution more closely resembles the normal distribution.

Eaxmple

Suppose you are fishing and you expect to get a fish once every 0.5 hour.

```
In [49]: beta <- 2 # 1/0.5

         # Compute the probability of catching 5 fishes in under 2 hours
         alpha <- 5
         pgamma(2, shape=alpha,scale=beta, lower.tail=TRUE) #  nearly impossible

         # Compute the probability that you will have to wait between 2 to 4 hours before you ca
         alpha <- 4
         pgamma(4, shape=alpha,scale=beta, lower.tail=TRUE) - pgamma(2, shape=alpha,scale=beta,l

         #Let's visualise it for alpha = 4 and beta = 2
         x <- 1:10
         y <- dgamma(x, alpha, beta)
         plot(x, y)
```

0.00365984682734371
0.123888382625299



5.2.3 Uniform distribution

The uniform distribution is used for random variables whose possible values are equally likely over an interval. If the interval is (a, b), then the uniform probability density function (PDF) $f(x)$ is flat for all values in that interval and 0 everywhere else.

$$X \sim Uniform(\alpha, \beta)$$

$$f(x \mid \alpha, \beta) = \frac{1}{\alpha - \beta} I_{\alpha \geq x \geq \beta}(x)$$

$$E[X] = \frac{\alpha + \beta}{2}$$

$$Var[X] = \frac{(\beta - \alpha)^2}{12}$$

The standard uniform distribution is obtained when a = 0 and b = 1.

Example Select ten random numbers between one and three.

```
In [50]: runif(10, min=1, max=3)
```

1. 1.13250032672659 2. 1.52847378887236 3. 2.13511152938008 4. 1.89702115161344 5. 1.038373043295 6. 2.72564477985725 7. 1.96969000948593 8. 2.14560577506199 9. 2.16961722914129 10. 1.72979898145422

5.2.4 Beta distribution

The beta distribution is used for random variables which take on values between 0 and 1. For this reason, the beta distribution is commonly used to model probabilities.

$$X \sim Beta(\alpha, \beta)$$

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{\{0<x<1\}}(x)$$

$$E[X] = \frac{\alpha}{\alpha+\beta}$$

$$Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where $\Gamma()$ is the gamma function introduced with the gamma distribution. Note also that > 0 and > 0. The standard Uniform(0, 1) distribution is a special case of the beta distribution with $=$ $= 1$.

5.2.5 Normal Distribution

The normal, or Gaussian distribution is one of the most important distributions in statistics. It arises as the limiting distribution of sums (and averages) of random variables. This is due to the Central Limit Theorem. Because of this property, the normal distribution is often used to model the "errors," or unexplained variation of individual observations in regression models.

The standard normal distribution is given by

$$X \sim Z(0,1)$$

$$f(z) = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{z^2}{2})$$

$$E[X] = 0$$

$$Var[X] = 1$$

Now consider $X = \sigma Z + \mu$ where $\sigma > 0$ and μ is any real constant. Then $E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \sigma 0 + \mu = \mu$ and $Var(X) = Var(\sigma Z + \mu) = \sigma^2 Var(Z) + 0 = \sigma^2 1 = \sigma^2$.

Then, X follows a normal distribution with mean μ and variance σ^2 (standard deviation σ) denoted as:

$$X \sim N(\mu, \sigma^2)$$

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

The normal distribution is symmetric about the mean μ, and is often described as a "bell-shaped" curve. Although X can take on any real value (positive or negative), more than 99% of the probability mass is concentrated within three standard deviations of the mean.

The normal distribution has several desirable properties.

One is that if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Consequently, if we take the average of n independent and identically distributed (iid) normal random variables,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ for $i = 1, 2, ..., n$ then

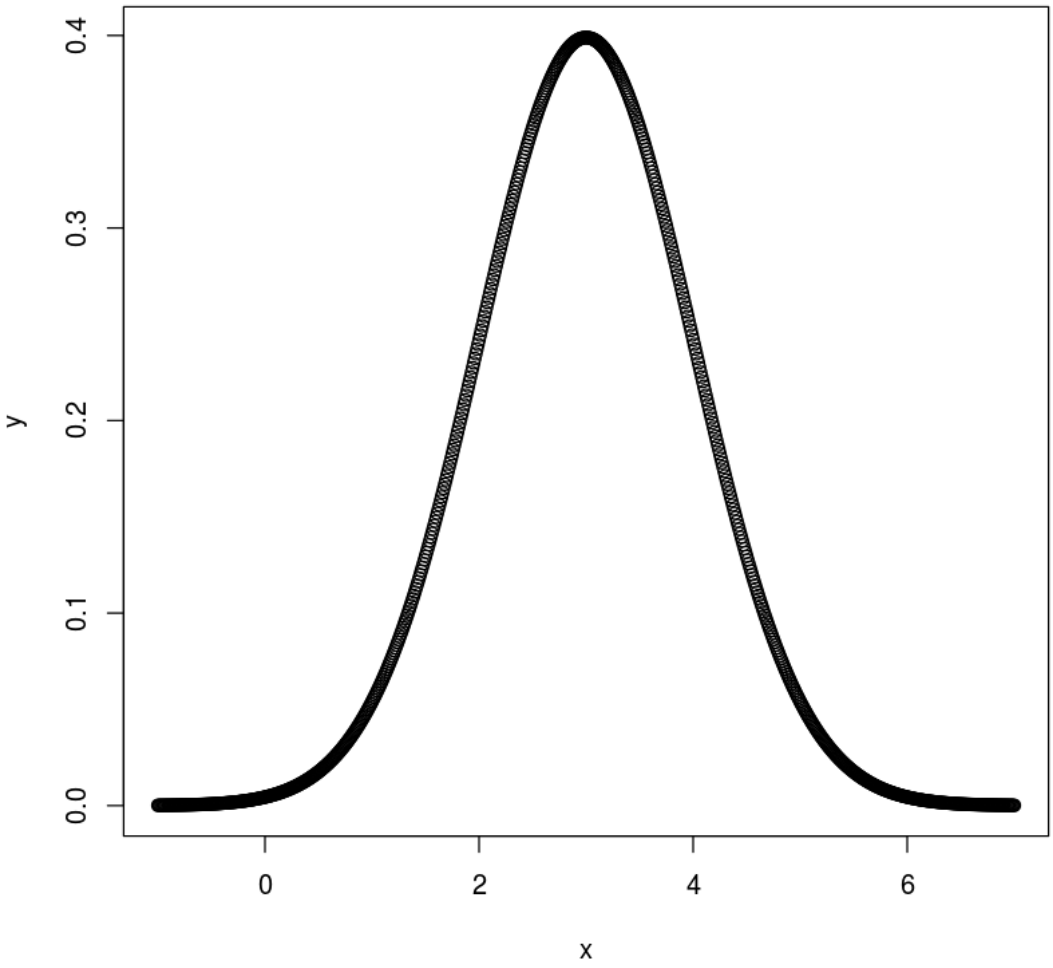
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

** Example **

Let $X \sim N(3, 1)$ Compute the following:

```
In [51]: m <- 3
         s <- 1

         # P(X < 2)
         pnorm(2, m, s) # 15.8%

         # P(2 <= X <= 4)
         pnorm(4, m, s) - pnorm(2, m, s) # 68.2%

         # Find the quantilies 0.025
         qnorm(0.025, m, s) # 1.04

         # Find the quantiles 0.5
         qnorm(0.5, m, s) # 3 makes sense, doesn't it?

         # Find the quantiles 0.975
         qnorm(0.975, m, s) # 4.95

         #Let's visualise the distribution
         x <- seq(-1,7,0.01)
         y <- dnorm(x, m, s)
         plot(x, y)
```

0.158655253931457
0.682689492137086
1.04003601545995
3
4.95996398454005



5.2.6 t Distribution

In real life, even though we may have normal data it is higly unlikely to know the standard deviation of the distribution (σ). Therefore, we make an estimate of it, and it is calculated as : $\frac{\sum_i (X_i - \bar{X})^2}{n-1}$. This distribution is not called normal, but t distribution with $= n1$ degrees of freedom.

$$Y \sim t$$

$$f(y) = \frac{\Gamma(\frac{+1}{2})}{\Gamma(\frac{-}{2})\sqrt{\pi)}}(1 + \frac{y^2}{\_\_})^{-(\frac{+1}{2})}$$

$$E[Y] = 0, \; if \; > 1$$

$$Var[Y] = \frac{}{-2}, \; if \; > 2$$

The t distribution is symmetric and resembles the normal distribution, but with thicker tails. As the degrees of freedom increase, the t distribution looks more and more like the standard normal distribution.

Example

Let T_8 distribution. Compute the following:

```
In [52]: df <- 8

         # Find the P(T_8 < 1)
         pt(1, df)

         # Find the P(T_8 > 2)
         1 - pt(2, df)

         # Find the P(-1 < T_8 < 1)
         pt(1, df) - pt(-1, df)

         # Find the quantile 0.025
         qt(0.025, df)

         # Find the quantile 0.5
         qt(0.5, df)

         # Find the quantile 0.975
         qt(0.975, df)
         x <- seq(-3, 3, 0.01)
         y <- dt(x, df)
         plot(x, y)
```

0.826703246456333
0.0402581189786313
0.653406492912666
-2.30600413520417
0
2.30600413520417