Probability Models

Bachelor in Computer Science and Engineering

2020/21

1. Objectives

- To represent probability/density and distribution functions of different models of continuous/discrete random variables.
- To compute probability using different distributions.
- To interpret and compare distribution graphs.
- To model real situations by means of probability distributions.

2. Probability models

The `stats` package of `R` has implemented the most used probability models. The following list is not exhaustive but contains all the models that will be used during this course.

- Beta distribution, `dbeta`.
- Binomial (including Bernoulli), `dbinom`.
- Cauchy distribution, `dcauchy`.
- Chi-squared distribution, `dchisq`.
- Exponential distribution, `dexp`.
- Fisher's F distribution, `df`.
- Gamma distribution, `dgamma`.
- Geometric distribution, `dgeom`.
- Hypergeometric distribution, `dhyper`.
- Log-normal distribution, `dlnorm`.
- Multinomial distribution, `dmultinom`.
- Negative binomial distribution, `dnbinom`.
- Normal distribution, `dnorm`.
- Poisson distribution, `dpois`.
- Student's t distribution, `dt`.
- Uniform distribution, `dunif`.
- Weibull distribution, `dweibull`.

As can be deduced from the list, the name of the functions is formed by the letter `d` plus the `name` or an abbreviation of the distribution's name. This rule is common to all distributions and the letters `d`, `p`, `q` and `r` are used to denote density or mass function, cumulative distribution function, quantile function and random variate generation, respectively.

First, we will illustrate its use with the uniform model, which is one of the simplest. Then, in the following sections, we will study in detail those models that are most frequently used to solve phenomena found in Engineering.

The uniform distribution, $U(a,b)$, has the following density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if} \quad x \in [a,b] \\ 0 & \text{if} \quad x \notin [a,b] \end{cases}$$

Let's assume that $a = -1$ and $b = 1$, then $\frac{1}{b-a} = 1/2$ if $x \in [-1,1]$ and zero otherwise. This can be verified with the following code

```
x_below_the_interval = seq(-2,-1.1,.1)
dunif(x_below_the_interval, min = -1, max = 1)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0
```

```
x_in_the_interval = seq(-1,1,.1)
dunif(x_in_the_interval, min = -1, max = 1)
```

```
##  [1] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
## [20] 0.5 0.5
```

```
x_above_the_interval = seq(1.1,2,.1)
dunif(x_above_the_interval, min = -1, max = 1)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0
```

Of course, the above is just an illustration not a formal proof.

The distribution function of an uniform random variable, $U(a,b)$, is given by

$$F(x) = \begin{cases} 0 & \text{if} \quad x < a \\ \frac{x-a}{b-a} & \text{if} \quad x \in [a,b) \\ 1 & \text{if} \quad x \geq b \end{cases},$$

which can be illustrated, for $a = -1$ and $b = 1$, with the following code

```
punif(x_below_the_interval, min = -1, max = 1)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0
```

```
punif(x_in_the_interval, min = -1, max = 1)
```

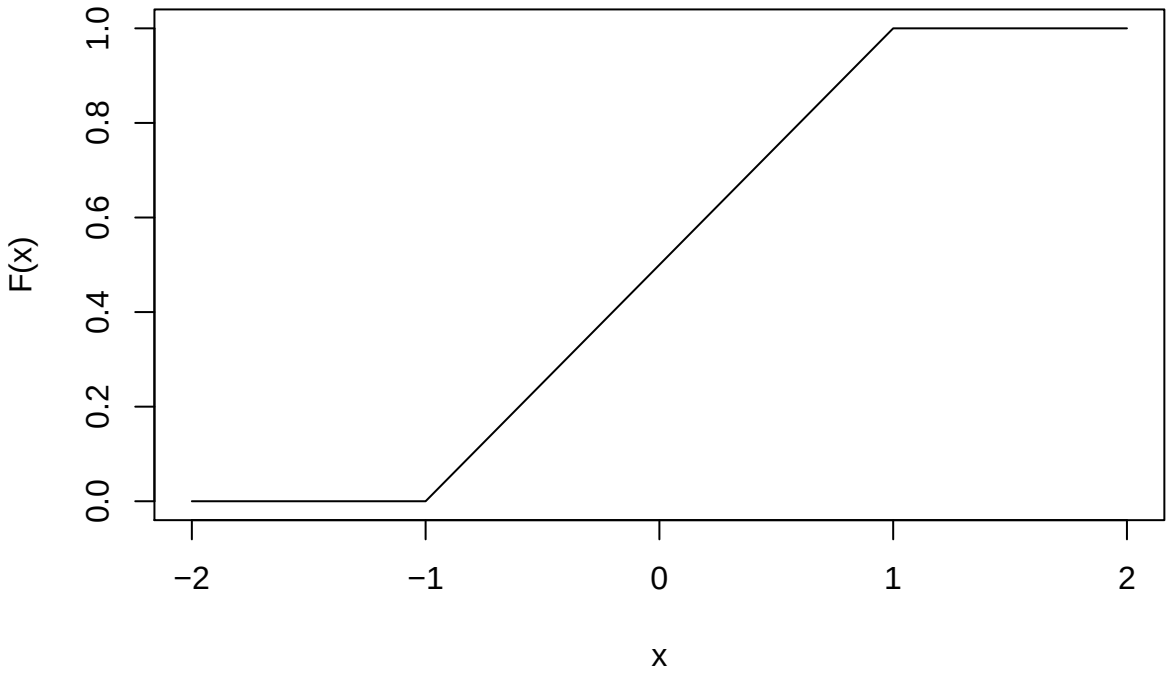
```
##  [1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70
## [16] 0.75 0.80 0.85 0.90 0.95 1.00
```

```
punif(x_above_the_interval, min = -1, max = 1)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1
```

Also, we can obtain the representation of this distribution function by

```
x = c(x_below_the_interval, x_in_the_interval, x_above_the_interval)
Fx = punif(x, min = -1, max = 1)
plot(x, Fx, type = "l", ylab = "F(x)")
```



The quantile function of an uniform random variable, $U(a, b)$, is defined by

$$Q(p) = a + p(b - a) \text{ if } p \in [0, 1].$$

For instance, if we want to calculate the quartiles of an $U(a, b)$ we should evaluate the above expression at the points $p = 0.25$, 0.5 and 0.75 for the first, second (median) and third quartiles, respectively. Let's look at the code for $a = -1$ and $b = 1$

```
p = c(0.25, 0.5, 0.75)
qunif(p, min = -1, max = 1)
```

```
## [1] -0.5  0.0  0.5
```

The quartiles of an $U(-1, 1)$ are -0.5, 0 and 0.5.

Of course, the quantile function is not defined outside de interval $[0, 1]$ and R returns `NaN`, *Not a Number*.

```
p = c(-0.25, 1.25)
qunif(p, min = -1, max = 1)
```

```
## Warning in qunif(p, min = -1, max = 1): NaNs produced
```

```
## [1] NaN NaN
```

2.1. Discrete distributions: Binomial and Poisson.

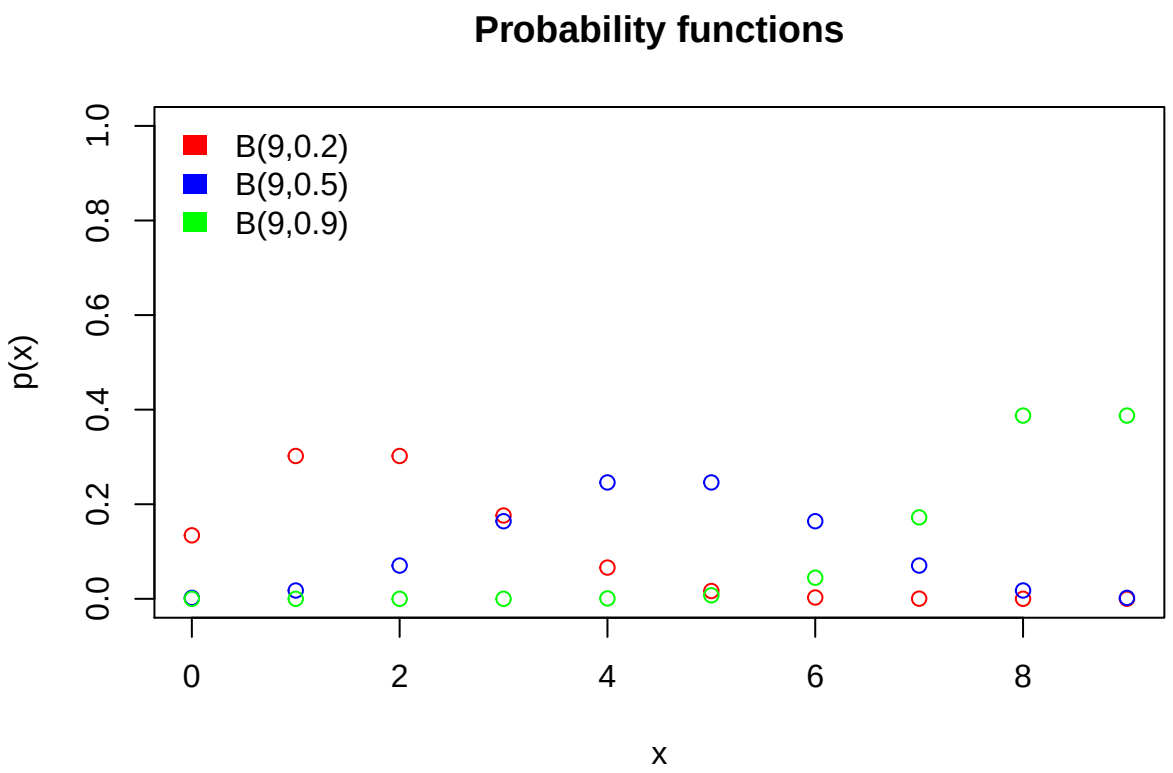
2.1.1. Binomial distribution, $X \sim B(n,p)$

We remind that a binomial distribution with parameters n and p represents a random variable that counts the number of successes that we could obtain by repeating a Bernoulli experiment, i.e. an experiment whose only results are 1 (success) and 0 (no success). n is the number of times we repeat the Bernoulli experiment (number of trials) and p is the success probability (event probability) and the experiments are assumed independent.

Graphical representation of the probability and distribution functions:

In the following example we use three different binomial distributions B(9,0.2), B(9,0.5) and B(9,0.9).

```
n = 9
p1 = 0.2
p2 = 0.5
p3 = 0.9
x = 0:n
Px1 = dbinom(x, n, prob = p1)
Px2 = dbinom(x, n, prob = p2)
Px3 = dbinom(x, n, prob = p3)
plot(x, Px1, xlim = c(0,9), ylim = c(0,1), col = "red", main = "Probability functions", ylab = "p(x)")
points(x, Px2, xlim = c(0,9), ylim = c(0,1), col = "blue")
points(x, Px3, xlim = c(0,9), ylim = c(0,1), col = "green")
legend('topleft',c('B(9,0.2)', 'B(9,0.5)', 'B(9,0.9)'), fill = c("red", "blue", "green"),
       bty = 'n', border = NA)
```

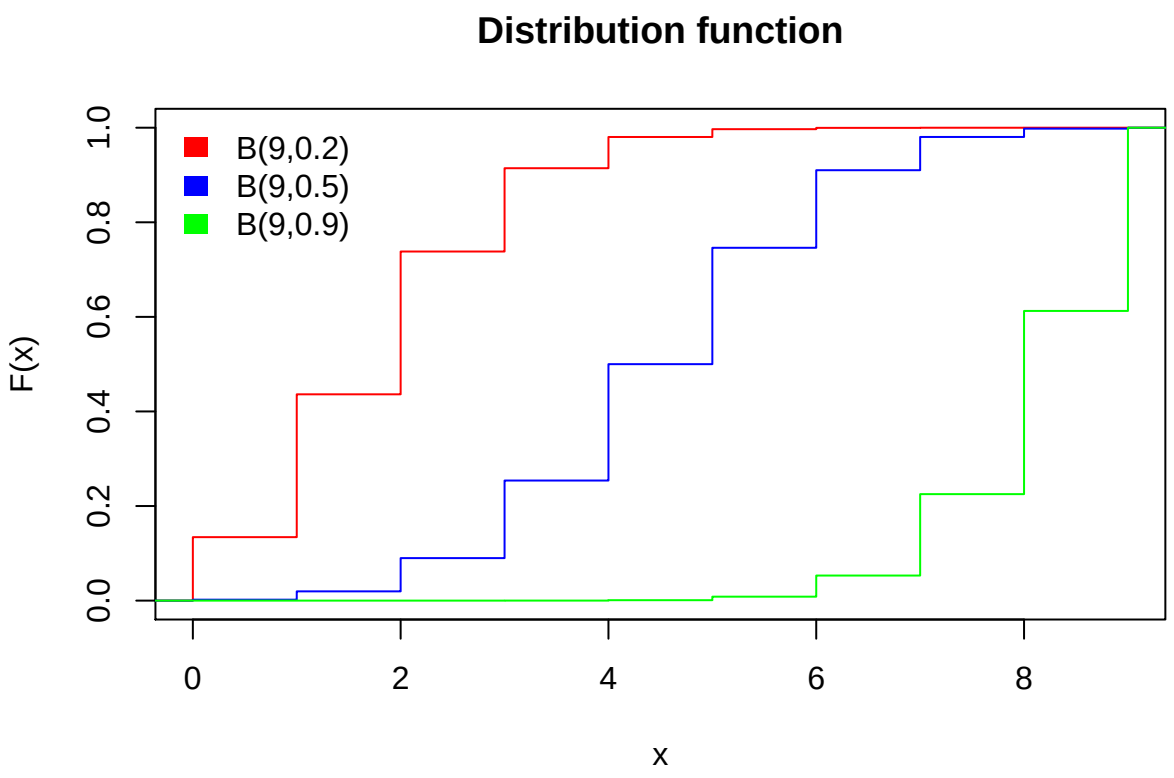


We can notice that:

- when $p = 0.5$ (in the figure above shown in blue) the distribution is symmetric,
- when $p < 0.5$ (in the figure above shown in red, corresponding to $p = 0.2$), the distribution is asymmetric towards the right, it means that the random variable is positive asymmetric,
- when $p > 0.5$ (in the figure above shown in green, corresponding to $p = 0.9$), the distribution is asymmetric towards the left, it means that the random variable is negative asymmetric.

If we want to show the distribution functions, just need to change `dbinom` by `pbinom`

```
x = c(-1, x, 10)
Fx1 = pbinom(x, n, prob = p1)
Fx2 = pbinom(x, n, prob = p2)
Fx3 = pbinom(x, n, prob = p3)
plot(x, Fx1, xlim = c(0,9), ylim = c(0,1), col = "red", main = "Distribution function",
     type = "s", ylab = "F(x)")
lines(x, Fx2, xlim = c(0,9), ylim = c(0,1), col = "blue", type = "s")
lines(x, Fx3, xlim = c(0,9), ylim = c(0,1), col = "green", type = "s")
legend('topleft',c('B(9,0.2)', 'B(9,0.5)', 'B(9,0.9)'), fill = c("red", "blue", "green"),
       bty = 'n', border = NA)
```



It should be noted that we add two additional values at `x = c(-1, x, 10)`, -1 and 10, which are outside the set $\{0, 1, ..., 9\}$ where the $B(9, p)$ takes values. The reason is to get a complete graph of the distribution function. Also notice that we used `lines` instead of `points` and we used `type = "s"` to get a step function.

Computing probabilities

Assume that we have a random variable $X \sim B(12, 0.4)$ and we want to compute the following probabilities:

- $\Pr(X = 7)$
- $\Pr(X > 3)$
- $\Pr(X \leq 8)$
- $\Pr(X < 5)$

The solutions, in R, are

```
dbinom(7, 12, prob = 0.4)
```

```
## [1] 0.1009024
```

```
1-pbinom(3, 12, prob = 0.4) # since Pr(X > 3) = 1 - Pr(X =< 3)
```

```
## [1] 0.7746627
```

```
pbinom(8, 12, prob = 0.4)
```

```
## [1] 0.9847327
```

```
pbinom(5, 12, prob = 0.4) # since Pr(X < 5) = Pr(X =< 4)
```

```
## [1] 0.6652086
```

Computing the distribution percentiles

Lets suppose that we want to compute the percentiles of a given random variable X. A percentile is a number that is not reached by the corresponding percentage of individuals of a population. In practice we select a value p and the function "`q+name`" will return the value, a, such that $\Pr(X \leq a) = p$.

For instance, assume that $X \sim B(4, 0.5)$. We want to compute the value, a, such that $\Pr(X \leq a) = 0.3125$.

```
qbinom(0.3125, 4, prob = 0.5)
```

```
## [1] 1
```

We can check the above result by calculating

```
pbinom(0:4, 4, prob = 0.5)
```

```
## [1] 0.0625 0.3125 0.6875 0.9375 1.0000
```

where we can see that $\Pr(X \leq 1) = 0.3125$.

It should be noted that for a discrete random variable the equation $\Pr(X \leq a) = p$ may have no solution. For example, in the output above, we see that there is no value that satisfies $\Pr(X \leq a) = 0.1$. However, the function `qbinom` returns a value

```
qbinom(0.1, 4, prob = 0.5)
```

```
## [1] 1
```

which in fact satisfies the following inequality $\Pr(X \leq a) \geq p$.

> **Practical example**:
>
> A metro traveler goes every morning at the same time to the metro platform. 18% of the times the train is already there and the rest of times he has to wait for the train.
>
> a) Considering seven consecutive days, what is the probability that only one over the seven days he does not have to wait for the train?

b) Considering fifteen consecutive days, what is the probability that at most three days he does not have to wait for the train?

c) Considering eighteen consecutive days, what is the probability that he does not have to wait for the train for more than five days?

Define X as the number of days that the traveler does not have to wait for the train. Then $X \sim B(n, p)$, i.e. it is distributed as a Binomial with event probability $p = 0.18$. We obtain that

a) $X \sim B(7, 0.18)$; $\Pr(X = 1) = 0.3830484$ using `dbinm(1, 7, 0,18)`.

b) $X \sim B(15, 0.18)$; $\Pr(X \leq 3) = 0.7218051$ using `pbinm(1, 15, 0,18)`.

c) $X \sim B(18, 0.18)$; $\Pr(X > 5) = 0.08893546$ using `1-pbinom(5, 18, 0.18)` or `pbinom(5, 18, 0.18, lower.tail = FALSE)`.

2.1.2. Poisson distribution, $X \sim Poisson(\lambda)$

A Poisson random variable, $X \sim Poisson(\lambda)$, may represent the number of independent events that can occur in a time unit when the underlining process is a Poisson process with constant parameter λ. The only parameter λ represents the average number of events in the time unit, (that can also be length, surface, volume or whatever the chosen continuous measurement unit is).

For the Poisson model, we can do the same calculations as for the binomial model, so we will go directly to a practical example.

Practical example:

The number of users that access to a network server is, in average, 3000 per hour. The network can service with optimal performance up to 100 accesses per minute. Assuming that the accesses are produced in an independent way and at constant rate, we want to compute the probability that in a given minute there are

a) exactly 40 users accessing,

b) between 40 and 50 users accessing,

c) more than 100 accesses, and therefore there are delays in the network communications.

Let X = number of accesses per minute, that means $X \sim Poisson(\lambda = 50)$. The required probabilities are given by

a) $\Pr(X = 40) = 0.02149963$ using `dpois(40, lambda = 50)`.

b) $\Pr(40 \leq X \leq 50) = \Pr(X \leq 50) - \Pr(X \leq 39) = 0.4729463$ using `ppois(50, lambda = 50) - ppois(39, lambda = 50)`.

c) $\Pr(X > 100) = 1.569746e - 10$ using `ppois(100, lambda = 50, lower.tail = FALSE)`. Therefore the network is well dimensioned for the actual amount of traffic.

2.2. Continuous Distributions: Normal and Exponential

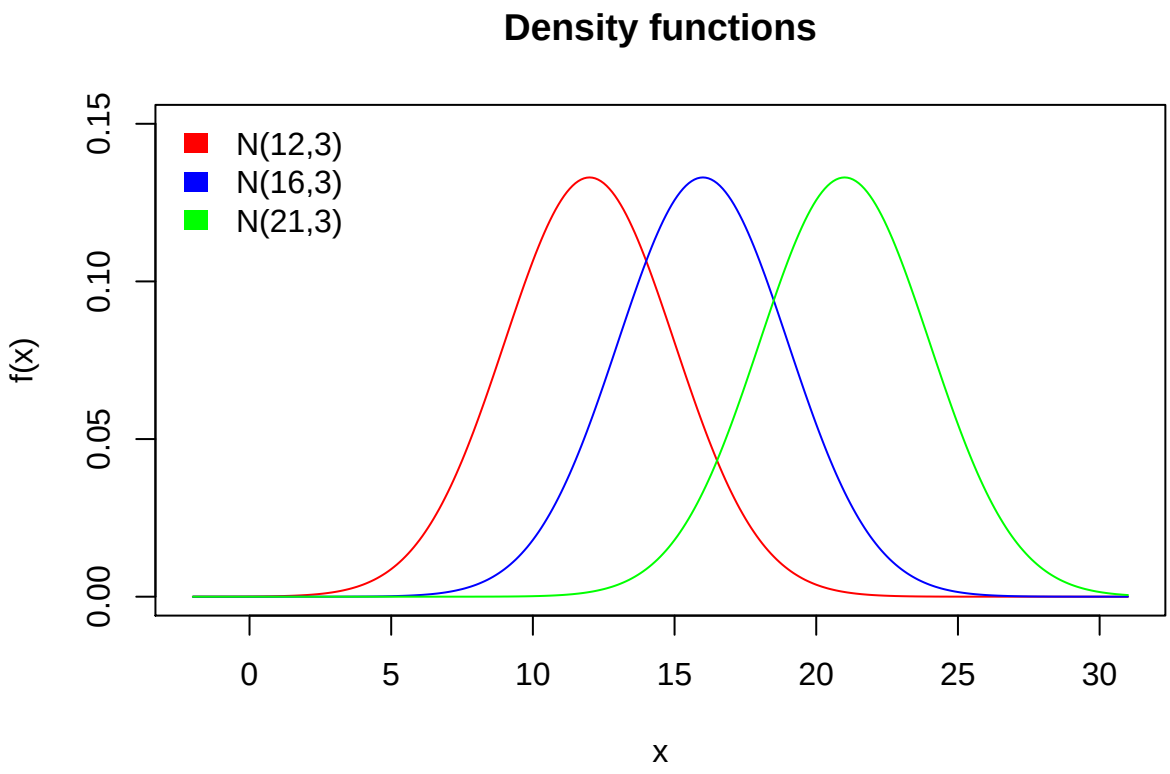
2.2.1. Normal Distribution, $X \sim \mathcal{N}(\mu, \sigma)$

The Normal distribution is symmetric. The mean, μ coincides with the mode and the median. The density function is bell-shaped and its form is usually called as the "Gauss' bell".

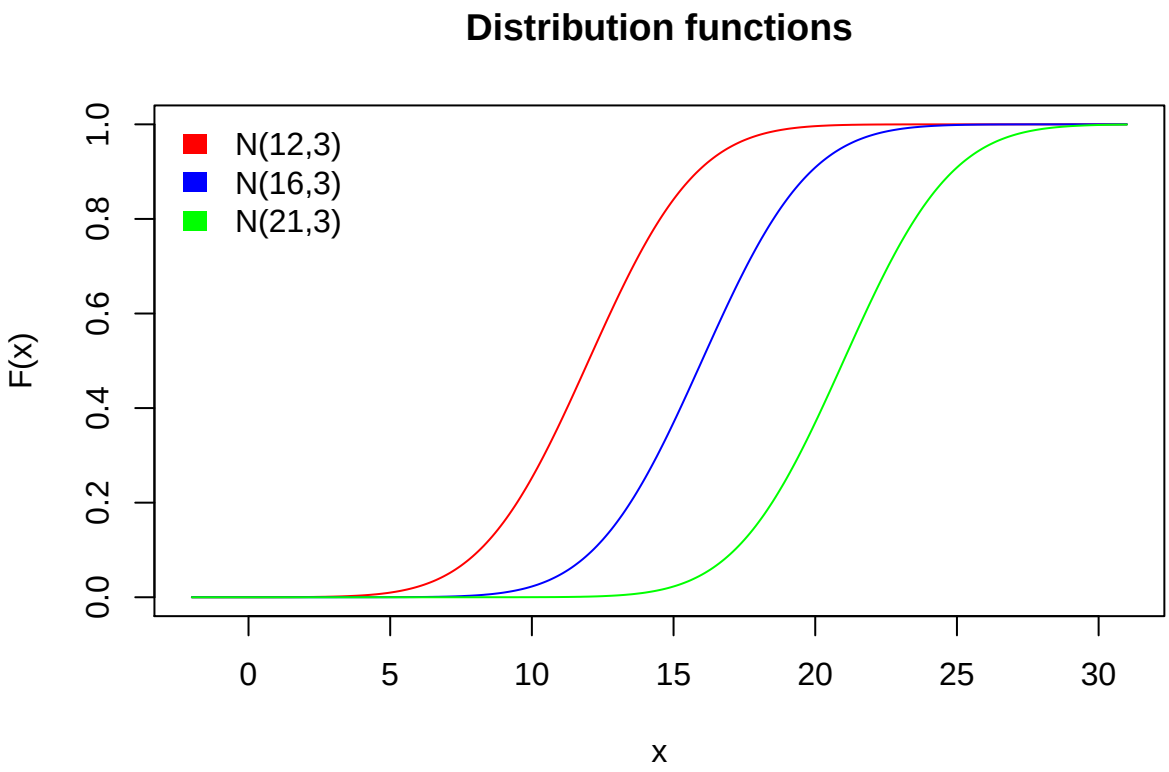
Comparing the density and the distribution functions

We draw the density and the distribution functions of three different Normal random variables, with equal variance $\sigma^2 = 9$ and different μ: $\mathcal{N}(12, 3)$, $\mathcal{N}(16, 3)$ and $\mathcal{N}(21, 3)$. We can see how the bell moves its center along the real axis without changing its size.

```r
x = seq(-2, 31, .1)
gx1 = dnorm(x, mean = 12, sd = 3)
gx2 = dnorm(x, mean = 16, sd = 3)
gx3 = dnorm(x, mean = 21, sd = 3)
plot(x, gx1, xlim = c(-2,31), ylim = c(0,.15), col = "red", main = "Density functions",
     type = "l", ylab = "f(x)")
lines(x, gx2, xlim = c(-2,31), ylim = c(0,.15), col = "blue")
lines(x, gx3, xlim = c(-2,31), ylim = c(0,.15), col = "green")
legend('topleft',c('N(12,3)', 'N(16,3)', 'N(21,3)'), fill = c("red", "blue", "green"),
       bty = 'n', border = NA)
```

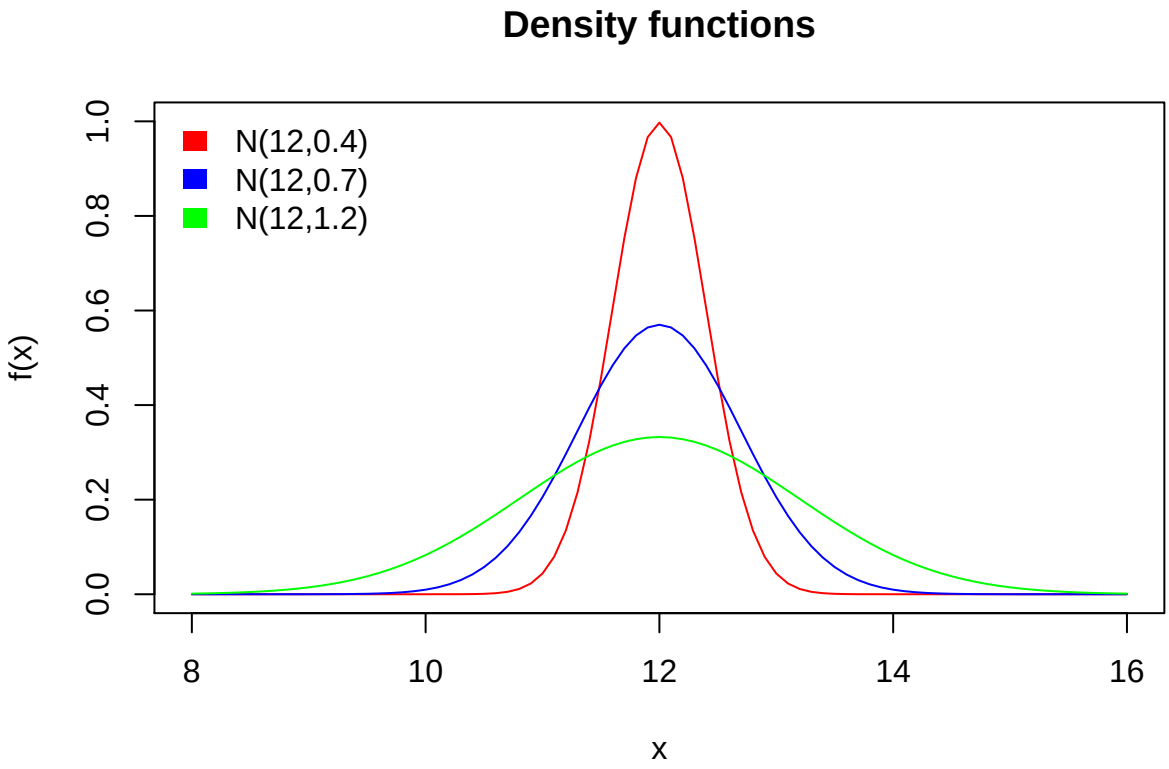



```r
Gx1 = pnorm(x, mean = 12, sd = 3)
Gx2 = pnorm(x, mean = 16, sd = 3)
Gx3 = pnorm(x, mean = 21, sd = 3)
plot(x, Gx1, xlim = c(-2,31), ylim = c(0,1), col = "red", main = "Distribution functions",
     type = "l", ylab = "F(x)")
lines(x, Gx2, xlim = c(-2,31), ylim = c(0,1), col = "blue")
lines(x, Gx3, xlim = c(-2,31), ylim = c(0,1), col = "green")
legend('topleft',c('N(12,3)', 'N(16,3)', 'N(21,3)'), fill = c("red", "blue", "green"),
       bty = 'n', border = NA)
```

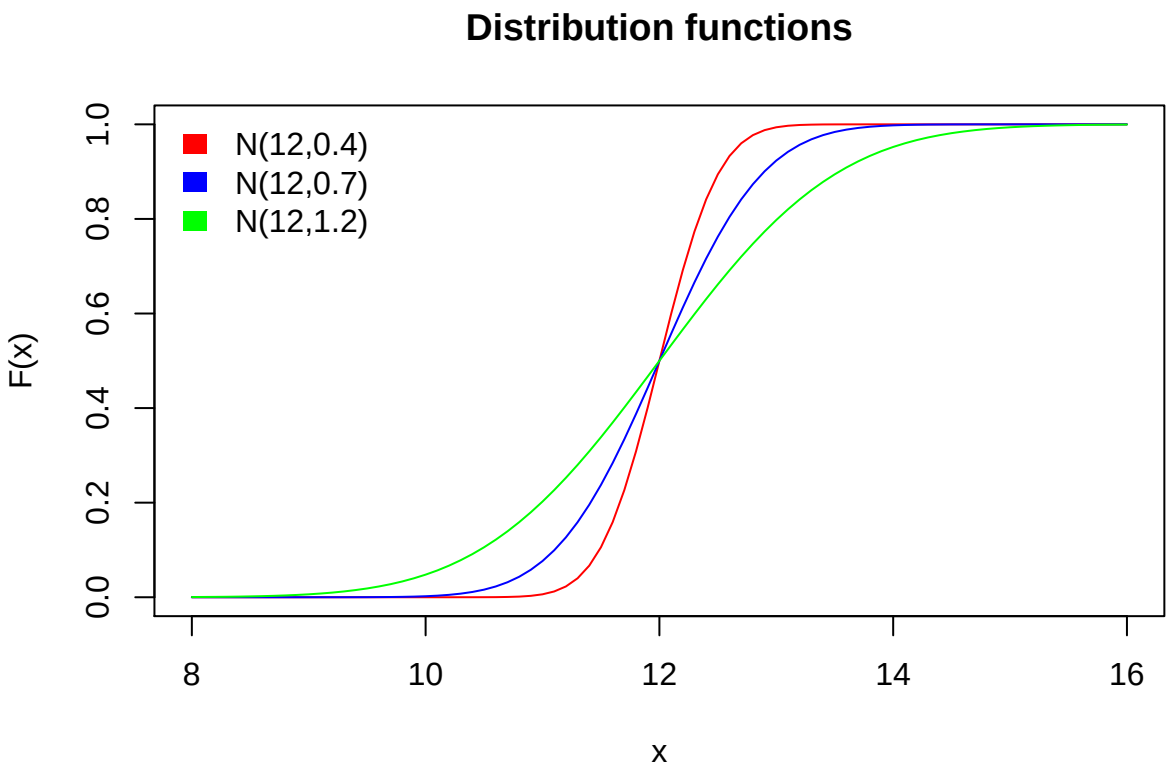


Now we draw the density and the distribution functions of three Normal random variables with equal mean $\mu = 12$ and different standard deviation σ: $\mathcal{N}(12, 0.4)$, $\mathcal{N}(12, 0.7)$ and $\mathcal{N}(12, 1.2)$. We can see now that the bells have the same center but different sizes (the dispersion of the values changes).

```
x = seq(8, 16, .1)
gx1 = dnorm(x, mean = 12, sd = 0.4)
gx2 = dnorm(x, mean = 12, sd = 0.7)
gx3 = dnorm(x, mean = 12, sd = 1.2)
plot(x, gx1, xlim = c(8,16), ylim = c(0,1), col = "red", main = "Density functions",
     type = "l", ylab = "f(x)")
lines(x, gx2, xlim = c(8,16), ylim = c(0,1), col = "blue")
lines(x, gx3, xlim = c(8,16), ylim = c(0,1), col = "green")
legend('topleft',c('N(12,0.4)', 'N(12,0.7)', 'N(12,1.2)'),
       fill = c("red", "blue", "green"), bty = 'n', border = NA)
```




```
Gx1 = pnorm(x, mean = 12, sd = 0.4)
Gx2 = pnorm(x, mean = 12, sd = 0.7)
Gx3 = pnorm(x, mean = 12, sd = 1.2)
plot(x, Gx1, xlim = c(8,16), ylim = c(0,1), col = "red", main = "Distribution functions",
     type = "l", ylab = "F(x)")
lines(x, Gx2, xlim = c(8,16), ylim = c(0,1), col = "blue")
lines(x, Gx3, xlim = c(8,16), ylim = c(0,1), col = "green")
legend('topleft',c('N(12,0.4)', 'N(12,0.7)', 'N(12,1.2)'),
       fill = c("red", "blue", "green"), bty = 'n', border = NA)
```



Computing probabilities

According to the definition of density function, to compute a given probability is equivalent to computing the value of an integral. Indeed the probability that a random variable X takes values in a given interval is equal to the value of the integral of the density function over that interval.[1]

Example: Given $X \sim \mathcal{N}(8, 2.6)$ compute $\Pr(X > 11.3)$, $\Pr(X < 7.9)$, $\Pr(-1 < X < 4)$ and $\Pr(X \geq 18)$.

a) $\Pr(X > 11.3) = 0.1021794$ using `pnorm(11.3, 8, 2.6, lower.tail = FALSE)`.

b) $\Pr(X < 7.9) = 0.4846598$ using `pnorm(7.9, 8, 2.6)`.

c) $\Pr(-1 < X < 4) = \Pr(X < 4) - \Pr(X < -1) = 0.06169935$ using `pnorm(4, 8, 2.6) - pnorm(-1, 8, 2.6)`,

d) $\Pr(X \geq 18) = 5.999322e-05$ using `pnorm(18, 8, 2.6, lower.tail = FALSE)`.

Computing percentiles

To compute the percentiles we have to follow the same steps as it was done for the discrete case. For instance, suppose that we want to calculate the 90%, 95%, 97.5% and 99% percentiles of the standard normal distribution, $Z \sim \mathcal{N}(0, 1)$

```
p = c(0.9, 0.95, 0.975, 0.99)
qnorm(p, 0, 1)
```

```
## [1] 1.281552 1.644854 1.959964 2.326348
```

[1] Since X is a continous random variable, $\Pr(X = x) = 0$ for all x, therefore $\Pr(X \leq x) = \Pr(X < x)$.

These values are usually denoted by $z_{0.1} = 1.281552$, $z_{0.05} = 1.644854$, $z_{0.025} = 1.959964$ and $z_{0.01} = 2.326348$.

In addition, in the specific case of the Normal random variable it is interesting to know what is the probability that the random variable $X \sim \mathcal{N}(\mu, \sigma)$ takes values distant the amount of 1, 2 or 3 standard deviations σ from the center μ. In other words we could be interested in the probabilities that X falls in the interval $(\mu - k\sigma, \mu + k\sigma)$, where $k = 1, 2$ or 3.

$$\Pr(\mu - k\sigma < X < \mu + k\sigma) = \Pr\left(-k < \frac{X - \mu}{\sigma} < k\right) = \Pr(-k < Z < k),$$

where Z is the usual notation for the standard normal distribution, $\mathcal{N}(0, 1)$. The above probabilities can be calculated by

```
pnorm(3,0,1) - pnorm(-3,0,1)
```

```
## [1] 0.9973002
```

```
pnorm(2,0,1) - pnorm(-2,0,1)
```

```
## [1] 0.9544997
```

```
pnorm(1,0,1) - pnorm(-1,0,1)
```

```
## [1] 0.6826895
```

So,

- $\Pr(X \in (\mu - \sigma, \mu + \sigma)) = 0.6826895$,
- $\Pr(X \in (\mu - 2\sigma, \mu + 2\sigma)) = 0.9544997$,
- $\Pr(X \in (\mu - 3\sigma, \mu + 3\sigma)) = 0.9973002$.

2.2.2. Exponential Distribution

The exponential distribution is interesting for its application to different phenomena:

- The waiting time up the first event in a Poisson process (the even could be an arrival, a phone call, or whatever success over a continuous support). If we count the number of these events in the time unit we would have a Poisson(λ), where λ is the mean number of events in the unit of time. The assumption is that events arrive independently and at constant rate.
- The time spent from a given instant of time up to the next occurrence of an event.

It is important to remember that the exponential random variable has the memory-less property.

> **Practical example**:
>
> The number of users that access to a network server is, in average, 3000 per hour. The network can service with optimal performance up to 100 accesses per minute. Assuming that the accesses are produced in an independent way and at constant rate, we want to compute the probability the time that spends between two accesses is at least 5 seconds.
>
> Solution: The assumption of independence and the constant rate imply that the number of accesses per unit of time is distributed according a Poisson random variable, and therefore the time between two successive accesses is an Exponential random variable, $T \sim Exp(\lambda = 3000/3600\ accesses/second)$. The mean of T is $1/\lambda = 3600/3000 = 1.2\ seconds$. The required probability is given by
>
> $$\Pr(T > 5) = 0.01550385$$
>
> obtained by `pexp(5, rate = 3000/3600, lower.tail = FALSE)`.