STATISTICS WITH R

Prof. Jorge Pineiro
Mathematics and Computer Science Department
Bronx Community College CUNY
New York, 07/07/2019

Class 0: Introduction

R is a language and environment for statistical computing and graphics. The steps to follow before we can start working with "R" and learning commands are:

(1) Download and install "R".

(2) Download and install the script editor "RStudio", which is the one we will be working with.

(3) Go to File and open a new "R script".

(4) Go to Session and set the working directory correctly.

Statistics is the science of collecting and analyzing data and the **The "R" language** will be a valuable tool in working and visualizing this data. The "R" language is going to help in doing both **descriptive statistics** and **inferential statistics**.

Descriptive Statistics: describe quantitatively features of a sample data. It aims to summarize or describe the sample using **statistics**. Descriptive Statistics can be univariate when studies features of just one variable or multivariate when aims to relate features of two or more variables.

Inferential Statistics: aims to use sample data to learn about the whole population. It is based heavily on **the theory of Probability**.

First "R" commands and examples:

Basic arithmetic and assignments:

The default amount of digits for a number is seven, but we can control how many digits we work with. Let us do the work with the numbers π and $\sqrt{2}$.

```
pi
```

```
## [1] 3.141593
```

```
options(digits = 20)
pi
```

```
## [1] 3.1415926535897931159
```

```
2^{.5}
```

```
## [1] 1.4142135623730951455
```

We can perform our usual mathematical operations with the use of R. Remember though to use * for the multiplication.

```
2*3
```

```
## [1] 6
```

```
7+2^5
```

```
## [1] 39
```

```
(7+2)^5
```

```
## [1] 59049
```

```
log10(10000000)
```

```
## [1] 7
```

```r
log(4^14, base=4)
```

```
## [1] 14
```

```r
# The Euler number e
exp(1)
```

```
## [1] 2.7182818284590450908
```

Assignment of a variable can be donde with leftward, rightward or equal:

```r
x <- 17
x
```

```
## [1] 17
```

```r
t = -2
t
```

```
## [1] -2
```

```r
x+t
```

```
## [1] 15
```

Logical variables and decision making:

Comparing two numbers. Determining if a number is equal, less than or simply different to another number. Using our findings with if-then commands.

```r
2==4
```

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

```r
2!=4
```

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

```r
2<=4
```

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

```r
if(2 <= 2 & 1 > 3) {print("I do not like Math")} else {print("I love Math")}
```

```
## [1] "I love Math"
```

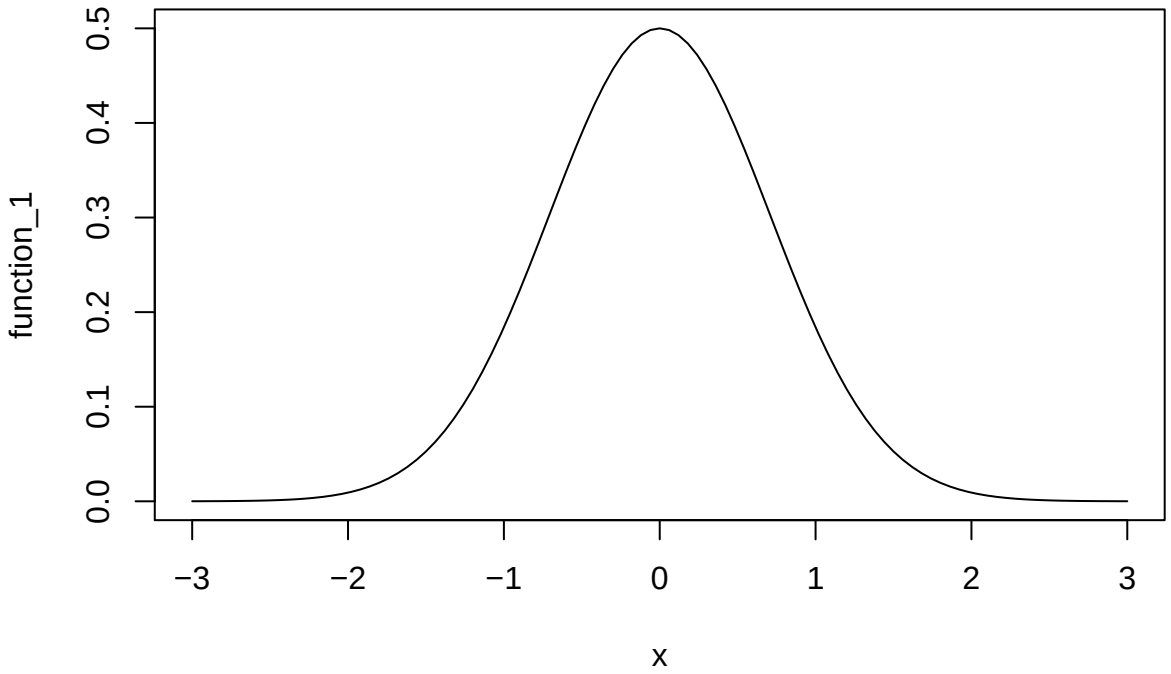
Functions and plots:

We can define and plot mathematical functions. We are going to define and plot a function of singular importance in Probability and Statistics: The Gauss bell of equation $f(x) = -\frac{e^{x^2}}{2}$, where e is representing the Euler number.

```r
# Definition of our function
function_1 <- function(x) {exp(-x^2)/2}
# Evaluation at a number as a test
function_1(0)
```

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

```r
# Now the graph
plot(function_1,-3,3, main="The graph of the Gauss bell")
```



We can also define and plot together several linear, quadratic and trigonometric functions:

```
y <- function(x) {sin(x)}
fun2 <- function(x) {-3*x}
fun3 <- function(x) {x^2}

# Some evaluations of our functions:

fun2(0)
```

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

```
fun3(1)
```

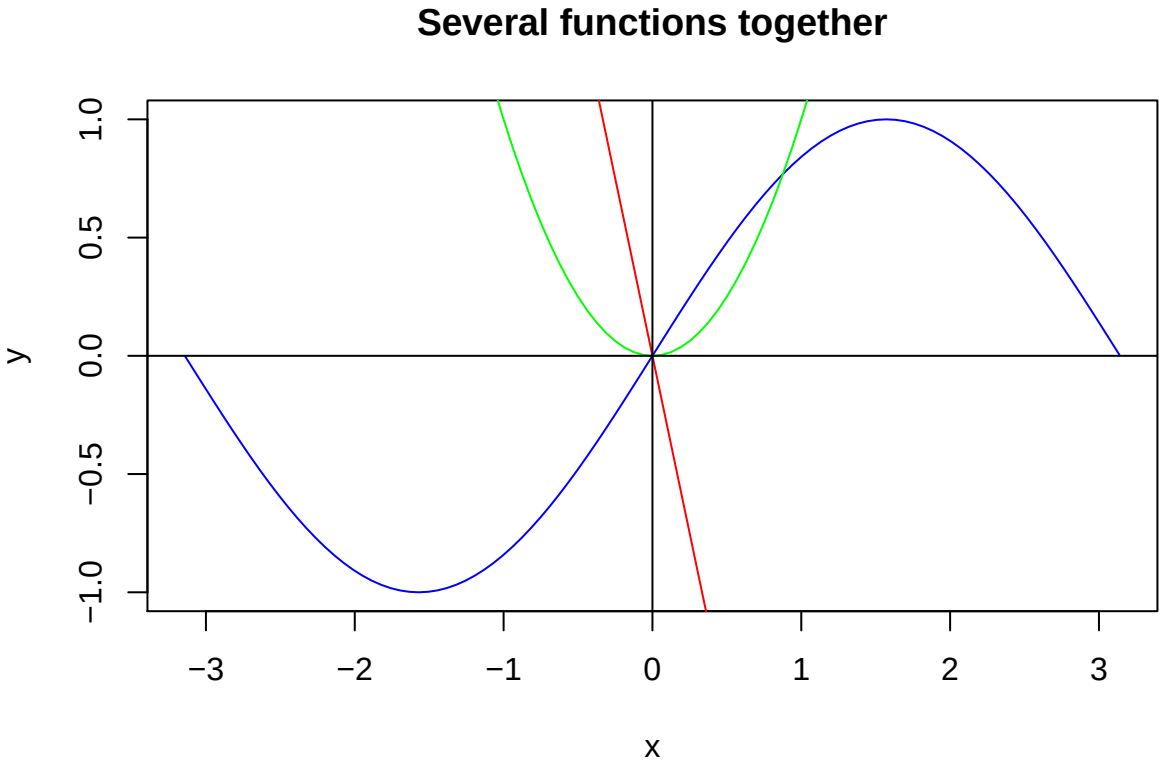
```
## [1] 1
```

```
# The command add=TRUE allows to put together the graph of several functions.

plot (y, -pi, pi,col="blue", main="Several functions together")
plot (fun2, -pi, pi, add=TRUE, col="red")
plot (fun3, -pi, pi, add=TRUE, col="green")

# We add the coordinates axis x and y:

abline(h=0)
abline(v=0)
```



Vectors:

We can create a vector of names in the following way:

```
color1 <- "red"
color2 <- "yellow"
color3 <- "green"
colors <- c(color1,color2,color3)
```

There are different types of object in R, among them: integers, doubles, characters and logical. We can inquire what type of data we have in our hands:

```
is(colors)
```

```
## [1] "character"           "vector"              "data.frameRowLabels"
## [4] "SuperClassMethod"
```

```
colors
```

```
## [1] "red"    "yellow" "green"
```

```
typeof(colors)
```

```
## [1] "character"
```

```
typeof(pi)
```

```
## [1] "double"
```

```
typeof(-1)
```

```
## [1] "double"
```

We can create and work with numerical vectors. The operator [ ] will allow us to access the components of the vector. Sometimes we use colom : to indicate a sequence of numbers.

```r
a=c(2,4,5,7,10,12)
a
```

```
## [1]  2  4  5  7 10 12
```

```r
a[1]
```

```
## [1] 2
```

```r
a[2:3]
```

```
## [1] 4 5
```

```r
seq(1, 9, by = 2)
```

```
## [1] 1 3 5 7 9
```

```r
sum(a)
```

```
## [1] 40
```

```r
length(a)
```

```
## [1] 6
```

You can also create a vector by typing the entries one by one:

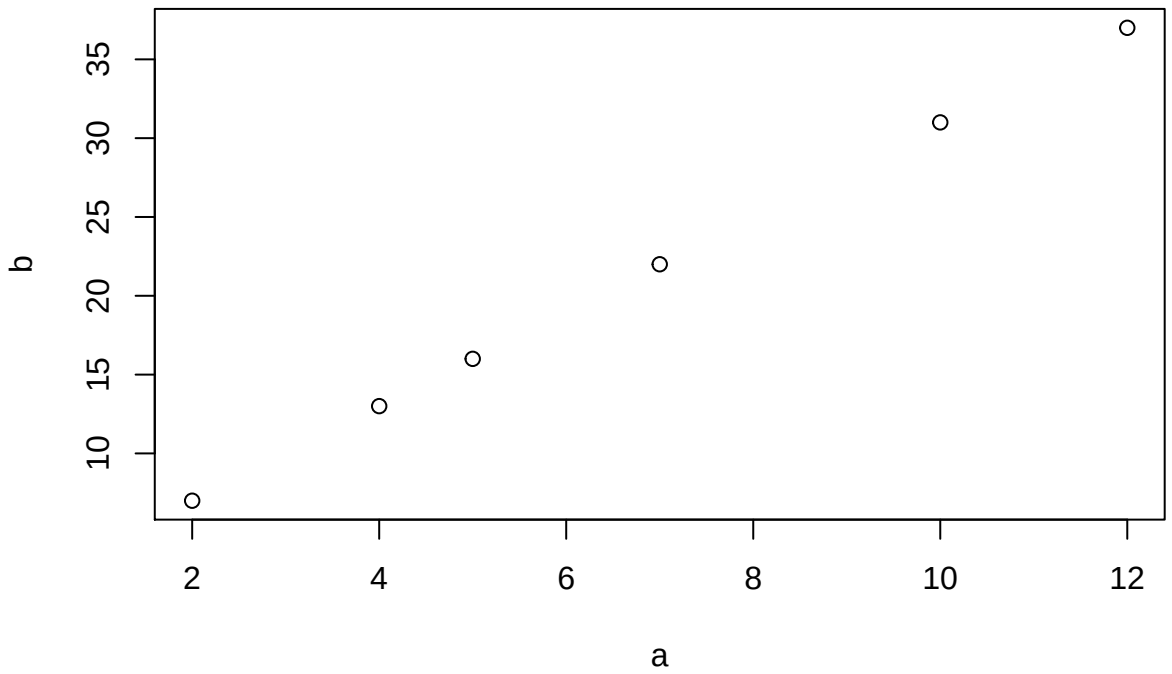
```r
v <- scan()
```

You can create a vector from another vector and plot both together as ordered pairs (x,y):

```r
b=3*a+1
b
```

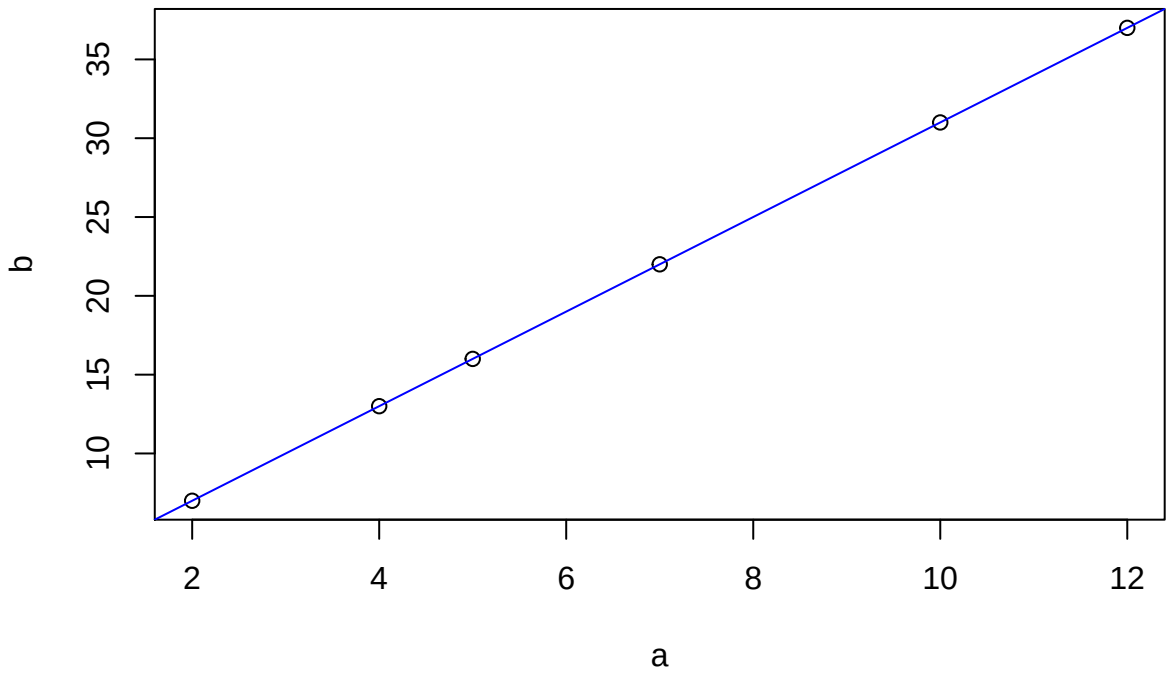
```
## [1]  7 13 16 22 31 37
```

```r
plot(a,b)
```



You can add a line to the plot with given slope and intercept, for example slope=3 and y-intercept=1:

```r
plot(a,b)
abline(1,3,col="blue")
```



Similar to how we compared numbers we can compare vectors:

```r
a==b
```

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

Loops:

Hee is an example of a loop with the command "for". In this case we will compute the sum of the squares of our previoulsy created vector a:

```r
print("This loop calculates the sum of the squares of a vector")
```

```
## [1] "This loop calculates the sum of the squares of a vector"
```

```r
a
```

```
## [1]  2  4  5  7 10 12
```

```r
sum <- 0
for(i in (1:4)) {
  sum <- sum +a[i]^2
}

print(sum)
```

```
## [1] 94
```

Data frames and vectors:

A data frame is a table in which each column contains values of one variable, in such a way that the column names are not empty and the row names are unique. We will create now our first data frame from two vectors.

```r
v1 <- 1:10
v2 <- 20: 21
v1/v2
```

```
##  [1] 0.05000000000000002776 0.09523809523809523808
##  [3] 0.14999999999999994449 0.19047619047619046617
##  [5] 0.25000000000000000000 0.28571428571428568425
##  [7] 0.34999999999999977796 0.38095238095238093234
##  [9] 0.45000000000000011102 0.47619047619047616042
```

```r
data <- data.frame(x= v1, y =v2, z = rep(c("M", "F"),5))
data
```

```
##     x  y z
## 1   1 20 M
## 2   2 21 F
## 3   3 20 M
## 4   4 21 F
## 5   5 20 M
## 6   6 21 F
## 7   7 20 M
## 8   8 21 F
## 9   9 20 M
## 10 10 21 F
```

```r
dim(data)
```

```
## [1] 10  3
```

```r
str(data)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ x: int  1 2 3 4 5 6 7 8 9 10
##  $ y: int  20 21 20 21 20 21 20 21 20 21
##  $ z: Factor w/ 2 levels "F","M": 2 1 2 1 2 1 2 1 2 1
```

If we decide to change the names of the variables:

```r
names(data)
```

```
## [1] "x" "y" "z"
```

```r
names(data) <- c("a","b","c")
head(data)
```

```
##   a  b c
## 1 1 20 M
## 2 2 21 F
## 3 3 20 M
## 4 4 21 F
## 5 5 20 M
## 6 6 21 F
```

Very important: different ways to access the data in your data frame:

```r
colors[3]
```

```
## [1] "green"
```

```
data[5,2]
```

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

```
colors[c(1,3)]
```

```
## [1] "red"   "green"
```

Accesing elements of a vector that satisfy certain conditions

```
a < 5
```

```
## [1]  TRUE  TRUE FALSE FALSE FALSE FALSE
```

```
a[a<5]
```

```
## [1] 2 4
```

```
a[which(a<5)]
```

```
## [1] 2 4
```

```
data[2,2]
```

```
## [1] 21
```

```
data[ ,2]
```

```
##  [1] 20 21 20 21 20 21 20 21 20 21
```

```
data[1, ]
```

```
##   a  b c
## 1 1 20 M
```

```
data[ ,2][3]
```

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

```
data
```

```
##     a  b c
## 1   1 20 M
## 2   2 21 F
## 3   3 20 M
## 4   4 21 F
## 5   5 20 M
## 6   6 21 F
## 7   7 20 M
## 8   8 21 F
## 9   9 20 M
## 10 10 21 F
```

```
log(data[ ,2][3])
```

```
## [1] 2.9957322735539908543
```

```
data$b
```

```
##  [1] 20 21 20 21 20 21 20 21 20 21
```

```
data$b[3:6]
```

```
## [1] 20 21 20 21
```

```
data[,"b"]
```

```
##  [1] 20 21 20 21 20 21 20 21 20 21
```

```
data[data$a >=4, 1]
```

```
## [1]  4  5  6  7  8  9 10
```

Working with .csv files:

In R, we can read data from files stored outside the R environment. We can also write data into files for future use. One of the formats of files we can read and write is .csv. We use for that the commands read.csv(file_name) and write.csv(file_name).

How to get help:

The command "help.start()" will give you access to some introductory packages. If you know the specific name of the function your are looking for, you can use the comand "help()".

Packages:

R packages are collections of R functions that we may need to do specific tasks. To install a package we use the command install.package("package-name") and then we can use library(pacakge.name) or require(package.name) to be able to use it in our session.

Questions:

(1) Use R to evaluate the expressions:

(a) $\frac{(-3-2)(2-1)}{(2+3)(3-4)}$ (b) $-4(3-2)^3-10$ (c) $3(2^{3-6})+\frac{6-7}{2^3}$.

(2) Suppose that you recieve a number x. Use the if-then commands in a script that prints the sentence "positive" is the number x is greater than zero and "negative" if it is less than zero.

(3) Read the file "Stats_Grades1.csv" and count students with final grade A or $A+$.

(4) Read the file "Stats_Grades1.csv" and count students with final grade above 90.

(5) Create tables of early intervention with students in the classes "Stats_Grades1.csv" and "Stats_Grades2.csv" with midterm grades below 70.

(6) Use the built-in data base "cars" and create tables satisfying:

(a) the speed is exactly 10.
(b) the speed is at least 20.
(c) the distance is between 40 and 60.

(7) Plot together in the same set of axis, the function $f(x)=\sin(x)$, $g(x)=\cos(x)$ and $h(x)=e^x$, each of them with a different color.

(8) Plot in the same set of axis the sets of data given by the first seven squares, the first seven prime numbers and the first seven powers of 2. Use different symbolos for the points in different sets.

Class 1: Organizing the Data

Different ways to present the data: Frequency tables, histograms, bar graphs, time series and circle graphs.

A frequency table: aims to present the data in a more clear way by partitioning the data into classes or bins of equal width. A frequency table is used to display quantitative data.

The class width: for a data made of integers we calculate the class width as the smallest integer greater or equal to

$$\frac{Largest\,value - Smallest\,value}{Number\,of\,classes}.$$

The lower class limit: is the smallest value within a class. **The upper class limit:** is the highest data value that can fit in a class. **The class width:** is the difference between **The upper class limit** and **The lower class limit**.

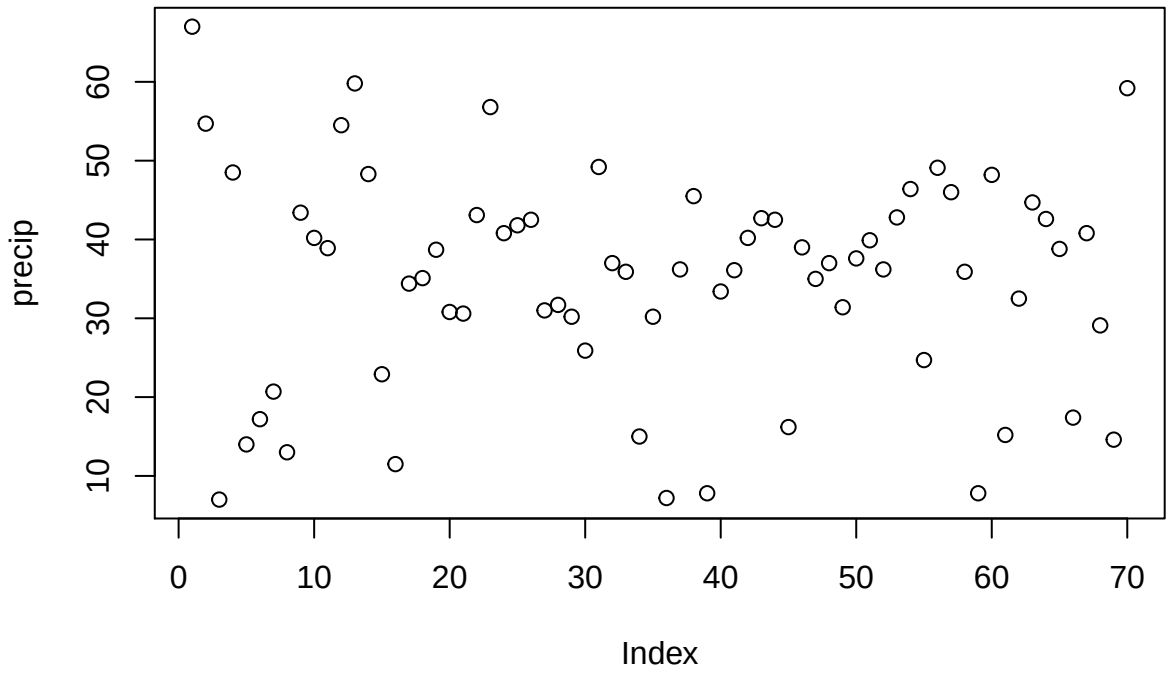
One of the main features of R is the possbility of working with **data frames**. A data frame is a two dimensional table containing a list of variables of the same number of rows or observations. In a data frame, the name of a column is not empty and the name of each row is unique. In many situations we will recieve information written in a data frame, we will proceed to read and process the information (organize it in more suitable way) and then return the results written in a new data frame. For this porpuse we will use .csv files.

A circle graph: or pie chart is used to display qualitative data. The circle graphs show the proportion of a population that shares a given quality. **Bar graphs** can be used for numerical as well as qualitative data and **Time series** represent the evolution of a measurement in time.

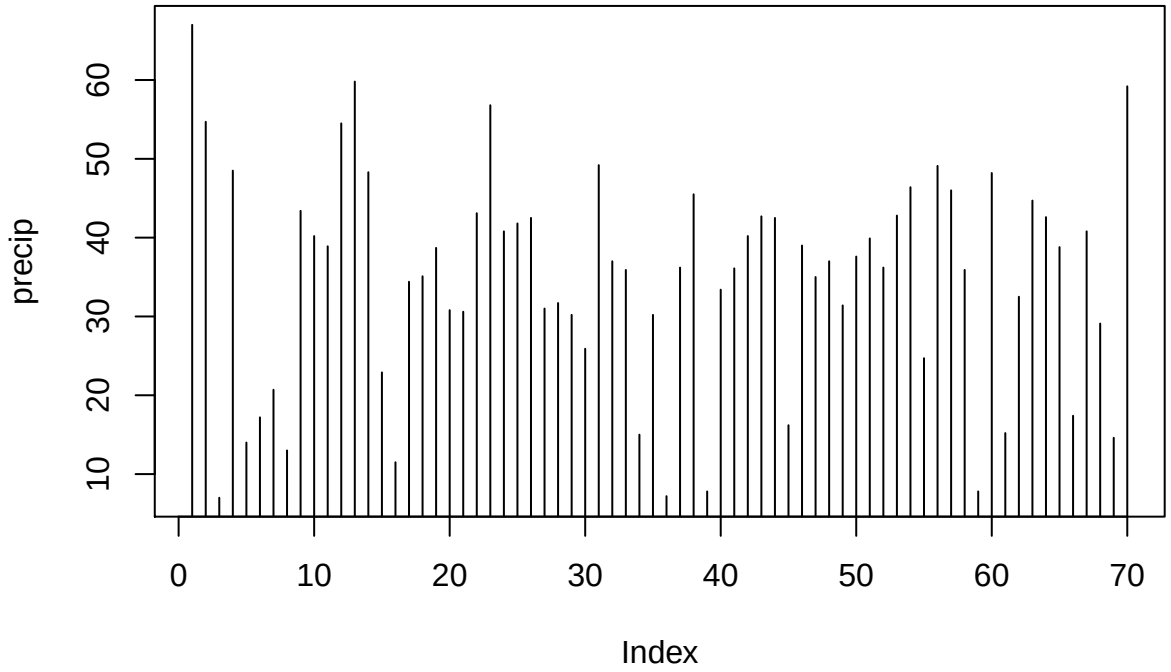
Examples:

(1) Comparing different forms of display the data. The vector precip contains average amount of rainfall (in inches) for each of 70 cities in the United States and Puerto Rico. We are going to display the data with different tools. The first three are determined by the "type", that can be l=line, p=point, h=histogram, et. The last two are frequency histograms determined by the amount of bins or classes.

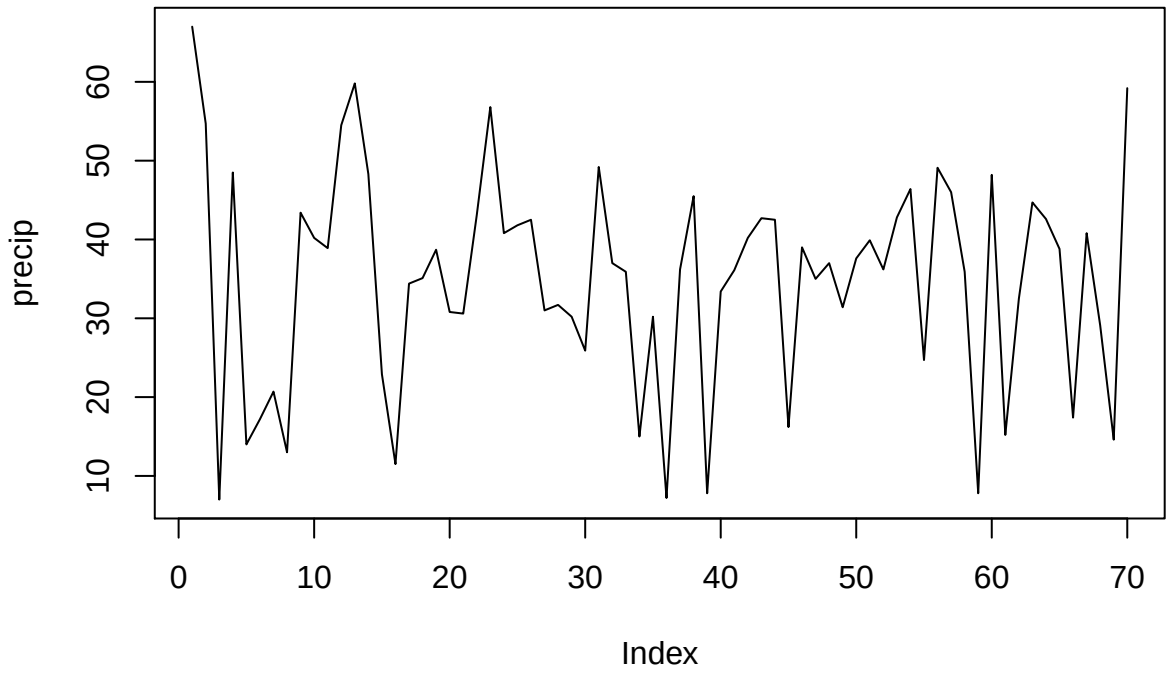
```
plot(precip, type = "p")
```



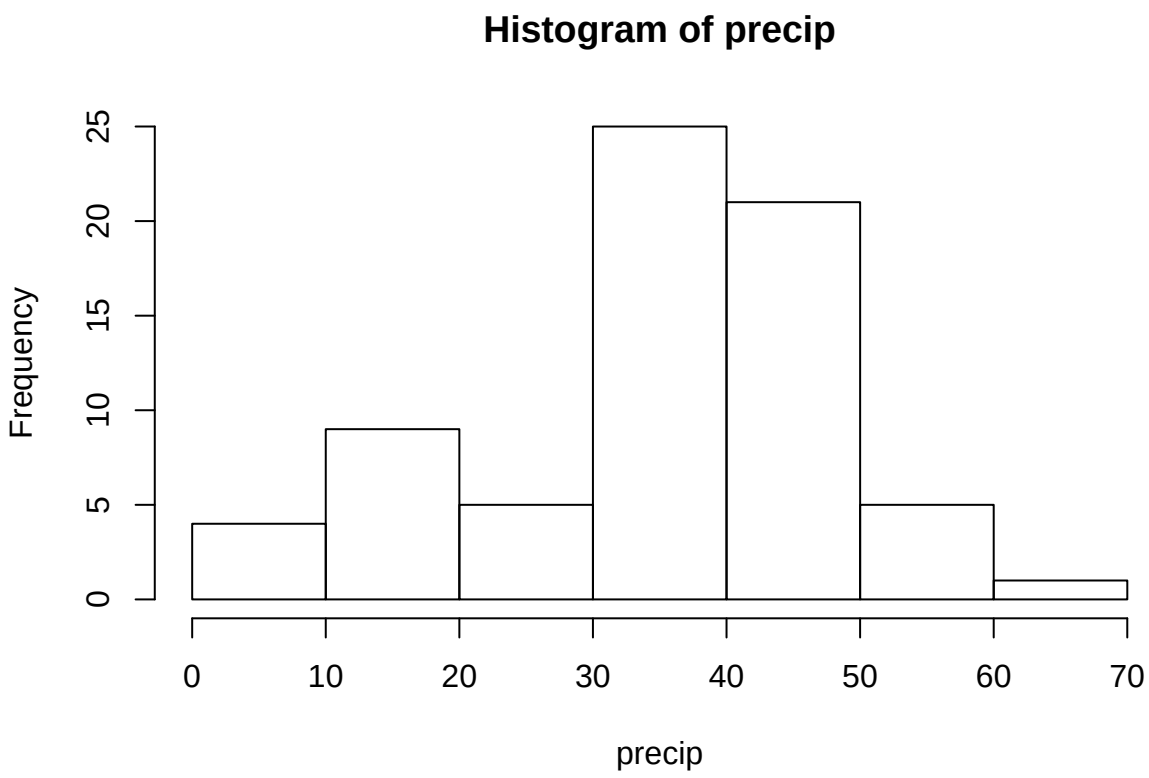

```
plot(precip, type = "h")
```



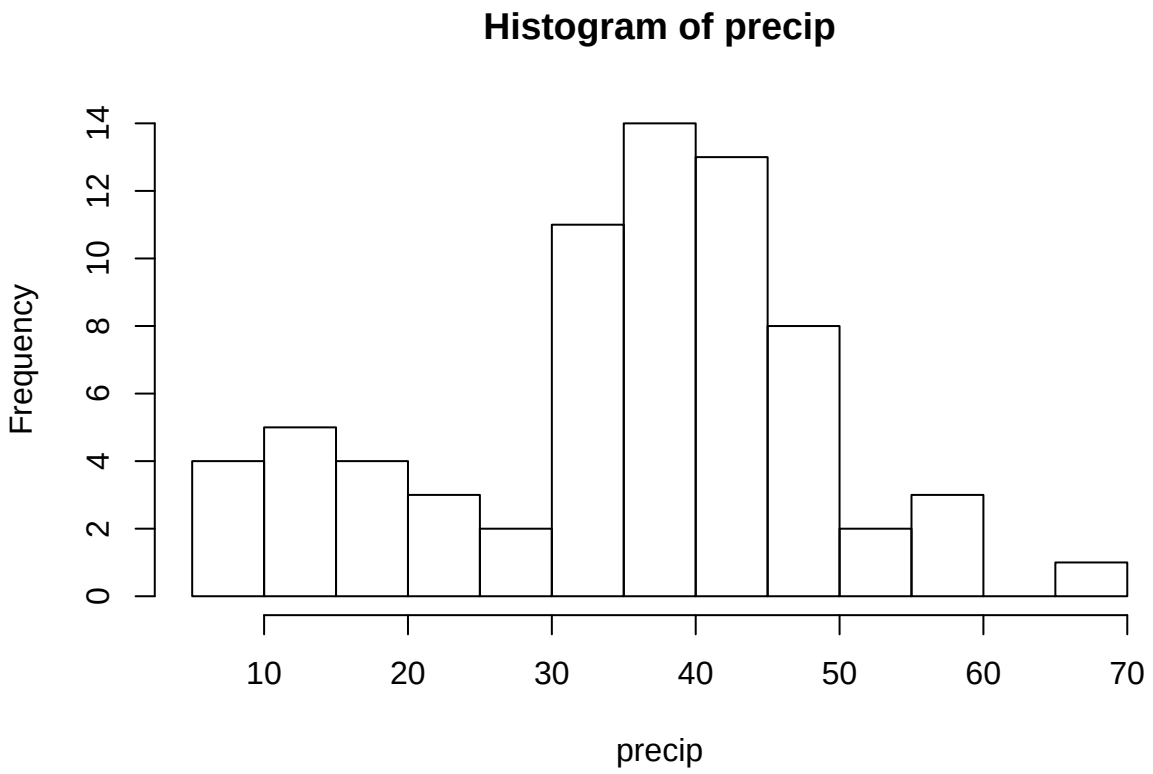

```
plot(precip, type= "l")
```




```
hist(precip)
```



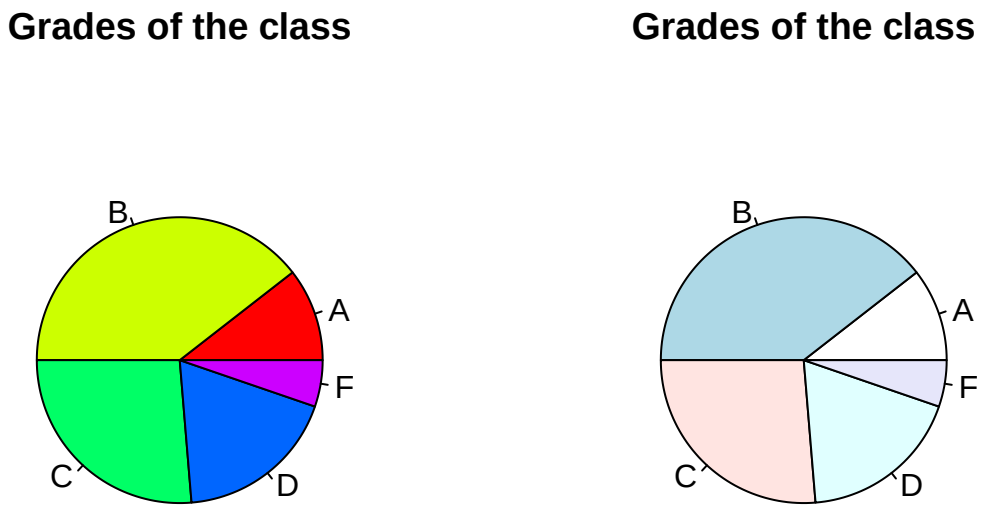

```
hist(precip, breaks = 20)
```



(2) In this second example we are going to present the grades of students in a class using a pie chart with different color pallets:

```
# Create data for the graph.
x <- c(4, 15, 10, 7, 2)
labels <- c("A", "B", "C", "D", "F")

# Plot the two charts with title and different color pallet.
par(mfrow = c(1,2))
pie(x, labels, main = "Grades of the class", col = rainbow(length(x)))
pie(x, labels, main = "Grades of the class")
```



(3) In this example, we are going to use a file named "emissions.csv"" to study the emission produced when gasoline is burned in internal combustion engines. The file contains emission levels for three major

polutans for a sample of 46 light duty engines.

```
Emission<-read.csv("Emission.csv",h=T)
```

We first learn about our data with. The structure of the data frame can be seen by using the str() function. The first several rows of our data frame can be seen by using the head() function.

```
str(Emission)
```

```
## 'data.frame':    46 obs. of  4 variables:
##  $ Engine..   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Hydrocarbon: num  0.5 0.65 0.46 0.41 0.41 0.39 0.44 0.55 0.72 0.64 ...
##  $ CO         : num  5.01 14.67 8.6 4.42 4.95 ...
##  $ NOx        : num  1.28 0.72 1.17 1.31 1.16 1.45 1.08 1.22 0.6 1.32 ...
```

```
head(Emission)
```

```
##   Engine.. Hydrocarbon    CO  NOx
## 1        1        0.50  5.01 1.28
## 2        2        0.65 14.67 0.72
## 3        3        0.46  8.60 1.17
## 4        4        0.41  4.42 1.31
## 5        5        0.41  4.95 1.16
## 6        6        0.39  7.24 1.45
```

We could calculate a whole summary of the data using summmary(). We do instead only the range of the data:

```
range <- max(Emission$Hydrocarbon)- min(Emission$Hydrocarbon)
range
```

```
## [1] 0.76
```

Then we divide in n=8 classes and draw the histogram of the organized data.

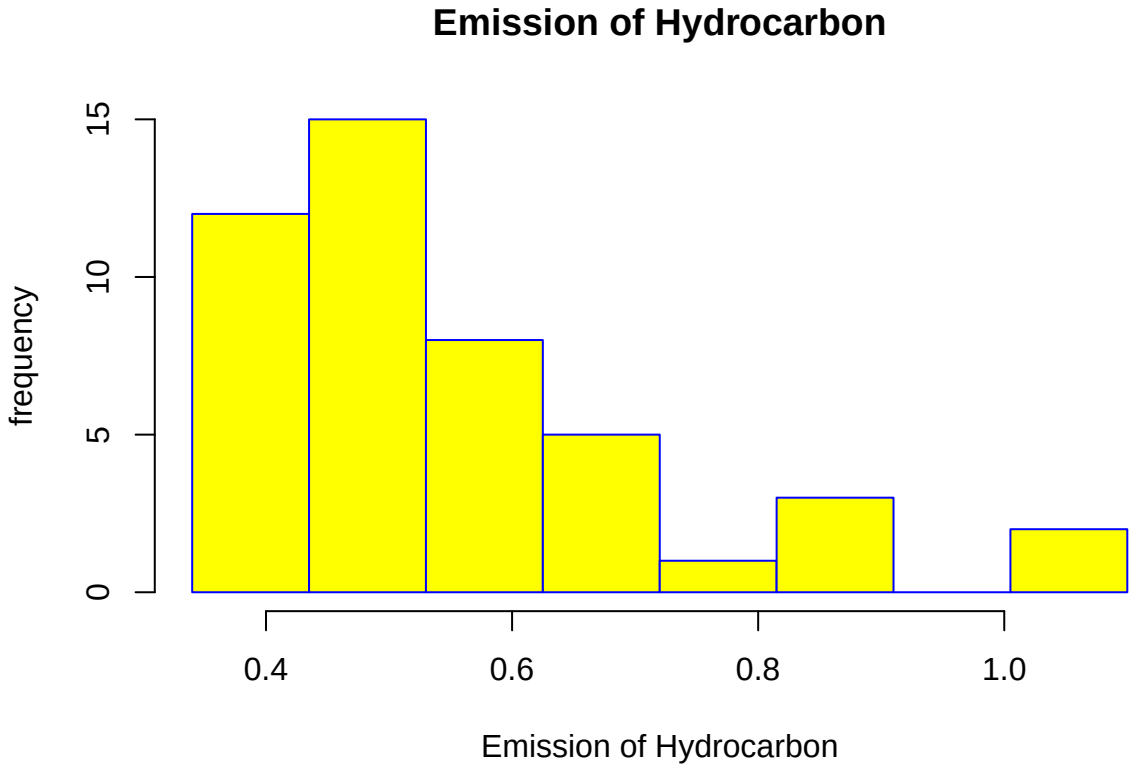
```
n <- 8
range/n
```

```
## [1] 0.095
```

```
bin <- seq(min(Emission$Hydrocarbon), max(Emission$Hydrocarbon), by= range/n)
bin
```

```
## [1] 0.340 0.435 0.530 0.625 0.720 0.815 0.910 1.005 1.100
```

```
hist(Emission$Hydrocarbon,
     breaks=bin,
     main = "Emission of Hydrocarbon",
     xlab = "Emission of Hydrocarbon",
     ylab="frequency",
     col = "yellow",
     border = "blue")
```

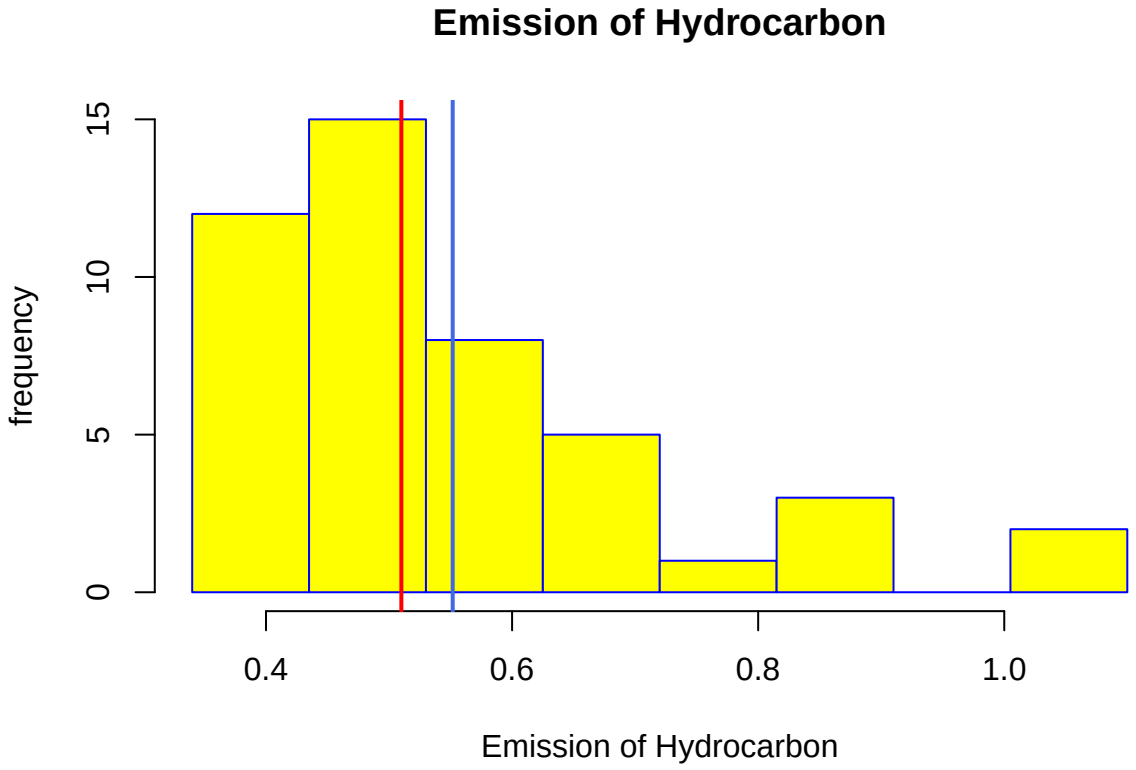


Later in this course we will learn various measures of central tendency. R has build-in functions for some of them that we can use just by calling the function. The mean, for example, represents the average of our set of data and the median will stand for the data at the center (once the data is in order). It is always a good idea to add the mean (royalblue) and the median (red) to the visualization of our data:

```
hist(Emission$Hydrocarbon,
     breaks=bin,
     main = "Emission of Hydrocarbon",
     xlab = "Emission of Hydrocarbon",
     ylab="frequency",
     col = "yellow",
     border = "blue")
abline(v = mean(Emission$Hydrocarbon),
       col = "royalblue",
       lwd = 2)
mean(Emission$Hydrocarbon)
```

```
## [1] 0.5517391
```

```
 abline(v = median(Emission$Hydrocarbon),
         col = "red",
         lwd = 2)
```




```
median(Emission$Hydrocarbon)
```

```
## [1] 0.51
```

Now, we are going to create a new table (ntable) with the information by groups. In the same way we recieved the data we will create a .csv file. The command list.files() at the end will let us check that we in fact have a new file in our directory.

```
Intervals <- cut(Emission$Hydrocarbon,bin)
table(Intervals)
```

```
## Intervals
## (0.34,0.435] (0.435,0.53] (0.53,0.625] (0.625,0.72] (0.72,0.815]
##           11           15            8            5            1
## (0.815,0.91]  (0.91,1.01]   (1.01,1.1]
##            3            0            2
```

```
freq_table <- transform(table(Intervals))
freq_table
```

```
##      Intervals Freq
## 1 (0.34,0.435]   11
## 2 (0.435,0.53]   15
## 3 (0.53,0.625]    8
## 4 (0.625,0.72]    5
## 5 (0.72,0.815]    1
## 6 (0.815,0.91]    3
## 7  (0.91,1.01]    0
## 8   (1.01,1.1]    2
```

```
ntable <-
  transform(freq_table, Rel_Freq=prop.table(Freq), Cum_Freq=cumsum(Freq))
```

```
ntable
```

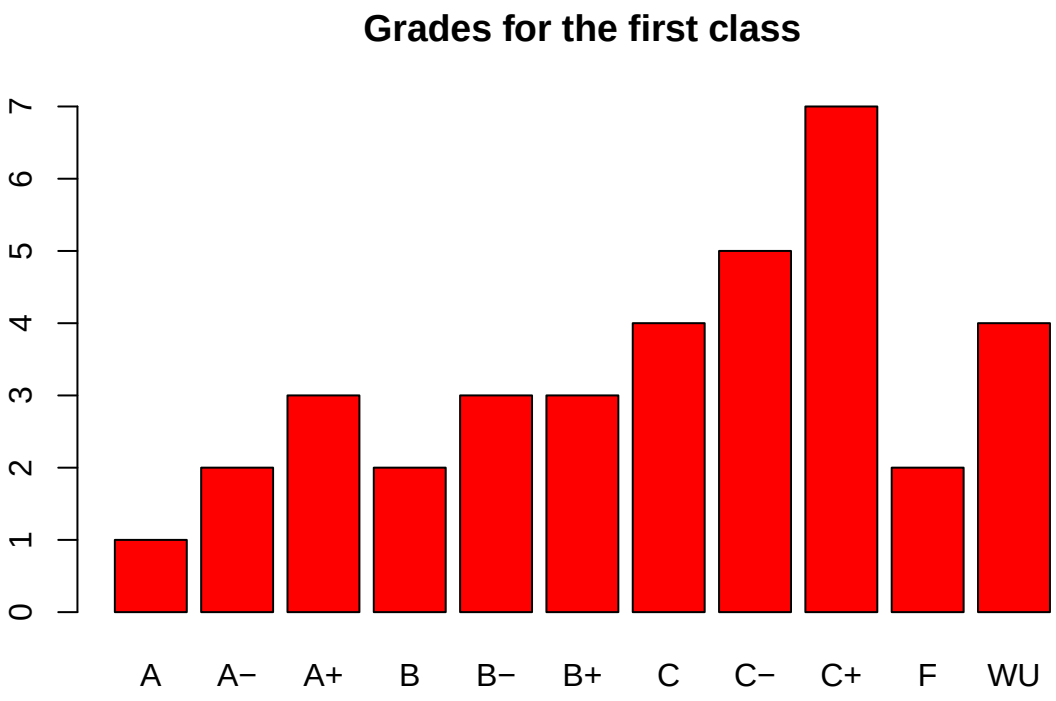
```
##      Intervals Freq   Rel_Freq Cum_Freq
## 1 (0.34,0.435]   11 0.24444444       11
## 2 (0.435,0.53]   15 0.33333333       26
## 3 (0.53,0.625]    8 0.17777778       34
## 4 (0.625,0.72]    5 0.11111111       39
## 5 (0.72,0.815]    1 0.02222222       40
## 6 (0.815,0.91]    3 0.06666667       43
## 7  (0.91,1.01]    0 0.00000000       43
## 8   (1.01,1.1]    2 0.04444444       45
```

```
write.csv(ntable, "New_table.csv")
list.files()
```

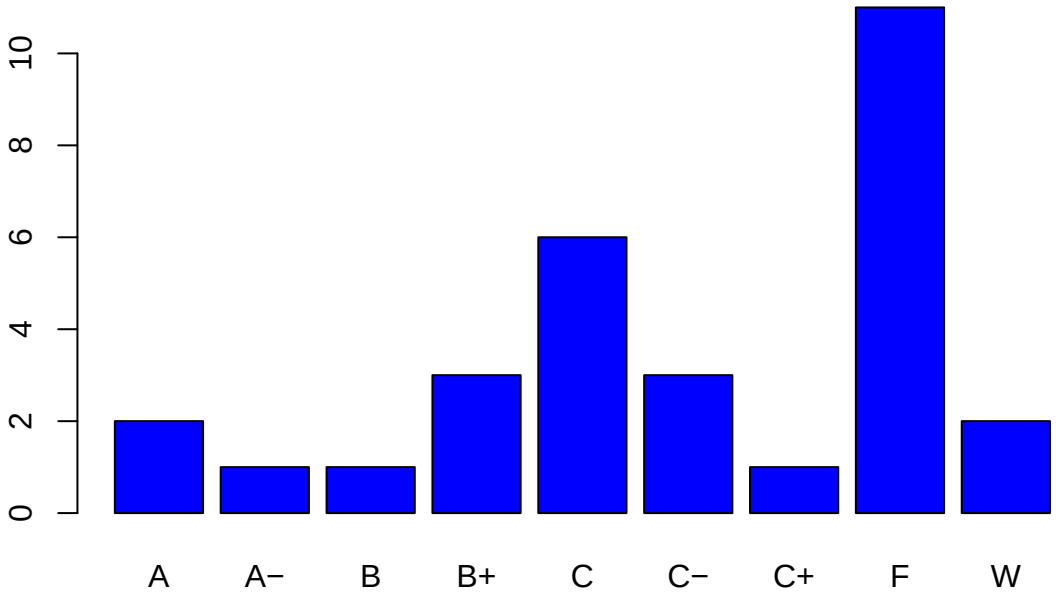
```
##  [1] "Emission_of_hydrocarbon_hist.pdf" "Emission_of_hydrocarbon_hist.R"
##  [3] "Emission.csv"                     "Grades.pdf"
##  [5] "Grades.R"                         "Grades1.R"
##  [7] "New_table.csv"                    "Organizing_the_data_files"
##  [9] "Organizing_the_data.pdf"          "Organizing_the_data.Rmd"
## [11] "p.R"                              "Stats_Grades1.csv"
## [13] "Stats_Grades2.csv"                "Untitled.R"
```

(3) In this example, we are going to read, the final grades of two clases and compare the results. The grades are read from the files "Stats_Grades1.csv" and "Stats_Grades2.csv". Statistics of grades for the first class in red and for the second class in blue.

```
Stats_Grades1=read.csv("Stats_Grades1.csv", h=T)
Freq_Grades1=table(Stats_Grades1$Final.Letter.Grade)
barplot(Freq_Grades1, col='red', main="Grades for the first class")
```

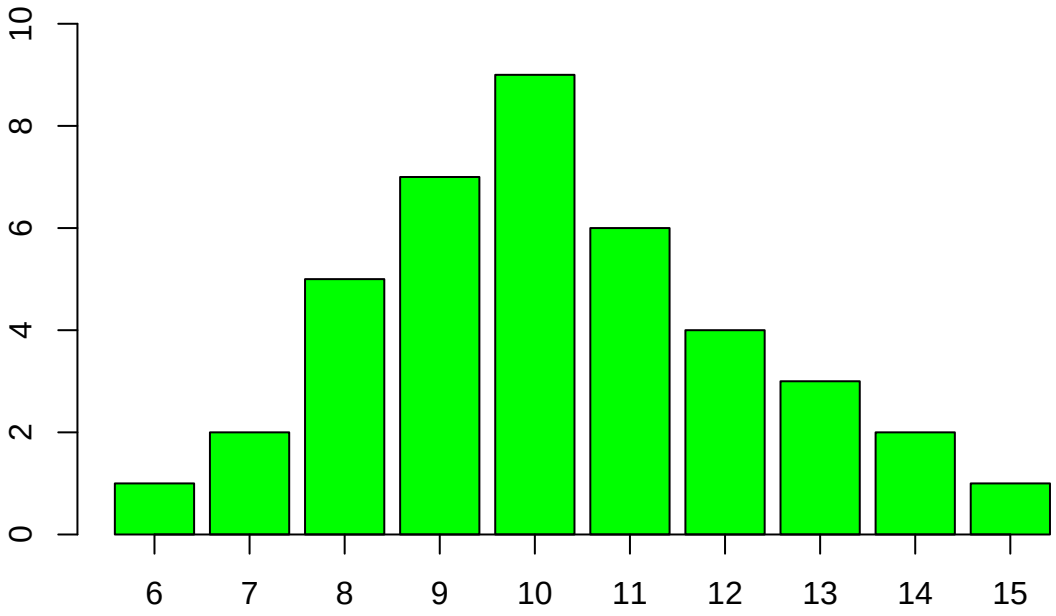



```
Stats_Grades2=read.csv("Stats_Grades2.csv", h=T)
Freq_Grades2=table(Stats_Grades2$GRADE)
barplot(Freq_Grades2, col='blue', main="Grades for the second class")
```



(3) In the next example, we recieved a frequency table, written directly on the program and not taken from a file. As before we work in .csv format.

```
dat = read.table(text = 'Var1 Freq
                 6    1
                 7    2
                 8    5
                 9    7
                 10    9
                 11    6
                 12    4
                 13    3
                 14    2
                 15    1', header = TRUE)
t = barplot(dat$Freq, ylim = c(0,10), col="green")
axis(1, at=t, labels=dat$Var1)
```



Questions:

(1) Explore the data bases "rivers" and "cars". Find appropriate ways to display the data.

(2) A group of 25 people were observed regarding their TV habits and were found to spend the following number of hours per week watching television:

30	32	34	36	36
37	39	39	41	41
42	42	43	43	44
45	45	45	46	47
47	49	49	52	53

In order to display the data in clearer form,

(a) determine the class width for four (4) classes,

(b) construct a frequency distribution showing the class limits for the four classes,

(c) in the table, show the class boundaries and the class marks,

(d) construct a histogram, labeling the class boundaries. Is the graph symmetrical, skewed left or skewed right?

(3) The following data represents the outcome of a scientific study:

15	16	18	18	22
27	28	29	29	30
32	32	33	33	34
35	35	35	36	38

In order to display the data in clearer form,

(a) determine the class width for three (3) classes,

(b) onstruct a frequency distribution showing the class limits for the four classes, in the table, show the class boundaries and the class marks,

(c) construct a histogram, labeling the class boundaries. Is the graph symmetrical, skewed to the left or skewed to the right?

(4) Construct a histogram, using 4 classes for the data in final exam grades in the file "Stats_Grades1.csv". Compare your result with the similar histogram for "Stats_Grades2.csv".

(5) Find the five numbers summary for the final exam results of students in the files "Stats_Grades1.csv" and "Stats_Grades1.csv".

Class 2: Statistics summary and measures of central tendency

Measures of central tendency: are different ways to indicate the typical or central value in a distribution of data. There are three main measures: the mode, the median and the mean.

The mode: is the single data that occurs most frequently.

The median: is the middle value of the data once the data has been arranged in order.

The mean: is the average, that is, the sum of all data values devided by the number of data values.

Measures of variation: are measures of the dispersion of the data.

The variance: σ^2 is "the average squared deviation from the mean". We use squares to prevent cancelations. To find **the standard deviation** σ we use square root to go back to the original units of measurements. In the sample standard deviation, to be able to use s as an "unbiased" estimation of σ, the sum of the squares of the deviations is divided by one less than the sample size.

Parameter	Defining formula	Computational formula
Population mean	$\mu = \frac{\sum x}{N}$	$\mu = \frac{\sum x}{N}$
Population standard deviation	$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$	$\sigma = \sqrt{\frac{\sum x^2 - (\sum x)^2/N}{N}}$

Statistic	Defining formula	Computational formula
sample mean	$\bar{x} = \frac{\sum x}{n}$	$\bar{x} = \frac{\sum x}{n}$
sample standard deviation	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$	$s = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}}$

The quartiles: Divide the data in four equal parts. **The interquartile range IQR:** is the difference $Q_3 - Q_1$ between the third and first quartiles. It defines how spread out is the center 50 % of the data.

Examples:

(1) R provides generic functions (build-in functions) for the most important statistical functions. We are going to develope a general example of the summary of numbers for a sample of data. We compute the mean, median, mode, quartiles and standard deviation for a sample vector x.

```
x <- c(3,5,12,17,23,48,5,12,12)
```

The mean $\bar{x}$ of the vector x is:

```
mean(x)
```

```
## [1] 15.22222
```

The median can be obtained simply by doing:

```
sort(x)
```

```
## [1]  3  5  5 12 12 12 17 23 48
```

```
median(x)
```

```
## [1] 12
```

The quartiles can be obtained as particular way to define a vector of probabilities using the function "quantile"

```
quantile(x)
```

```
##   0%  25%  50%  75% 100%
##    3    5   12   17   48
```

```
quantile(x, probs = c(0, 0.25, 0.50,.75))
```

```
##  0% 25% 50% 75%
##   3   5  12  17
```

The sample standard deviation s is obtained as:

```
sd(x)
```

```
## [1] 13.81826
```

The variance s^2

```
var(x)
```

```
## [1] 190.9444
```

The interquartile range as measure of the central portion of the data:

```
IQR(x)
```

```
## [1] 12
```

R does not have a build-in function for the mode. We are going to create and use the function "getmode". Look for help to find the meaning of the command "unique".

```
getmode <- function(x) {
  uniqx <- unique(x)
  uniqx[which.max(tabulate(match(x, uniqx)))]
}
```

Obtain the mode:

```
getmode(x)
```

```
## [1] 12
```

(2) The use of the command summary. We could have obtained some of the information above by using the command "summary":

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    5.00   12.00   15.22   17.00   48.00
```

(3) We can obtain the mean and standard deviation of each row of a given matrix. In the following example, we construct a 10×20 matrix of random numbers and compute the mean and standar deviations of the rows. The results are directly written to a pair of tables at the end.

```
m = 10;  n = 20;  mu = 0;  sigma = 15
x = rnorm(m*n, mu, sigma)
MAT = matrix(x, nrow=m)
x.bar = rowMeans(MAT);  s = apply(MAT, 1, sd)
X.bar=table(x.bar)
S=table(s)
write.csv(X.bar, "table_of_means.csv")
write.csv(S,"table_of_standard_deviations")
list.files()
```

```
## [1] "Grades.R"                    "Stats_Grades1.csv"
## [3] "Stats_Grades2.csv"           "Summary_Statistics.pdf"
## [5] "Summary_Statistics.R"        "Summary_Statistics.Rmd"
## [7] "table_of_means.csv"          "table_of_standard_deviations"
```

Questions:

(1) Calculate the range, mean, median, first and third quartiles, interquartile range, mode, variance, and standard deviation for the following population data.

47 59 50 56 56 51 53 57 52 49

(2) Find the mean, the range, and the standard deviation for the following set of sample data.

10 9 12 11 8 15 9 7 8 6

(3) Giving the data in the files "Stats_Grades1.csv" and "Stats_Grades2.csv". Compare the mean final grade in both classes. Find also the standard deviation for each of the two sets?

(4) Determine the range and the sample standard deviation of the following data:

x	f
10.3	7
22	12
38.5	5
43.2	2

(a) Include a frequency bar plot of the data.

(5) A consumer testing service obtained the following mileage (in miles per gallon) in five test runs for three different types of compact cars:

	First Run	**Second Run**	**Third Run**	**Fourth Run**	**Fifth Run**
Car A	28	32	28	34	30
Car B	31	31	29	29	31
Car C	32	29	28	32	30

(a) If the manufacturer of Car A wants to advertise that their car performed the best in this test, which measure of central tendency (mean, median or mode) should be used to support their claim?

(b) Which measure should the manufacturer of Car B use to claim that their car performed best, mean median or mode?

(c) Which measure should the manufacturer of Car C use to support a similar claim?

(6) In a class of 40 students, the grade of a particular student is the 90-th percentile. How many students score similar or more? Can she be sure that she passed the class?

(7) In a set of data, what percent of the data is between Q_1 and Q_2 approximately?

(8) Find the quartiles and the interquartile range for the data files "Stats_Grades1.csv" and "Stats_Grades2.csv". Write the results together on a table of two columns.

Class 3: Correlation and Regression

A scatter diagram: is a graph in which data pairs are plotted as individual points in a system of Cartesian coordinates. The variable x is called the explanatory or independent variable and the variable y is the response variable or dependent variable.

A linear regression: Finds a model of the response variable y as a linear function of the independent variable x.

High or Strong correlation: when the points are close to a straight line.

Positive linear correlation: The variables x and y are said to have positive linear correlations if low values of x are associated to low values of y and high values of x correspond to high values of y.

Negative linear correlation: The variables x and y are said to have negative linear correlations if low values of x are associated to high values of y and high values of x correspond to low values of y.

The correlation coefficient or **Pearson's correlation coefficient** is a numerical measure of the linear relation between two variables. It admits a geometric interpretation as the cosine of the angle between the vectors $x - \bar{x}$ and $y - \bar{y}$ and is therefore a number r such that $-1 \leq r \leq 1$. The $r = 1$ indicates **perfect positive correlation** (the points on the plot lie on a line of positive slope) and $r = -1$ is an indication of **perfect negative correlation** (points (x, y) are on a line of negative slope). On the other hand $r \approx 0$ will be an indication of **little or no linear correlation** whatsoever between x and y.

Statistic	Defining formula	Computational formula
Coefficient of linear correlation r	$r = \frac{1}{n-1} \sum \frac{(x-\bar{x})}{s_x} \frac{(y-\bar{y})}{s_y}$	$r = \frac{n \sum xy - (\sum x)(\sum (y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$

Least squares line: the line such that the sum of squares of the difference of the y-values between the line and the points is as small as possible.

Fact: The least squares lines of equation $y = bx + a$ may or may not pass by any of the points, but it always contains the point

$$(\bar{x}, \bar{y}) = \left(\frac{\sum x}{n}, \frac{\sum y}{n} \right).$$

On the other hand the slope b is giving by the formula:

$$b = \frac{n \sum xy - (\sum x)(\sum (y)}{n \sum x^2 - (\sum x)^2},$$

and we have an equation for the least squares line of the form $y - \bar{y} = b(x - \bar{x})$.

The coefficient of determination: is the square r^2 of the coefficient of correlation r. It reflects what portion of the variance of the response variable y can be explained by the variance of the independent variable x and the model $\hat{y} = a + bx$. The proportion $1 - r^2$ of the variance cannot be explained using the model.

Examples using R:

(1) In our first example, we are going to study the relation between grades of students in their midterm and final exams.

We start by reading the file with the points. Our first line, list all the files in the directory, just to make sure we are in the correct directory and we have a file to read.

```
list.files()
```

```
##  [1] "a.R"                                "b.pdf"
##  [3] "Correlation_and_Regression.pdf"     "Correlation_and_Regression.Rmd"
##  [5] "Grades_linear_regression.R"         "Plotting math functions.R"
##  [7] "Points1.csv"                        "Points2.csv"
##  [9] "Points3.csv"                        "Polynomial_fitting_of_points.pdf"
## [11] "Polynomial_fitting_of_points.Rmd"   "Polynomial_fitting_points.html"
## [13] "Polynomial_fitting_points.R"        "Polynomial_fitting_points.Rmd"
## [15] "Polynomial_fitting_points3.R"       "Regression_Line.R"
## [17] "Stats_Grades1.csv"                  "Stats_Grades2.csv"
## [19] "Vector_and_Regression_Line.R"
```

```
points3<- read.csv("Points3.csv", h=T)
str("Points3.csv")
```

```
##  chr "Points3.csv"
```

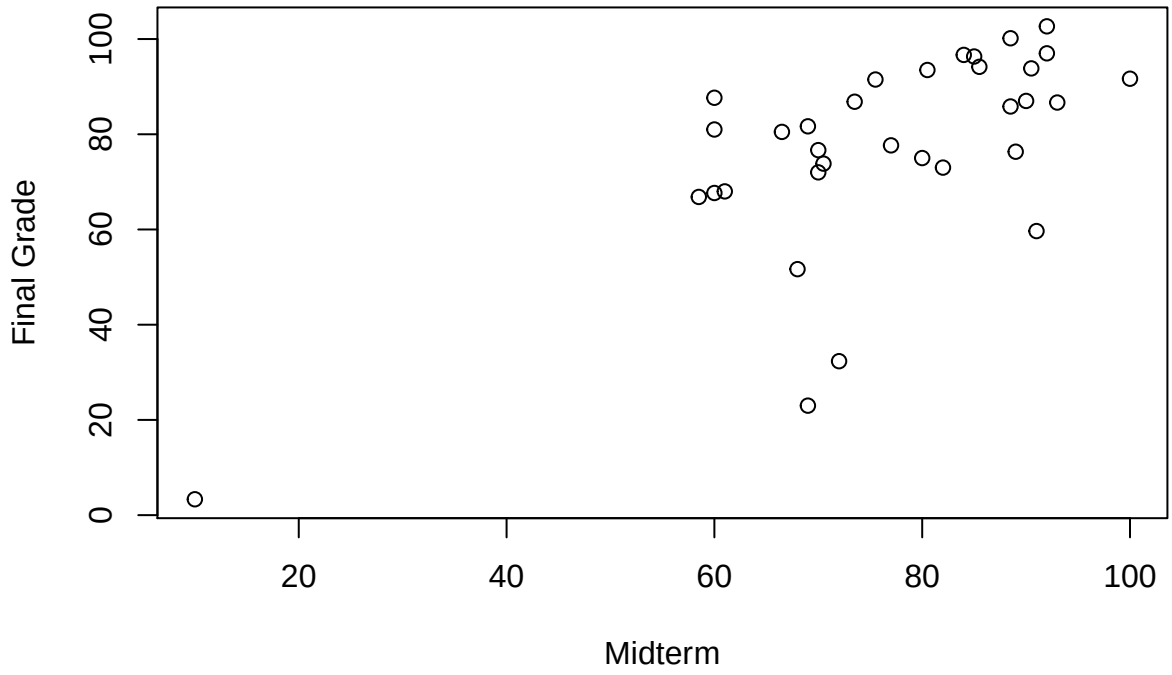
```
points3 <- as.data.frame(points3)
points3$MIDTERM
```

```
##  [1]  60.0  73.5  88.5  80.5  70.5  90.0  82.0  92.0  77.0  85.0  84.0
## [12]  58.5  61.0  90.5  92.0  80.0  70.0  91.0  10.0  66.5  85.5  60.0
## [23]  89.0  68.0  70.0 100.0  69.0  72.0  88.5  75.5  93.0  60.0  69.0
```

```
points3$Final.Grade
```

```
##  [1]  67.666667  86.833333 100.166667  93.500000  73.833333  87.000000
##  [7]  73.000000 102.666667  77.666667  96.333333  96.666667  66.833333
## [13]  68.000000  93.833333  97.000000  75.000000  76.666667  59.666667
## [19]   3.333333  80.500000  94.166667  87.666667  76.333333  51.666667
## [25]  72.000000  91.666667  23.000000  32.333333  85.833333  91.500000
## [31]  86.666667  81.000000  81.666667
```

```
plot (points3$MIDTERM,points3$Final.Grade,
      main="Midterm vs. Final Grade",
      xlab="Midterm",
      ylab="Final Grade")
```

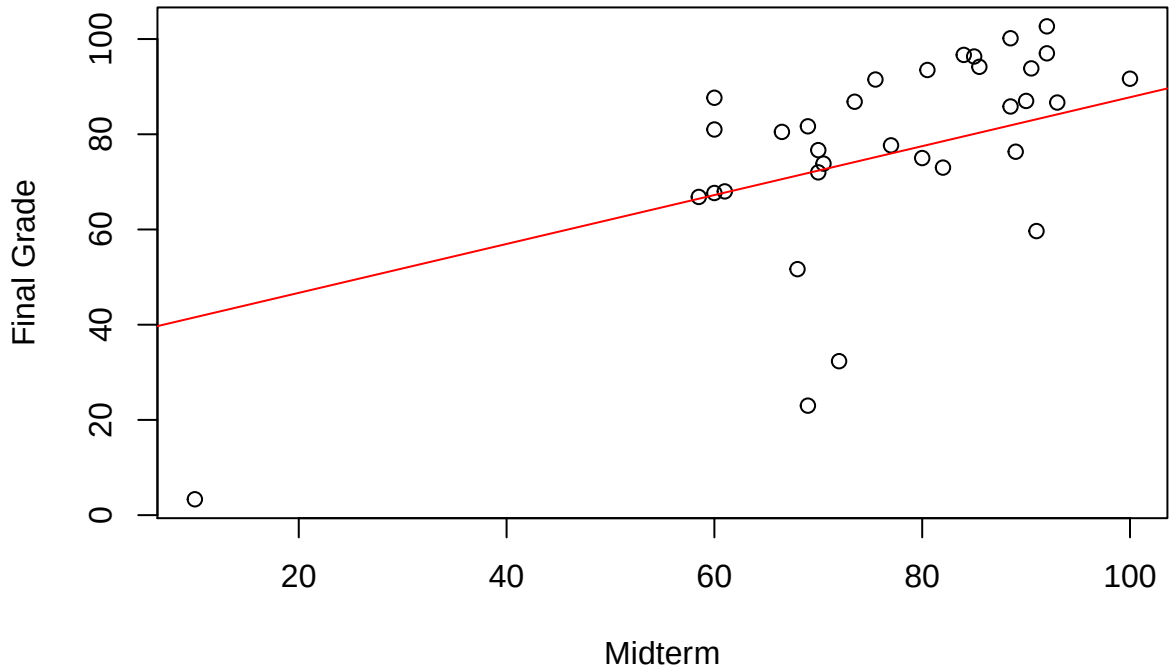


And finally, the line that best approximate:

```
plot (points3$MIDTERM,points3$Final.Grade,
      main="Midterm vs. Final Grade",
      xlab="Midterm",
      ylab="Final Grade")
lm(points3$MIDTERM ~ points3$Final.Grade)
```

```
##
## Call:
## lm(formula = points3$MIDTERM ~ points3$Final.Grade)
##
## Coefficients:
##         (Intercept)  points3$Final.Grade
##             36.4169               0.5136
```

```
abline(lm(points3$MIDTERM ~ points3$Final.Grade),col = "red")
```




```
cor(points3$MIDTERM,points3$Final.Grade)
```

```
## [1] 0.6863553
```

Interpretation: we observe a moderate positive correlation between the grades of students in the Midterm exam and the fina exam.

(2) As our second example we are given two vectors and we obtain the covariance, the coefficient of linear correlation, the plot of the points as well as the coefficients m,b of the least squares line.

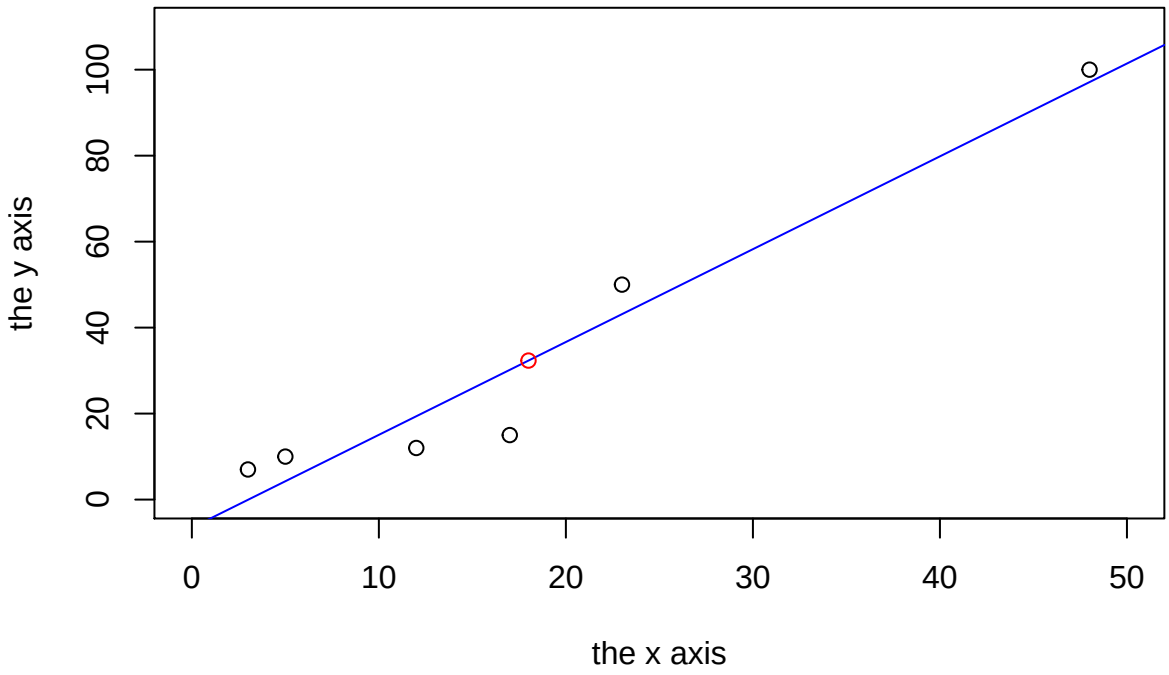
```
x <- c(3,5,12,17,23,48)
y <- c(7,10,12,15,50,100)
cov(x,y)
```

```
## [1] 585.6
```

```
cor(x,y)
```

```
## [1] 0.968143
```

```
plot(x,y, xlab="the x axis", ylab="the y axis", xlim=c(0,50),ylim=c(0,110))
abline(lsfit(x, y), col=4)
points(mean(x), mean(y), col="red")
```




```
lsfit(x,y)$coef
```

```
## Intercept         X
## -6.533923  2.159292
```

Interpretation: In this example the correlation is strong and positive.

(3) As our third example we compute the regression line and linear correlation between the height and bodymass of several individuals.The plot also shows how the line passes by the point (mean bm, mean h)

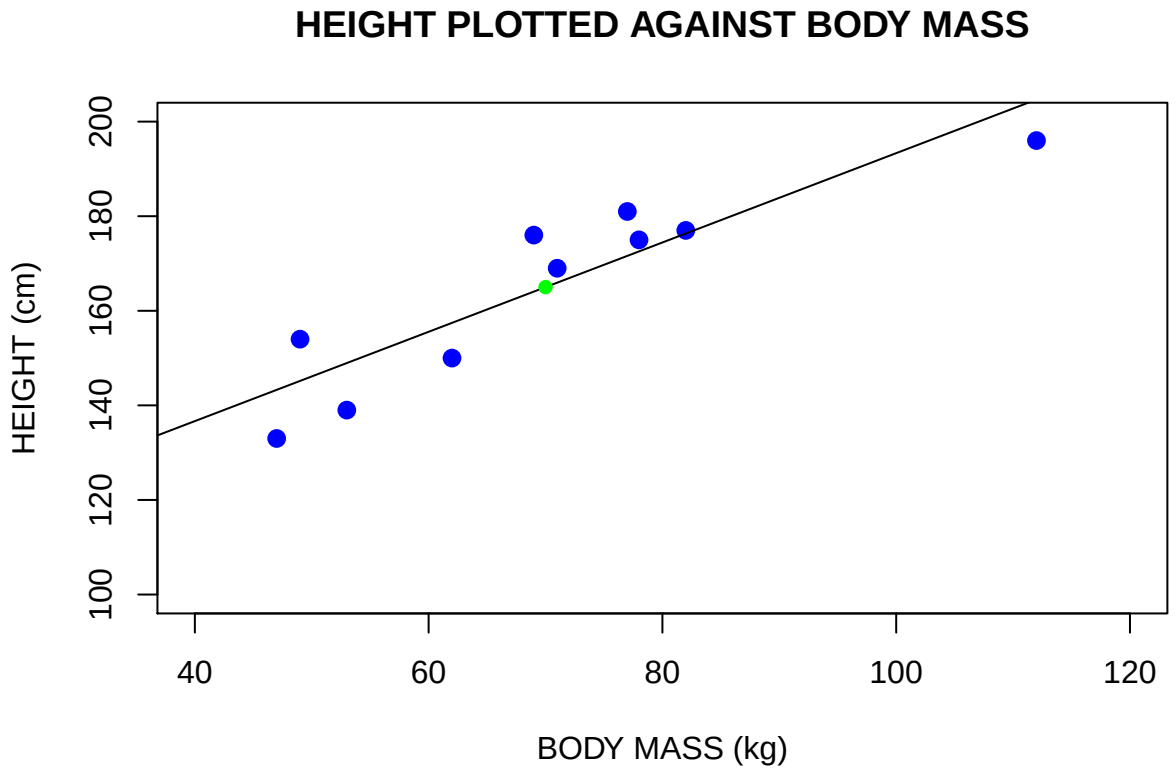
```
height <- c(177, 154, 139, 196, 133, 176, 181, 169, 150, 175)
bodymass <- c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
```

We plot the points together with the least squares line obtained with the linar model "lm" comand. We will include in the graph, the point $(\bar{x}, \bar{y})$ that always belong to the regression line.

```
plot(bodymass, height,
     pch = 16,
     cex = 1.3,
     col = "blue",
     main = "HEIGHT PLOTTED AGAINST BODY MASS",
     xlab = "BODY MASS (kg)",
     ylab = "HEIGHT (cm)",
     xlim=c(40,120), ylim=c(100,200))
lm(height ~ bodymass)
```

```
##
## Call:
## lm(formula = height ~ bodymass)
##
## Coefficients:
## (Intercept)     bodymass
##     98.8912       0.9444
```

```
abline(lm(height ~ bodymass))
hm=mean(height); bmm=mean(bodymass)
points(bmm,hm , pch=16, col="green")
```




```
cor(bodymass,height)
```

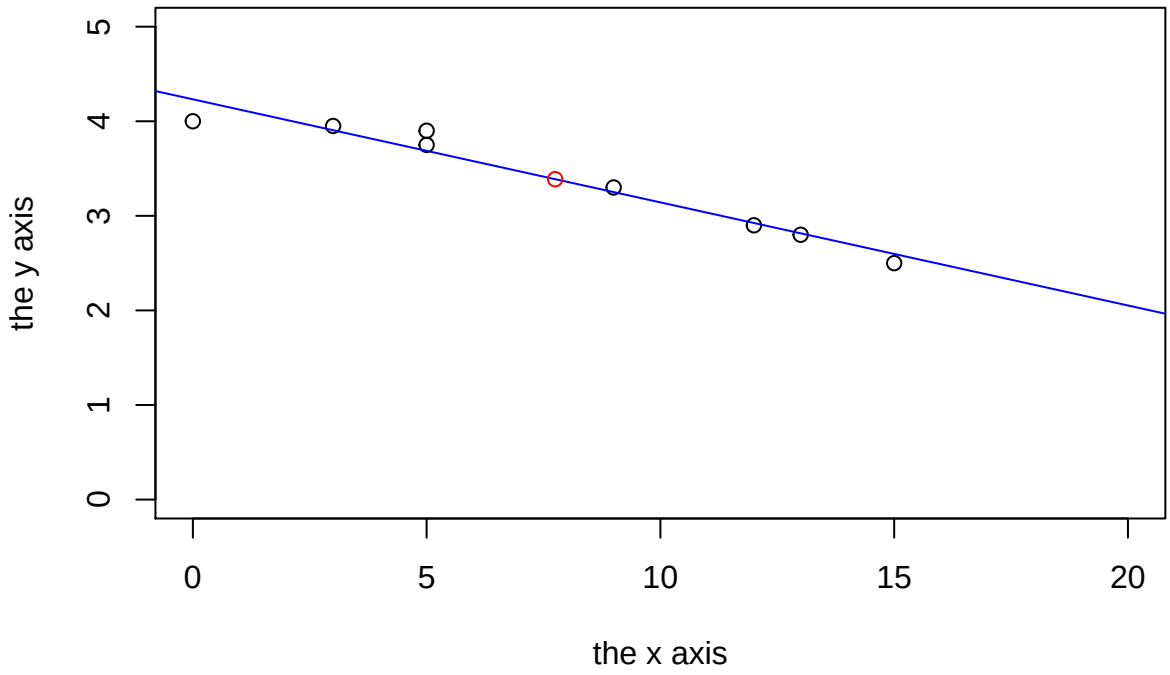
```
## [1] 0.9049716
```

Interpretation: in this third example, the correlation between the bodymass and the height is still strong, but not as strong though as our second example. The correlation in the three examples is positive. High values of x correspond mostly to high values of y and viceversa. The cloud of points around the point $(\bar{x}, \bar{y})$ looks "thin" and with "positive slope".

(4) To visualize negative correlation, consider the following table relating amount of hours per week wathching TV and the GPA for a group of eight students:

Hours	0	15	3	5	5	13	12	9
GPA	4	2.5	3.95	3.90	3.75	2.80	3.90	3.30

```
x <- c(0,15,3,5,5,13,12,9)
y <- c(4,2.5,3.95,3.90,3.75,2.80,2.90,3.30)
plot(x,y, xlab="the x axis", ylab="the y axis", xlim=c(0,20),ylim=c(0,5))
abline(lsfit(x, y), col=4)
points(mean(x), mean(y), col="red")
```




```
cor(x,y)
```

```
## [1] -0.9758011
```

Interpretation: We can see now that high values of x correspond to smaller values of y and our cloud of points is "decreasing" or close to form a line with negative slope. We have a strong negative correlation in this situation. The point in red represents, as in the second example, the point $(\bar{x}, \bar{y})$ always contained in the least squares line.

Questions

(1) Provide graphical examples of points (x, y) illustrating:

(a) A strong positive linear correlation.
(b) A strong negative linear correlation.
(c) Not linear correlation whatsoever.

(2) Compute for each of the examples in question (1) the coefficient r of linear correlation.

(3) The manager of a salmon cannery suspects that the demand for her product is closely related to the disposable income of her target region. To test out this hypothesis she collected the following data for five different target regions, where x represents the annual disposable income for a region in millions of dollars and y represents sales volume in thousands of cases.

x	y
10	1
20	3
40	4
50	5
30	2

(a) Draw the scatter graph of this set of data.
(b) Compute the correlation coefficient r.
(c) Compute the coefficient of determination r^2.

(d) Find and graph the least square line.
(e) If a region has disposable annual income $\$25,000,000$ what is the predicted sales volume?

(4) The following table represents two sets of data:

x	y
3	4.2
5	4
12	3.5
17	3.8
23	2.4
48	.5

(a) Draw the scatter graph of this set of data.
(b) Based on the graph do you expect the correlation coefficient to be positive, negative or close to zero?
(c) Compute the coefficient of linear correlation r.
(d) Compute the coefficient of determination r^2.
(e) Find and graph the least square line.
(f) What will be the y predicted by the model for $x = 30$?

(5) Do a regression analysis for the variables Midterm Grade vs. Final Grade for the data in the files "Stats_Grades1.csv" and "Stats_Grades2.csv". Which of the two sections shows a stronger correlation between the two variables.

Class 4: Introduction to Probability Theory

A statistical experiment: is any random activity that results in a definite outcome.

An event: is a set of one or more outcomes of a statistical experiment or observation. **A simple event:** is one particular outcome of a statistical experiment.

The sample space: is the set Ω of all simple events. The set of events $\mathcal{F}$ is a collection of subsets of Ω.

Probability: is a numerical measure, denoted $P(A)$, between 0 and 1 that describes the likelihood that an event A will occur. The higher the probability of an event, the more certain that the event will occur. If $P(A) = 1$, the event A is certain to occur and if $P(A)$=0, the event A is certain not to occur (impossible).

The complement of the event A: is the event that A will not occur. It is denoted by A^c

Mutually exclusive events: Two events are mutually exclusive if they **cannot** occur together. That is when

$$P(A\,and\,B) = 0.$$

Addition rule for mutually exclusive events: states that for A and B mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B).$$

General addition rule: For any events (not necessarily mutually exclusive) we have:

$$P(A\,\text{or}\,B) = P(A) + P(B) - P(A\,and\,B).$$

We choose the collection $\mathcal{F}$ in such a way that we can always take **complements** and **unions** (**or**). Also, the whole sample space Ω is always in $\mathcal{F}$. A collection $\mathcal{F}$ of sets with these properties is called a **"tribe"** and it represents the collection of measurable events.

A probability space: Is a triple $(\Omega, \mathcal{F}, P)$ consisting of a sample space Ω, a space of measurable events $\mathcal{F}$ and a probability assignment $P\colon \mathcal{F} \to [0, 1]$ in such a way that

1. The probability of the total space $P(\Omega) = 1$.
2. For any event A, the probability of the complement is $P(A^c) = 1 - P(A)$.
3. For any two mutually exclusive events A, B, $P(A\,\text{or}\,B) = P(A) + P(B)$.

A probability assignment based on equally likely outcomes: uses the formula

$$P(A) = \frac{number\,of\,favorable\,outcomes}{total\,number\,of\,outcomes}.$$

Two events are independent: if the occurrence or nonoccurrence of one event does not change the probability that the other event will occur.

The multiplication rule for independent events: states that for A and B independent

$$P(A \text{ and } B) = P(A).P(B).$$

The conditional probability $P(A|B)$: denotes the probability that event A will occur given that event B already occurred. For events A, B that are not independent we have $P(A|B) \neq P(A)$ or $P(B|A) \neq P(B)$.

In general $P(A|B) \neq P(B|A)$.

The general multiplication rule: For any events A and B (not necessarily independent) we have:

$$P(A \text{ and } B) = P(A).P(B|A), \quad P(A \text{ and } B) = P(B).P(A|B).$$

The conditional probability $P(A|B)$: when $P(A) \neq 0$ can be found using the formula:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

Baye's theorem: The conditional probability of an event can be expressed in terms of prior knowledge of conditions that may be related to the event:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

The complement of the union: can be found with the formula:

$$P(\text{neither } A \text{ nor } B) = 1 - P(A) - P(B) + P(A \text{ and } B).$$

Total probability formula: For a partition of the sample space in events $B_1, B_2, \ldots, B_n$, we have

$$P(A) = \sum_{i=1}^{n} P(A \text{ and } B_i) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n).$$

The study of probability in R is done easier with the help of the package "prob". We need to download the package "prob" and then use the command "require" to be able to use it.

Examples:

(1) We can simulate rolling a fair die or tossing a coin:

```
require(prob)
```

```
## Loading required package: prob
```

```
## Loading required package: combinat
```

```
##
## Attaching package: 'combinat'
```

```
## The following object is masked from 'package:utils':
##
##     combn
```

```
## Loading required package: fAsianOptions
```

```
## Loading required package: timeDate
```

```
## Loading required package: timeSeries
```

```
## Loading required package: fBasics
```

```
## Loading required package: fOptions
```

```
##
## Attaching package: 'prob'
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, union
```

```
tosscoin(1)
```

```
##   toss1
## 1     H
## 2     T
```

```
rolldie(1)
```

```
##   X1
## 1  1
## 2  2
## 3  3
## 4  4
## 5  5
## 6  6
```

(2) We can work with cards, using the command "cards" and find the union and intersection of sets (events). The optional parameter "makespace" will define the probaility space associated with our experiment. To find the associated probability R uses the formula for the probability assignment based on equally likely outcomes.

```
S <- cards(makespace=TRUE)
A <- subset(S, suit == "Heart")
B <- subset(S, rank %in% 6:8)
C=union(A,B)
D=intersect(A,B)
Prob(A)
```

```
## [1] 0.25
```

```
Prob(B)
```

```
## [1] 0.2307692
```

```
Prob(C)
```

```
## [1] 0.4230769
```

```
Prob(D)
```

```
## [1] 0.05769231
```

(3) The computation of conditional probabilities can be used to show for example that $P(A|B)$ is in general not equal to $P(B|A)$.

```
S <- rolldie(2, makespace = TRUE)
head(S)
```

```
##   X1 X2      probs
## 1  1  1 0.02777778
## 2  2  1 0.02777778
## 3  3  1 0.02777778
## 4  4  1 0.02777778
## 5  5  1 0.02777778
## 6  6  1 0.02777778
```

```
Prob(S, X1==X2, given = (X1 + X2 >= 6) )
```

```
## [1] 0.1538462
```

```
Prob(S, X1+X2 >= 6, given = (X1==X2) )
```

```
## [1] 0.6666667
```

(4) We place 15 balls in an urn, among them, 8 are red, 4 are yellow and the rest are green. What is the probability of selecting a green ball? What is the probability of selecting two green balls in a row if we work without replacement?

```
L <- rep(c("red","yellow","green"), times = c(8,4,3))
M1 <- urnsamples(L, size = 1, replace = FALSE, ordered = TRUE)
N1 <- probspace(M1)
Prob(N1, isrep(N1, "green", 1))
```

```
## [1] 0.2
```

```
L <- rep(c("red","yellow","green"), times = c(8,4,3))
M2 <- urnsamples(L, size = 2, replace = FALSE, ordered = TRUE)
N2 <- probspace(M2)
Prob(N2, isrep(N2, "green", 2))
```

```
## [1] 0.02857143
```

Questions

(1) Given $P(E^C) = 0.3$, $P(F) = 0.35$, and $P(F|E) = 0.25$ find

(a) $P(E \text{ and } F)$
(b) $P(E \text{ or } F)$
(c) $P(E|F)$.

(2) Two dice are rolled. Find the probability of the following events:

(a) Both numbers are 6.
(b) The first dice gives 5 and the second 6.
(c) There is one 5 and one 6.
(d) The sum is equal to 10.
(e) Both are 6 or the sum 10.
(f) The sum is more than 5 but less than 8.
(g) Both numbers are even.

(3) An urn contains three yellow, four green, and five blue balls. Two balls are randomly drawn without replacement. Find the probability of the following events:

(a) Both balls are blue.
(b) The first ball is green and the second yellow.
(c) There is one green and one yellow ball.

(4) Repeat the previous exercise but now assume that the balls are drawn with replacement.

(5) Three cards are randomly drawn from a standard 52 card deck without replacement. Find the probability of the following events:

(a) All cards are red.
(b) There are two red and one black card.
(c) All cards are spades.
(d) There is one spade, one club, and one diamond.
(e) All cards are aces.
(f) Two cards are aces and one card is a king.

(6) Most of the time, a medical test is able to correctly indicate if a person has a condition. However, some of the time, there are false positives (it indicates the condition is present when it is not) or false negatives (it indicates the condition is not present when it is there). Use the table below to determine the probabilities for a randomly selected person from the population.

	condition present	condition not present	row total
Test Result +	125	10	135
Test Result −	15	50	65
column total	140	60	200

(a) What is the probability of a false positive?
(b) What is the probability of either a false positive or a false negative?
(c) What is the probability of a positive test result given that the condition is present?
(d) What is the probability that the condition is present given a positive test result?

(7) One college found that during one semester $1,259$ students in its four most popular majors had the following class distributions. Use the table below to determine the probabilities for a randomly selected student in this group.

	first year	sophomore	junior	senior	row total
Business	115	90	105	111	421
Psycology	88	95	91	96	370
Nursing	85	81	79	76	321
Biology	63	45	25	14	147
column total	351	311	300	297	1259

(a) What is the probability of being a business major?
(b) What is the probability of not being a biology major?
(c) What is the probability of being a senior and majoring in psychology?
(d) What is the probability of being a senior or sophomore?
(e) What is the probability of being a senior or biology major?
(f) What is the probability of being a junior, given being a nursing major?
(g) What is the probability of being a nursing major, given being a junior?

(8) An island is a habitat for 208 species of birds. 82 of these species are found only on this particular island. 75 species are seabirds. 12 are a species of seabird and are found only on this particular island. One species of bird is chosen at random.

(a) What is the probability it is a seabird or unique to this island?
(b) What is the probability it is neither a seabird nor unique to this island?

(9) A company is looking hire more sales staff. The human resources department accepts only the 45% of the submitted resumes that meet the hiring criteria. The managers then select 20% of the applicants with accepted resumes to come in for an interview. What is the probability that an applicant selected at random will have her resume accepted and be granted an interview?

(10) In one high school, the athletic director found that 4% of the varsity athletes had concussions while playing at the school and 18% had severe sprains and 1% had experienced both. What is the probability that a randomly selected varsity athlete has either had a concussion or a severe sprain?

Class 5: Discrete random variables and discrete probability distributions

A random variable: is a quantitative variable X that takes random outcomes. It can be though of as a function from the sample space to the real numbers, in such a way that we can always measure the probability of X being on a given interval.

A discrete random variable: can take at most countable many values.

A continuous random variable: takes all the values of an interval in the real line.

The probability distribution or density function ρ for a discrete random variable X: is an assignment of a probability to each value taken by a discrete random variable in such a way that **the sum of all probabilities is always 1**.

The expected value or mean of a discrete probability distribution is:

$$E(X) = \mu(X) = \sum xP(x).$$

The expected value is a linear operator:

$$E(aX + bY) = aE(X) + bE(Y).$$

The standard deviation and variance of a discrete probability distribution are:

$$\sigma(X) = \sqrt{\sum (x-\mu)^2 P(x)}, \qquad \sigma^2(X) = \sum (x-\mu)^2 P(x).$$

The variance satisfies the formula:

$$\sigma^2(X) = E(X^2) - E(X)^2.$$

And therefore:

$$\mu(ax+b) = a\mu(x) + b \qquad \sigma^2(aX+b) = a^2\sigma^2(X).$$

Some discrete probability distributions			
Distribution	Density function	mean	The variable X represents:
Poisson	$\rho(k) = P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$	$\mu = \lambda$	The number of events on an interval.
Geometric (type I)	$\rho(k) = P(X=k) = (1-p)^{k-1}p$	$\mu = \frac{1}{p}$	The number of Bernoulli trials for the first success.
Binomial	$\rho(k) = P(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$	$\mu = np$	The number of successes in n Bernoulli trials.

We have the following commands for any discrete probability distribution:

(1) d for the "probability distribution or density function".

(2) p for the "probability cumulative distribution" that allows to compute the left tail area.

(3) q for "quantile" or inverse of the the cumulative distribution function.

(4) r for "random" generating a vector of numbers of a given distribution.

Examples:

(1) Finding the mean and standard deviation of the discrete probability distribution given by the table:

x	0	1	2	3
f	1/9	1/3	4/9	1/9

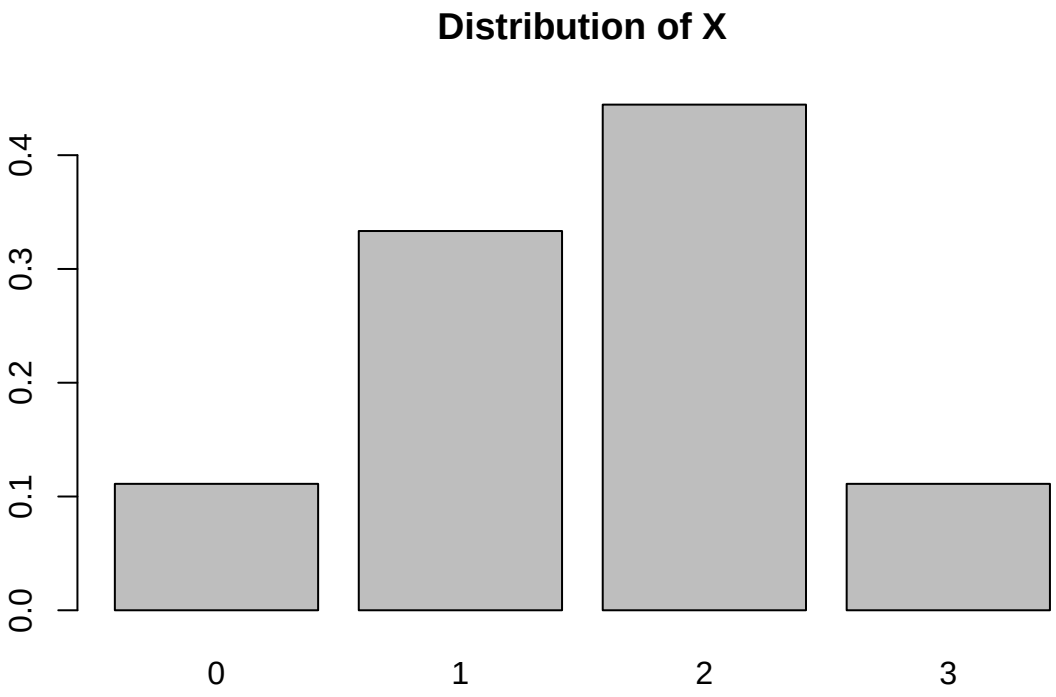
```
 x <- c(0,1,2,3)
p <- c(1/9, 3/9, 4/9, 1/9)
 mu <- sum(x * p)
 mu
```

```
## [1] 1.555556
```

```
  sigma2 <- sum((x-mu)^2 * p)
  sigma <- sqrt(sigma2)
  sigma
```

```
## [1] 0.8314794
```

```
barplot(p,main="Distribution of X",
                  names.arg=c(0:3))
```



Questions:

(1) Complete the table in such a way that we have a discrete probability distribution.

x	2	3	4	5	6
$P(x)$	.25	.1	.3	.2	

(a) Sketch the graph of this distribution and calculate its expected value and standard deviation.

(2) A fair coin is tossed 7 times. Sketch the graph of the resulting binomial distribution.

(3) (Bernoulli trials) Consider a random variable X with two positive outcomes, success (1) with probability p and failure (0) with probability $1 - p$. Find expected value and standard deviation for X.

(4) Consider the following discrete probability distribution:

x	2	3	4	5	6
$P(x)$	.2	.1	.35	.2	.15

Sketch the graph of this distribution and calculate its expected value and standard deviation.

(5) The following probability distribution represents the claim sizes (x) for an auto insurance policy.

x	1	2	3	4
P(x)	.1	.2	.25	

(a) Complete table to be a discrete probability distribution.
(b) Sketch the bar graph of the distribution.
(c) Calculate the expected value of the distribution.
(d) Calculate the standard deviation.

Class 6: Binomial distribution

A binomial experiment: is an experiment with a fixed number **n** of independent trials, each of which can only have two possible outcomes (Bernoulli trials), and the probability of each outcome remains constant on each trial.

The probability of success: will be the probability **p** of one of the two outcomes on each trial. **The probability of failure:** will be the probability **q=1-p** of the other outcome.

Main question: What is the probability (for **r=0,1, . . . n**) of getting exactly **r** successful outcomes in **n** trials?

Answer: The probability of getting exactly r successes in n trials is

$$P(X = r) = C_{n,r}p^r(1-p)^{n-r} = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}.$$

To probability $P(X = r)$ can be found using the Binomial Probability Distribution table.

As sum of independent Bernoulli events, the mean and standard deviation of a binomial probability distribution are:

$\mu = np$	$\sigma = \sqrt{np(1-p)}$

In the binomial distribution: The closer p is to .5 and the larger the number of sample observations n, the more symmetric the distribution becomes.

Examples

(1) To introduce the density function for the binomial, we compute with $n = 6$ and $p = .2$ the probabilities that we get exactly $r = 3$ successes and and at most $r \leq 3$ successes. We have the following commands for the binomial probability distribution:
   (1) dbinom for the density function.
   (2) pbinom for the probability cumulative distribution.
   (3) qbinom for quantile or inverse of cumulative distribution function.
   (4) rbinom for random generating a vector of numbers ditributed binomial.

```
dbinom(3,6,prob=.2)
```

```
## [1] 0.08192
```

```
pbinom(3,6,prob=.2)
```

```
## [1] 0.98304
```

(2) We present here different barplots of binomial distributions, for various values of n and p. A portion of the graph is being shaded using red color. We can find also the shaded area. The labels for the x-axis are taken from the names of the data so you simply need to define a data vector with appropriate names. We have choosen for the plot the values of $n = 100$ and $p = .65$, also we want to shade in red the columns between the values of $r = 50$ and $r = 75$.

```
n <- 100
P <- 0.65
mean=n*P
mean
```

```
## [1] 65
```

```r
sd=sqrt(n*P*(1-P))
sd
```

```
## [1] 4.769696
```

We include here a visualization of our data, beyond what you can observe in any table (if we can actually get any information from those numbers??).

```r
data <- dbinom(x=0:n,size=n, prob=P)
names(data) <- 0:n
data
```

```
##            0            1            2            3            4
## 2.551552e-46 4.738597e-44 4.356124e-42 2.642715e-40 1.190166e-38
##            5            6            7            8            9
## 4.243791e-37 1.247877e-35 3.112052e-34 6.718697e-33 1.275486e-31
##           10           11           12           13           14
## 2.155571e-30 3.275349e-29 4.511403e-28 5.671478e-27 6.545348e-26
##           15           16           17           18           19
## 6.969238e-25 6.875900e-24 6.309649e-23 5.403263e-22 4.330736e-21
##           20           21           22           23           24
## 3.257332e-20 2.304507e-19 1.536837e-18 9.679208e-18 5.767194e-17
##           25           26           27           28           29
## 3.255993e-16 1.744282e-15 8.878304e-15 4.298730e-14 1.982074e-13
##           30           31           32           33           34
## 8.711689e-13 3.653289e-12 1.462946e-11 5.598462e-11 2.048849e-10
##           35           36           37           38           39
## 7.175152e-10 2.405954e-09 7.728778e-09 2.379650e-08 7.025634e-08
##           40           41           42           43           44
## 1.989760e-07 5.407710e-07 1.410787e-06 3.533998e-06 8.502248e-06
##           45           46           47           48           49
## 1.964964e-05 4.363196e-05 9.309920e-05 1.909088e-04 3.762517e-04
##           50           51           52           53           54
## 7.127282e-04 1.297684e-03 2.270948e-03 3.819600e-03 6.174009e-03
##           55           56           57           58           59
## 9.589759e-03 1.431125e-02 2.051637e-02 2.824792e-02 3.734470e-02
##           60           61           62           63           64
## 4.739220e-02 5.771416e-02 6.742184e-02 7.552469e-02 8.108790e-02
##           65           66           67           68           69
## 8.340469e-02 8.214099e-02 7.741219e-02 6.976855e-02 6.009051e-02
##           70           71           72           73           74
## 4.942138e-02 3.878136e-02 2.900907e-02 2.066400e-02 1.400205e-02
##           75           76           77           78           79
## 9.014655e-03 5.507073e-03 3.187768e-03 1.745682e-03 9.028303e-04
##           80           81           82           83           84
## 4.401298e-04 2.018232e-04 8.684725e-05 3.497807e-05 1.314652e-05
##           85           86           87           88           89
## 4.595758e-06 1.488659e-06 4.448866e-07 1.220549e-07 3.056271e-08
##           90           91           92           93           94
## 6.937250e-09 1.415765e-09 2.572120e-10 4.109071e-11 5.682758e-12
##           95           96           97           98           99
## 6.665490e-13 6.447275e-14 4.937530e-15 2.807051e-16 1.053150e-17
##          100
## 1.955851e-19
```

The colors can be passed as a vector of colour names. You can create a vector filled with “grey” and then

just insert “red” for the bars that you want to have red:

```
cols <- rep("grey", n + 1)
r <-c(50:75)
r
```

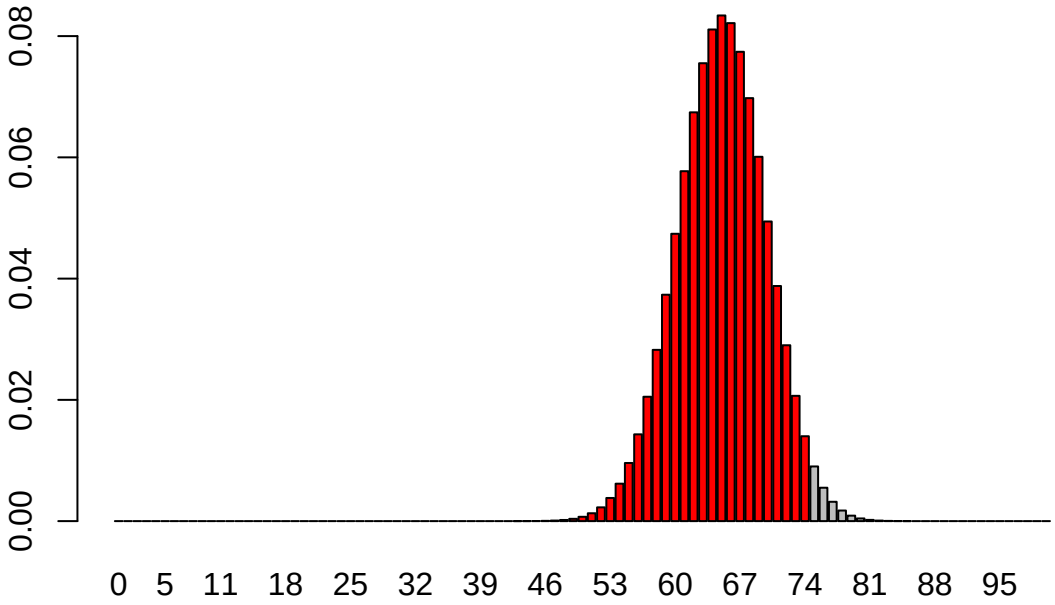
```
##  [1] 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
## [24] 73 74 75
```

```
cols[r] <- "red"
cols
```

```
##   [1] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey"
##  [11] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey"
##  [21] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey"
##  [31] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey"
##  [41] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "red"
##  [51] "red"  "red"  "red"  "red"  "red"  "red"  "red"  "red"  "red"  "red"
##  [61] "red"  "red"  "red"  "red"  "red"  "red"  "red"  "red"  "red"  "red"
##  [71] "red"  "red"  "red"  "red"  "red"  "grey" "grey" "grey" "grey" "grey"
##  [81] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey"
##  [91] "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey" "grey"
## [101] "grey"
```

And finally the plot:

```
barplot(data, col = cols)
```



Answer: We can observe in the graph that since the mean is 65, most of the area is higher that 50, for the are to the right of the mean, we do observe some gray columns. The value of the area $P(50 \leq r \leq 75)$ can be obtained with the command:

```
sum(dbinom(50:75, 100, 0.65))
```

```
## [1] 0.9871352
```

(3) As our mext example, we present the bar plot for the binomial with $n = 20$ and $p = .65$ and compute $P(13 \leq r \leq 16)$, where is easy to distinguish the colored columns.

```
n <- 20
P <- 0.65
mean=n*P
mean
```

```
## [1] 13
```

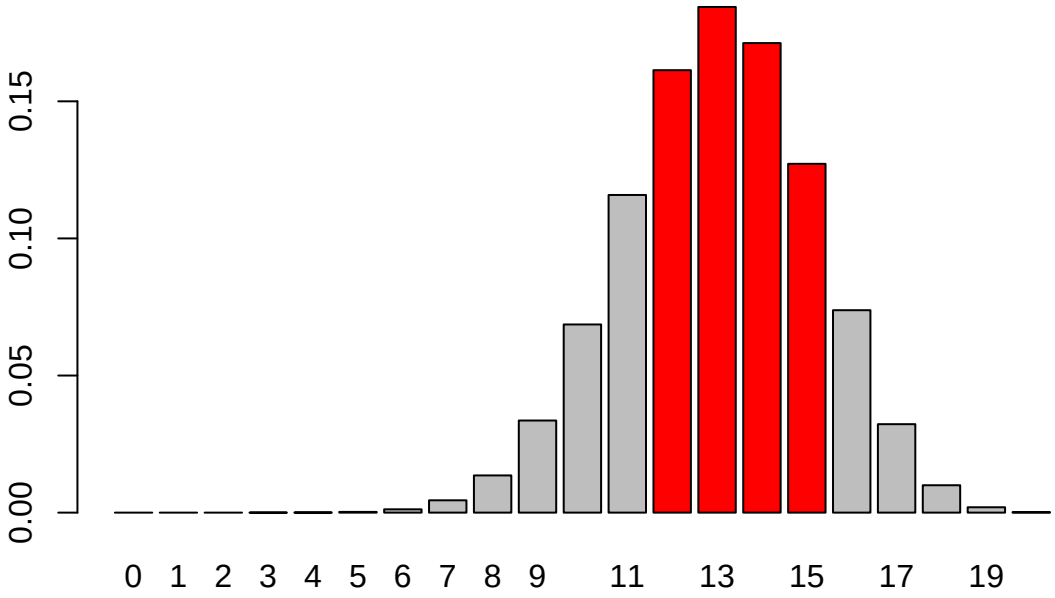
```
sd=sqrt(n*P*(1-P))
sd
```

```
## [1] 2.133073
```

```
data <- dbinom(x=0:n,size=n, prob=P)
names(data) <- 0:n
cols <- rep("grey", n + 1)
r <-c(13:16)
r
```

```
## [1] 13 14 15 16
```

```
cols[r] <- "red"
barplot(data, col = cols)
```



Answer: as before the value of the red are is:

```
sum(dbinom(13:16, 20, 0.65))
```

```
## [1] 0.556651
```

Questions

(1) If 30% of the people in a community use the Library in one year, find the probability that in a random sample of 15 people

(a) At most 7 use the Library,
(b) Exactly 7 use the Library,
(c) At least 5 use the Library,
(d) No more than 2 use the Library,
(e) Not less than 10 use the Library.

(f) Sketh the graphs for parts (a),(b),(c),(d) and (e).

(2) A basketball player makes 70% of the free throws he shoots. What is the probability that he will make more than 7 throws

(a) If he tries 15 free throws?
(b) If he tries 10 free throws?
(c) Compare the two results graphically.

(3) Approximately 5% of the eggs in a store are cracked. Suppose you buy a dozen eggs from the store.

(a) What is the probability that no more than one of your eggs is cracked?
(b) What is the probability that fewer than 3 eggs are cracked?
(c) Find the expected value and standard deviation of the number of cracked eggs.

(4) A surgery has a success rate of 75%. Suppose that the surgery is performed on six patients. Find the expected value and the standard deviation of the number of successes.

(5) One-third of all deaths are caused by heart attacks. If three deaths are chosen randomly, find the probability that none resulted from heart attack.

(6) Explain what we understand by Binomial experiment. Give an example of a variable following a binomial distribution.

(7) Suppose that the probability of a hurricane in a calendar year is $p = .05$. Find the probability that, in a 10-year period, we have:

(a) Exactly 1 hurricane.
(b) At least 3 hurricanes.
(c) At most 2 hurricanes.

(8) Assuming that the probability of having a daughter is of the 50%.

(a) What is the probability of having exactly 1 daughter in a family with 5 kids?
(b) What is the expected value of the number of daughters in families with 5 kids?
(c) What is the standard deviation ?

Class 7: 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 rate and independently of the time since the last event. In other words, a variable following a Poisson distributions satisfies the following conditions:

(1) The number of successes in two disjoint time intervals is independent.

(2) The probability of a success during a small time interval is proportional to the entire length of the time interval.

When X follows a Poisson with mean λ: the probability of X is successes is determined by the formula:

$$P(X) = \frac{e^{\lambda}\lambda^{X}}{X!},$$

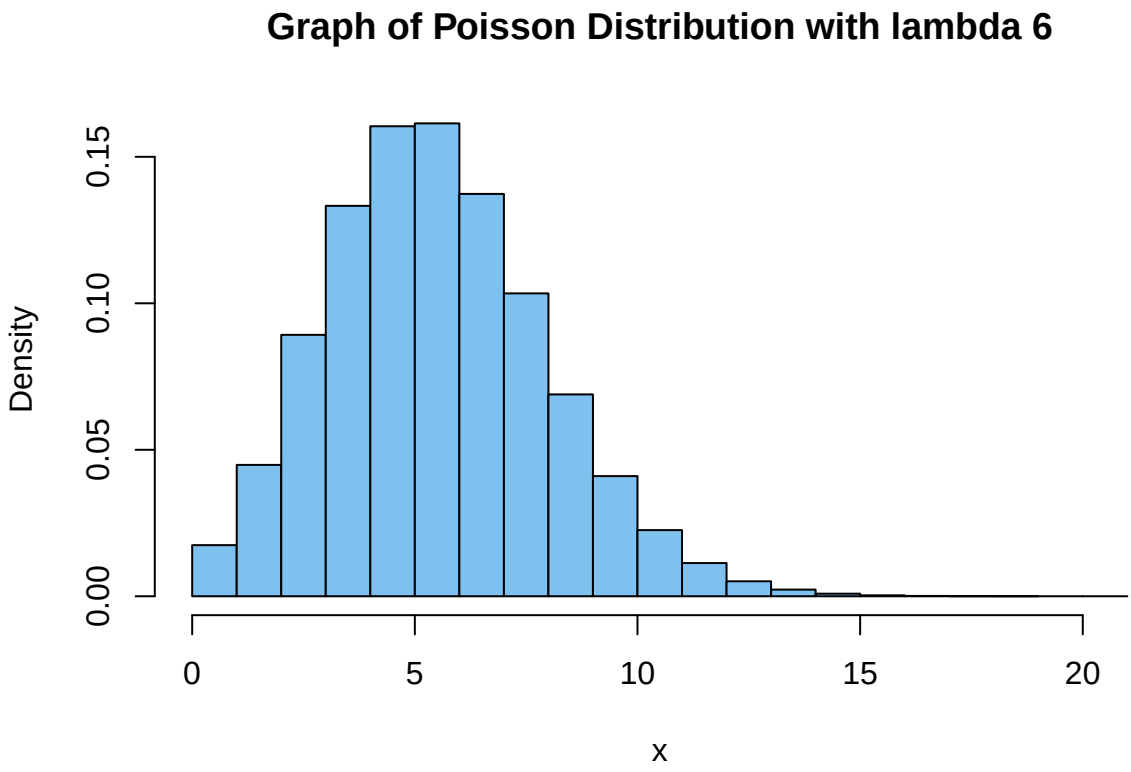
where λ is the mean number of independent successes in a unit of time or space.

The Poisson distribution can be used to determine whether events or objects occur randomly in space or time. When events are random in time or space (not clumped or disperse) it is reasonable to think that they will follow a Poisson distribution.

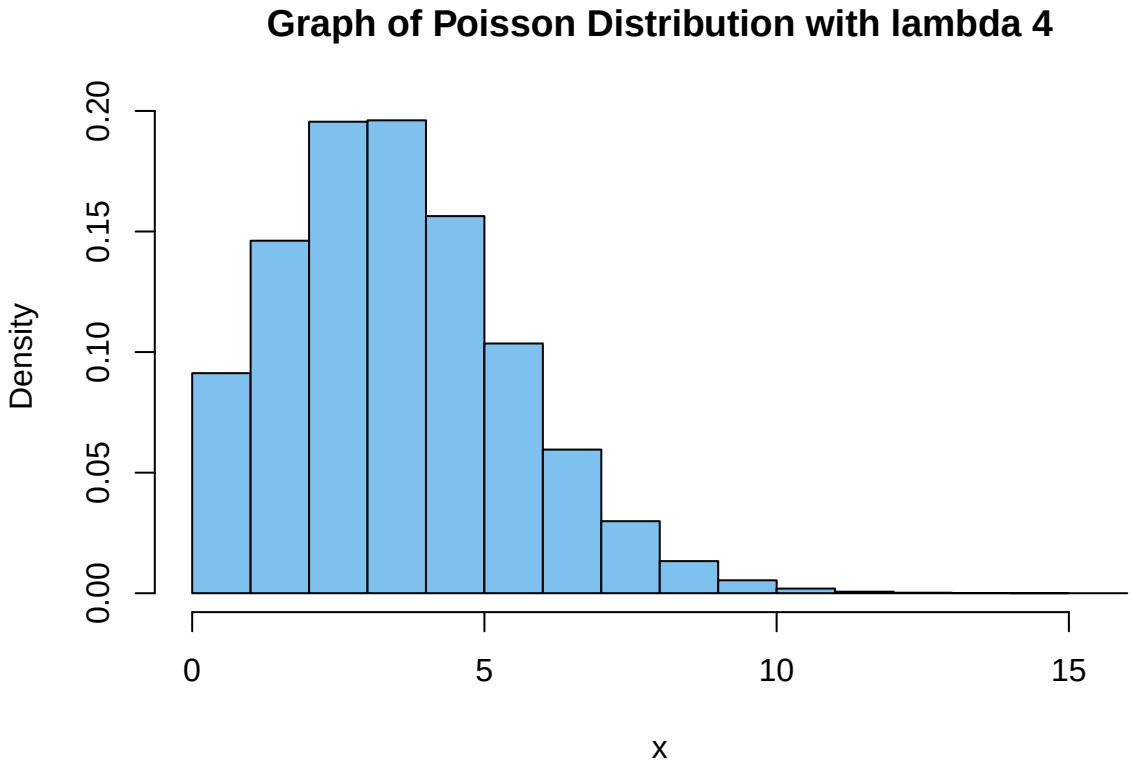
Examples:

(1) We begin by representing the graphs of several Poisson distributions. As expected "rpois" generates a vector of number distributed poisson with a given lambda, while "ppois" and "dpois" represent the cumulative distribution function and density function respectively.

```
lambda <- 6
y = rpois(10^6, lambda);  up=max(y)
hist(y, prob=T, col="skyblue2", xlab="x",
     main="Graph of Poisson Distribution with lambda 6")
```




```
lambda <- 4
y = rpois(10^6, lambda);  up=max(y)
hist(y, prob=T, col="skyblue2", xlab="x",
     main="Graph of Poisson Distribution with lambda 4")
```



(2) Problem: Suppose that there are twelve cars crossing a bridge per minute on average, find the probability

of having seventeen or more cars crossing the bridge in a particular minute.

Solution: The probability of having sixteen or less cars crossing the bridge in a particular minute is given by the function ppois that computes the distribution function of Poisson. On the probability of having seventeen or more cars crossing the bridge in a minute is in the upper tail of the probability density function.

```
ppois(16, lambda=12)   # lower tail < = 16
```

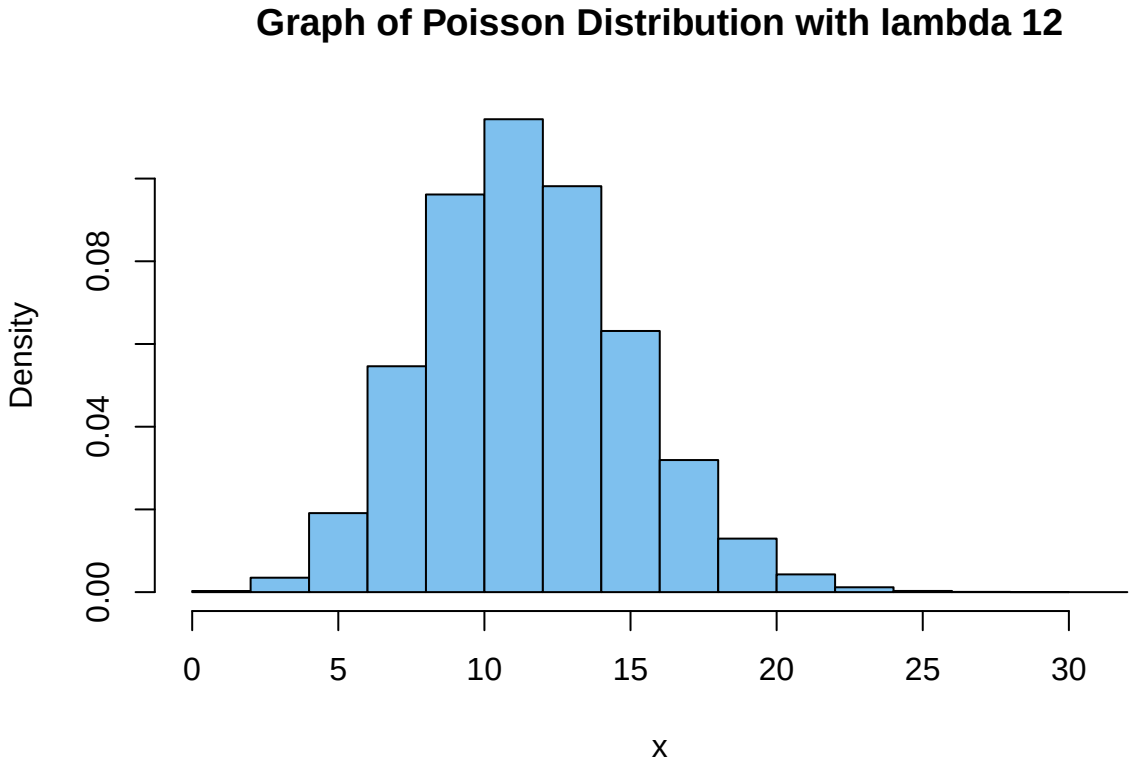
```
## [1] 0.898709
```

```
ppois(16, lambda=12, lower=FALSE)   # upper tail > = 16
```

```
## [1] 0.101291
```

And the graph of the Poisson with mean 12 can be seen:

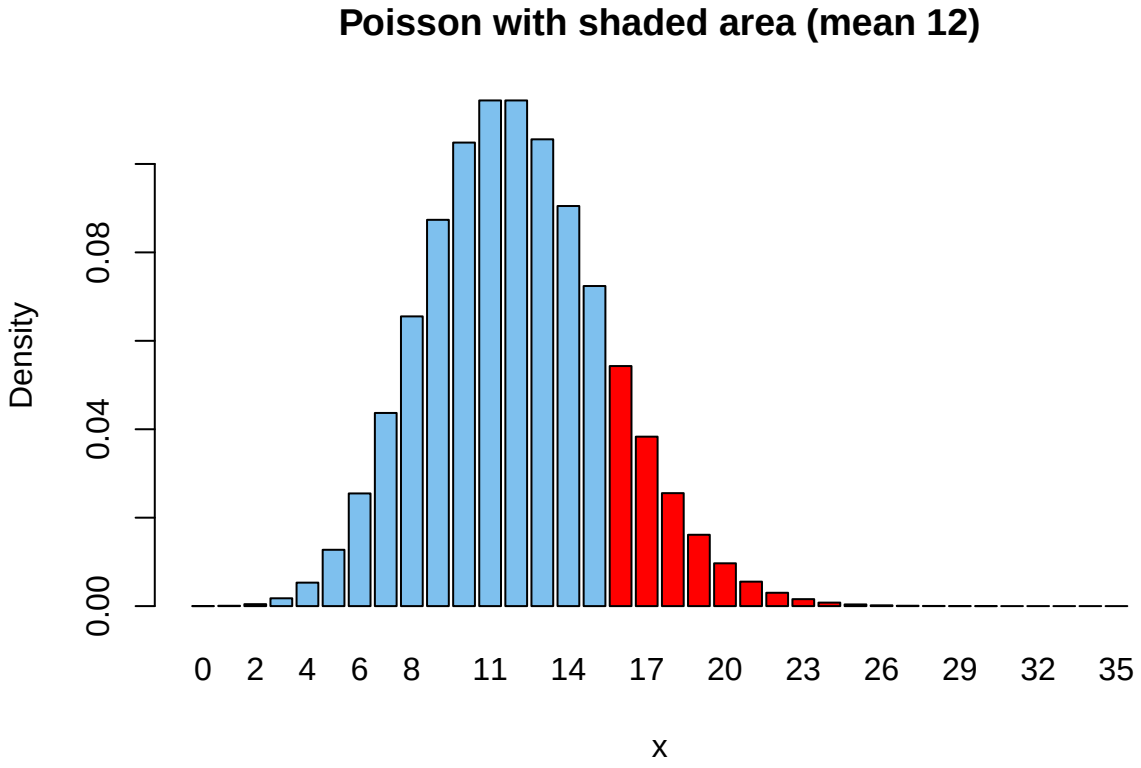
```
lambda <- 12
y = rpois(10^6, lambda);  up=max(y)
hist(y, prob=T, col="skyblue2", xlab="x",
     main="Graph of Poisson Distribution with lambda 12")
```



Where we can shade the area to the right of 17:

```
cols <- rep("skyblue2", 35)
r <-c(17:35)
cols[r] <- "red"

barplot(dpois(0:35,12),
        xlab="x",
        main="Poisson with shaded area (mean 12)",
        ylab="Density",
        names.arg=0:35,
        col=cols)
```



Questions:

(1) Sketch the graphs of the Poisson distributions for values of $\lambda = 1, 5, 10$.

(2) The mean value for an event X to occur is $\lambda = 2$ in a day. Find the probability of event X to occur three times in a given day.

(3) If there are ten cars crossing a bridge per minute on average, find the probability of having eight or less cars crossing the bridge in a particular minute.

(4) Test the claim that the numbers in the table with the given frequencies follow a Poisson distribution with mean $\lambda = 2.44$ (this is equivalent to test for the randomness of the numbers and frequencies in the table) Use $\alpha = .05$.

Number	Frequency
0	7
1	6
2	2
3	3
4	4
5	6
6	1
Total	29

(5) The number of calls coming per minute into a hotels reservation center is Poisson random variable with mean $\lambda = 3$.

(a) Find the probability that no calls come in a given 1 minute period.
(b) Assume that the number of calls arriving in two different minutes are independent. Find the probability that at least two calls will arrive in a given two minute period.
(c) Sketch the areas representing the regions in (a) and (b).

Class 8: Continuous probability distributions

A continuous random variable X: takes all values on a whole interval of the real line.

The probability distribution ρ for a continuous random variable X: is an assignment of probability to each interval of the values taken by the variable X, in such a way that the total area under the curve given by

$$A = \int_{-\infty}^{\infty} \rho(x)dx = 1.$$

Some continuous probability distributions			
Distribution	Density function	mean	Meaning or Relevance
Normal	$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$	μ	It is an approximation to the sampling distribution of $\bar{X}$ for large n.
Student's t-distribution (df=ν)	$\rho(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$\mu = 0$	Distribution of the sample mean of n observations from a normal distribution relative to the true mean.
Chi-square	$\rho(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	$\mu = k$	Sum of the squares of independent normal standard variables.

The expected value and standard deviation of the distribution are:

$$\mu(X) = E(X) = \int_{-\infty}^{\infty} x\rho(x)dx, \qquad \sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2\rho(x)dx}.$$

The geometric interpretation of the probability $P(a < X < b)$: for a continuous random variable with distribution function ρ, the probability $P(a < X < b)$ is the area under the curve $y = \rho(x)$ and above the x-axis when $a < x < b$. Notice that a or b or maybe both can be equal to ∞ or $-\infty$.

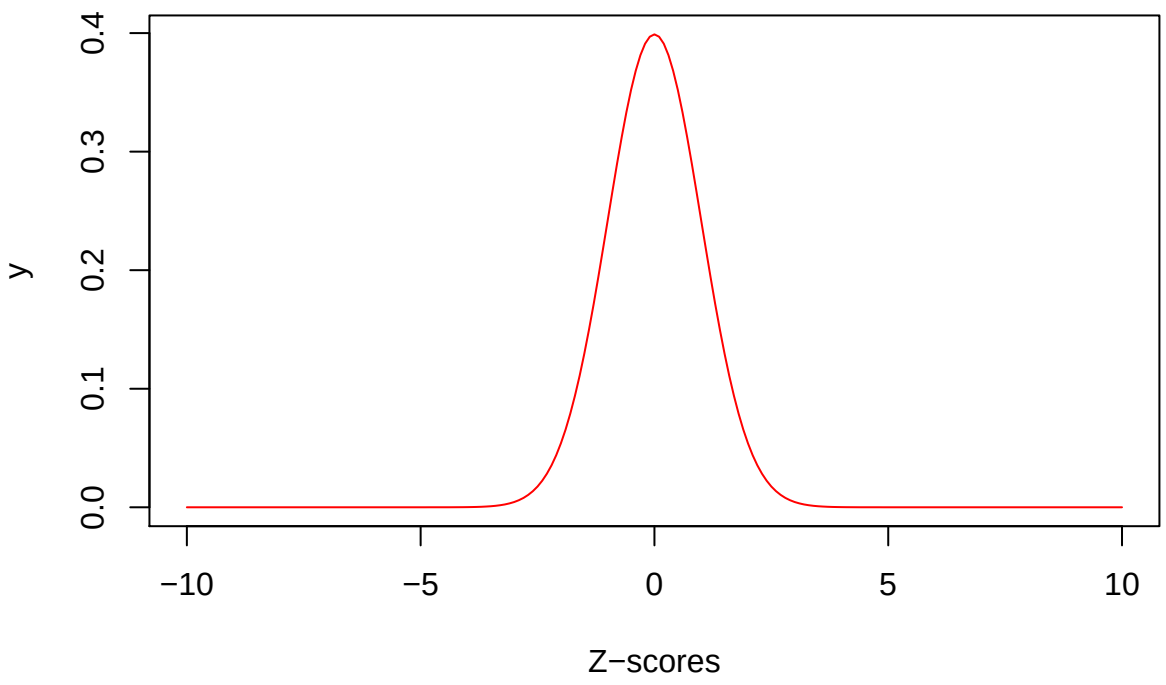
To find the probability that X fall in a given interval we use:

$$P(a < X < b) = \int_a^b \rho(x)dx.$$

Examples:

(1) The normal distribution. Let us use the probability distribution function "dnorm" to create a plot of the standard normal with mean. Also we use the "pnorm" to find the total area under the curve. The graph will be done with plot(), later we will see an easier alternative with the command curve().

```
z_scores <- seq(-10, 10, by = .1)
y <- dnorm(z_scores, mean = 0, sd = 1)
plot(z_scores, y, type = "l",
     main = "Distribution of the Standard Normal",
     xlab= "Z-scores", col="red")
```




```
Total_Area=pnorm(Inf,0,1)
Total_Area
```

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

Similar to the treatment in the discrete case, we have the following commands for any continouos probability distribution:

(1) d for "probabilitity distribution or density function".

(2) p for the "probability cumulative distribution" that allows to compute the left tail area.

(3) q for "quantile" or inverse of the cumulative distribution function.

(4) r for "random" generating a vector of numbers of a given distribution.

In the particular case of the normal distribution, the command "$dnorm(x, \mu, \sigma)$" will give the value of the distribution function $\rho(x)$ for $N(\mu, \sigma)$. The function "$pnorm(x, \mu, \sigma)$" is the cumulative function of $N(\mu, \sigma)$ that will allow us to compute the area to the left of the graph. The command "$qnorm(x, \mu, \sigma)$" will be the inverse function of "$pnorm(x, \mu, \sigma)$" and "rnorm(x)" will generate a vector of random numbers.

```
dnorm(0)
```

```
## [1] 0.3989423
```

```
pnorm(0)
```

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

```
pnorm(5,5,2)
```

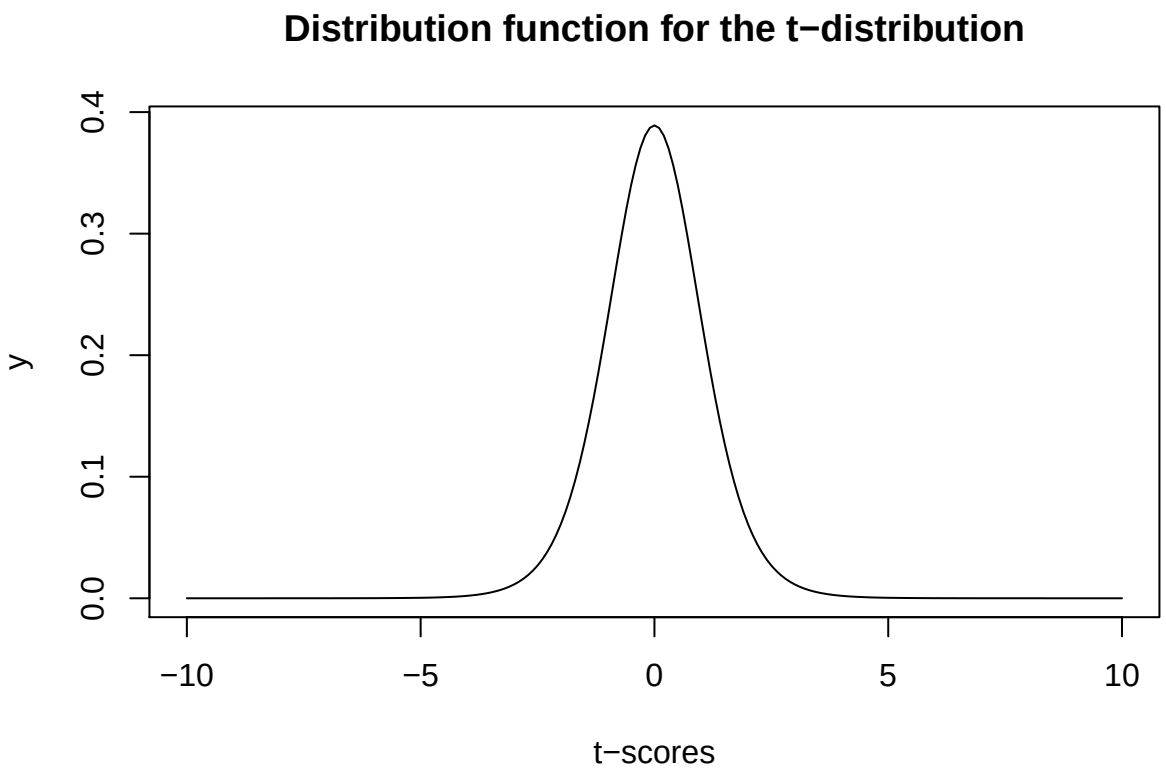
```
## [1] 0.5
```

```
qnorm(.5,5,2)
```

```
## [1] 5
```

(2) The t-distribution. We do a similar graph using the t-distribution with degrees of freedom d.f=10. The graph is suppose to show like the normal but with "fatter" tails (we will see the two graphs together in a bit). The total area enclosed by the curve should be equal 1 as computed with the comand "pt".

```
t_scores <- seq(-10, 10, by = .1)
y <- dt(t_scores, df=10)
plot(t_scores, y, type = "l",
     main = "Distribution function for the t-distribution",
     xlab= "t-scores")
```

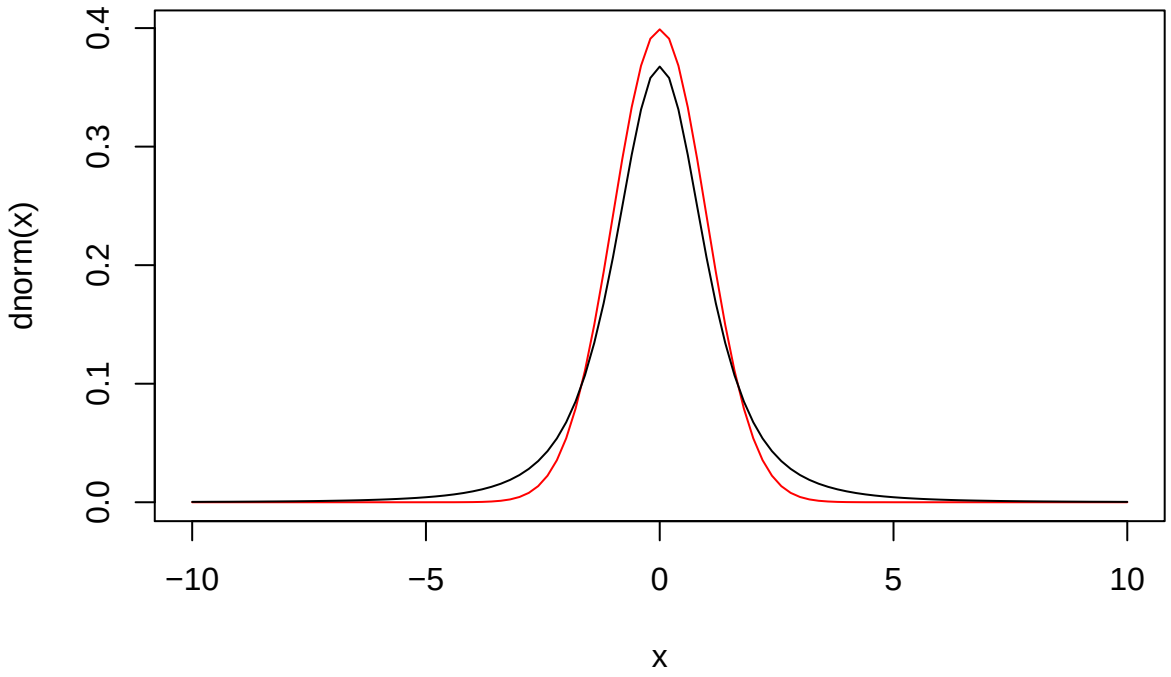



```
Total_Area=pt(Inf,df=10)
Total_Area
```

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

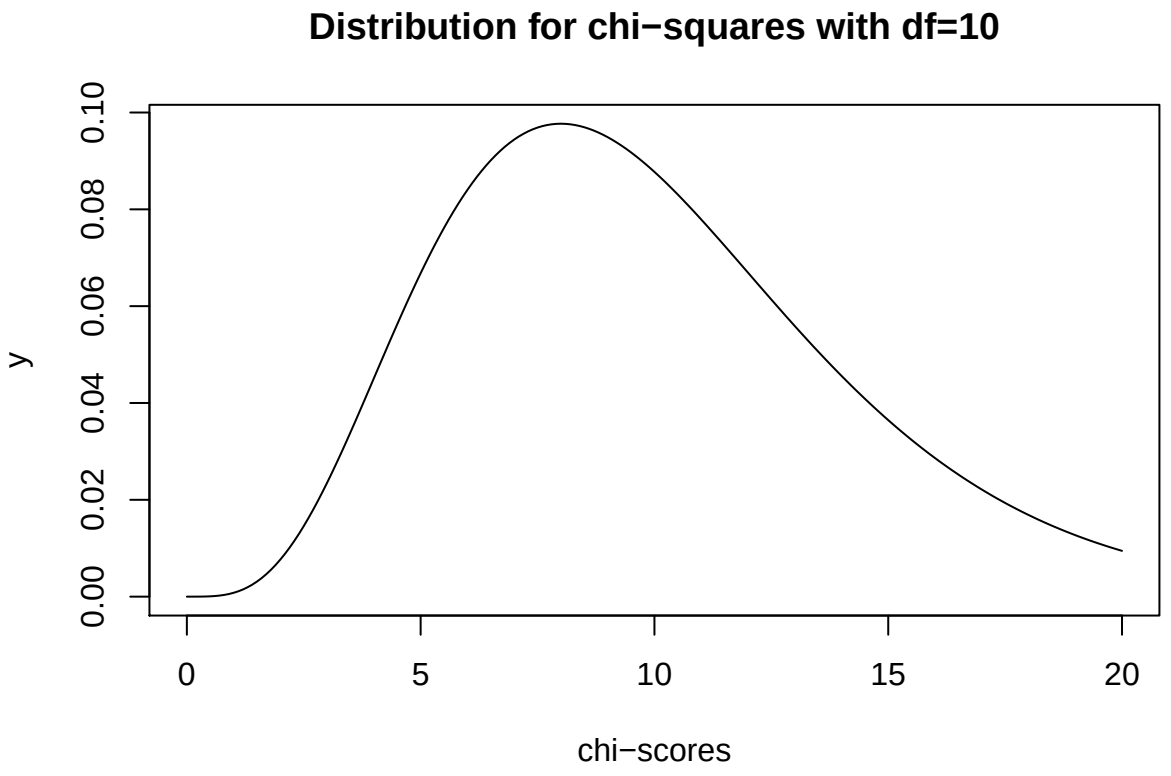
We can compare two t-distribution and standar normal directly using the density functions dnorm and dt. We use for that the function "curve". The command add=TRUE allows to put together the graphs of several functions.

```
curve(dnorm(x), -10, 10, col = "red")
curve(dt(x, df = 3), add = TRUE)
```



(3) The χ^2-ditribution. Again for the χ^2 distribution, we produce the graph and check that the total area under the curve is actually equal to 1. Note the use of the comand "dchisq".

```
x <- seq(0, 20, by = .1)
y <- dchisq(x, df=10)
plot(x,y, type = "l",
     main = "Distribution for chi-squares with df=10",
     xlab= "chi-scores")
```

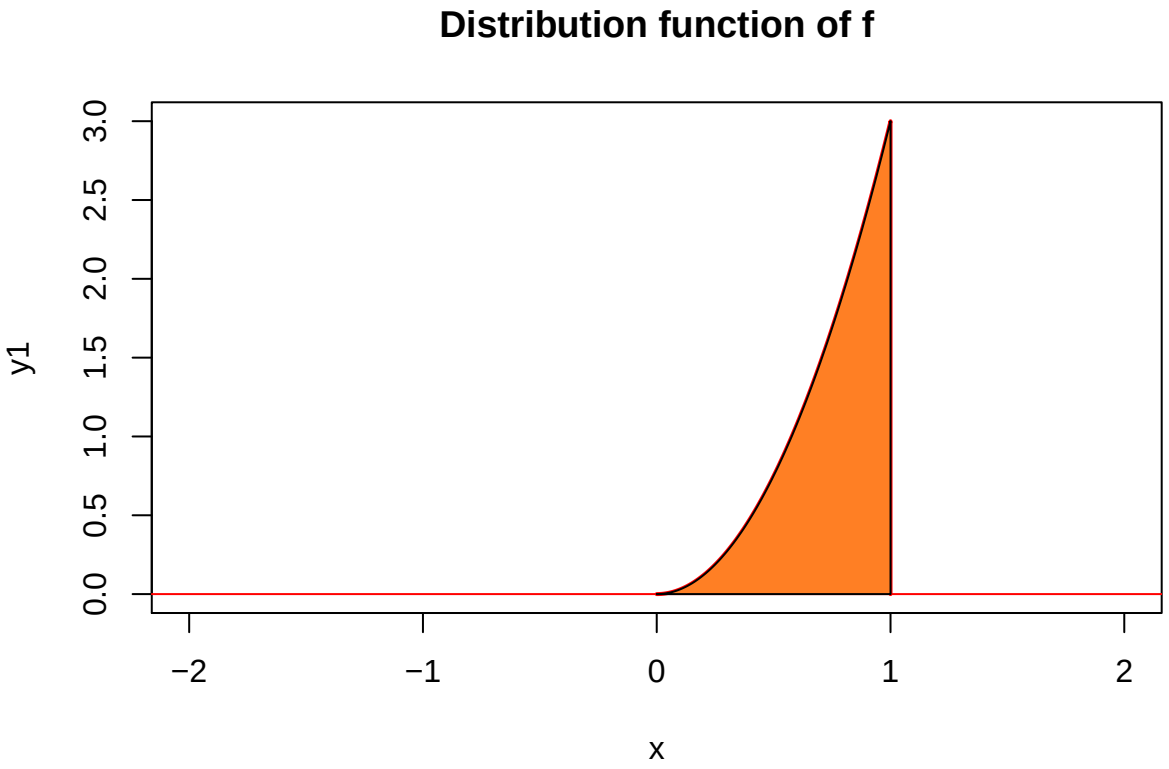


```
Total_Area=pchisq(Inf,df=10)
Total_Area
```

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

(4) Suppose that we want to work with an arbitrary probability distribution, that is not one of the most commonly used: normal, t-dist, etc. Consider the function $f(x)$ given by the formula $f(x) = 3x^2$ in the interval $[0, 1]$ and defined as $f(x) = 0$ outside that interval. We can show the graph and compute the area under the curve:

```
x=c(0,seq(0,1,.01),1)
f <-function(x) 3*x^2
y1=c(0,f(seq(0,1,.01)),0)
plot(x,y1,type="l",lwd=2,col="red",xlim=c(-2,2), , main="Distribution function of f")
abline(h=0,col="red")
polygon(x,y1, col="chocolate1")
```




```
integrate(f, lower = 0, upper = 1)
```

```
## 1 with absolute error < 1.1e-14
```

Interpretation: We can check that the total area is in fact one and our function is cleary positive. We have then a probability distribution function.

Questions:

(1) Suppose that the random variable X follows a continuous probability distribution.

(a) What is the probability $P(X = 1)$?
(b) If the probability $P(X > 3) = .3$, what is the probability $P(X < 3)$? What is the probability $P(X \leq 3)$?
(c) What is the probability $P(-\infty < X < \infty)$?

(2) Graph the distribution t of student with $d.f. = 20$ in the interval $(-5, 5)$.

(3) Find and sketch the area under the standard normal when $-\infty < x < 1$.

(4) Compare the area and the graph in (3) with a similar graph for the t-distribution with degrees of freedom $d.f. = 4$ and $d.f = 25$.

(5) Sketch the function defined as $g(x) = 4x^3$ in the interval $[0, 1]$ and $g(x) = 0$ otherwise. Show that we have a probability distribution function.

Class 9: Areas under the normal curve and central limit theorem

The density function for the normal probability distribution satisfies the differential equation

$$\frac{dy}{dx} = -\frac{1}{\sigma^2}(x-\mu)y,$$

for some real numbers μ and σ ($\sigma > 0$). The solution to the equation is the family of functions

$$y = y(x) = Ke^{\frac{(x-\mu)^2}{2\sigma^2}},$$

and we are looking for the solution such that $\int_{-\infty}^{\infty} y(x)dx = 1$. Using the fact $\int_0^{\infty} e^{x^2/2} = \sqrt{2\pi}/2$ and the change of variable $x' = \frac{x-\mu}{\sigma}$, we get $K = \frac{1}{\sqrt{2\pi}\sigma}$ and the density function as

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x-\mu)^2}{2\sigma^2}}.$$

The normal distribution is also called Gaussian distribution in honor of the german mathematician K.F. Gauss. The expected value and standard deviation of the distribution are:

$$\mu = E(X) = \int_{-\infty}^{\infty} x\rho(x)dx, \qquad \sigma = \sqrt{E(X^2) - E(X)^2} = \sqrt{\int_{-\infty}^{\infty}(x-\mu)^2\rho(x)dx}.$$

The maximum of the functions is at the point $(\mu, \frac{1}{\sqrt{2\pi}\sigma})$ and the inflexion points at $x = \pm\sigma$. The normal curve with $\mu = 0$ and $\sigma = 1$ is called **standard normal distribution**. A random variable that follows a standard normal distribution is usually denoted with the letter Z and probabilities $P(a < Z < b)$ can be found in the **Standard Normal Distribution Table**.

The Standard Normal Distribution Table gives areas to the left, that is $P(Z < z)$. To find areas to the right and between two scores, you use:

$$P(Z > z) = 1 - P(Z < z) \quad \text{and} \quad P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1).$$

We have the raw score: $X = \sigma Z + \mu$.

The standard score: $Z = \dfrac{X-\mu}{\sigma}$.

Using a change a variable we can relate the raw and standard score proving that:

$$P(a < Z < b) = P(\frac{a-\mu}{\sigma} < X < \frac{b-\mu}{\sigma}).$$

A sampling distribution: is the probability distribution of a sample statistic based on all possible simple random samples of the same size from the population. For example the distribution of the sample mean $\bar{X}$ based on random samples of size n.

Whenever X is normally distributed with mean μ and s.dev.σ: the variable $\bar{X} = \dfrac{\sum x}{n}$ that averages random samples of size n is also normally distributed with mean and standard deviation given by the formulas:

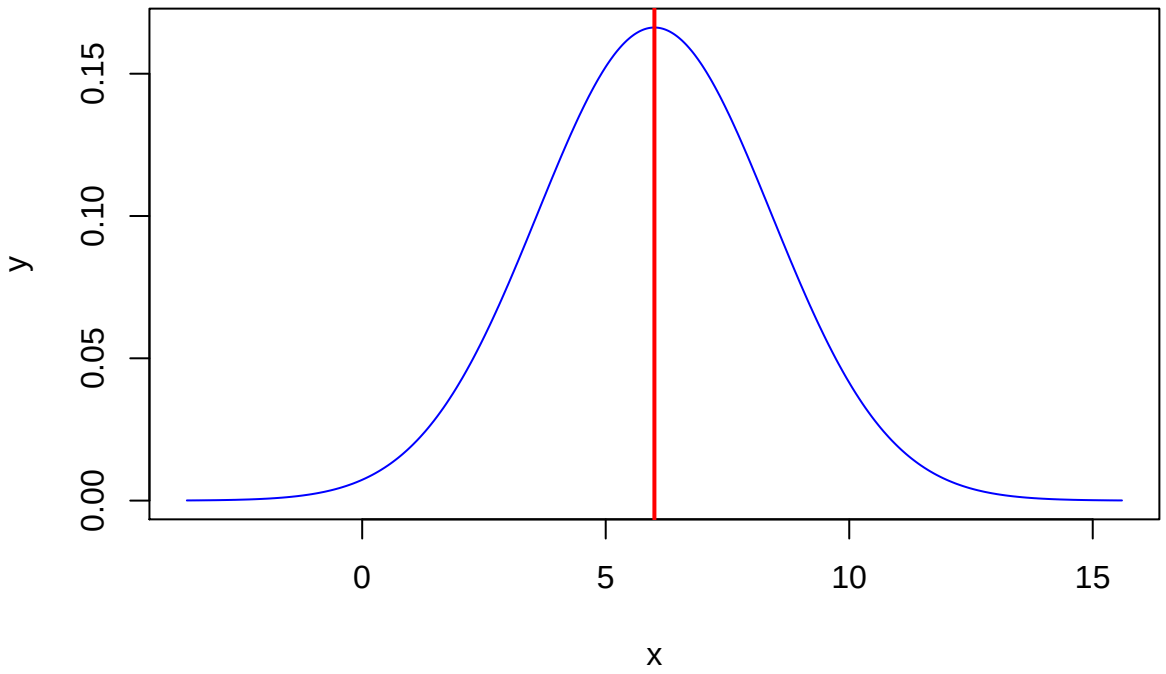
$$\mu_{\bar{x}} = \mu_x, \qquad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}.$$

Central limit Theorem: Regardless of the distribution followed by X with mean μ and standard deviation σ, the sequence of random variables $X_n = \bar{X}$ is close, for n large, to a normal distribution with mean μ and standard deviation $\dfrac{\sigma}{\sqrt{n}}$. For practical considerations $n \geq 30$ is usually sufficient.

Examples:

(1) Let us see an example of the plot of a normal curve with given mean and standard deviation. We are using a sequence of 1000 points, four standard deviations above and below the mean, this means, in the interval $(\mu - 4\sigma, \mu + 4\sigma)$. In our example $\mu = 6$ and $\sigma = 2.4$.

```
mean=6
sd=2.4
x <- seq(mean-4*sd,mean+4*sd,length=1000)
y <- dnorm(x,mean, sd)
plot(x,y, type="l", lwd=1, col="blue")
abline(v = mean, col = "red", lwd = 2)
```

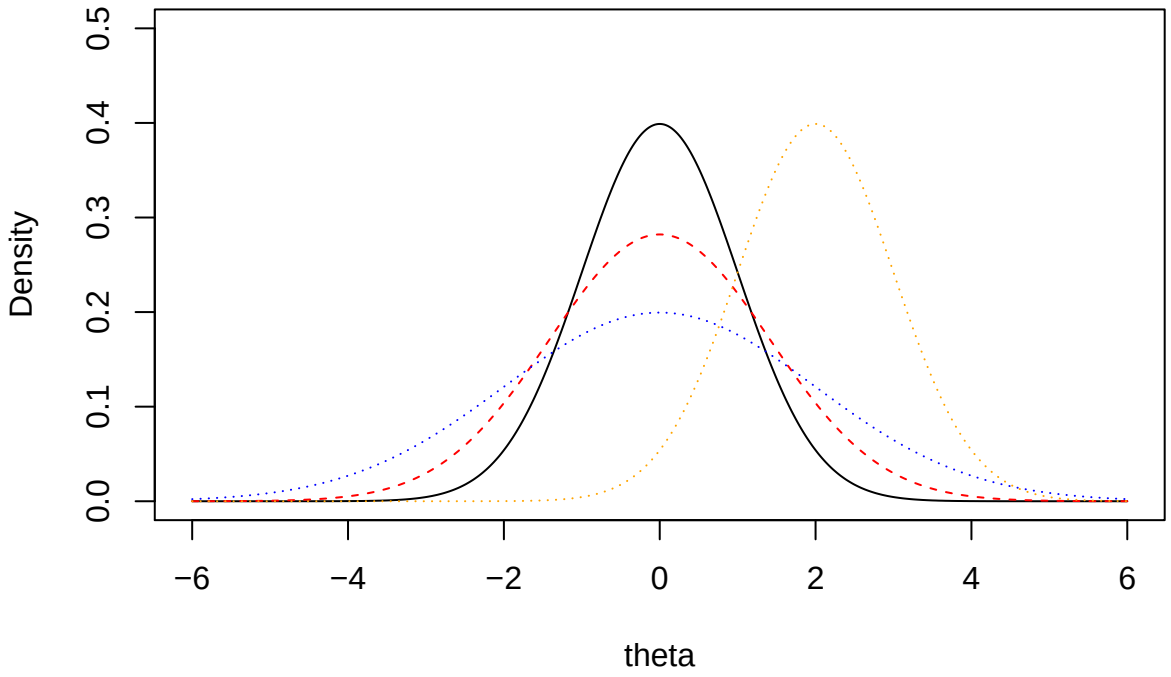


(2) Suppose that you want to plot several Gaussian distributions on the same graph, e.g.,

$$\{N(0,1), N(0,2), N(0,4), N(0,.5), N(2,1)\},$$

where as usual $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

```
dgauss <- function( theta, mu, sigma2 )
  {exp( - ( theta - mu )^2 / ( 2 * sigma2 ) ) / sqrt( 2 * pi * sigma2 ) }
theta <- seq( -6, 6, length = 500 )
plot(theta,dgauss(theta,0,1),
     type='l',
     xlab="theta",
     ylim=c(0,.5),
     ylab="Density",
     main="Several Normal Distributions")
lines(theta, dgauss( theta,0, 2 ), lty = 2, col="red" )
lines(theta, dgauss(theta,0,4), lty=3, col="blue")
lines(theta, dgauss(theta,2,1), lty=3, col='orange')
```



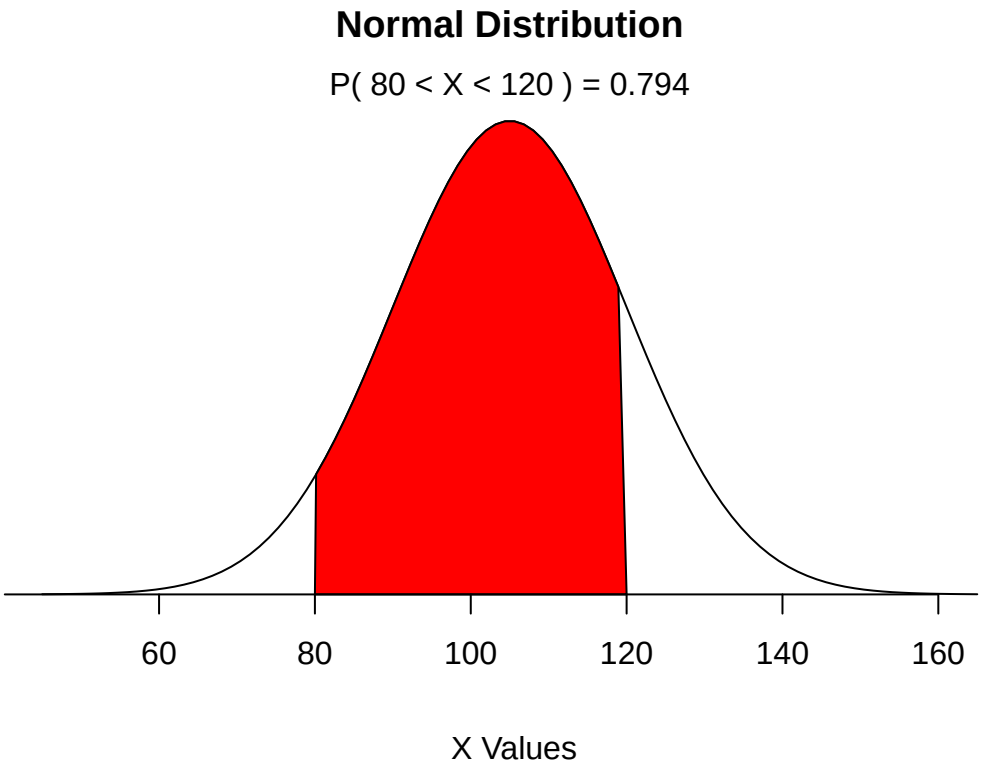
(3) Now we are going to compute and show graphically the area under the normal distribution with certain mean and standard deviation located between the given numbers lb and ub.

```
mean=105; sd=15
lb=80; ub=120
x <- seq(-4,4,length=100)*sd + mean
hx <- dnorm(x,mean,sd)

plot(x, hx, type="n", xlab=" X Values", ylab="",
     main="Normal Distribution", axes=FALSE)

i <- x >= lb & x <= ub
lines(x, hx)
polygon(c(lb,x[i],ub), c(0,hx[i],0), col="red")

area <- pnorm(ub, mean, sd) - pnorm(lb, mean, sd)
result <- paste("P(",lb,"< X <",ub,") =",
                signif(area, digits=3))
mtext(result,3)
axis(1, at=seq(40, 160, 20), pos=0)
```

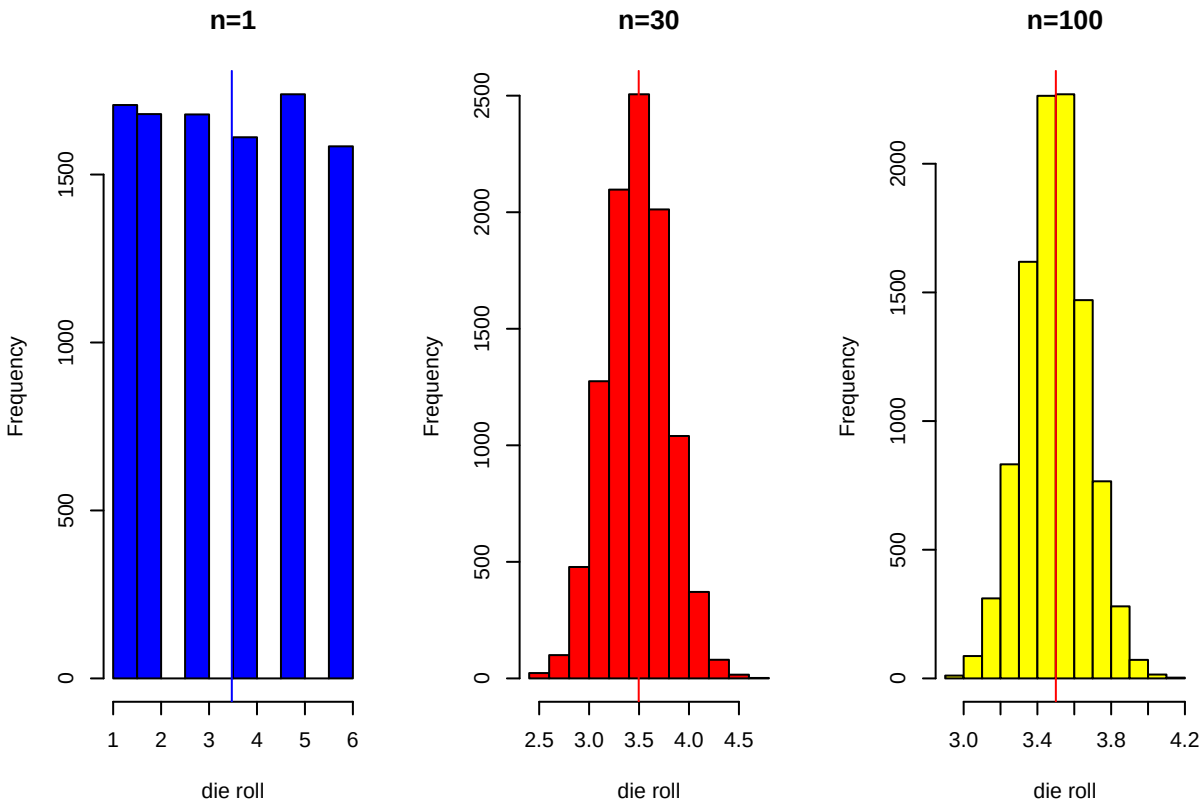


(4) We model sampling distribution for samples of size 1, 30 and 100 from a population following a unifrom distribution with mean $\mu = 3.5$. More specifically we are working with a fair die with values from 1 to 6. To model the population we draw 10,000 samples of each given size. The Central limit theorem predicts our data to look more and more like the a normal curve with mean $\mu = 3.5$.

```
x1 <- c()
x30 <- c()
x100 <- c()
k =10000
for ( i in 1:k){
  x1[i] = mean(sample(1:6,1, replace = TRUE))
  x30[i] = mean(sample(1:6,30, replace = TRUE))
  x100[i] = mean(sample(1:6,100, replace = TRUE))
}
par(mfrow=c(1,3))
hist(x1, col ="blue",main="n=1",xlab ="die roll")
abline(v = mean(x1), col = "blue")

hist(x30, col ="red", main="n=30",xlab ="die roll")
abline(v = mean(x30), col = "red")

hist(x100, col ="yellow",main="n=100",xlab ="die roll")
abline(v = mean(x100), col = "red")
```



Questions

(1) Let z have the standard normal distribution. For each of the following probabilities, draw an appropriate diagram, shade the appropriate region and then determine the value:

(a) $P(0 < z < 1.74)$
(b) $P(0.62 < z < 2.48)$
(c) $P(z > 2.1)$
(d) $P(-1.31 < z < 1.07)$.

(2) Let z have the standard normal distribution. For each of the following probabilities, draw an appropriate diagram, shade the appropriate region and then determine the value of z_c:

(a) $P(0 < z < z_c) = 0.4573$
(b) $P(z_c < z < 0) = 0.3790$
(c) $P(z < z_c) = 0.1190$
(d) $P(-z_c < z < z_c) = 0.8030$.

(3) Let x be a normally distributed random variable with $\mu = 70$ and $\sigma = 8$. For each of the following probabilities, draw an appropriate diagram, shade the appropriate region and then determine the value:

(a) $P(70 < x < 80.4)$
(b) $P(61.2 < x < 85.2)$
(c) $P(x < 58)$
(d) $P(x > 76)$.
(e) $P(68 < \bar{x} < 72)$, if a random sample of size $n = 49$ is drawn.
(f) $P(\bar{x} > 71)$, if a random sample of size $n = 81$ is drawn.

(4) Find z so that:

(a) 98% of the area under the standard normal curve lies between $-z$ and z.

(b) 97.5% of the area under the standard normal curve lies to the left of z.
(c) 46% of the area under the standard normal curve lies to the right of z.

(5) Find the area under the standard normal curve

(a) between $z = -2.74$ and $z = 2.33$.
(b) between $z = -2.47$ and $z = 1.03$.

(6) The lifetime of a certain type TV tube has a normal distribution with a mean of 80.0 and a standard deviation of 6.0 months. What portion of the tubes lasts between 62.0 and 95.0 months?

(7) The scores in a standardized test are normally distributed with $\mu = 100$ and $\sigma = 15$. Find the percentage of scores that will fall below 112. A random sample of 10 tests is taken. What is the probability that their mean score $\bar{x}$ is below 112?

(8) The weights (in pounds) of metal discarded in one week by households are normally distributed with a mean of 2.22,lb. and a standard deviation of 1.09,lb. If one household is randomly selected, find the probability that it discards more than 2.00,lb. of metal in a week. Find a weight p_{30} so that the weight of metal discarded by 70% of the houses is above x.

(9) If the salary of computer technicians in the United States is normally distributed with the mean of $\$32,550$ and the standard deviation of $\$2,000$, find the probability for a randomly selected technician to earn More than $\$35,000$. Between $\$31,500$ and $\$35,000$. What is the probability that the mean salary of a random sample of 4 technicians is more than $\$35,000$?

(10) The lifetime of a AAA battery is normally distributed with mean $\mu = 28.5$ hours and standard deviation $\sigma = 5.3$ hours.

(a) For a battery selected at random, what is the probability that the lifetime will be more than 30 hours.
(b) For a sample of three batteries, what is the chance that all three last more than 30 hours?
(c) For a sample of three batteries, what is the probability that their mean lifetime $\bar{x}$ is more than 30 hours? (d)What is the probability that the mean lifetime $\bar{x}$ of batteries from a package of 12 will be less than 27 hours?

(11) The weekly amount a family spends on groceries follows (approximately) a normal distribution with mean $\mu = \$200$ and a standard deviation $\sigma = \$15$.

(a) If \$220 is budgeted for next week's groceries what is the probability that the actual cost will exceed the budget?
(b) How much should be budgeted for weekly grocery shopping so that the probability that the budgeted amount will be exceeded is only 0.05?

Class 10: Normal approximation to the binomial distribution

The Normal distribution can be used to approximate the **Binomial distribution** under certain conditions for n and p. More precisely if n and p satisfy $np > 5$ and $n(1-p) > 5$, we can approximate the binomial distribution $B(n,p)$ (a discrete distribution) by the normal curve $N(\mu,\sigma) = N(np, \sqrt{np(1-p)})$ (a continuous distribution). As the Binomial is sum of Bernouli events, this is nothing but an application of the central limit theorem.

Examples:

(1) Let us consider a binomial with $n = 100$ and $p = .60$. First, we find the mean and standard deviation. Then we use 1000 points in the interval made of four standar deviations before and after the mean, to build our contiuous normal curve $N(\mu, \sigma^2)$.

```
n <- 100
p <- 0.60
n*p
```

```
## [1] 60
```

```
sd=sqrt(n*p*(1-p))
sd
```

```
## [1] 4.898979
```

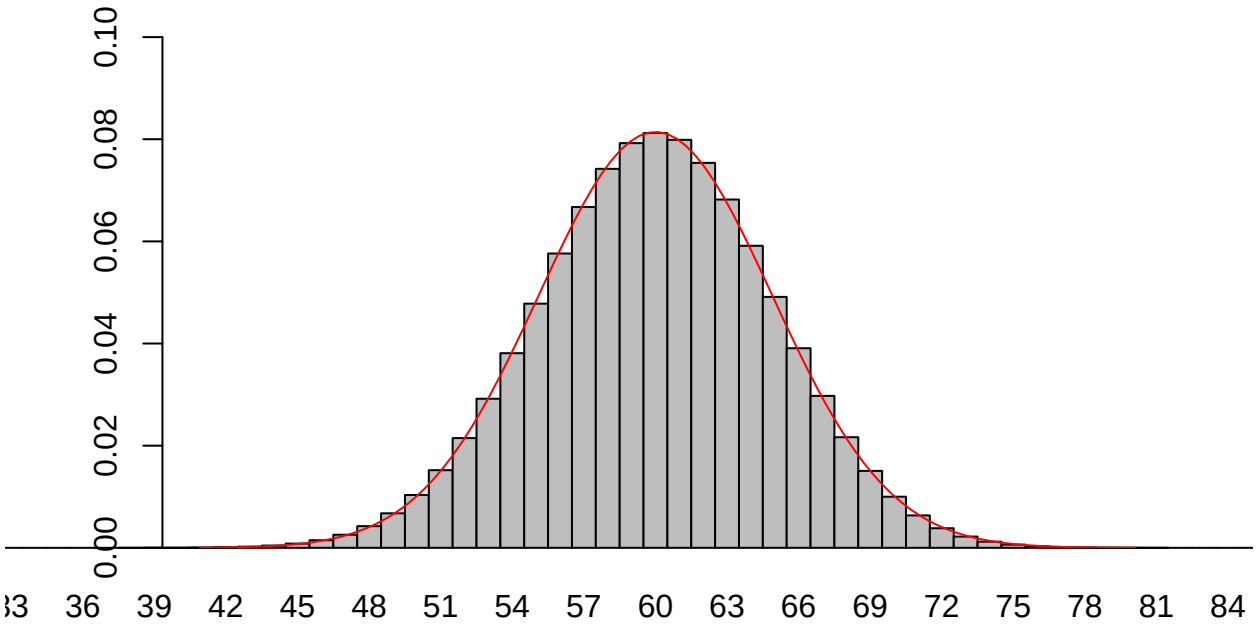
```
m=n*p-4*sd
m
```

```
## [1] 40.40408
```

```
M=n*p+4*sd
M
```

```
## [1] 79.59592
```

```
probs2 = dbinom(1:n, size=n, prob=p)
x <-probs2
barplot(probs2, names.arg=c(1:n), space=0, xlim=c(m,M), ylim=c(0,0.1))
curve(dnorm(c(x+.5), mean=n*p, sd=sqrt(n*p*(1-p))), from=m, to=M, xlim = c(m:M), add=T, col="red")
```



(2) For $n = 125$ and $p = .75$, let us compare the value of the approximation with the actual value if we were to evaluate the binomial for such a high number. First we find the exact value and second we present the approximation:

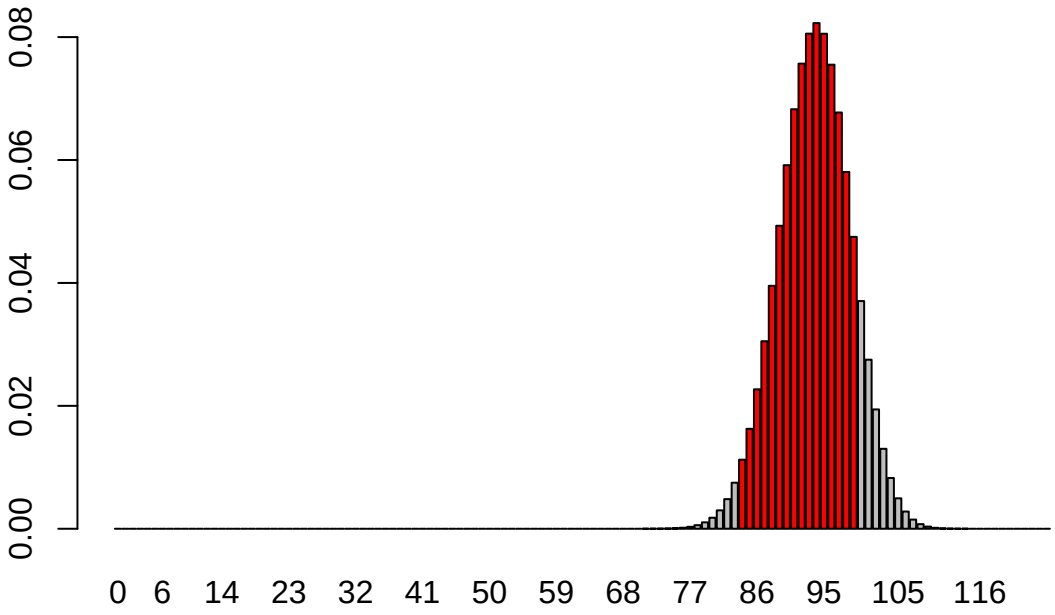
```
n <- 125
P <- 0.75
lb=85; ub=100
mean=n*P
mean
```

```
## [1] 93.75
```

```
sd=sqrt(n*P*(1-P))
sd
```

```
## [1] 4.841229
```

```
data <- dbinom(x=0:n,size=n, prob=P)
names(data) <- 0:n
cols <- rep("grey", n + 1)
r <-c(lb:ub)
cols[r] <- "red"
barplot(data, col = cols)
```




```
sum(dbinom(lb:ub, n, P))
```

```
## [1] 0.8906046
```

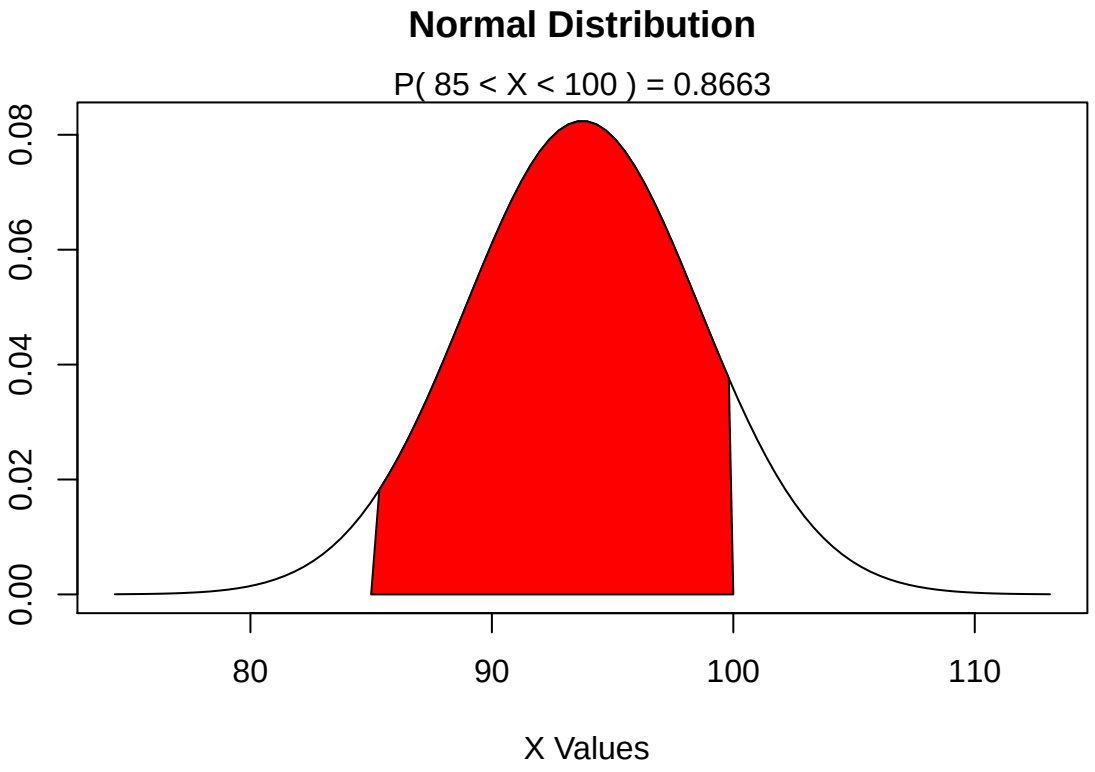
With the use of the normal distribution:

```
x <- seq(-4,4,length=100)*sd + mean
hx <- dnorm(x,mean,sd)

plot(x, hx, type="n", xlab=" X Values", ylab="",
     main="Normal Distribution", axes=TRUE)

i <- x >= lb & x <= ub
lines(x, hx)
polygon(c(lb,x[i],ub), c(0,hx[i],0), col="red")

area <- pnorm(ub, mean, sd) - pnorm(lb, mean, sd)
result <- paste("P(",lb,"< X <",ub,") =",
                signif(area, digits=4))
mtext(result,3)
```



Questions

(1) In Jennifer's Fall 2014 history class, 14 of 34 students passed the class. If you assume a professor's passing rates are constant, would it be appropriate to use a normal curve approximation to the binomial distribution to estimate the mean passing rate for the same professor's Spring 2015 semester class of 28 students? Explain your answer.

(2) According to the Vision Council of America, 75 percent of the U.S. adult population wears some form of glasses to correct their vision. In a random sample of 950 adults, what is the probability that fewer than 700 people wear glasses?

(3) An environmental group did a study of recycling habits in a California community. It found that 70 percent of aluminum cans sold in the area were recycled. If 400 cans are sold in one day, what is the probability that between 260 and 300 will be recycled?

(4) Suppose that a class has 400 and any student drops the class independenly with a probability of $p = .06$.

(a) Check that we can use the normal to approximate the binomial distribution. What are the values of μ and σ?
(b) What is the probability that less than 30 students drop the class?
(c) What is the probability that betweeen 30 and 60 students drop the class?

(5) Suppose that a sample of $n = 1,600$ equipments of the same type are obtained at random from an ongoing production process in which 8% of them are defective. What is the probability that in such a sample 150 or fewer equipments will be defective?

Class 11: The Chi-square distribution

The chi-square distribution $\chi^2(k)$, with $d.f. = k$ degrees of freedom: is the distribution that follows the sum of the squares of k independent standard normal random variables.

The chi-square distribution $\chi^2(k)$ is used primarily for: the χ^2 test of independence in contingency tables and the χ^2 test of goodness of fit of observed data to hypothetical distributions.

In a test of independency: the χ^2 is used to determine if there is a significant relationship between two nominal (categorical) variables. The data can be presented in a contingency table, where the number of observations of type i is denoted by O_i. The null hypothesis is in this case H_0 that there is no relation between the two variables.

The expected frequency E_i of type i is given by

$$E_i = \frac{(\text{Row Total for } i)(\text{Colum Total for } i)}{\text{Sample size}}.$$

and the test statistic

$$\sum_k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2((R-1)(C-1)),$$

where R represents the number of columns and C the number of rows. The null hypothesis H_0 is the assumption that there is no relation between the variables. The alternative hypothesis H_1 is the hypothesis that there is certain relation between the two variables.

The χ^2 goodness of the fit test: determines how well a theoretical distribution (such as normal, binomial, Poisson or simply a prescribed distribution) fits an empirical distribution. In the goodness of the fit test, the population is divided in categories and a theoretical probability or frequency is assigned to each category. Then we get a random sample of size n and count the amount of observed values n_i in each category. The null hypothesis H_0 is that the observed values fit our empirical distribution.

The observed frequency of the category i: is denoted by $O_i = n_i/n$ and the expected frequency by $E_i = np_i$, where n is the sample size and p_i is theoretical probability of the category i. With this notation we can express our hypothesis:

$$H_0 : O_i = E_i, \qquad H_1 : O_i \neq E_i.$$

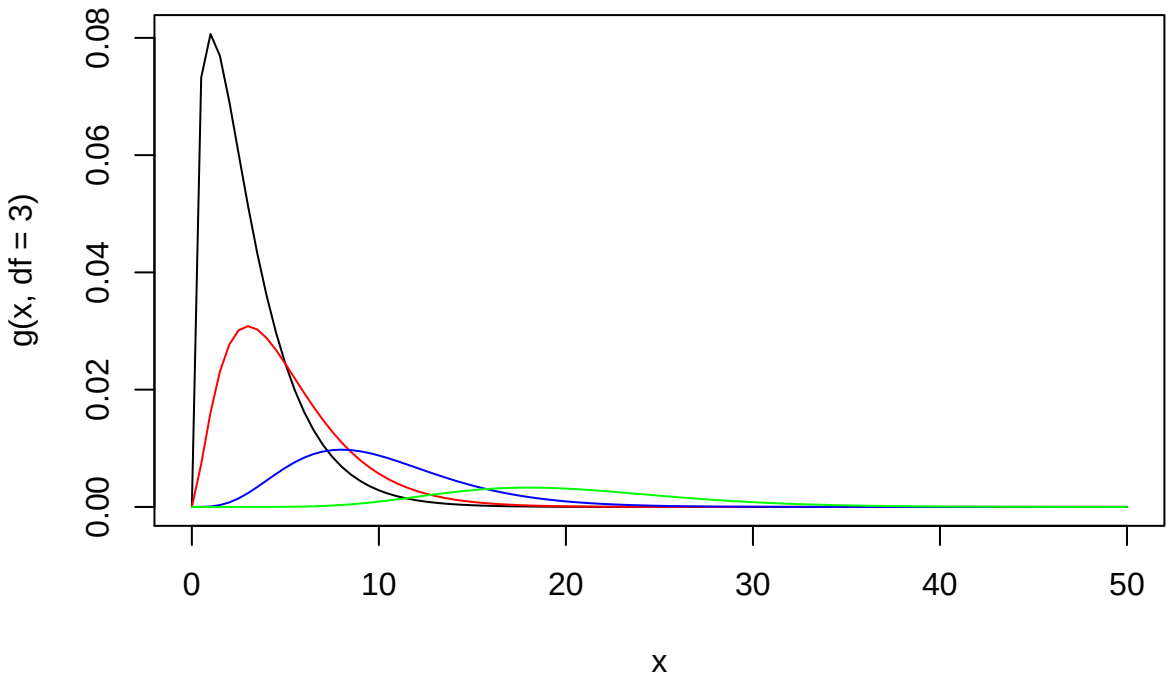
The statistic:

$$\sum_k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(n-1),$$

Examples:

(1) For various degrees of freedom we present, the graph of the Chi-square distribution.

```
g <- function(x, df)  dchisq(x, df)/df
curve(g(x, df = 3), 0, 50)
curve(g(x, df = 5), 0, 50, col="red", add= TRUE)
curve(g(x, df = 10), 0, 50, col = 'blue', add = TRUE)
curve(g(x, df = 20), 0, 50, col = 'green', add = TRUE)
```

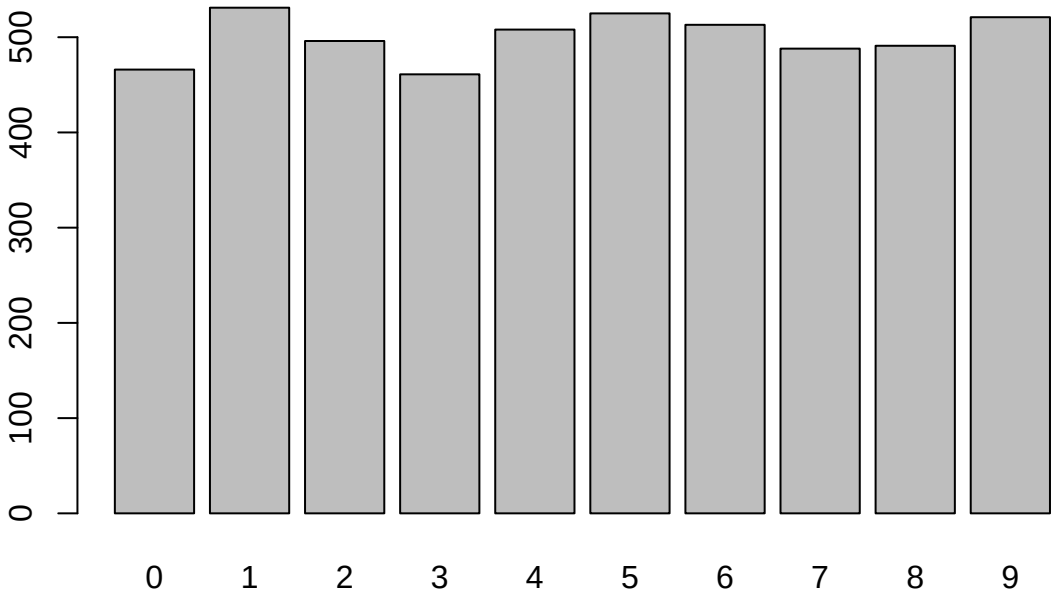


(2) The χ^2 test to determine if the digits of the number π in decimal notation follow a uniform distribution.

```
pi <- read.table('http://www.itl.nist.gov/div898/strd/univ/data/PiDigits.dat',skip=60)
t.pi <- table(pi)
t.pi
```

```
## pi
##   0   1   2   3   4   5   6   7   8   9
## 466 531 496 461 508 525 513 488 491 521
```

```
barplot(t.pi)
```




```
exp <- rep(0.1,10)
chisq.test(t.pi, p=exp)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  t.pi
## X-squared = 10.356, df = 9, p-value = 0.3224
```

Interpretation: The p-value obtained is cosiderable hight .3224 (for example higher than $\alpha = .05$ or even $\alpha = .1$), so we do not have enogh information to reject the null hypothesis H_0 that the all digits are represented in equal proportion.

Questions:

(1) Test the independency of the factors A and B with the factors α, β and γ

	A	B	Row total
α	62	45	107
β	68	94	162
γ	186	220	406

(2) A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of 6	Number of rolls
0	48
1	35
2	15
3	3

(a) Elabore a hypothesis testing experiment to determine if the dice are fair using $\alpha = .05$.

(3) The following table show the results of vacination for 184 individuals with the objective to control a given disease. The total of individuals were divided in two groups to measure the effect of the treatment.

Outcome	Vacinated	Non-vacinated	Row total
Sick with disease	23	5	28
Sich other disease	8	10	18
Non-sick	61	77	138
Column Sum	92	92	184

(a) Perform a χ^2 test to determine if the vacination is actually having an effect on the disease. Use $\alpha = .05$.

Class 12: Confidence intervals and sample size

Estimation: is the process of inferring an unknown parameter using sample data. **A point estimation** for a parameter of the population is given by a single value of a statistic.

Given a population parameter θ and a sample statistic t representing a point estimate for θ, we will like to create an interval estimate with a high confidence of containing the actual parameter θ.

Let c be a real number $0 < c < 1$. The c-confidence interval for θ is an interval $[t - E_c, t + E_c]$ around t such that we will be $100c\%$ confident that it will cover the parameter θ of the entire population. **The statistic E_c is called the margin of error.** The critical value at level c for a continuous random variable X is a number x_c such that

$$P(-x_c < X < x_c) = c.$$

In other words $P(X < -x_c) = \frac{1-c}{2}$ or equivalently $P(X < x_c) = \frac{1+c}{2}$.

Confidence interval for the mean μ: In case we want to estimate the mean μ of the population using the statistic $\bar{x}$, the margin of error takes the shape:

Margin of Error when σ is known	$E_c = z_c \frac{\sigma}{\sqrt{n}}$
Margin of Error when σ is unknown	$E_c = t_c \frac{s}{\sqrt{n}}$

For samples of size n from a normal distribution of mean μ: the quotient of random variables

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}},$$

follows a t-distribution. **The t-distribution depends on one parameter: the degrees of freedom $(d.f.)$.** If we take a sample of n observations from a normal distribution, then the t-distribution with $d.f. = \nu = n - 1$ degrees of freedom can be defined as the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation.

As the degrees of freedom grow larger and larger samples drawn from a normal population resemble more and more the whole population and the t-distribution get closer and closer to a normal standard distribution.

The t-distribution is used to: estimate the μ of a **normal** distribution when the standard deviation σ is **unknown**.

Examples:

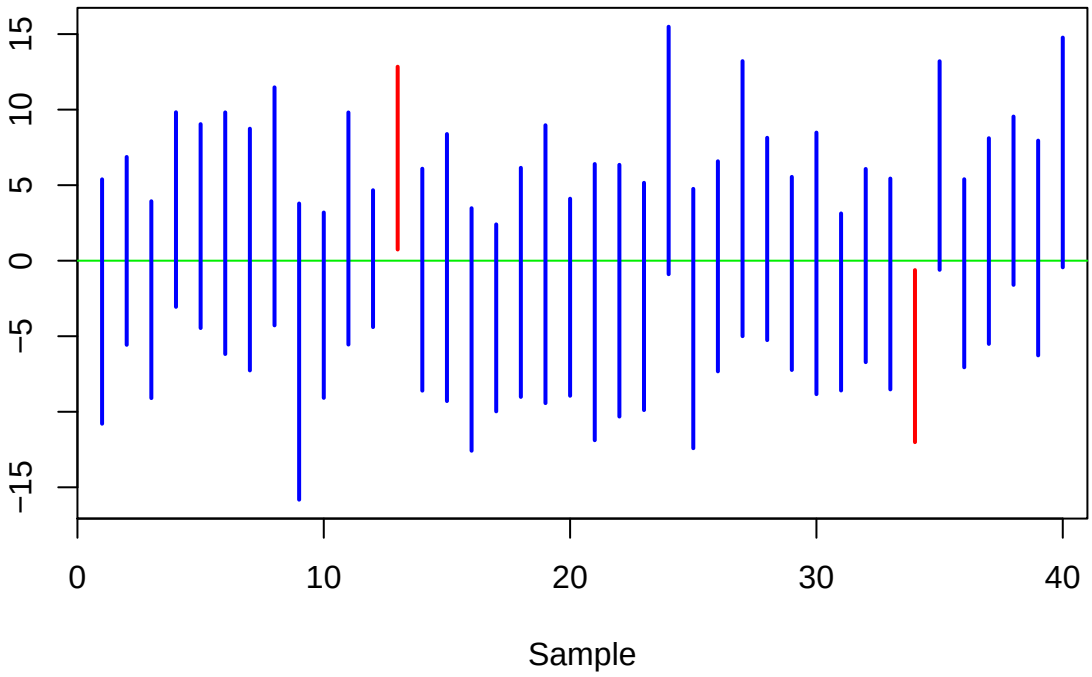
(1) In a simulation of a large amount of samples of the same size from a population, we expect a fraction c of the confidence intervals constructed to contain the actual μ of the population. Let us illustrate this with an experiment with $m = 40$ samples of size $n = 20$ and the 95% confidence interval for mean.

```
set.seed(1776)
m = 40;  n = 20;  mu = 0;  sigma = 15;  conf.level = .95
x = rnorm(m*n, mu, sigma)
MAT = matrix(x, nrow=m)  # m x n matrix: each row a sample of size n
x.bar = rowMeans(MAT);  s = apply(MAT, 1, sd)

t.crit = qt(1-(1-conf.level)/2, n-1)
LCL = x.bar - t.crit*s/sqrt(n);  UCL = x.bar + t.crit*s/sqrt(n)
cover = LCL < mu & UCL > mu
HI = max(UCL); LO = min(LCL)  # to set dimensions of the plot

plot(c(0,m+1), c(HI, LO), col="white", ylab="", xlab="Sample", main="", xaxs="i")
```

```
abline(h=mu, col="green2")
for(i in 1:m) {bar="blue"; if (cover[i]==F) {bar="red"}
lines(c(i,i), c(UCL[i], LCL[i]), col=bar, lwd=2) }
```



(2) Suppose that a random sample of $n = 50$ individuals of a population gives a a sample mean $a = 5.3$. If the standard deviation of the whole population is known to be $\sigma = .4$. Find a 95% confidencd interval for the mean μ of the whole population.

```
a <- 5.3
c <- .95
sigma <- .4
n <- 50
error <- qnorm((c+1)/2)*sigma/sqrt(n)
left <- a-error
right <- a+error
left
```

```
## [1] 5.189128
```

```
right
```

```
## [1] 5.410872
```

(2) Suppose that we are in a similar situation as in the first example, except that we do not have any information about the standard deviation σ of the population, but only know that the sample standard deviation $s = .42$. What is the inteval that we find now for the μ?

```
a <-5.3
c <- .95
s <- .42
n <- 50
error_t <- qt((c+1)/2,df=n-1)*s/sqrt(n)
left_t <- a-error_t
right_t <- a+error_t
left_t
```

```
## [1] 5.180637
```

```
right_t
```

```
## [1] 5.419363
```

(3) Computing the sample size for a known value of sigma: Suppose that the standard deviation σ of the population is known to be $\sigma = 7$ and we want to be 95% confident that the margin of error is less than $E = 2$. How big a sample we need to consider?

```
zstar = qnorm(.975)
sigma = 7
E = 2
zstar^2*sigma^2/E^2
```

```
## [1] 47.05787
```

Interpretation: The result we get is that we need at least 48 individuals in our sample.

Questions

(1) A study is being planned to estimate the mean number of semester hours taken by students at a college. The population standard deviation is assumed to be $\sigma = 4.7$ hours. How many students should be included in the sample to be 99% confident that the sample mean $\bar{x}$ is within one semester hour of the population mean μ for all students at this college?

(2) To determine the mileage of a new model automobile, a random sample of 36 cars was tested. A sample with a mean of 32.6 mpg and a standard deviation of 1.6 mpg was obtained. Construct the 90% confidence interval for the actual mean mpg of the population of this model automobile.

(3) A random sample of 12 employees was taken and the number of days each was absent for sickness was recorded (during a one-year period). If the sample had a mean $\bar{x}$ of 5.03 days and standard deviation s of 3.48 days, create a 95% confidence interval for the population mean days absent for sickness, assuming the distribution of absences is normal.

(4) Computer Depot is a large store that sells and repairs computers. A random sample of 110 computer repair jobs took technicians an average of $\bar{x} = 93.2$ minutes per computer. Assume that σ is known to be 16.9 minutes. Find a 99% confidence interval for the population mean time μ for computer repairs.

(5) The following data represent a sample of the number of home fires started by candles. Assuming that the number of home fires started by candles is approximately normally distributed find a 95% confidence interval for mean number of home fires started by candles each year.

5400 5860 6070 6210 7360 8450 9960

(6) A random sample of 41 NBA players gave a standard deviation $s = 3.32$ inches for their height. How many more NBA players have to be included in the sample to make 95% sure that the sample mean $\bar{x}$ of their height is within 0.75 inch of the mean μ of the height of the population of all NBA players.

(7) A sample of 15 test-tubes tested for number of times they can be heated on Bunsen burner before they cracked gave a sample mean $\bar{x} = 1,230$ hours and a sample standard deviation $s = 270$. Construct 99% confidence interval for the mean μ of the whole population. Assume that the distribution is approximately normal.

(8) Suppose that you want to estimate the mean systolic blood pressure of adults in a country with 95% confidence and a margin of error no larger than 2 mmHg, how many individuals are required? Assume a population variance of $\sigma = 100$.

(9) A 99% confidence interval for the mean number μ of televisions per American household is $(.92, 4.97)$. For each of the following statements about the above confidence interval, choose true or false and explain your answer:

(a) The probability that μ is between .92 and 4.97 is .99.
(b) We are 99% confident that the true mean number μ of televisions per American household is between .92 and 4.97.
(c) 99% of all samples should have $\bar{x}$ between .92 and 4.97.
(d) 99% of all American households have between .92 and 4.97 televisions.
(e) Of many intervals calculated the same way (99% intervals), we expect 99% of them to capture the population mean μ.
(f) Of many intervals calculated the same way (99% intervals), we expect 100% of them to capture the sample mean $\bar{x}$.

Class 13: Testing statistical hypothesis

Hypothesis testing: is a statistical test to decide whether or not there is enough evidence in a sample data to infer that some conclusion is true for the whole population.

H_0**:** The null hypothesis. The statement under investigation, that is usually a statement of "no effect" or "no difference". It represents a statement that we expect to reject.

H_1**:** The alternate hypothesis. An alternate to the null hypothesis that we expect to adopt if the evidence is enough to reject H_0.

The p**-value:** is the probability that we observe results as extreme as the test statistic observed if the null hypothesis H_0 were to be true.

Error of type I: Is the probability α that we reject H_0 when it was in fact true. It represents our willingness of rejecting a true null hypothesis. The number α is also called **the significance level** of the test. An outcome will be considered "unlikely" if its probability is less than α.

Error of type II: Is the probability β of accepting H_0 when it was in fact false.

The probability of rejecting H_0 **when it was in fact false:** is the quantity $1-\beta$ and is call the power of the test.

By increasing the significance level α**:** we are more likely to reject the null hypothesis. This means that we are less likely to accept the null hypothesis when it is false; i.e., less likely to make a Type II error. Hence, the power of the test is increased.

Sample Test Statistics of Test of Hypothesis:

Test Statistic for μ **(**σ **known):** $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Test Statistics for μ **(**σ **unknown):** $t = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$

The rejection zone: Is the portion of the x-axis that represents values as extreme as the level of significance α. If test statistic falls in the rejection zone it means that the probability of observing such an extreme result when H_0 is correct is less than α (p-value $< \alpha$) and we conclude that H_0 should be rejected. Otherwise if the test statistics does not fall in the rejection zone or critical zone (equivalently p-value $> \alpha$), we conclude that there are not enough evidence to reject H_0.

Examples:

(1) Sample problem on left tail: Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is $\sigma = 0.25$ grams. At $\alpha = .05$ significance level, can we reject the claim on food label?

Solution The null hypothesis is $H_0 : \mu \leq 2$ vs the alternate hypothesis $H_1 : \mu > 2$. We begin with computing the test statistic.

```
xbar = 2.1
mu0 = 2
sigma = 0.25
n = 35
z = (xbar-mu0)/(sigma/ sqrt(n))
z
```

```
## [1] 2.366432
```

We then compute the critical value at level of significance .05.

```
alpha = .05
z.alpha = qnorm(1-alpha)
z.alpha
```

```
## [1] 1.644854
```

Answer The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 significance level, we reject the claim that there is at most 2 grams of saturated fat in a cookie.

(2) Sample problem on right tail: Suppose the manufacturer claims that the mean lifetime of a light bulb is more than $10,000$ hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is $\sigma = 120$ hours. At $\alpha = .05$ significance level, can we reject the claim by the manufacturer?

Solution The null hypothesis is that $H_0 \colon \mu \geq 10,000$ and the alternative is $H_1 \colon \mu < 10,000$. We begin with computing the test statistic.

```
xbar = 9900
mu0 = 10000
sigma = 120
n = 30
z=(xbar-mu0)/(sigma/sqrt(n))
z
```

```
## [1] -4.564355
```

We then compute the critical value at $\alpha = .05$ significance level.

```
alpha = .05
z.alpha = qnorm(1-alpha)
-z.alpha
```

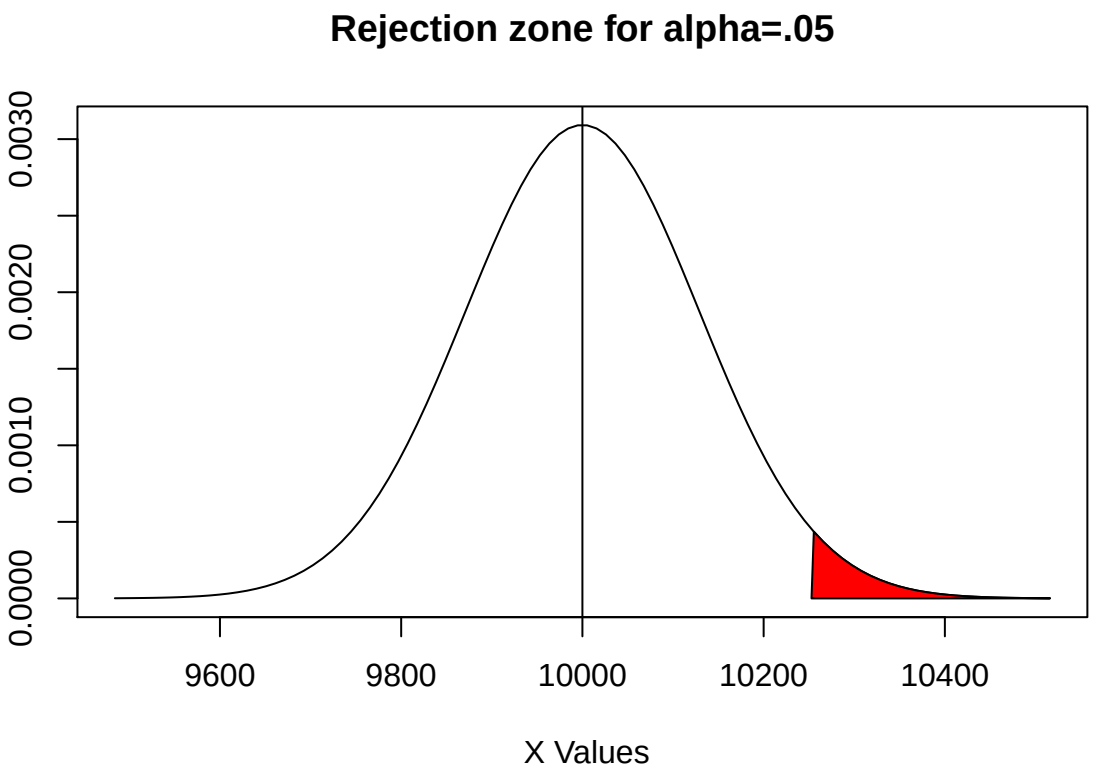
```
## [1] -1.644854
```

Answer: The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at $\alpha = .05$ significance level, we reject the claim that mean lifetime of a light bulb is above $10,000$ hours. The graph of the rejection zone can be done with the code:

```
mean=10000;
sd=129
lb=mean + qnorm(1-.05/2)*sd
ub=mean + 4*sd
x <- seq(-4,4,length=100)*sd + mean
hx <- dnorm(x,mean,sd)

plot(x, hx, type="n", xlab=" X Values", ylab="",
     main="Rejection zone for alpha=.05", axes=TRUE)

i <- x >= lb & x <= ub
lines(x, hx)
polygon(c(lb,x[i],ub), c(0,hx[i],0), col="red")
abline(v=mean)
```



(3) A problem that uses the t-distribution: Suppose the manufacturer claims that the mean lifetime of a light bulb is more than $10,000$ hours. In a sample of 30 light bulbs, it was found that they only last $9,900$ hours on average. Assume the sample standard deviation is 125 hours. At .05 significance level, can we reject the claim by the manufacturer?

Solution: The null hypothesis is that $H_0\colon \mu \geq 10,000$ and the alternate is $H_1\colon \mu < 10,000$. We begin with computing the test statistic.

```
xbar = 9900
mu0 = 10000
s = 125
n = 30
t = (xbar-mu0)/(s/sqrt(n))
t
```

```
## [1] -4.38178
```

We then compute the critical value at .05 significance level.

```
alpha = .05
t.alpha=qt(1-alpha, df=n-1)
-t.alpha
```

```
## [1] -1.699127
```

Answer: The test statistic -4.3818 is less than the critical value of -1.6991. Hence, at $\alpha = .05$ significance level, we can reject the claim that mean lifetime of a light bulb is above $10,000$ hours.

(4) A sample problem using the two-tails: Suppose the mean weight of King Penguins found in an antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

Solution The null hypothesis is that $H_0: \mu = 15$ vs the alternate hypothesis $H_1: \mu \neq 15$. We begin with computing the test statistic.

```
xbar = 14.6
mu0 = 15.4
sigma = 2.5
n = 35
z=(xbar-mu0)/(sigma/sqrt(n))
z
```

```
## [1] -1.893146
```

We then compute the critical values at .05 significance level.

```
alpha = .05
z.half.alpha = qnorm(1-alpha/2)
c(-z.half.alpha, z.half.alpha)
```

```
## [1] -1.959964  1.959964
```

Answer: The test statistic -1.8931 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year.

Questions

(1) Gregor Mendel was a pioneer in the theory of genetics. His idea was to assign probabilities to significant population traits of plants or animals, like eye color, based on "dominant" or "recessive" traits. For example, he studied peas with green pods (a dominant trait) or yellow pods (a recessive trait). He predicted that the probability that a hybrid ("offspring) of a green pea with a yellow pea will have a yellow pod is $p = 0.25$. Mendel conducted an experiment of green-yellow hybrids. In one experiment, 428 offspring had green pods and 152 offspring had yellow pods. Use a level of significance of $\alpha = 0.01$ to test the claim that Mendel's claim that $p = 0.25$ is wrong.

(2) A teacher has developed a new technique for teaching which he wishes to check by statistical methods. If the mean of a class test turns out to be 60 (or less), the results will be considered unsuccessful. Alternatively, if the mean is greater than 60, the results will be considered successful. The results of the test with a class of 36 students had a mean $\bar{x} = 66.2$ with a standard deviation of $s = 24.0$. Test whether the results were successful at the $\alpha = 5\%$ level of significance. (Use 1-tail test.) State the null and the alternate hypothesis and include diagrams.

(3) The average annual salary of employees at a retail store was $28,750$ last year. This year the company opened another store. Suppose a random sample of 18 employees had an average annual salary of $\bar{x} = \$25,810$ with sample standard deviation of $s = 4230$. Use a level of significance $\alpha = 1\%$ to test the claim that the average annual salary for all employees is different from last years average salary. Assume salaries are normally distributed.

(4) A machine in the lodge at a ski resort dispenses a hot chocolate drink. The average cup of hot chocolate is supposed to contain $\mu = 7.75$ ounces. We may assume that x has a normal distribution with $\sigma = 0.3$ ounces. A random sample of 16 cups of hot chocolate from this machine had a mean content of $\bar{x} = 7.62$ ounces. Use a $\alpha = 0.05$ level of significance and test whether the mean amount of liquid is different than 7.75 ounces.

(5) A group of 100 resistors have an average of 102 Ohms. Assuming a population standard deviation $\sigma = 8$ Ohms, test whether the population mean is $\mu = 100$ Ohms at a significance level of $\alpha = 0.05$. Do the two-tailed test. What will be result if we choose instead $H_1: \mu > 100$ and do the right-tail test?

(6) Suppose that we have a population with distribution approximately normal and draw a sample of $n = 10$ individuals. Sketch the graph of the rejection region for a right tail hypothesis testing with

unknown variance σ and level of significance $\alpha = .05$. Compare with the graph of the rejection zone for the same α if we know the σ of the entire popultion.

(7) It is claimed that a vacuum cleaner consumes 46 kWh per year. A random sample of 12 homes indicates that vacuum cleaners expend an average of 42 kWh per year with (sample) standard deviation 11.9 kWh. At a 0.05 level of significance, does this suggest that, on average, vacuum cleaner expend less than 46 kWh per year? Assume the population to be normally distributed. Does the result of the test changes if we gain extra information that the population standard deviation is in fact $\sigma = 12$? Include a graph of the rejection region in each case.

(8) What is the effect of increasing the sample size in the type I and type II errors?

(9) A Type II error is made when

(a) the null hypothesis is accepted when it is false.
(b) the null hypothesis is rejected when it is true.
(c) the alternate hypothesis is accepted when it is false.
(d) the null hypothesis is accepted when it is true.
(e) the alternate hypothesis is accepted when it is true.

(10) A Type I error is made when

(a) the null hypothesis is accepted when it is false.
(b) the null hypothesis is rejected when it is true.
(c) the alternate hypothesis is accepted when it is false.
(d) the null hypothesis is accepted when it is true.
(e) the alternate hypothesis is accepted when it is true.

(11) How many Kleenex should a package of tissues contain? Researchers determined that 60 tissues is the average number of tissues used during a cold. Suppose a random sample of 100 Kleenex users yielded the following data on the number of tissues used during a cold: $\bar{x} = 52$, $s = 22$. Using the sample information provided, calculate the value of the test statistic t.

(12) A pharmaceutical company claims that its weight loss drug allows women to lose in average of $\mu = 8lb$ after one month of treatment. If we want to conduct an experiment to determine if the patients are losing less weight than advertised, what would be the null H_0 and alternative hypothesis H_1?

(13) Suppose our p-value is .047. What will our conclusion be at alpha levels of $\alpha = .10$, $\alpha = .05$ and $\alpha = .01$? Explain your selection.

(a) We will reject H_0 at $\alpha = .10$, but not at $\alpha = .05$
(b) We will reject H_0 at $\alpha = .10$ or .05, but not at $\alpha = .01$
(c) We will reject H_0 at $\alpha = .10$, .05, or .01
(d) We will not reject H_0 at $\alpha = .10$, .05, or .01

(14) Suppose the p-value for a test is .02. Which of the following is true? Explain your selection.

(a) We will not reject H_0 at $\alpha = .05$
(b) We will reject H_0 at $\alpha = .01$
(c) We will reject H_0 at $\alpha = 0.05$
(d) We will reject H_0 at alpha equals 0.01, 0.05, and 0.10
(e) None of the above is true.

(15) A survey was conducted to get an estimate of the proportion of smokers among the graduate students. Report says 35% of them are smokers. Lida doubts the result and thinks that the actual proportion is much less than this. Choose the correct choice of null and alternative hypothesis Lida wants to test. Explain your selection.

(a) $H_0 : p = .35$ versus $H_1 : p \neq .35$.
(b) $H_0 : p = .35$ versus $H_1 : p > .35$.
(c) $H_0 : p = .35$ versus $H_1 : p < .35$.

(d) None of the above

(16) The null hypothesis $H_0 : \mu = .5$ against the alternative $H_1 : \mu > .5$ was rejected at level $\alpha = 0.01$. Pete wants to know what the test will result at level $\alpha = 0.10$. What will be his conclusion? Explain your selection.

(a) Reject H_0.
(b) Fail to Reject H_0.
(c) No conclusion can be made.
(d) Reject H_1.

(17) The null hypothesis $H_0 : \mu = 5$ against the alternative $H_1 : \mu > 5$ was rejected at certain level of significance. What will be the conclusion for testing $H_0 : \mu = 5$ against the alternative $H_1 : \mu \neq 5$ at the same level? Explain your selection.

(a) Fail to Reject H_0.
(b) Reject H_0.
(c) No conclusion can be made.
(d) Reject H_1.

(18) A researcher wanted to test the null hypothesis $H_0 : \mu = 10$ vs. $H_1 : \mu > 10$. She obtained that a sample statistic $\bar{x} = 10.5$ with a sample size of $n = 20$ did not provide enough evidence to reject H_0 at a significance level $\alpha = .01$. What can we say about the conditional probability

$$p = Pr(\bar{x} \geq 10.5 \mid \mu = 10)?$$

Explain your answer.

(a) $p < .01$
(b) $p > .01$
(c) Both (a) and (b) can occur.
(d) $p = .01$

(19) A null hypothesis was rejected at level $\alpha = 0.10$. What will be the result of the test at level $\alpha = 0.05$? Explain your answer.

(a) Reject H_0.
(b) Fail to Reject H_0.
(c) No conclusion can be made.
(d) Reject H_1.

Class 14: Inferences about differences

Two samples are dependent: if each data value in one sample can be paired with a corresponding value of the other sample.

Consider a random sample of n pairs: assuming that the differences d between the first and second members of each pair follow an approximately normal distribution (this would be true for n big), then the random variable

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

will follows a t distribution with $n - 1$ degrees of freedom.

Two samples are independent: if the selection of sample data from one population is completely unrelated to the selection from the other population.

Differences of the means when σ_1, σ_2 are known values: Suppose that x_1 and x_2 are normally distributed with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. If we take independent random samples of size n_1 and n_2 respectively from the x_1 and the x_2 distributions, the variable $\bar{x}_1 - \bar{x}_2$ will follow a normal distribution with mean $\mu = \mu_1 - \mu_2$ and standard deviation $\sigma = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Confidence interval for $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2 - E < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + E$, where $E = z_c\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Differences of the means when σ_1, σ_2 are unknown: Suppose that x_1 and x_2 are normally distributed with means μ_1 and μ_2. If we take independent random samples of size n_1 and n_2 respectively from the x_1 and the x_2 distributions obtaining sample standard deviations s_1 and s_2 for our samples, the sample test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

will follow approximately a t distribution with degrees of freedom approximately equals to the minimun between $n_1 - 1$ and $n_2 - 1$.

Examples:

(1) Here is an example of how to use the t-test for two independent samples of data. The test aim to determine whether or not there is a significant difference between the mean of the two groups. By default we always have a 95% confidence interval. The null hypothesis is H_0, the mean of the two groups is the the same. The alternate hypothesis is H_1, saying that the two groups have different mean.

```
x <- c(60,41,24,62,73,81,55,56,26,20,18,46,71,44,73,47,50,34,51,56,77,19,70,91,57,51,77)
y <- c(78,83,85,43,43,64,36,37,84,71,48,83,45,86,75,64,66,59,71,76,68,63,59,70,9,91,52,80,62)
t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -2.0757, df = 52.832, p-value = 0.0428
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.3638242  -0.3654222
## sample estimates:
## mean of x mean of y
##  52.96296  63.82759
```

Answer: We compare the p-value with $\alpha = .05$ and decide to reject the H_0 and accept H_1.

(2) Let us compare two results for unpaired data, in the first one we will have enough information to reject $H_0 : \mu_1 = \mu_2$ (with p-value= .00001855), while in the second case we do not have enough evidence agains H_0 with a p-value= .1245.

```
t.test(1:10, y = c(7:20))
```

```
##
##  Welch Two Sample t-test
##
## data:  1:10 and c(7:20)
## t = -5.4349, df = 21.982, p-value = 1.855e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.052802  -4.947198
## sample estimates:
## mean of x mean of y
##       5.5      13.5
```

```
t.test(1:10, y = c(7:20, 200))
```

```
##
##  Welch Two Sample t-test
##
## data:  1:10 and c(7:20, 200)
## t = -1.6329, df = 14.165, p-value = 0.1245
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -47.242900   6.376233
## sample estimates:
## mean of x mean of y
##   5.50000  25.93333
```

(3) In this example, we have paired data and use the t-test again to determine difference in the mean. The data is give as two vectors x and y.

```
x <- c(121,93,105,115,130,98,142,118,125)
y <- c(76,93,64,117,82,80,79,67,89)
t.test(x,y, alternative="greater", paired=TRUE,conf.level = .99)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = 4.3623, df = 8, p-value = 0.001203
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  11.20073      Inf
## sample estimates:
## mean of the differences
##                33.33333
```

Answer: In this second example again we compare the p-value with $\alpha = .01$ and decide to reject the H_0 and accept H_1. Alternatively we can observe that the test statistic is not part of the confidence interval.

Questions:

(1) For a random sample of 36 data pairs, the sample mean of the differences is 0.8 and the standard deviation of the differences is 2. Test the claim that the population mean of the differences is different from 0 at a 5% level of significance.

(2) We have a data consisting of $n = 9$ pairs of observation with sample mean and standard deviation of $\bar{d} = 33.3$ and $s = 22.9$. At the level of significance $\alpha = .01$, test the claim that the mean of the differences is positive.

(3) Two samples of size $n_1 = 10$ and $n_2 = 12$ from two normally distributed populations of unknowns means μ_1 and μ_2 gave sample statistics $\bar{x} = 11$ and $\bar{x}_2 = 10$. Assume that the standard deviations are known though to be $\sigma_1 = 2.5$ and $\sigma_2 = 3$. Can we say that there are significative difference between the means of the populations with $\alpha = .05$? Build the 95% confidence interval for $\mu_1 - \mu_2$.

(4) Given the following data collected in young adults and older adults

Older Adults	Younger Adults
45	34
38	22
52	15
48	27
25	37
39	41
51	24
46	19
55	26
46	36

(a) Determine whether or not, there are significant difference between the mean of the two groups using $\alpha = .05$.