A Short Introduction to R

Table of contents

R programming language

Main reasons for using R:

- R is an open source, high-level language and an environment for statistics
- widely spread in literature (you need it in order to understand the articles).
- widely used in many domains like machine learning, statistics, data mining for medicine, economy, biology, social sciences etc.
- quality back-up and support available
- it has a command line interface, therefore it is very flexible

A little bit of history

- R is a programming language for statistical computing, graphics, numerical simulation, data analysis and handling, statistical models.
- R was initially developed by Ross Ihaka and Robert Gentleman (Department of Statistics of the University of Auckland in New Zealand) in the '90s.
- R was developed as the open source version of the statistical program S (developed by John Chambers at Bell Labs in the '80s).

Resources

- G. James, D. Witten, T. Hastie, R. Tibshirani - An Introduction to Statistical Learning with Applications in R, Springer, 2013.
- T. Fischetti - Data Analysis with R, Packt Publishing, 2015.
- W. J. Braun, D. J. Murdoch - A first course in statistical programming with R, Cambridge University Press, 2007.
- J. M. Chambers - Software for Data Analysis. Programming with R, Springer, 2008.
- Y. Zhao - R and Data Mining: Examples and Case Studies, http://www.rdatamining.com

Installing R

How to install R:

- go to the link https://www.r-project.org/
- choose a mirror website from https://cran.r-project.org/mirrors.html
- the R programming language is available for Windows, Linux, etc.
- choose "install R for the first time" (you want the base package) $\rightarrow$ download R 4.3.1 for (Windows) - click on the icon that appears on the desktop once the Windows setup program has been downloaded from CRAN and you will be guided through the process
- an R icon will be placed on the user's desktop and the R system can be started by double-clicking on that icon.

Current directory:

- identification: getwd( )
- changing the directory: setwd(...) or File $\rightarrow$ Change dir...
- objects in the workspace: objects()

Instructions:

- are written after the Command prompt (>)
- are executed after typing Enter
- an unfinished command can be continued on the next line using +
- commands can be written and saved in a script and run with Ctrl+R
- comments begin with #

Getting help in R

Documentation in R: https://cran.r-project.org/manuals.html

- The simplest way to get help in R: the Help menu.
  - R functions (text)...
  - Search Help... which opens the html Help page in your browser.
- In the console:
  - help.start()
  - help.search("data input") #get help about a certain subject
  - help(solve) #get help about a specific function
  - ?solve
  - ??solve #if you are not sure how the function is named
  - example(lm) # to see a worked example

Packages in R

- "Packages" section on the CRAN home page
- see "Table of available packages, sorted by name".

Table 1.1. Libraries used in this book that come supplied as part of the base package of R.

`lattice`	lattice graphics for panel plots or trellis graphs
`MASS`	package associated with Venables and Ripley's book entitled *Modern Applied Statistics using S-PLUS*
`mgcv`	generalized additive models
`nlme`	mixed-effects models (both linear and non-linear)
`nnet`	feed-forward neural networks and multinomial log-linear models
`spatial`	functions for kriging and point pattern analysis
`survival`	survival analysis, including penalised likelihood

- to use one of the built-in libraries: library(MASS)
- to install packages: install.packages("...")

In R we can do:

- simple calculations: $+, -, *, /, \hat{}, \%\%, \%/\%$
- use built-in constants: LETTERS, letters, month.abb, month.name, pi
- use special constants: Inf, -Inf, NaN, NA, Null
- round numbers: floor(5.3), ceiling(5.3), round(5.3,0), round(pi,3)
- change the no. of significant digits printed: options(digits=x)
- compute logarithms: $\log(10)$, $\log 10(6)$, $\log(9, 3)$
- compute exponentials: exp(1)
- use mathematical functions
- assign variables: $\leftarrow$, $=$
- use operators: logical (&, |, !), relational ($==$, $<$,$>$, $<=$, $>=$, $!=$)

Mathematical functions in R

Mathematical functions used in R.

Function	Meaning
`log(x)`	log to base e of x
`exp(x)`	antilog of x (e^x)
`log(x,n)`	log to base n of x
`log10(x)`	log to base 10 of x
`sqrt(x)`	square root of x
`factorial(x)`	$x! = x \times (x-1) \times (x-2) \times \cdots \times 3 \times 2$
`choose(n,x)`	binomial coefficients $n!/(x!\,(n-x)!)$
`gamma(x)`	$\Gamma(x)$, for real x $(x-1)!$, for integer x
`lgamma(x)`	natural log of $\Gamma(x)$
`floor(x)`	greatest integer less than x
`ceiling(x)`	smallest integer greater than x
`trunc(x)`	closest integer to x between x and 0, e.g. trunc(1.5) = 1, trunc(–1.5) = –1; trunc is like floor for positive values and like ceiling for negative values
`round(x, digits=0)`	round the value of x to an integer
`signif(x, digits=6)`	give x to 6 digits in scientific notation
`runif(n)`	generates n random numbers between 0 and 1 from a uniform distribution
`cos(x)`	cosine of x in radians
`sin(x)`	sine of x in radians
`tan(x)`	tangent of x in radians
`acos(x), asin(x), atan(x)`	inverse trigonometric transformations of real or complex numbers
`acosh(x), asinh(x), atanh(x)`	inverse hyperbolic trigonometric transformations of real or complex numbers
`abs(x)`	the absolute value of x, ignoring the minus sign if there is one

Variables

- assigning variables: $a < -3$, $-6.2- > b$, $c = 7$
- removing a variable: rm(a); rm(a,b)
- a variable name can be created using letters, digits, periods, and underscores
- R is case sensitive
- reserved words cannot be used as variable names (q, c, T, F, etc).

Objects

Everything in R is an object!

R has 6 basic data types (the ones listed below and *raw*):

- character: "a", "swc"
- numeric (real numbers): 12, 23.5
- integer: 4L
- complex: 2-3i
- logical: TRUE (T), FALSE (F)

Attributes

R provides many functions to examine features of objects:

- class() - what kind of object is it?
- typeof() - what is the object's data type?
- length() - what is the size of the object?
- attributes() - does it have any metadata?

Vectors

- The most basic object in R is an **atomic vector** (atomic means all the elements have the same type).
- There is also the object **list** which is represented as a vector but can contain objects of different types.
- Create a vector: "vector()" or "c()"

- $x \leftarrow c(0.5, 0.6)$
- $y = c(TRUE, FALSE)$
- $z = c("a","b","c")$
- $x1 = 9 : 26$
- $x2 = c(1 + 2i, 2 - 3i)$
- c(x1,x2) $\rightarrow$ concatenate vectors

- Coercion - from lower to higher types, from logical to integer to double to character: $v \leftarrow c(1, 5.4, TRUE, "hello")$
- Other ways of creating vectors:
  - $x \leftarrow scan()$ #another way of creating a vector
    1:-3
    2:6
    3:
  - $x \leftarrow seq(2, 0, -0.5)$ # creates the sequence form 2 to 0 with a step of -0.5
  - $x \leftarrow seq(1, 5, length.out = 4)$
  - $x \leftarrow numeric(20)$ # creates a vector of length 20 with elements equal to 0
  - $x \leftarrow rep("a", 5)$ # repeats the element "a" five times
  - $x \leftarrow rep(1:6, 2)$ # repeats the sequence 1:6 twice
  - $x \leftarrow rep(c(-1, 2, 3), c(2, 4, 1))$

- Operations with vectors are done element by element:
  - $x \leftarrow c(6, 3, -2, 4); y \leftarrow c(-1, 5, 2, 3)$
  - $x + 2$
  - $2 * x$
  - $x + y$
  - $x * y$
  - $x\hat{}(-3)$
  - $x\hat{}y$
- When the vectors do not have the same length, the shorter one is recycled.
- Vector comparison is done element by element and the result is a vector of logical values.

Functions for vectors:

- length(x) $\rightarrow$ gives the number of elements of a vector
- mean(x) $\rightarrow$ average of the elements of a vector
- max(x) $\rightarrow$ maximum value of a vector
- min(x) $\rightarrow$ minimum value of a vector
- sum(x) $\rightarrow$ sums the elements of a vector
- diff(x)
- range(x) $\rightarrow$ range of the vector
- sort(x) $\rightarrow$ sorts the vector in an ascending order
- sort(x,decreasing=T)

Extracting elements from a vector: using square brackets; subscripts and logical subscripts.

- $x \leftarrow 16:1$
- $x[3]$
- $x[3:5]$
- $x[c(1,3,9)]$
- $x[-1] \rightarrow$ drops first element
- $x[-length(x)] \rightarrow$ drops last element
- $x[x > 6] \rightarrow$ elements of x larger than 6 (logical subscripts)
- $x[x\%\%3 == 0] \rightarrow$ elements of x that are divisible with 3
- naming elements of the vector
  $colors \leftarrow c("white","green","red")$
  $y \leftarrow 1:3$
  $names(y) = colors$
  $y["red"]$

Factors, Arrays and Data Frames

Table of contents

Lists

A **list** in R:

- is an object consisting of an ordered collection of objects;
- is an one-dimensional, heterogeneous data structure;
- similar to a vector, but the elements can have different types;
- is created with the function **list()**.

Creating lists:

- a = c(2, 3, 5)
- $char = c("aa","bb","cc","dd","ee")$
- $b = c(TRUE, FALSE, TRUE, TRUE)$
- $list1 = list(a, char, b, 3)$

Member referencing - using "[[...]]"

- list1[[2]] $\rightarrow$ the user gets a copy of the vector "char"
- list1[[2]][3] $\rightarrow$ the third element in the second element of the list

List slicing:

- we retrieve a list slice with the single square bracket "[]" operator;
- the following is a slice containing the second member of *list1*, which is a copy of *char*
  - > list1[2]
    [[1]]
    [1] "aa" "bb" "cc" "dd" "ee"
  - > list1[c(2,4)] → slicing multiple members

Naming the elements of a list:

- names(list1)←c("e1", "e2", "e3", "e4")
- students←list(Id=1:5,name=c("John", "Alan", "Susan", "Cathy", "Sam"), grade=c(8,9,7,8,10), nrstud=5)
- students["name"]
- students[["grade"]]
- students[[c("name","grade")]][1]
- students$grade

- Deleting elements from a list:
  - > list1[-2]
  - > list1[[3]][-1]
- replacing an element in a list:
  - > list1[[1]]=5:1
  - > list1[[2]][3]="hh"
  - > students$name[5]="Tom"
- concatenating two lists:
  - > list2=list(seq(2,10,2))
  - > list3=c(list1,"elem4"=list2)
- adding an element to a list: list3[["elem5"]]=c(T, T, F, F, F)
- transforming a list into a vector: vec=unlist(list1)

Factors

A factor:

- represents the values of a **categorial** variable as a vector with integer elements $\{1, 2, 3, ...k\}$ (k is the number of categories) and an internal vector of characters strings representing the levels asociated to the k integers.

Types of variables:

- **quantitative**: discrete or continuous $\rightarrow$ are represented by numeric vectors
- **qualitative**: binomial (0 or 1), nominal (names), ordinal (orders) $\rightarrow$ are represented using **factors**.

Example: consider a survey that has data on 71 females and 72 males:

> gender $\leftarrow$ c(rep("female",71), rep("male",72))
> class(gender) $\rightarrow$ character
> gender_f$\leftarrow$ factor(gender)
> class(gender_f) $\rightarrow$ factor
> levels(gender_f)

We can change the levels in a factor:

- > levels(genderf)$\leftarrow$ c("F","M")
- > gender_f
- > as.numeric(gender_f) $\rightarrow$ internally the factor gender_f is represented by a numeric vector 1,1,...,2,2..., where 1 represents "female" and 2 "male"

We can create ordered factors for ordinal variables (the levels can be compared):

- > size$\leftarrow$ c("small", "large", "small", "medium", "large", "medium")
- > sizef $\leftarrow$ factor(size)
- > sizef_ordered $\leftarrow$ factor(size, levels=c("small","medium","large"), ordered=T)

Arrays

An array:

- is a multi-dimensional object (a vector is an one-dimensional array);
- has the attribute **dim** which gives the maximal indices in each dimension.

Creating an array:

> x← 1:24

> dim(x) ← c(2,4,3) → the array consists of 3 matrices with dimensions 2 by 4

```
, , 1

     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8

, , 2

     [,1] [,2] [,3] [,4]
[1,]    9   11   13   15
[2,]   10   12   14   16

, , 3

     [,1] [,2] [,3] [,4]
[1,]   17   19   21   23
[2,]   18   20   22   24
```

Extracting elements from an array:

> x[2,1,3] → the element in the third table, on the second line and first column

Matrices

A matrix

- is a two-dimensional array;
- is created in several ways:
  - using **matrix()**;
  - using **rbind()** or **cbind()**;
  - by changing the dimensions attribute of a vector
- by default the elements are stored by column.

Examples:

```
> M1 ← matrix(1:9, nrow=3)
> M2 ← matrix(1:12, ncol=4)
> v ← c(1:4,4:1); M3 ← matrix(v, nrow=2, byrow=T)
> class(M1)
> attributes(M1)
```

Other ways to create a matrix:

```
> dim(v) <- c(2,4)
> is.matrix(v)
> x <- c(1, 0, 0, 0)
> y <- 4:1
> M4 <- rbind(x, y)
> M5 <- cbind(x, y)
```

We can also add rows and columns to matrices with these functions.

Naming the rows and columns of matrices

- using the functions **rownames()** and **colnames()**;
- at first, matrices have numbers naming their rows and columns.

```
> m<-matrix(1:12,nrow=3)
> m
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> rownames(m)<-c("r1","r2","r3")
> colnames(m)<-c("c1","c2","c3","c4")
> m
   c1 c2 c3 c4
r1  1  4  7 10
r2  2  5  8 11
r3  3  6  9 12
```

Extracting an element from a matrix

```
> m[1,2]#element on the first row, second column
[1] 4
> m[2,]#second line
c1 c2 c3 c4
 2  5  8 11
> m[,1]#first column
r1 r2 r3
 1  2  3
> m["r2",]
c1 c2 c3 c4
 2  5  8 11
> m[,"c3"]
r1 r2 r3
 7  8  9
> m[-1,]#first row is left out
   c1 c2 c3 c4
r2  2  5  8 11
r3  3  6  9 12
> m[,-2]# second column is left out
   c1 c3 c4
r1  1  7 10
r2  2  8 11
r3  3  9 12
```

Operations with matrices I

- add and multiply two matrices
- multiply a matrix with a number
- compute the transpose of a matrix
- compute the determinant of a square, nonsingular matrix
- find the inverse of a nonsingular matrix
- compute the sum on rows or columns

Operations with matrices II

```
> m1<-matrix(1:9,ncol=3,byrow=T)
> m2<-matrix(c(-9,3,5,4,8,6,-1,-4,-6),nrow=3)
> m1
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
> m2
     [,1] [,2] [,3]
[1,]   -9    4   -1
[2,]    3    8   -4
[3,]    5    6   -6
> 2*m1
     [,1] [,2] [,3]
[1,]    2    4    6
[2,]    8   10   12
[3,]   14   16   18
> m1+m2
     [,1] [,2] [,3]
[1,]   -8    6    2
[2,]    7   13    2
[3,]   12   14    3
> m1%*%m2
     [,1] [,2] [,3]
[1,]   12   38  -27
[2,]    9   92  -60
[3,]    6  146  -93
```

```
> det(m1)
[1] 6.661338e-16
> solve(m1)
Error in solve.default(m1) :
  system is computationally singular: reciprocal condition number = 2.59052e-18
> det(m2)
[1] 230
> solve(m2)
             [,1]       [,2]        [,3]
[1,] -0.104347826 0.07826087 -0.03478261
[2,] -0.008695652 0.25652174 -0.16956522
[3,] -0.095652174 0.32173913 -0.36521739
> m2^(-1)
           [,1]      [,2]       [,3]
[1,] -0.1111111 0.2500000 -1.0000000
[2,]  0.3333333 0.1250000 -0.2500000
[3,]  0.2000000 0.1666667 -0.1666667
```

```
> #Calculations on rows or columns of the matrix
> mean(m1[,2])
[1] 5
> sum(m1[3,])
[1] 24
> rowSums(m1)
[1]  6 15 24
> colSums(m1)
[1] 12 15 18
> rowMeans(m2)
[1] -2.000000  2.333333  1.666667
> colMeans(m2)
[1] -0.3333333  6.0000000 -3.6666667
```

Data frames

- a **data frame** is a two-dimensional list containing (potentially a mix of) numbers, text or logical variables in different columns
- a data frame is a more general object than a matrix, in the sense that different columns may have different types
- all elements of any column must, however, have the same mode, i.e., all numeric, or all factor, or all character, or all logical
- all the columns must have the same length
- **data.frame()** creates a dataframe in R:

```
> df=data.frame(x=c(1,2,3),y=c("a","b","c"))
> df
  x y
1 1 a
2 2 b
3 3 c
```

In the package MASS there is a dataset called **bacteria**:

```
> library(MASS)
> head(bacteria)
  y ap hilo week  ID     trt
1 y  p   hi    0 X01 placebo
2 y  p   hi    2 X01 placebo
3 y  p   hi    4 X01 placebo
4 y  p   hi   11 X01 placebo
5 y  a   hi    0 X02   drug+
6 y  a   hi    2 X02   drug+
> colnames(bacteria)#rownames(bacteria)
[1] "y"    "ap"   "hilo" "week" "ID"   "trt"
> bacteria[1:3, 2:3] # Rows 1-3 and columns 2-3
  ap hilo
1  p   hi
2  p   hi
3  p   hi
> bacteria[, 2:3] # Columns 2-3 (all rows)
   ap hilo
1   p   hi
2   p   hi
3   p   hi
4   p   hi
5   a   hi
```

Functions for data frames:

- **subset()** → offers an alternative way to extract rows and columns:
  > subset(bacteria, ap="p", select=c(hilo, trt))
- **names()** → allows us to change the names of the columns of a data frame.
- **attach()** → allows us to extract columns of a data frame using just the name of the variable (column):
  > bacteria$hilo
  > attach(bacteria)
  > hilo
- **detach(bacteria)** → detaches the data frame
- **str(bacteria)** → information about the data
- **dim(bacteria)** → number of rows and columns
- **edit(bacteria)** → opens the data frame in a spreadsheet format to change values

Some more resources

- Peter Dalgaard, Introductory Statistics with R, 2nd Edition, Springer, 2008
- John Verzani, Using R for Introductory Statistics, 2nd Edition, CRC Press,2014
- Hadley Wickham, Garrett Grolemund, R for Data Science. Import, Tidy, Transform, Visualize and Model Data, O'Reilly, 2017.

Programming with R. Data input.

Table of contents

Conditional execution: if statements I

- the **if** function tests whether a given statement is true; if the statement is true, the succeeding expression is evaluated.
- an **else** can be added to provide an alternative expression to be evaluated in the case where the given statement is false.

Syntax:

> if (condition) {commands when TRUE}

> if (condition) {commands when TRUE} else {commands when FALSE}

> ifelse(condition, true.value, false.value)

Conditional execution: if statements II

- if you put the else part on a new line, you will get an error;
- a common convention for typing **if ... else** is to put the else on the same line as the previous closing brace.

Example

```
> x = 3
> if (x > 2) y = 2 * x else y = 3 * x
```

Example

```
> x = -2
> if (x > 0) s = 1 else
+     if(x == 0) s = 0 else s = -1
```

For loop I

- R programming language allows conditional execution as well as looping constructs.
- The **for()** statement allows one to specify that a certain operation should be repeated a fixed number of times.

Syntax:

- for (name in vector) { commands }
  - this sets a variable called *name* equal to each of the elements of vector, in sequence;
  - for each value, whatever commands are listed within the curly braces will be performed;
  - if there is only one command to execute, the braces are not needed.

For loop II

Example

Some plots:

$x = seq(0, 1, .05)$

$plot(x, x, ylab = "y", type = "l")$

$for(j\ in\ 2:8)\ lines(x, x\hat{}j)$

Example

The factorial n! counts how many ways n different objects could be ordered. It is defined as

$$n! = n * (n-1) * \cdots * 1.$$

Use a **for** *loop in R to compute* 7!.

For loop III

Example

The Fibonacci sequence is a famous sequence in mathematics. The first two elements are defined as $1, 1$. *Subsequent elements are defined as the sum of the preceding two elements. For example, the third element is* $2 = 1 + 1$, *the fourth element is* $3 = 1 + 2$, *the fifth element is* $5 = 2 + 3$, *and so on.*

Use a **for** *loop in R to compute the first 20 terms of the Fibonacci sequence.*

While, repeat loops I

- we may want to repeat statements, but the pattern of repetition isn't known in advance (no of repetitions is unknown)
- we need to do some calculations and keep going as long as a condition holds
- the **while()** statement accomplishes this.

Syntax:

- while (condition) { statements }

The condition is evaluated, and if it evaluates to FALSE, nothing more is done. If it evaluates to TRUE the statements are executed, condition is evaluated again, and the process is repeated.

While, repeat loops II

Example

Suppose we want to list all Fibonacci numbers less than 1000. We don't know beforehand how long this list is, so we wouldn't know how to stop the for() loop at the right time, but a **while()** *loop is perfect. Find these numbers in R.*

What if we don't want to put the condition (test) at the top of the loop?

- use a **repeat** loop
- this loop executes until a **break** statement is executed

While, repeat loops III

Syntax:

- repeat { statements }

This causes the statements to be repeated endlessly. The statements should normally include a break statement, typically in the form

- if (condition) break

but this is not a requirement of the syntax.

- the **break** statement causes the loop to terminate immediately
- break statements can also be used in for() and while() loops
- the **next** statement causes control to return immediately to the top of the loop; it can also be used in any loop

While, repeat loops IV

Example

Newton's method is a popular numerical method to find a root of an algebraic equation $f(x) = 0$. If $f(x)$ has derivative $f'(x)$, then the following iteration should converge to a root of the above equation if started close enough to the root:

$$x_0 \text{ - initial guess}$$

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}.$$

The idea is based on the Taylor approximation

$$f(x_n) \approx f(x_{n-1}) + (x_n - x_{n-1})f'(x_{n-1}).$$

Find the root of the function $f(x) = x^3 + 2x^2 - 7$ with an error of 10^{-6} and an initial guess of 1.

Implicit loops I

apply(), sapply(), lapply(), tapply() functions

Function	Arguments	Objective	Input	Output
apply	apply(x, MARGIN, FUN)	Apply a function to the rows or columns or both	Data frame or matrix	vector, list, array
lapply	lapply(X, FUN)	Apply a function to all the elements of the input	List, vector or data frame	list
sapply	sappy(X FUN)	Apply a function to all the elements of the input	List, vector or data frame	vector or matrix

$tapply(cats\$Bwt, cats\$Sex, mean)$

Functions I

- functions are self-contained units of R code with a well-defined purpose
- in general, functions take inputs, do calculations (possibly printing intermediate results, drawing graphs, calling other functions, etc.), and produce outputs.

The definition of a function normally has the following structure:

- the word **function**
- a pair of round parentheses () which enclose the argument list (the list may be empty)
- a single statement or a sequence of statements enclosed in curly braces {}.

Functions II

Examples:

```
roll=function(){
  die=1:6
  dice=sample(die,size=2,replace=T)
  s=sum(dice)
  return(s)
}
roll()
roll2=function(die){
  dice=sample(die,size=2,replace=T)
  s=sum(dice)
  return(s)
}
roll2(1:12)
```

Data input I

- Reading from external files: **read.table()**, **read.csv()**
- the **scan()** function
- more details on importing and exporting data into R: see the R Data Import/Export manual
- for the **read.table()** the file has to be created in a plain text editor (NotePad)
- the first line usually contains the name of the variables

Price	Floor	Area	Rooms	Age	Cent.heat
52.00	111.0	830	5	6.2	no
54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	no
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes
...					

> HousePrice = read.table("houses.txt", header=TRUE)

- **read.table()** returns a data frame

Data input II

- it expects to find data in a corresponding layout where each line in the file contains all data from one subject (or house, etc.) in a specific order, separated by blanks or, optionally, some other separator
- consider the built-in data set in the ISwR package ("thuesen" dataframe)
  > thuesen2 = read.table("D:/ISwR/thuesen.txt",header=T)
- there are special functions for handling CSV files: **read.csv** and **read.csv2**
- the former assumes that fields are separated by a comma, and the latter assumes an European format for numbers
- both formats have the option "header=T" as the default value
- further variants: **read.delim** and **read.delim2** for reading delimited files (by default, Tab-delimited files)

Data input III

- to edit a data frame, you can use the **edit()** function:
  $>$ aq $\leftarrow$ edit(airquality)

  This brings up a spreadsheet-like editor with a column for each variable in the data frame.

Random variables. Probability distributions.

Table of contents

Random variables I

- A **random variable** is a function defined on a **sample space** (Ω) with values in $\mathbb{R}$ (or $[a, b]$, $\mathbb{N}$, $\{1, 2, \ldots, n\}$, etc.) associated to an **experiment**.
- For example we toss a coin three times and count the number of times Heads appears.
  - experiment=tossing the coin three times (tossing three coins)
  - sample space: $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
  - random variable: **X** is the number of Heads we obtain, $X : \Omega \to \{0, 1, 2, 3\}$, $X(HHH) = 3$, etc.
- Random variables can be discrete or continuous.
  - discrete random variables: number of girls in a family with 3 children, number that appears when we roll a die, the number of jobs submitted to a printer, the number of failed computer components, number of meteor hits in a year, number of accidents on a highway, etc.

Random variables II

- continuous random variables: various times (software installation time, code execution time, connection time, waiting time, lifetime), or physical variables like weight, height, voltage, temperature, distance etc.

- Discrete random variables are described by their **distribution** and the **probability mass function** (**PMF**).
- $X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$ is the distribution of the random variable

$$\sum_{i=1}^{n} p_1 = 1$$

- The PMF is $f(x) = P(X = x)$

$$f(x_i) = p_i,\ i = \overline{1, n},\ \text{otherwise}\ f(x) = 0,\ x \neq x_i$$

Random variables III

- The **cumulative distribution function** (CDF) is defined:

$$F : \mathbb{R} \to \mathbb{R}, F(x) = P(X \leq x) = \sum_{x_i \leq x} p_i.$$

- The **expected value (mean)** and **variance** of a discrete random variable:

$$E[X] = \sum_{i=1}^{n} x_i * p_i,$$

$$D^2[X] = E[(X - E[X])^2] = E[X^2] - E^2[X].$$

Ex.: Determine the distribution for X in the above example and compute its PMF, CDF, expected value and variance.

Random variables IV

- Continuous random variables have values in intervals ($\mathbb{R}$, $\mathbb{R}_+$, $[a, b]$, etc.).
- The PMF shows no information in this case (it is equal to 0).
- The **probability density function** (PDF) of a continuous random variable is an integrable function

$$f : \mathbb{R} \to [0, 1], \int_{-\infty}^{\infty} f(x) = 1.$$

- The **CDF** is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

- **Mean** and **variance**

$$E[X] = \int_{\mathbb{R}} x * f(x)dx, \; D^2[X] = \int_{\mathbb{R}} (x - E^2[X])^2 * f(x)dx.$$

Binomial distribution

- models situations where we **count the number of successes** in a sequence of n independent trials (experiments with only two outcomes)
- number of defective computers in a shipment, the number of updated files in a folder, the number of girls in a family

- The distribution $X : \begin{pmatrix} 0 & 1 & 2 & \dots & n \\ p_1 & p_2 & p_3 & \dots & p_n \end{pmatrix}$

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

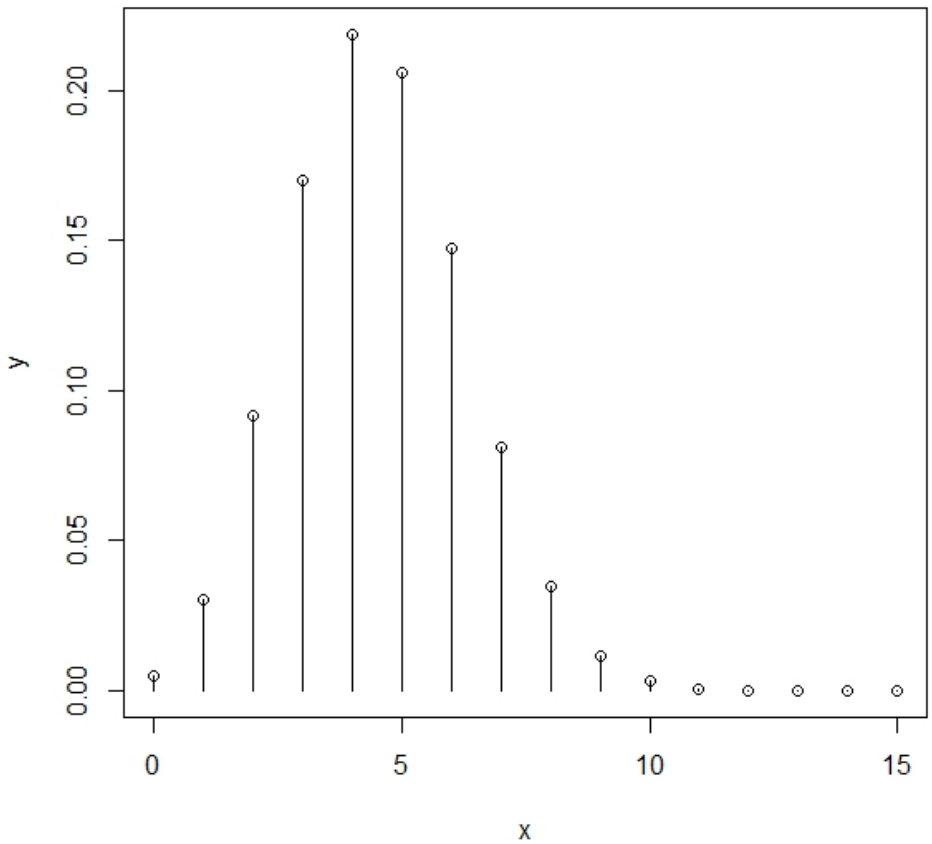
- $E[X] = n * p$, $D^2[X] = n * p * (1-p)$
- These probabilities can be computed using the **dbinom()** function.

Functions for the binomial distribution:

- dbinom(x, size, prob) $\rightarrow$ PMF
- pbinom(q, size, prob, lower.tail = TRUE) $\rightarrow$ CDF
- qbinom(p, size, prob, lower.tail = TRUE) $\rightarrow$ inverse of the CDF (quantile function)
- rbinom(n, size, prob) $\rightarrow$ generates random numbers

Example

Suppose we would like to know how well our student would do on a multiple choice test consisting of 20 questions. The test is passed if the student answers correctly at least half of the questions. What is the probability that he pases the test? What is the probability that he obtains 15p?



PMF of a binomial random variable with $n = 15$ and $p = 0.3$ (Bin(15,0.3)).

Geometric distribution I

- models situations where we **count the numbers of trials needed to get the first success**
- ex.: a search engine goes through a list of sites until it finds a given key phrase; the number of sites visited has a geometric distribution
- from a box of black and white balls we extract a ball with replacement until we find a white one

- The distribution: $X : \begin{pmatrix} 1 & 2 & \dots & k & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{pmatrix}$

$$p_k = P(X = k) = (1 - p)^{k-1} * p$$

- expected value and variance

$$E[X] = \frac{1}{p}, \; D^2[X] = \frac{1 - p}{p^2}$$

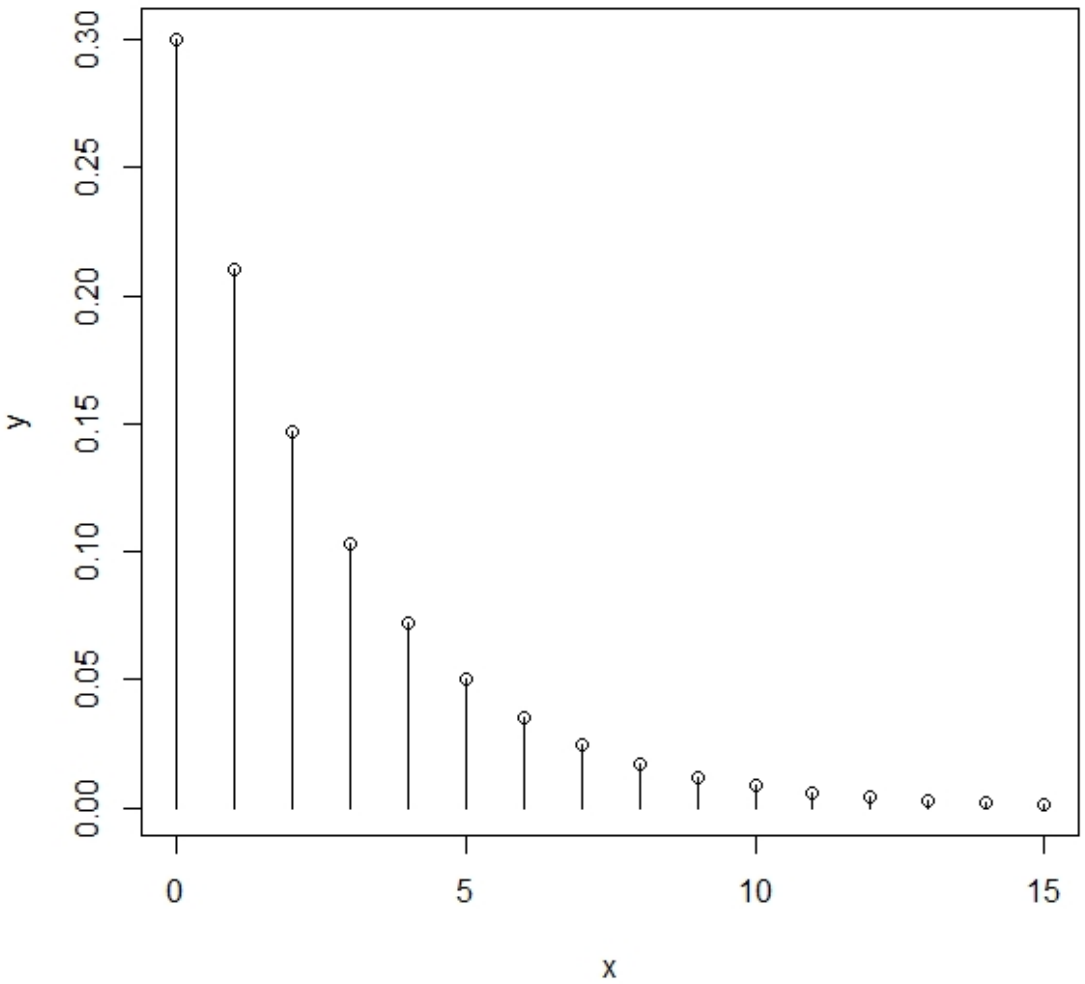
Geometric distribution II

Functions for the geometric distribution:

- dgeom(x, prob)
- pgeom(q, prob, lower.tail = TRUE)
- qgeom(p, prob, lower.tail = TRUE)
- rgeom(n, prob)

Example

What is the probability that the student has to go through 5 questions before he answers one correctly? What is the probability that he has to go through at least 7 questions until he answers one correctly?



PMF of a geometric r.v. with $p = 0.3$.

Poisson distribution I

- **rare events**, or **Poissonian events**: arrivals of jobs at a printer, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, and earthquakes
- counts the number of rare events (unlike to occur simultaneously) in a time period (days, weeks, etc.)
- is the **limit of a sequence of binomial distributions** with parameters n and p_n, where $n \to \infty$ and $p_n \to 0$, but the expected value $np_n \to \lambda$

- The distribution: $X : \begin{pmatrix} 0 & 1 & \dots & k & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{pmatrix}$

$$p_k = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \text{ where k is the rate of events}$$

- expected value and variance

$$E[X] = \lambda,\ D^2[X] = \lambda$$

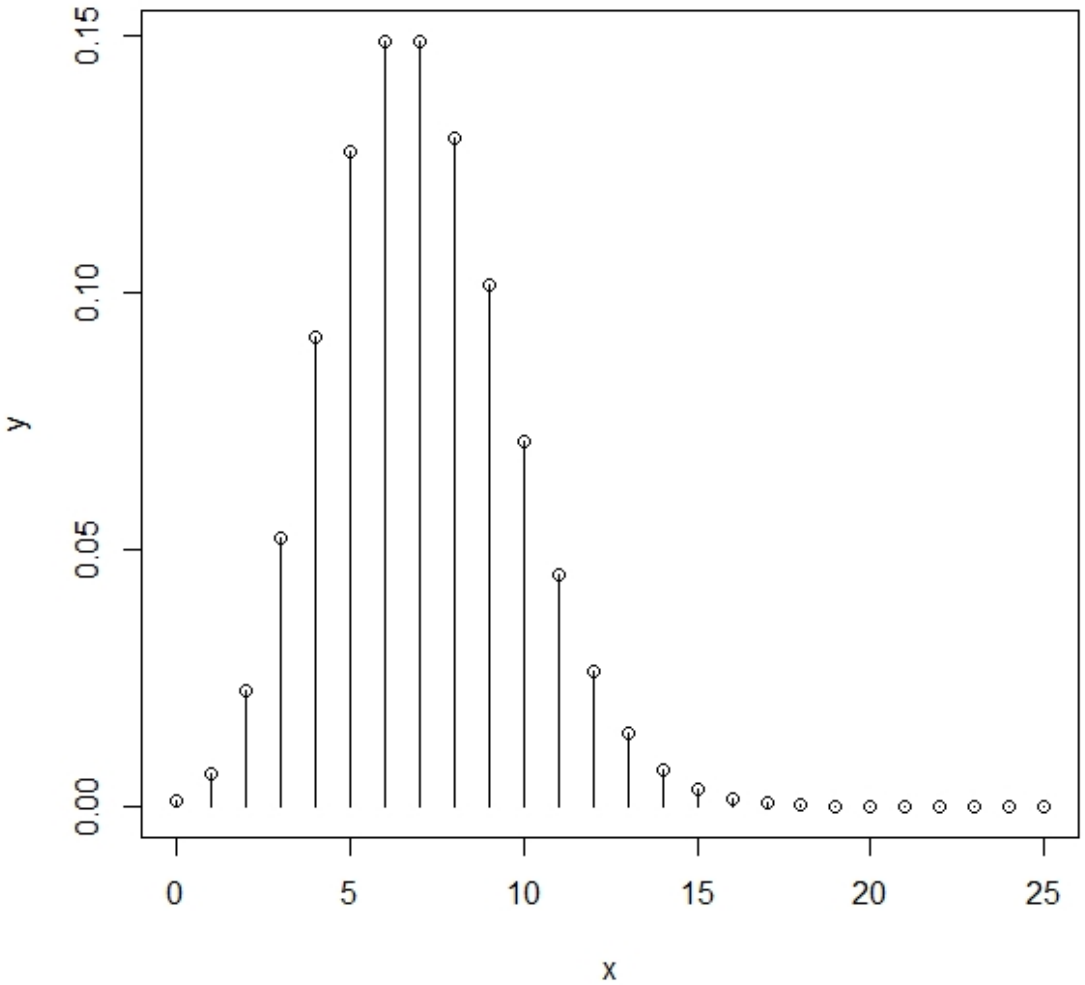
Poisson distribution II

Functions for the Poisson distribution:

- dpois(x, lambda)
- ppois(q, lambda)
- qpois(p, lambda)
- rpois(n, lambda)

Example

(New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day. (a) What is the probability that more than 8 new accounts will be initiated today? (b) What is the probability that more than 16 accounts will be initiated within 2 days?



PMF of a Poisson r.v. with $\lambda = 7$.

Discrete uniform distribution I

Rolling a die is an example of a discrete uniform distribution.

- The distribution: $X : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$.
- $E[X] = \frac{1}{n} \sum_{i=1}^{n} x_n$

Random generating numbers with this distribution:

- sample(min:max,size,replace=T)

Example

You play the following game: a die is thrown twice and if the sum is 7 you win, otherwise you lose. Simultate the game 1000 times and see what your chances are of winning.

Discrete uniform distribution II

Discrete distributions built in to R

Name	rXXXX	Parameters	Range	Mean	Kernel of probability
Binomial	rbinom	size $= n$, prob $= p$	$x = 0, \ldots, n$	np	$\binom{n}{x} p^x (1-p)^{n-x}$
Geometric	rgeom	prob $= p$	$x = 0, 1, \ldots$	$\frac{1-p}{p}$	$(1-p)^x$
Hypergeometric	rhyper	m, n, k	$x = \max(0, k-n), \ldots, \min(m, k)$	$\frac{km}{m+n}$	$\binom{m}{x}\binom{n}{k-x}$
Negative binomial	rnbinom	size $= n$, prob $= p$	$x = 0, 1, \ldots$	$n\frac{1-p}{p}$	$\binom{x+n-1}{x}(1-p)^x$
Poisson	rpois	lambda $= \lambda$	$x = 0, 1, \ldots$	λ	$\frac{\lambda^x}{x!}$

Other distributions: hypergeometric distribution, negative binomial distribution, multinomial distribution.

Uniform continuous distribution I

Uniform distribution: **U[a,b]**

- The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds: a, b.
- All intervals of the same length on the distribution's support are equally probable.

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

- expected value and variance: $E[X] = \frac{a+b}{2}$, $D^2[X] = \frac{(b-a)^2}{12}$

Functions for the uniform distribution $U[a,b]$:

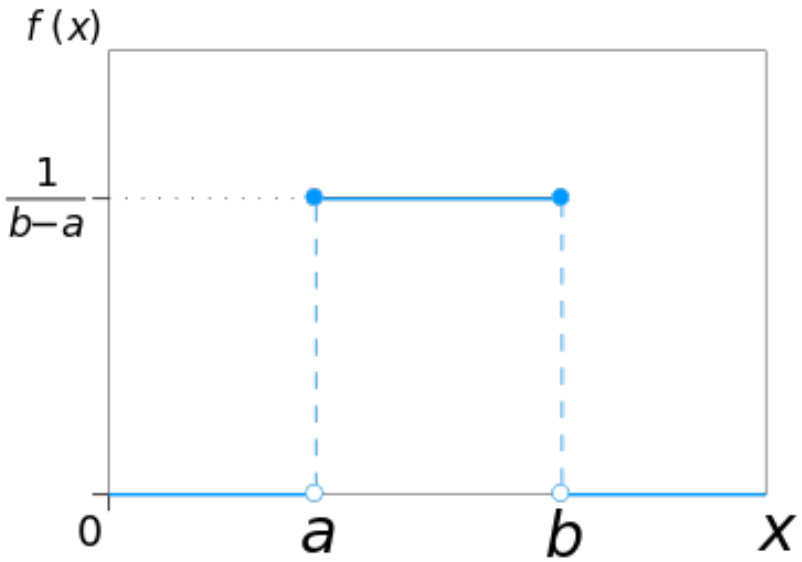
- dunif(d,a=0,b=1)
- punif(q, a=0,b=1)
- qunif(p, a=0,b=1)

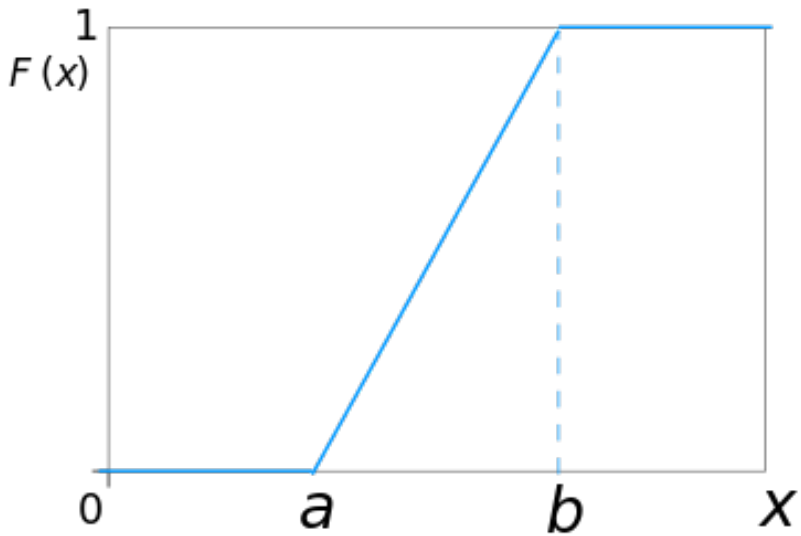
Uniform continuous distribution II

- runif(n, a=0,b=1)

Example

The arrival time of a flight has a uniform distribution on [4 : 50; 5 : 10]. Compute the probability that the flight does not arrive before 5:05 and the expected time of the arrival.





Normal distribution I

- good model for physical variables like **weight, height, temperature, voltage, pollution level**, and for instance, **household incomes or student grades**
- notation $N(\mu, \sigma^2)$, where the parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ^2 is its variance
- physical quantities that are expected to be the **sum of many independent processes**, such as measurement errors, often have distributions that are nearly normal

Normal distribution II

$$PDF: \ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

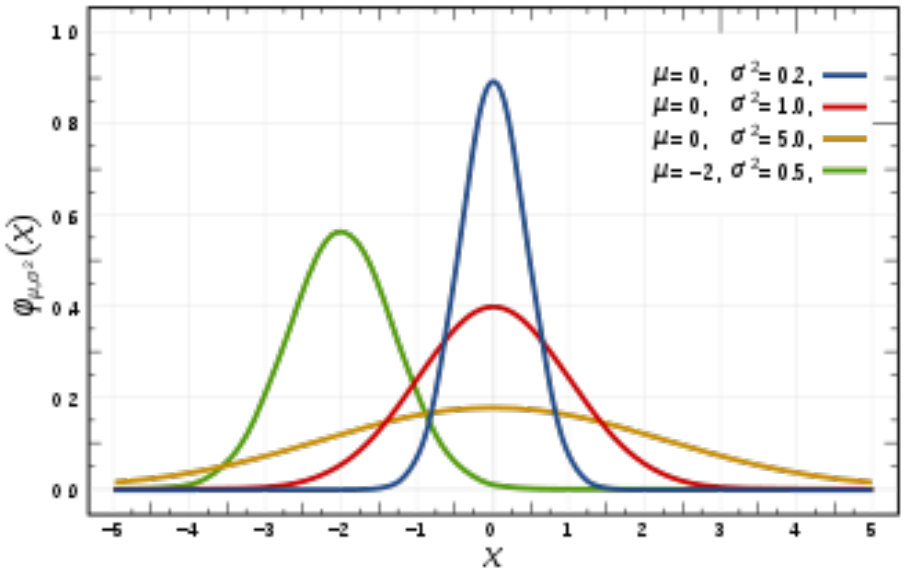
$$E[X] = \mu, \ D^2[X] = \sigma^2$$

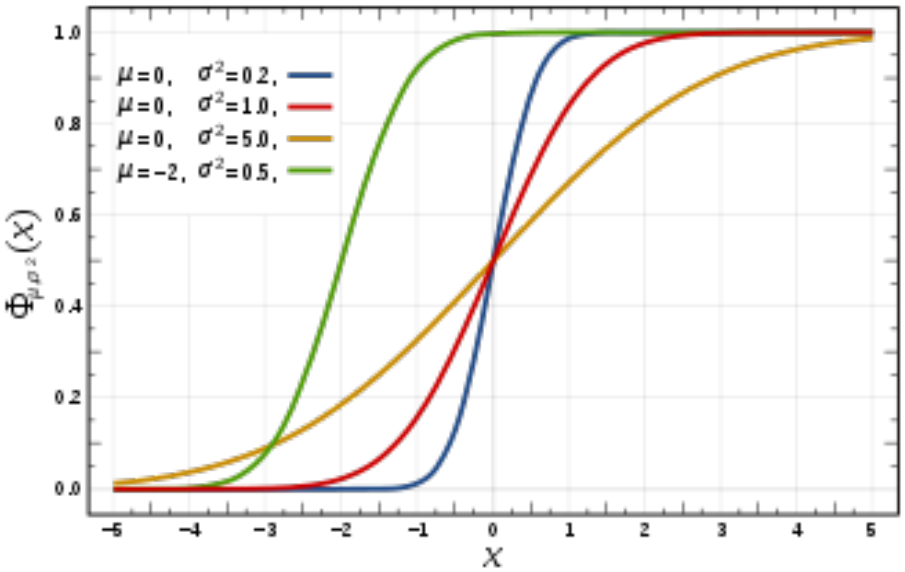
$$X \sim N(\mu, \sigma^2) \Leftrightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

Functions for the normal distribution:

- dnorm(x,mean=0,sd=1)
- pnorm(q, mean = 0, sd = 1)
- qnorm(p, mean = 0, sd = 1)
- rnorm(n, mean = 0, sd = 1)

Normal distribution III





Normal distribution IV

Example

Suppose that the average household income in some country is 900 coins, and the standard deviation is 200 coins. Assuming the Normal distribution of incomes, compute the proportion of "the middle class", whose income is between 600 and 1200 coins.

Student t distribution I

- arises when **estimating the mean** of a normally distributed population in situations where the **sample size is small** and the population standard deviation is unknown
- it was developed by William Sealy Gosset under the pseudonym Student
- this distribution is symmetric and bell-shaped, like the normal distribution, but has **heavier tails**, meaning that it is more prone to producing values that fall far from its mean
- the parameter ν is the **number of degrees of freedom**.

Student t distribution II

$$PDF: \ f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \ \forall x \in \mathbb{R}$$

$$T = Z\sqrt{\frac{\nu}{V}}, \text{ where } Z \sim N(0,1) \text{ and } V \sim \chi^2(\nu)$$

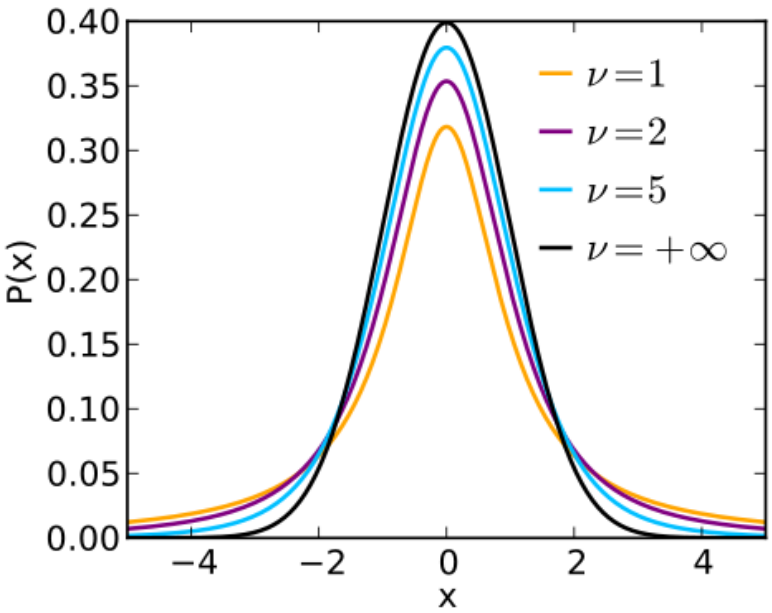
Functions for the Student t distribution:

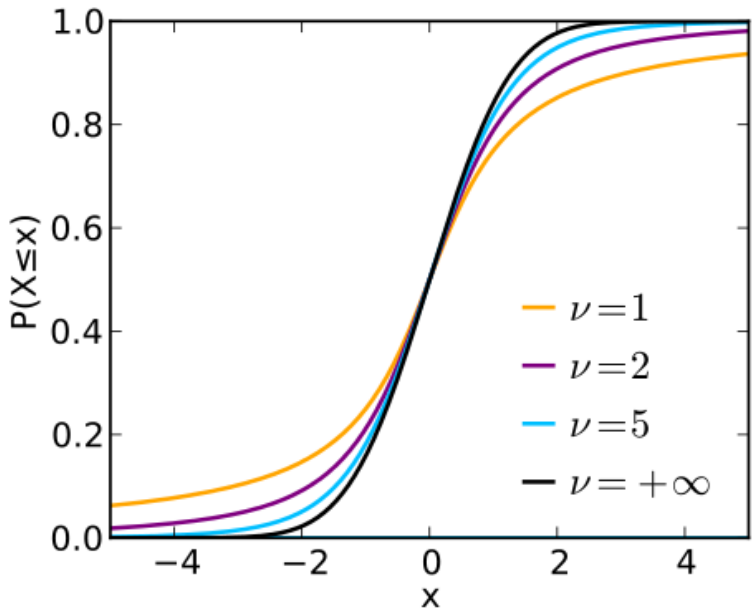
- dt(x, df, ncp)
- pt(q, df, ncp)
- qt(p, df, ncp)
- rt(n, df, ncp)

Student t distribution III

Example

Suppose scores on an IQ test are normally distributed, with a population mean of 100. Suppose 10 people are randomly selected and tested. The standard deviation in the sample group is 15. What is the probability that the average test score in the sample group will be at most 110?





Chi-square distribution I

The chi-squared distribution (χ^2-distribution) with k degrees of freedom is the distribution of a sum of squares of k independent standard normal random variables.

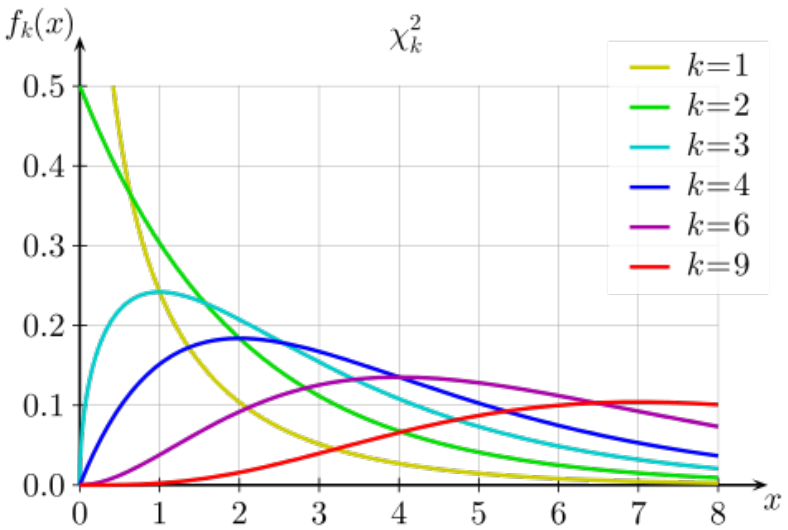
$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \forall x \in [0, \infty)$$

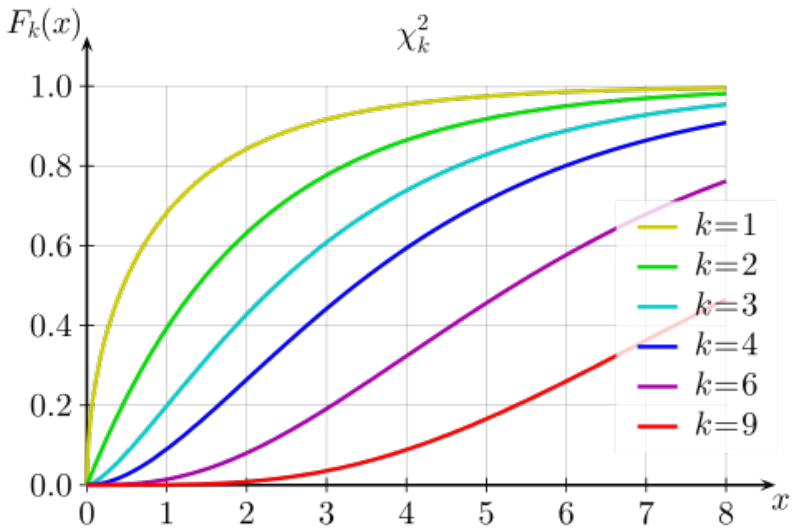
$$Q = \sum_{i=1}^{k} Z_k^2, \text{ where } Z_i \sim N(0,1)$$

Functions for the χ^2 distribution:

- dchisq(x, df)
- pchisq(q, df)
- qchisq(p, df)
- rchisq(n, df)

Chi-square distribution II





Example

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes. Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. What is the probability that the standard deviation in the test (of the sample) would be greater than 6 minutes? Use the random variable $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$.

F distribution I

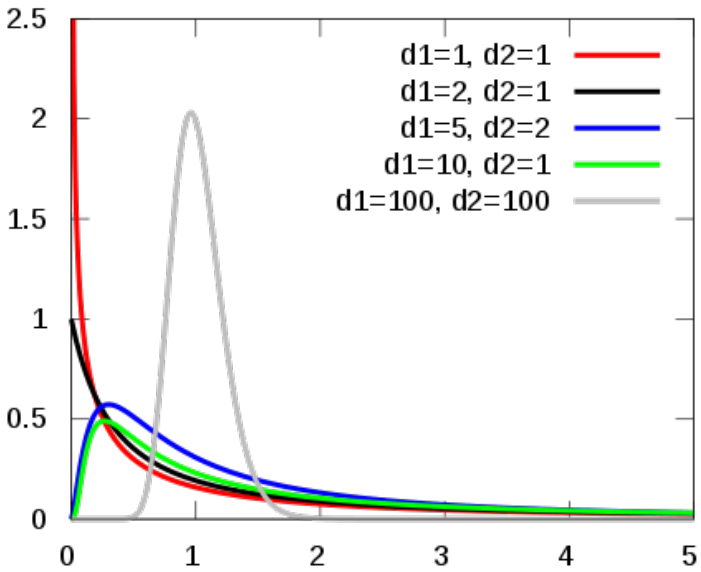
A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates: $F = \frac{U_1/d1}{U_2/d2}$, where U_1, U_2 are independent and have d_1, d_2 degrees of freedom.

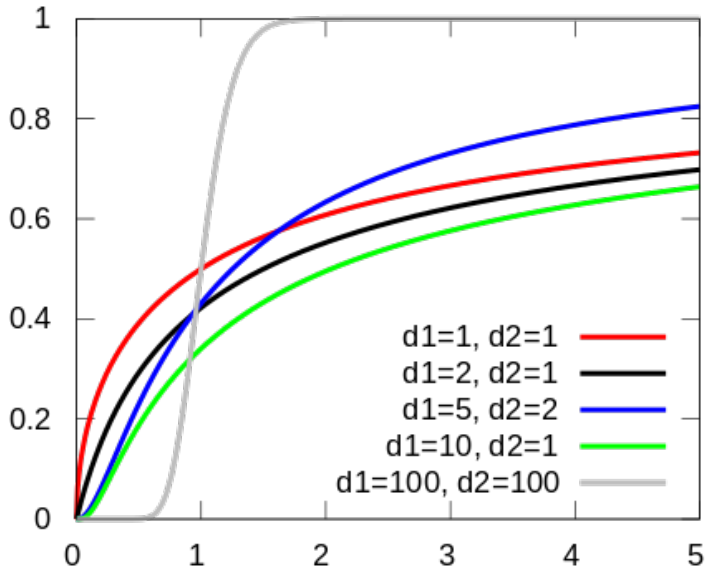
$$f(x) = \frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}} / \left(xB\left(\frac{d_1}{2}, \frac{d_2}{2}\right)\right), \forall x \in [0, \infty)$$

Functions for the F distribution:

- df(x, df1, df2)
- pf(q, df1, df2)
- qf(p, df1, df2)
- rf(n, df1, df2)

F distribution II





Example

Suppose we select 7 computer parts form company A and 12 from company B. The standard deviation of the lifetime of computer parts is 30(u) for A and 50(u) for B. Compute the probability that the ratio of standard deviations of computer parts is less than 0.5.

Descriptive Statistics

Table of contents

Descriptive statistics I

- consist of methods for **organizing** and **summarizing** information through **summary statistics**
- includes the construction of **graphs, charts, and tables**
- the calculation of various descriptive measures such as **averages, measures of variation, and percentiles**
- is distinguished from **inferential statistics** (or inductive statistics): descriptive statistics $\rightarrow$ **summarize a sample**, inferential statistics $\rightarrow$ learn about the population
- a sample often reveals features that lead to the choice of the appropriate inferential method to be later used.

Measures used in descriptive statistics I

- Commonly used to describe a data set are measures of
  - **central tendency**
  - **variability** or **dispersion**
  - **shape**
- Measures of central tendency include the
  - mean, median & mode
- Measures of variability include the
  - variance, standard deviation, range & interquartile range
- Measures of shape include the
  - kurtosis & skewness.

Central tendency

- One of the most basic features of a data set is its **center**
- the "center" of a dataset: number that represents a **middle** or **general tendency of the data**
- measurements often **cluster** around certain intermediate values $\rightarrow$ central tendency
- even if the data does not cluster round some central value, the parameters derived from repeated experiments almost inevitably do
- values that serve as a center: arithmetic mean, median, mode, geometric mean, weighted mean, etc.

Mean I

- One measure of central tendency (location) for a data set is the arithmetic mean (average or simply mean).
- The arithmetic mean (or mean or sample mean) is usually denoted by $\bar{x}$.

Definition

The arithmetic mean is the sum of all the observations $x_1, x_2, \ldots x_n$ divided by the number of observations:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The function in R: **mean()**.

Mean II

Example

Consider the data frame "birthwt" from the package "MASS" which contains 189 *observations and* 10 *variables representing potential risk factors associated with low infant birth weight. Compute the mean for the variable "bwt" representing the weight of the newborn babies.*

Median I

- One limitation of the mean is that it is oversensitive to extreme values (outliers) $\rightarrow$ use the median.
- The median is the middle value that separates the higher half from the lower half of the data set.
- The median and the mode are the only measures of central tendency that can be used for **ordinal** data, in which values are ranked relative to each other.

Definition

The median of a data set is

- *The* $\left(\frac{n+1}{2}\right)$*th largest observation if n is odd*
- *The average of the* $\left(\frac{n}{2}\right)$*th and* $\left(\frac{n}{2}+1\right)$*th largest observations if n is even*

Median II

Example

Compute the median for the same data set "birthwt" for all the numeric columns. Also compute the median of the weights of newborns with smoking mothers that have hypertension.

What can you say observing the mean and median of the "bwt" variable?

The relationship between the arithmetic mean and the sample median can be used to assess the **symmetry of a distribution**:

- symmetric distributions $\rightarrow$ arithmetic mean is approximately the same as the median;
- positively skewed distributions (skewed to the right), the arithmetic mean tends to be larger than the median;
- negatively skewed distributions (skewed to the left), the arithmetic mean tends to be smaller than the median.

Mode I

Definition

The mode is the most frequently occurring value among all the observations in a data set.

- This is the only central tendency measure that can be used with **nominal data**, which have purely qualitative category assignments.
- Some distributions have more than one mode: a distribution with one mode is called **unimodal**; two modes, **bimodal**; three modes, **trimodal**; and so forth (for example try mixing two different normal distributions).

Example

For the same data set "birthwt" compute the mode.

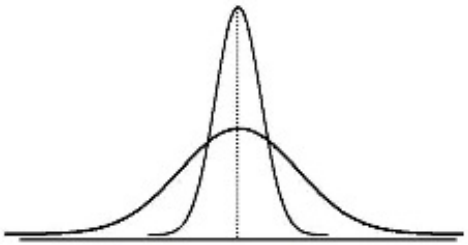
Mode II

- A common problem with the mode: it is not a useful measure of location if there is a large number of possible values, each of which occurs infrequently.
- In such cases the mode will be either far from the center of the sample or, in extreme cases, will not exist.

$x = rnorm(100)$
$which(table(x) == max(table(x)))$
$xx = cut(x, breaks = -6 : 6)$
$m = max(table(xx))$
$which(table(xx) == m)$

Dispersion

- Measures of variation give information on the **spread** of the data values.



Data sets with a large spread tend to cover a large interval of values, while data sets with small spread tend to cluster tightly around a central value.

Measures of dispersion:

- range
- variance
- standard deviation
- interquartile range
- coefficient of variation

Range

Several different measures can be used to describe the variability of a sample. Perhaps the simplest measure is the range.

Definition

The range of a set of data is the difference between the largest and smallest values.

Example

Compute the range of the data set "bwt" in R in two different ways. Compare it to the spread of the weight of newborns that have a mass smaller than 2500g.

Variance and standard deviation I

The variance measures how far a set of data is spread out from their average value (mean) and is computed as the average of the squares of the deviations from the mean.

Definition

The variance is defined as follows:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Variance and standard deviation II

The standard deviation is the square root of the variance, therefore also measures the spread of data from the mean. The advantage of using the standard deviation is that, unlike the variance, it is expressed in the same units as the data.

Definition

The standard deviation is defined as follows:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Example

Compute the variance and standard deviation of the weight of newborns whose mothers are younger than 25 years old and are hypertensive.

Interquartile range I

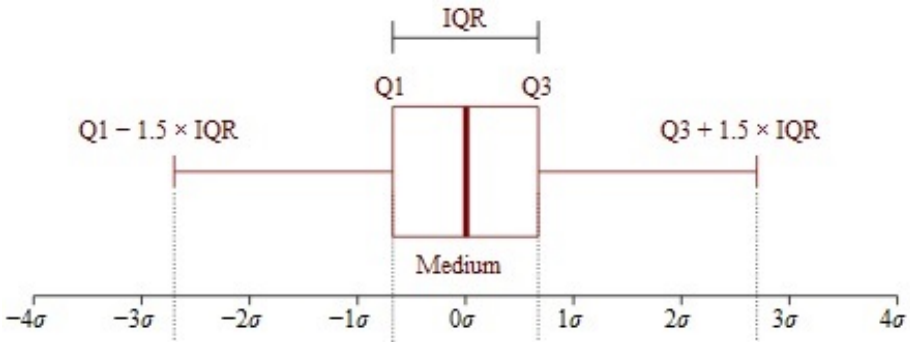
Definition

The interquartile range is a measure of spread equal to the difference of the upper and lower quartile or the 75th and 25th percentiles:

$$IQR = Q_3 - Q_1.$$

- The quartiles are the three points that divide the data set into four equal groups, each group comprising a quarter of the data.
- A quartile is a type of quantile.
- The IQR can be used to identify outliers: they are defined as observations that fall below $Q_1 - 1.5 * IQR$ or above $Q_3 + 1.5 * IQR$.

Interquartile range II



Example

Compute the IQR of the data set "bwt" in R. Are there any outliers?

Coefficient of variation

It is useful to relate the arithmetic mean and the standard deviation to each other.

Definition

The coefficient of variation (CV) is defined by

$$CV = \frac{s}{\bar{x}} \times 100.$$

Example

Compute the CV of the data in "bwt".

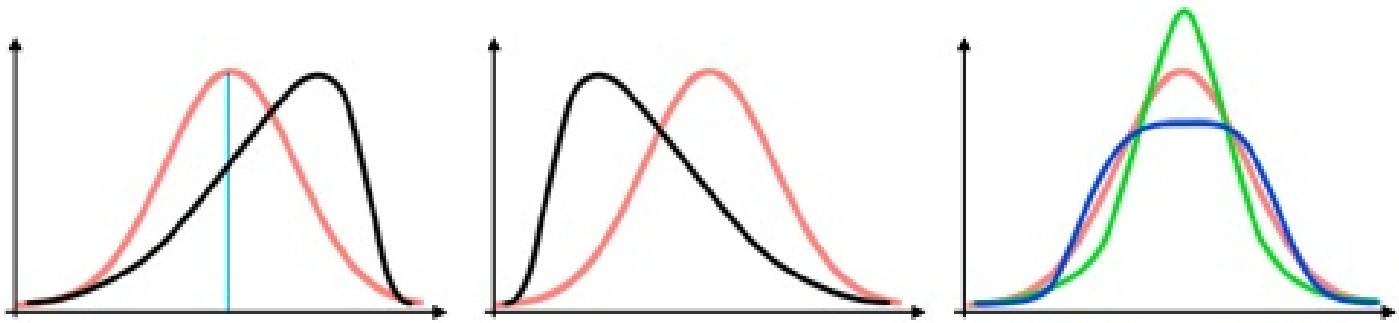
The CV is most useful in comparing the variability of several different samples, each with different arithmetic means.

Shape

- shape of a data set $\rightarrow$ shape of a graphical display, such as a histogram
- shows info about the underlying structure of the data
- helps decide which statistical procedure to use for further analysis

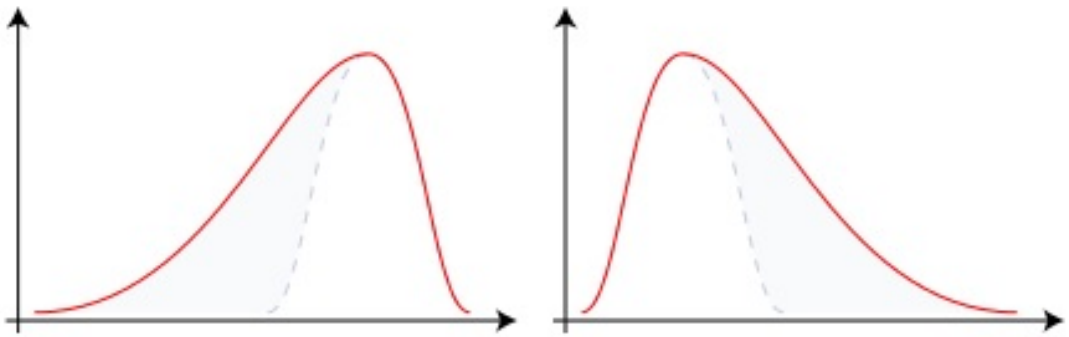
Measures of the shape of data:

- skewness
- kurtosis



Skewness I

- is a measure of the asymmetry of the distribution of a set of data about its mean
- a distribution is said to be right-skewed (or positively skewed) if the right tail seems to be stretched from the center
- left-skewed (or negatively skewed) distribution is stretched to the left side
- a symmetric distribution has a graph that is balanced about its center



Skewness II

Definition

The skewness of a set of data is computed as follows

$$g_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\frac{1}{n-1}(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}}.$$

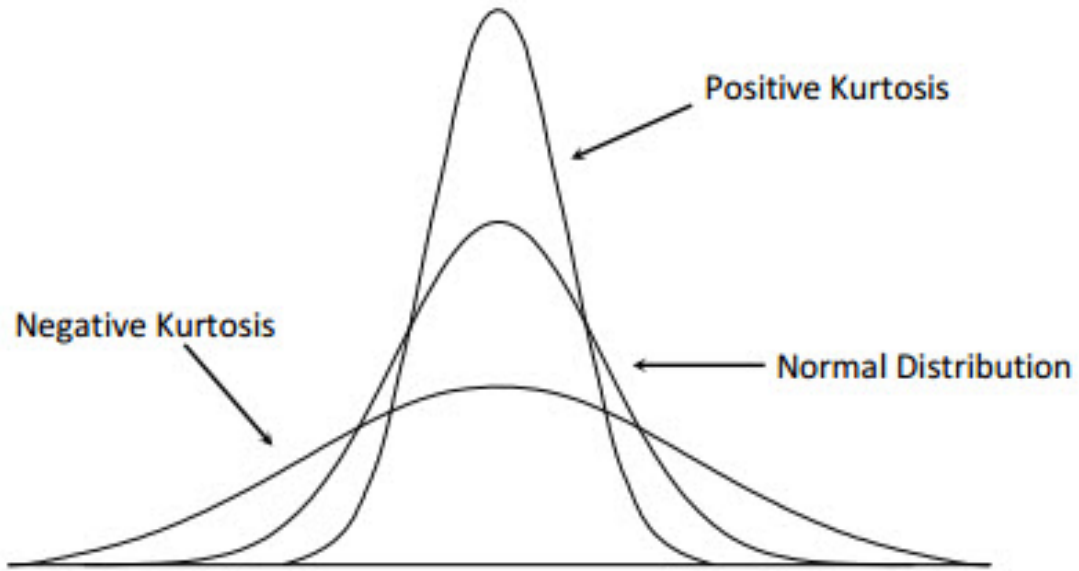
Example

Compute the skewness of "bwt" and say if the distribution of the data is symmetric or not. Use the appropiate function in the package "moments".

Kurtosis I

- Another component to the shape of a distribution is how "peaked" it is.
- Some distributions tend to have a flat shape with thin tails. These are called **platykurtic**, and an example of a platykurtic distribution is the uniform distribution.
- On the other end of the spectrum are distributions with a steep peak, or spike, accompanied by heavy tails; these are called **leptokurtic**. Examples of leptokurtic distributions are the Laplace distribution and the logistic distribution.
- In between are distributions (called **mesokurtic**) with a rounded peak and moderately sized tails. The standard example of a mesokurtic distribution is the famous bell-shaped curve, also known as the Gaussian, or normal, distribution.

Kurtosis II



Definition

The kurtosis of a set of data is computed as follows

$$g_2 = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\frac{1}{n-1}(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}.$$

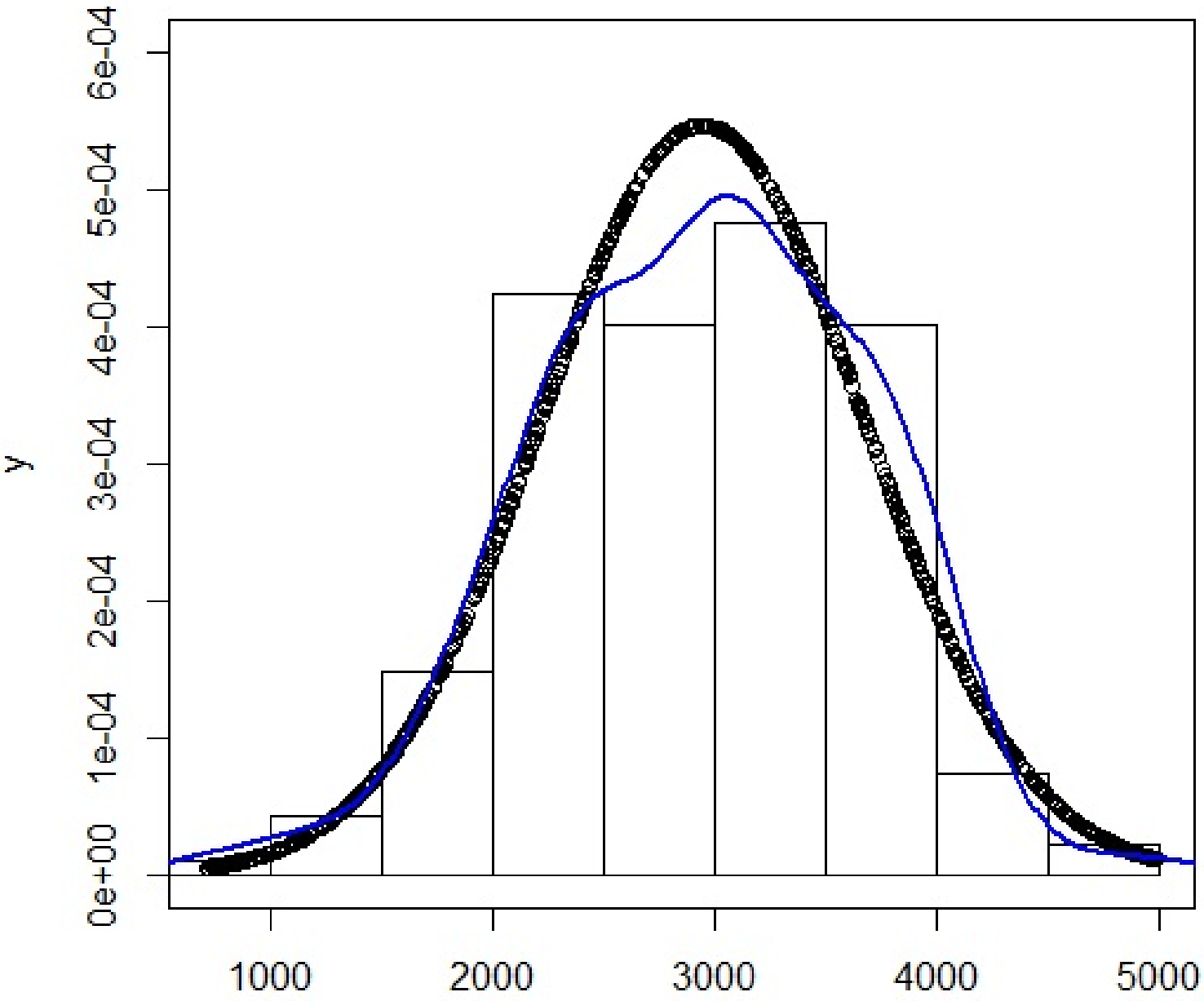
Kurtosis III

- $kurtosis < 3 \Rightarrow$ the distribution is platykurtic
- $kurtosis = 3 \Rightarrow$ the distribution is mesokurtic
- $kurtosis > 3 \Rightarrow$ the distribution is leptokurtic

Example

Compute the kurtosis of the "bwt" data set. How is the distribution of the data: platykurtic, mesokurtic or leptokurtic?

The excess of a data set is computed as $kurtosis - 3$.
The most platykurtic distribution of all is the Bernoulli distribution with $p = \frac{1}{2}$ (for example the number of times one obtains "heads" when flipping a coin once, a coin toss), the kurtosis being 1.


y
0e+00
1e-04
2e-04
3e-04
4e-04
5e-04
6e-04
1000
2000
3000
4000
5000

Histogram

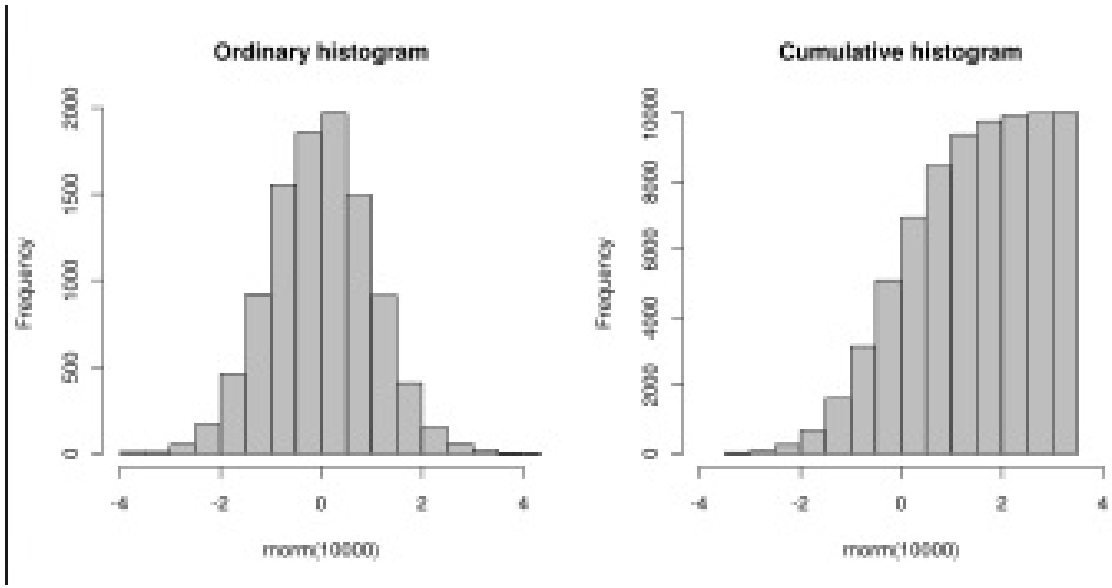
Definition

A histogram is an accurate graphical representation of the distribution of numerical data. It is an estimate of the probability distribution.

Steps for constructing a histogram:

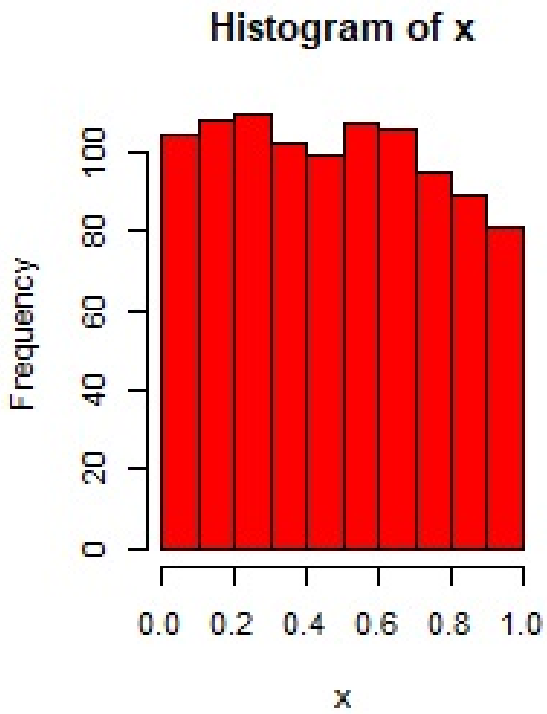
- compute the max and min values of the data
- choose a number of bins (Sturges' formula: $k = [\log_2 n] + 1$)
- compute the bin width $h = (max - min)/k$
- divide the entire range of values into sub-intervals (consecutive, non-overlapping): $min, min + h, min + 2h, \ldots, max$
- erect a rectangle over the bin with height equal to the frequency or relative frequency of the bin

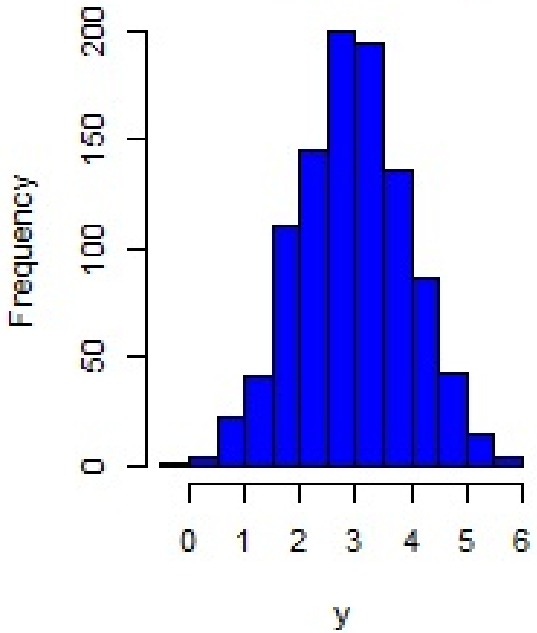
- bins need not be of equal width $\rightarrow$ rectangles have area proportional to the frequency of cases in the bin
- histograms give a rough sense of the density of the underlying distribution of the data

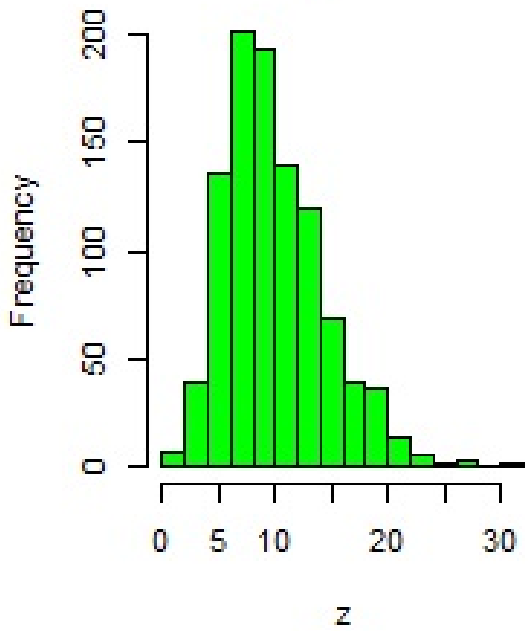


Example

Construct the histogram for 1000 random generated numbers with the following distributions: $U[0,1]$, $N(3,1)$, $\chi^2(10)$.







Bar chart, pie chart I

- a bar chart is similar to the histogram
- can also be used for qualitative variables (categorical variables)
- rectangles are not contiguous
- the bars can be plotted vertically or horizontally

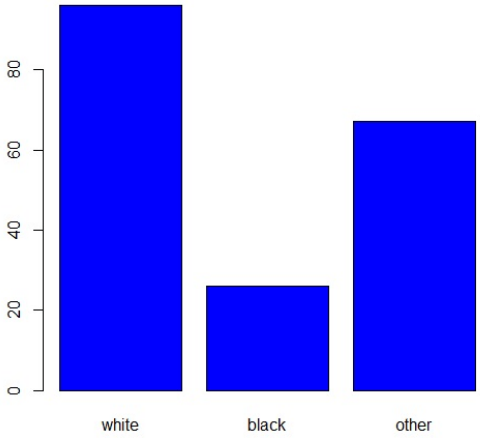
◇ a pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion

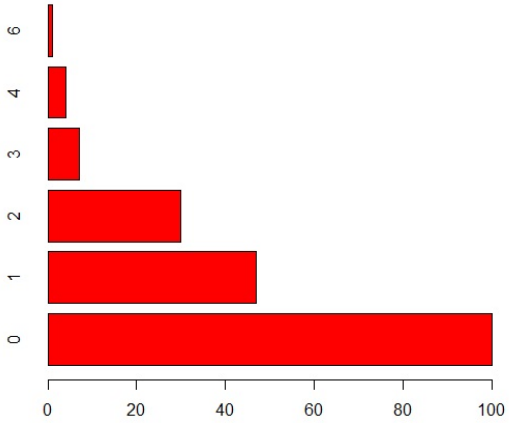
◇ the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents

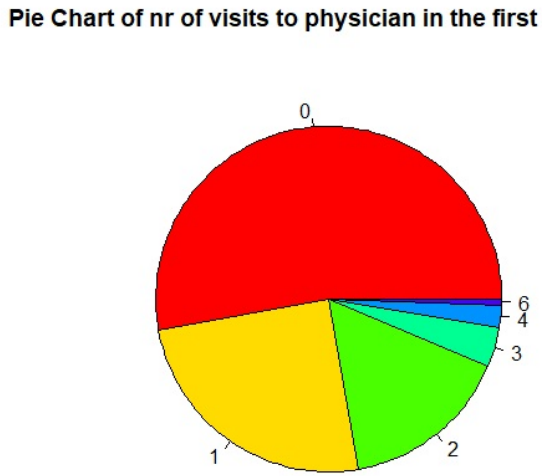
Bar chart, pie chart II

Example

Construct a bar chart using the data in "birthwt" for the variable "race" and a pie chart for the variable "ftv".

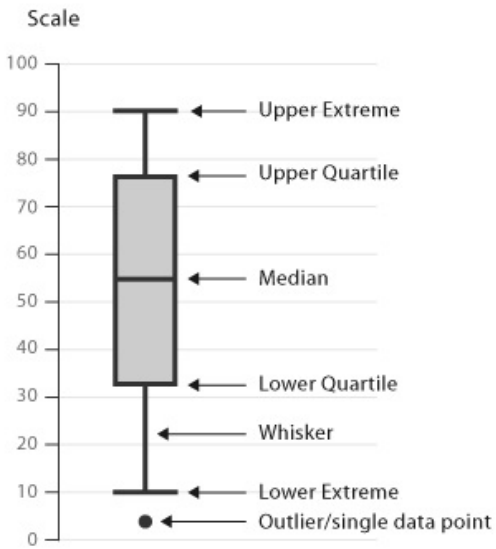


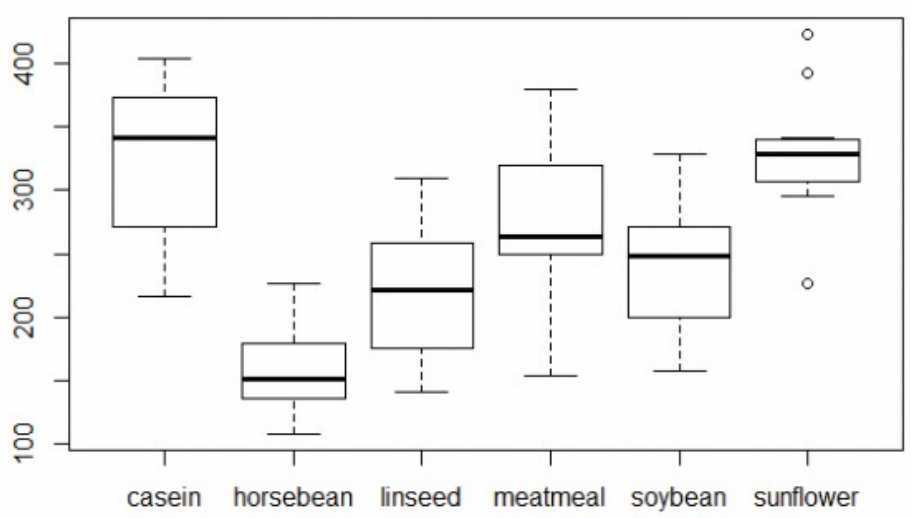




Boxplot I

- a boxplot is a method for graphically depicting groups of numerical data through their **quartiles**
- boxplot has lines extending vertically from the box (whiskers) indicating variability outside the upper and lower quartiles (box-and-whisker plot and box-and-whisker diagram)
- outliers are plotted as individual points.





Boxplot II

Example

Determine the boxplot of the data in "bwt" vs "smoke".

In R exista functia boxplot().

Multivariate Analysis

Table of contents

Multivariate analysis

- We are interested in the relationship between the following types of variables:
  - quantitative vs qualitative
  - qualitative vs qualitative
  - quantitative vs quantitative

Relationship between a quantitative and a qualitative variable

1) Let's consider the **iris** built-in data set from *datasets*. We will analyze the sepal and petal length and width, respectively, from 3 different species:

- compute numeric characteristics (mean, median, standard deviation, etc.)
- visual representation (boxplot of the variables with respect to the species)
  $boxplot(iris\$Petal.Length \sim iris\$Species)$

Relationship between a quantitative and a qualitative variable

2) Let's consider the data set **energy** from the **ISwR** package and the variables **expend** (the energy consumption for 22 women) and **stature** (lean or obese).

- boxplot on categories
- when the groups analyzed are small, it's best to plot the data as points:
  $stripchart(expend \sim stature, method = "jitter")$

Relationship between two or more qualitative variables

In this case we use **contigency tables (cross tabs)** to represent the data.

1) The contigency table is given: we analyze the preference for a certain beverage (coke, coffee and tea) on a sample of 105 people.

	coke	coffee	tea
F	10	12	28
M	20	31	4

The data is stored in a matrix and then transformed into a table with **as.table()**.

Relationship between two or more qualitative variables

R code:

```
beverage = matrix(c(10,12,28,20,31,4),byrow=T,nrow=2)

# naming the coloumns and rows
colnames(beverage)=c("coke","coffee","tea")
rownames(bauturi)=c("F","M")

beverage

# transforming into a table
beverage=as.table(beverage)

beverage
```

Relationship between two or more qualitative variables

- To a cross tab we can add:
  - marginal frequencies: frequencies for every qualitative variable:
    `margin.table(beverage,1)` $\rightarrow$ `computes the total frequency on each row`
    `margin.table(beverage,2)` $\rightarrow$ `computes the totals on columns`
- We can construct cross tabs for relative frequencies (proportions):
  - `prop.table(beverage,1)` $\rightarrow$ `proportions of each cell from the row total (relative frequencies for each level of the row variable)`
  - `prop.table(beverage,2)` $\rightarrow$ `proportion of each cell from the column total (relative frequencies for each level of the column variable)`
  - `prop.table(beverage)` $\rightarrow$ `proportions for each cell out of the sample size (table total)`

Relationship between two or more qualitative variables

2) We are given a dataframe and have to construct the cross tab ourselves, using R functions: **table()**, **xtabs()**.

- the data set **UCBAdmissions** contains information about the admissions at University of California, Berkeley in 1973.

```
UCBAdmissions
xtabs(Freq~Gender+Admit,data=UCBAdmissions)
prop.table(xtabs(Freq~Gender+Admit,data=UCBAdmissions),1) →
gender proportions for the students that were admitted
```

```
We can see that 44.5% of the male aplicants were admitted,
whereas only 30.4% of females.
```

Relationship between two or more qualitative variables

However if we analyze the situation on each of the six departments, we can see that the admission percentages for men and women are similar.

This happens because women applied to departments that had a higher rate of rejection.

This situation illustrates **the Simpson paradox**: a trend that appears for a combination of groups maybe not appear for each particular group or may be reversed.

Relationship between two or more qualitative variables

Consider the data set **bacteria** in the package **MASS** which contains information about the presence of the bacteria H. influenzae in children with otitis media in the Northern Territory of Australia.

Construct the contigency table for two categorical variables of your choice.

```
library(MASS)
attach(bacteria)
table(ap,trt)
add.margins(table(ap,trt))
prop.table(table(ap,trt),1)
```

Relationship between two quantitative variables

Assume we have two samples of size n representing two different characteristics (variables X and Y) measured on the same individuals:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

To study if there is a relationship between them we can:

- do a scatterplot with the **plot()** function
- compute the Pearson correlation coefficient:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Relationship between two quantitative variables

Properties of the Pearson correlation coefficient:

- $r \in [-1, 1]$
- if the value is nonnegative, it shows a linear correlation
- $r < 0 \rightarrow$ negative linear correlation
- $r > 0 \rightarrow$ positive linear correlation
- $r \approx 0 \rightarrow$ there is no linear relationship, but there could be any other type of correlation
- r is invariant to linear transformations, therefore the data can be standardized.

In R: `cor(x,y)`

Relationship between two quantitative variables

1) Let's analyze the relationship between height and weight for 15 women, ages between 15 and 39, using the data set *women*. Compute the correlation coefficient and draw the scatterplot.
2) The *mtcars* data set contains characteristics for 32 car models from the period 1973-1974. Analyze if there is a relationship between the weight of a car (*wt*) and the milage (*mpg*).
   - If the relationship is not linear, but is monotone, it is best to use the **Spearman correlation coefficient** ρ. It can be used for any rank data (like ordinal data).

In R: `cor(x, y, method="spearman")`

QQ Plots

If we have two data sets (two measurements of the same characteristic on two different groups), we can use **quantile-quantile plots** (qq-plots) to see if they have the same distribution.

A qq-plot is constructed as follows:

- the two data sets are sorted increasingly so that $x_1 < x_2 < \cdots < x_n$ and $y_1 < y_2 < \cdots < y_n$
- we plot the points (x_i, y_i)
- if the two data sets have different lengths, the bigger group will be reduced to the size of the smallest, keeping the range and choosing as new points equidistant quantiles.

If the two data sets have the same distribution, the points will be approximately on the first bisector.

In R: `qqplot(x,y)`

Exercises

1. Consider the data set **airquality**. Plot the temperatures as separate boxplots for each month.
2. In the data frame **twins** (found in the *UsingR* package) the IQ scores for 27 pairs of twins can be found. The twins were separated, one living with the biological family and the other with a foster family. Determine if there exists a correlation between the IQ scores of the two categories of twins.
3. In the **kid.weights** data frame from the *UsingR* package you can find the height and weight of 250 children, ages 0-12. Plot the weight vs height of the children and find the appropriate correlation coefficient.

Exercises

4. The data set **UNLifeExpectancy**, available at the web address https://instruction.bus.wisc.edu/jfrees/jfreesbooks/ Regression%20Modeling/BookWebDec2010/CSVData/UNLifeExpectancy.csv, contains information about life expectancy and other socio-economic factors worldwide.
   a) Import the data set in R as a data frame. Construct the histogram and the density for the variable *LIFEEXP*.
   b) Construct the histogram for health expenses (*HEALTHEXPEND*).
   c) For a country of your choice, determine if there exists a relationship between life expentancy and health expenses. Do the same for life expectancy and fertility.

Estimation and Hypothesis Testing

Table of contents

For better or worse, Null Hypothesis Significance Testing (NHST) is the most popular hypothesis testing framework in modern use.

NHST is a lot like being a prosecutor in the United States' or Great Britain's justice system. In these two countries - and a few others - the person being charged is presumed innocent, and the burden of proving the defendant's guilt is placed on the prosecutor. The prosecutor then has to argue that the evidence is inconsistent with the defendant being innocent. Only after it is shown that the extant evidence is unlikely if the person is innocent, does the court rule a guilty verdict. If the extant evidence is weak, or is likely to be observed even if the dependent is innocent, then the court rules not guilty. That doesn't mean the defendant is innocent (the defendant may very well be guilty!)- it means that either the defendant was guilty, or there was not sufficient evidence to prove guilt.

With **simple NHST**, we are testing two competing hypotheses: the null and the alternative hypotheses.

- The default hypothesis is called the **null hypothesis** (H_0) - it is the hypothesis that our observation occurred from chance alone. In the justice system analogy, this is the hypothesis that the defendant is innocent.
- The **alternative hypothesis** (H_1 or H_a) is the opposite (or complementary) hypothesis; this would be like the prosecutor's hypothesis.

The *null hypothesis* terminology was introduced by a statistician named R. A. Fischer in regard to the curious case of Muriel Bristol: a woman who claimed that she could discern, just by tasting it, whether milk was added before tea in a teacup or whether the tea was poured before the milk. She is more commonly known as the *lady tasting tea*.

The null and alternative hypothesis are mutually exclusive.
So, here's the basic idea behind **NHST** (hypothesis testing) as we know it so far:

- Assume the opposite of what you are testing.
- (Try to) show that the results you receive are unlikely given that assumption.
- Reject the assumption.

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect
- Determine the test statistic - some measure of the sample.

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect
- Determine the test statistic - some measure of the sample.
- Determine the sampling distribution of the test statistic

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect
- Determine the test statistic - some measure of the sample.
- Determine the sampling distribution of the test statistic will tell us which values of the test statistics are most likely to occur by chance (under the null hypothesis) with repeated trials of the experiment.
- Compute the p-value

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect
- Determine the test statistic - some measure of the sample.
- Determine the sampling distribution of the test statistic will tell us which values of the test statistics are most likely to occur by chance (under the null hypothesis) with repeated trials of the experiment.
- Compute the p-value

  Once we know what the sampling distribution of the test statistic looks like, we can tell what the probability of getting a result as extreme as we got is (this is called a p-value).
- **the test**

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect
- Determine the test statistic - some measure of the sample.
- Determine the sampling distribution of the test statistic will tell us which values of the test statistics are most likely to occur by chance (under the null hypothesis) with repeated trials of the experiment.
- Compute the p-value

  Once we know what the sampling distribution of the test statistic looks like, we can tell what the probability of getting a result as extreme as we got is (this is called a p-value).
- **the test**

  If it is equal to or below some pre-specified boundary, called a significance level (α level), we decide that the null hypothesis is a bad hypothesis and embrace the alternative hypothesis or fail to reject the null hypothesis.

Steps of NHST

- Formulate the null and alternative hypothesis
  - H_0 there is no effect
  - H_a there is an effect
- Determine the test statistic - some measure of the sample.
- Determine the sampling distribution of the test statistic will tell us which values of the test statistics are most likely to occur by chance (under the null hypothesis) with repeated trials of the experiment.
- Compute the p-value

  Once we know what the sampling distribution of the test statistic looks like, we can tell what the probability of getting a result as extreme as we got is (this is called a p-value).
- **the test**

  If it is equal to or below some pre-specified boundary, called a significance level (α level), we decide that the null hypothesis is a bad hypothesis and embrace the alternative hypothesis or fail to reject the null hypothesis.

Largely, as a matter of tradition, an alpha level of .05 is used most often, though other levels are occasionally used as well (0.01, 0.1). So, if the observed result would only occur 5% or less of the time (p-value $< .05$), we consider it a sufficiently unlikely event and reject the null hypothesis.

Example

We have a coin and want to determine if it fair or not. For this we flip the coin 30 times and observe 5 heads. Formulate the two hypothesis, determine the test statistic, compute the p-value and draw your conclusion.

- H_0 : the coin is fair (the probability of obtaining heads is $p = 0.5$)
- H_a : the coin is not fair (the probability of obtaining heads $p \neq 0.5$)

- Let's use the **number of heads** in our sample as the **test statistic**.
- What is the **sampling distribution** of this test statistic?

- Let's use the **number of heads** in our sample as the **test statistic**.
- What is the **sampling distribution** of this test statistic?
  In other words, if the coin were fair, and you repeated the flipping-30-times experiment many times, what is the relative frequency of observing particular numbers of heads? It's the binomial distribution. A binomial distribution with parameters n=30 and p=0.5 describes the number of heads we should expect in 30 flips.

- Let's use the **number of heads** in our sample as the **test statistic**.
- What is the **sampling distribution** of this test statistic?
  In other words, if the coin were fair, and you repeated the flipping-30-times experiment many times, what is the relative frequency of observing particular numbers of heads? It's the binomial distribution. A binomial distribution with parameters n=30 and p=0.5 describes the number of heads we should expect in 30 flips.
- What is the **p-value**?

- Let's use the **number of heads** in our sample as the **test statistic**.
- What is the **sampling distribution** of this test statistic?
  In other words, if the coin were fair, and you repeated the flipping-30-times experiment many times, what is the relative frequency of observing particular numbers of heads? It's the binomial distribution. A binomial distribution with parameters n=30 and p=0.5 describes the number of heads we should expect in 30 flips.
- What is the **p-value**? (0.0052)

- Let's use the **number of heads** in our sample as the **test statistic**.
- What is the **sampling distribution** of this test statistic?
  In other words, if the coin were fair, and you repeated the flipping-30-times experiment many times, what is the relative frequency of observing particular numbers of heads? It's the binomial distribution. A binomial distribution with parameters n=30 and p=0.5 describes the number of heads we should expect in 30 flips.
- What is the **p-value**? (0.0052)
- What is the **conclusion**?

- Let's use the **number of heads** in our sample as the **test statistic**.
- What is the **sampling distribution** of this test statistic?
  In other words, if the coin were fair, and you repeated the flipping-30-times experiment many times, what is the relative frequency of observing particular numbers of heads? It's the binomial distribution. A binomial distribution with parameters n=30 and p=0.5 describes the number of heads we should expect in 30 flips.
- What is the **p-value**? (0.0052)
- What is the **conclusion**? We reject H_0.

One-tailed and two-tailed tests

The previous example was one involving a **two-tailed test** because the alternative hypothesis was formulated as $p \neq 0.5$.
If we wanted to test whether the probability of obtaining a head is 0.05 or less, than the test would have been an one-tailed test: $H_a : p < 0.05$. The null hypothesis is the same.
The null hypothesis is always formulated using an equality!
In the one-tailed test, the p-value is 0.0026 and the same conclusion is drawn.
The function in R that computes the p-value for this test is **binom.test()**.

```
> binom.test(7,30,alternative="less")

        Exact binomial test

data:  7 and 30
number of successes = 7, number of trials = 30, p-value = 0.002611
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.3939474
sample estimates:
probability of success
             0.2333333
```

Type I and type II errors

Since we make judgments and inferences based on probabilities, mistakes happen. In particular, there are two types of mistakes that are possible in NHST: Type I errors and Type II errors.

- A Type I error is when a hypothesis test concludes that there is an effect (rejects the null hypothesis) when, in reality, no such effect exists
- A Type II error occurs when we fail to detect a real effect in the world and fail to reject the null hypothesis even if it is false.

Coin type	Failure to reject null hypothesis (conclude no detectable effect)	Reject the null hypothesis (conclude that there is an effect)
Coin is fair	Correct positive identification	Type I error (false positive)
Coin is unfair	Type II error (false negative)	Correct identification

Power of a test

- The probability of making a Type II error (fail to reject the null hypothesis if the alternative hypothesis were true) is defined as β .
- The power of a test is defined as $1 - \beta$ and represents the probability of correctly detecting a true effect if one exists.
- The power of a test depends on the type of test being performed, the sample size being used, and on the size of the effect that is being tested (**the effect size**).
- Greater effects, like the average difference in height between women and men, are far easier to detect than small effects, like the average difference in the length of earthworms in Carlisle and in Birmingham.
- Statisticians like to aim for a power of **at least** 80% (a beta level of .2).
- A test that doesn't reach this level of power (because of a small sample size or small effect size, and so on) is said to be **under-powered**.

Tests for the population mean and for proportions

- One sample Z-test
- Two sample Z-test
- One sample t-test
- Two sample t-test for unpaired data with the same variance
- Two sample t-test for unpaired data with different variances
- Two sample t-test for paired data
- Z-test for proportions
- Z-test for difference of proportions
- χ^2 test for variance/standard deviation
- F test for two variances/standard deviations

One sample Z-test I

- The population has a normal distribution and its variance is known. (This test can also be used if the population is not normally distributed and we have a large number of data, $n > 30$, because of the Central Limit Theorem.)
- Hypothesis:

$$\begin{aligned} H_0 : &\quad \mu = \mu_0 \\ H_a : &\quad \mu \neq \mu_0 \\ &\quad \mu < \mu_0 \\ &\quad \mu > \mu_0 \end{aligned}$$

- The statistic used:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$$

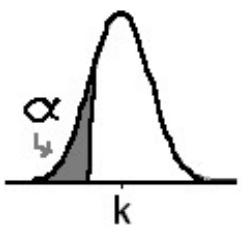
where $\bar{X}$ is the sample mean, σ is the population standard deviation and n is the sample size.

One sample Z-test II

The critical region:

$$\begin{array}{rll} H_a: & \mu \neq \mu_0 \rightarrow W = (-\infty, z_{\frac{\alpha}{2}}) \cup (z_{1-\frac{\alpha}{2}}, \infty) \\ & \mu < \mu_0 \rightarrow W = (-\infty, z_\alpha) \\ & \mu > \mu_0 \rightarrow W = (z_{1-\alpha}, \infty) \end{array},$$

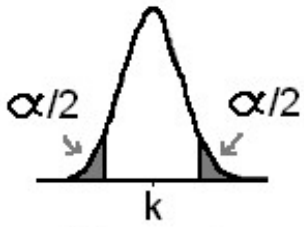
where $z_{\frac{\alpha}{2}}$ și z_α are the quantiles of the standard normal distribution $N(0,1)$.



$H_0: \mu = k$

$H_1: \mu < k$

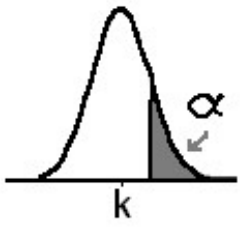
α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33



$H_0: \mu = k$

$H_1: \mu \neq k$

α	z critical
0.10	± 1.65
0.05	± 1.96
0.01	± 2.58



$H_0: \mu = k$

$H_1: \mu > k$

α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

One sample Z-test III

Example

Lets consider a sample of 25 male individuals in the USA that are smokers and have hypertension, ages ranging between 20 and 74. By measuring the cholesterol level for each individual, we obtain a sample mean of 228.85mg/100ml. Knowing that the average cholesterol level of the male population in the US is 211mg/100ml with a standard deviation of 46mg/100ml, does this data suggest that the smoking, hypertensive male population of the USA has a larger cholesterlor level?

Confidence intervals for the population mean

- There is a strong link between hypothesis testing and confidence intervals: the confidence interval is the same as the acceptance region; it contains all the possible values of the test statistic such that H_0 cannot be rejected.
- The $(1-\alpha)\%$ confidence interval for the population mean μ:

$$\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

- The test can be rephrased: if μ_0 belongs to the confidence interval, then H_0 cannot be rejected; otherwise H_0 is rejected

Example

Compute the 95% confidence interval for the population cholesterol level of the smoking hypertensive population in the US, using the data in the previous example.

One sample t test I

- The population has a normal distribution, but its variance is unknown. It can also be used instead of the Z test, for a large number of data ($n > 30$); in this case the two tests yield the same results.
- Hypothesis:

$$\begin{aligned} H_0 : &\quad \mu = \mu_0 \\ H_a : &\quad \mu \neq \mu_0 \\ &\quad \mu < \mu_0 \\ &\quad \mu > \mu_0 \end{aligned}$$

- The test statistic:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim T(n-1),$$

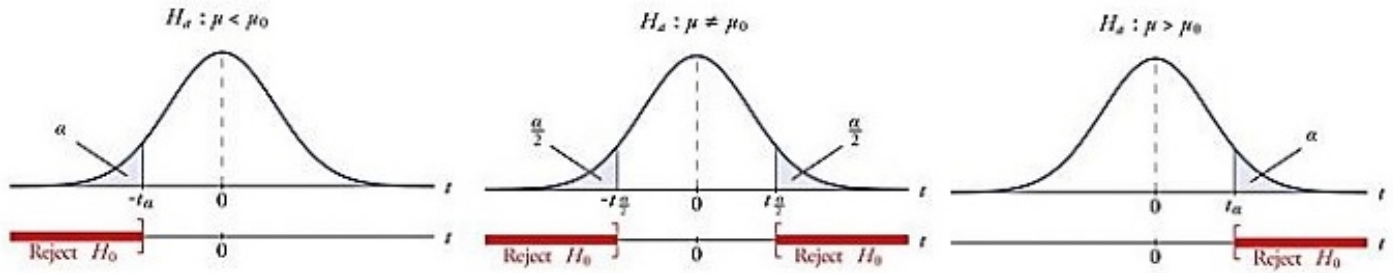
where $\bar{X}$ is the sample mean, s is the sample standard deviation and n the sample size.

One sample t test II

- The critical region:

$$
\begin{array}{lll}
H_a: & \mu \neq \mu_0 \rightarrow W = (-\infty, t_{\frac{\alpha}{2}}) \cup (t_{1-\frac{\alpha}{2}}, \infty) \\
& \mu < \mu_0 \rightarrow W = (-\infty, t_{\alpha}) \\
& \mu > \mu_0 \rightarrow W = (t_{1-\alpha}, \infty)
\end{array}
,
$$

where $t_{\frac{\alpha}{2}}$ and t_{α} are the quantiles of the Student t distribution with $n-1$ degrees freedom ($T(n-1)$):



One sample t test III

Example

In this example, we'll be using R's built-in "precip" data set that contains precipitation data from 70 US cities. Is the US' precipitation significantly different to the rest of the world's precipitation, knowing that the average world precipitation was 45.5 (in/year)? Use a level of significance $\alpha = 0.05$.

- The $(1-\alpha)\%$ confidence interval for the population mean μ if the variance is not known:

$$\left(\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right).$$

Example

Compute the 95% confidence interval for the precipitation level in the USA , using the data in the previous example.

Two sample t tests. Comparing the means of two populations. I

- The two populations have a normal distribution.
- Hypothesis:

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \\ H_a: &\quad \mu_1 \neq \mu_2 \\ &\quad \mu_1 < \mu_2 \\ &\quad \mu_1 > \mu_2 \end{aligned}$$

- The critical region is the same as for the one sample t test.
- T tests: unpaired data (equal/unequal variances), paired data.

Two sample t test, unpaired data, equal variances.

- The test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2),$$

where $\bar{X}_1, \bar{X}_2$ are the sample means, n_1, n_2 are the sample sizes, $s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$, s_1, s_2 are the sample standard deviation.

Example

In the data set "Achieve" from the BSDA package the first column represents the mathematics achiement score of 25 high school students (boys and girls). Determine if the boys had a significantly higher score than the girls, considering that the populations of scores are normally distributed with equal variances. Take $\alpha = 0.01$.

Two sample t test, unpaired data, unequal variances.

- Test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n_2}}},$$

  where $\bar{X}_1, \bar{X}_2$ are the sample means, s_1, s_2 are the sample standard deviations, n_1, n_2 are the sample sizes.

- The r.v. T has a Student t distribution $T(\nu)$, where $\nu = \frac{[(s_1/n_1)+(s_2/n_2)]^2}{(s_1/n_1)^2/(n_1-1)+(s_2/n_2)^2/(n_2-1)}$ is rounded.

Example

The dataset we will be using for this test is the "mtcars" dataset. Specifically, we are going to test the hypothesis that the mileage is better for manual cars than it is for cars with automatic transmission. Compare the means and produce a boxplot first.

Two sample t test, paired data I

- The data is paired; for example we measure some characteristic (blood pressure) before and after the treatment with a medicine.
- Hypothesis:

$$\begin{aligned} H_0: &\quad \delta = 0 \\ H_a: &\quad \delta \neq 0 \\ &\quad \delta < 0 \\ &\quad \delta > 0 \end{aligned}$$

  $\delta = \mu_1 - \mu_2$ is the difference of the two population means.

- The test statistic:

$$T = \frac{\bar{d} - \delta}{s_{\bar{d}}} \sim T(n-1),$$

  where $\bar{d} = \frac{\sum_{i=1}^{n} x_i^1 - x_i^2}{n}$ is the mean of the paired differences, $s_{\bar{d}}$ is the standard deviation of the variablei $\bar{d}$, n is thesample size.

Two sample t test, paired data II

Example

The total-body bone mineral content (TBBMC) of young mothers was measured during breast feeding and then in the postweaning period. Despite the calcium drain of breastfeeding and low calcium intake, did the mothers gain at least 25 grams of bone mineral content? The average mother's age is 16 years old, so the mothers' bodies are growing too. Use a significance level of 0.05. During breast feeding the TBBMC scores are 1928, 2549, 2825, 1924, 1628, 2175, 2114, 2621, 1843, 2541 and after 2126, 2885, 2895, 1942, 1750, 2184, 2164, 2626, 2006, 2627.

Tests for the proportions I

- Hypothesis:

$$H_0 : p = p_0$$

$$H_a : \quad p \neq p_0$$
$$p < p_0$$
$$p > p_0$$

- The test statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

has the standard normal distribution $N(0, 1)$, where $\hat{p} = \frac{x}{n}$, x is the number of cases that exhibit the studied characteristic, n is the sample size.

- Conditions: $n \times p \geq 5$ și $n \times (1 - p) \geq 5$.

Tests for the proportions II

Example

We study the effect of birth weight on the cognitive abilities of babies. In order to do this the IQ score of 33 randomly chosen children that were underweight at birth ($< 1500gr$) is measured and it is found that 8 of them have a score less than 70. In normal children, this proportion is 3.2%. Test if the low birth weight has a significant effect on the cognitive abilities of children.

Confidence intervals for proportions:
The $(1-\alpha)\%$ confidence interval for the proportion of a population is

$$(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}),$$

where $\hat{p} = x/n$, x number of succes in the sample, n is the sample size, $z_{\alpha/2}$ is the corresponding quantile of the standard normal distribution.

Tests for the proportions III

Example

Compute the 95% and 99% confidence intervals for the proportion of underweight newborns that have an IQ score less than 70.

Comparison of two proportions I

- Ipotezele statistice:

$$\begin{aligned} H_0: &\quad p_1 = p_2 \\ H_a: &\quad p_1 \neq p_2 \\ &\quad p_1 < p_2 \\ &\quad p_1 > p_2 \end{aligned}$$

- The test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

  has approximately the $N(0,1)$ distribution, where $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$ (proportion of succes in the two samples), n_1, n_2 sample sizes, $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$.

- Conditions: $n_1 \cdot \hat{p}, n_1 \cdot (1-\hat{p}), n_2 \cdot \hat{p}, n_2 \cdot (1-\hat{p}) > 5$.

Comparison of two proportions II

Example

A study is conducted to find out the factors that help spread tuberculosis among drug users. Two samples of 97 drug users that admittedly shared needles and 161 that didn't are chosen and tested. The reports show that 34 individuals in the first group and 28 in the second have tuberculosis. Test whether the proportion of TBC infected people is larger among drug users that share needles, with a 0.05 level of significance.

The χ^2 for variance I

- The population has a normal distribution.
- Hypothesis:

$$\begin{aligned} H_0: &\quad \sigma^2 = \sigma_0^2 \\ H_a: &\quad \sigma^2 \neq \sigma_0^2 \\ &\quad \sigma^2 < \sigma_0^2 \\ &\quad \sigma^2 > \sigma_0^2 \end{aligned}$$

- Test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$$

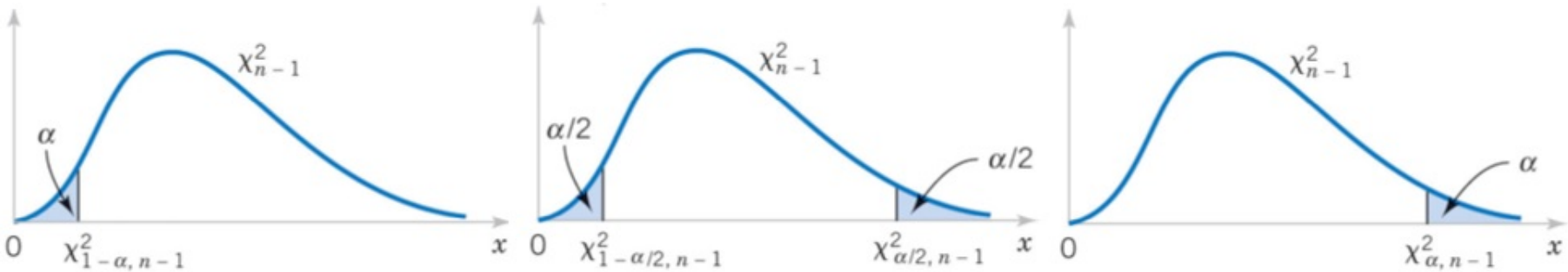
where s^2 is the sample variance, n is the sample size.

The χ^2 for variance II

- The critical region:

$$
\begin{aligned}
H_a: \quad & \sigma^2 \neq \sigma_0^2 \rightarrow W = (0, \chi^2_{\alpha/2}) \cup (\chi^2_{1-\alpha/2}, \infty) \\
& \sigma^2 < \sigma_0^2 \rightarrow W = (0, \chi^2_{\alpha/2}) \\
& \sigma^2 > \sigma_0^2 \rightarrow W = (\chi^2_{1-\alpha/2}, \infty)
\end{aligned}
$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are the quantiles of the $\chi^2(n-1)$ distribution.



Example

Test whether the variance of the data in the "weight" variable of the "PlantGrowth" data frame is equal to 0.5 of larger, $\alpha = 0.05$.

The χ^2 for variance III

- The $(1-\alpha)\%$ confidence intervals for variance/standard deviation:

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \text{ (varianță)}$$

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$$

Example

Compute the 95% confidence interval for the variance of the weight of plants in the "PlantGrowth" data frame.

F test for the variance of two populations I

- The two populations must have a normal distribution.
- Hypothesis:

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 \\ H_a: &\quad \sigma_1^2 \neq \sigma_2^2 \\ &\quad \sigma_1^2 < \sigma_2^2 \\ &\quad \sigma_1^2 > \sigma_2^2 \end{aligned}$$

- Test statistic:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

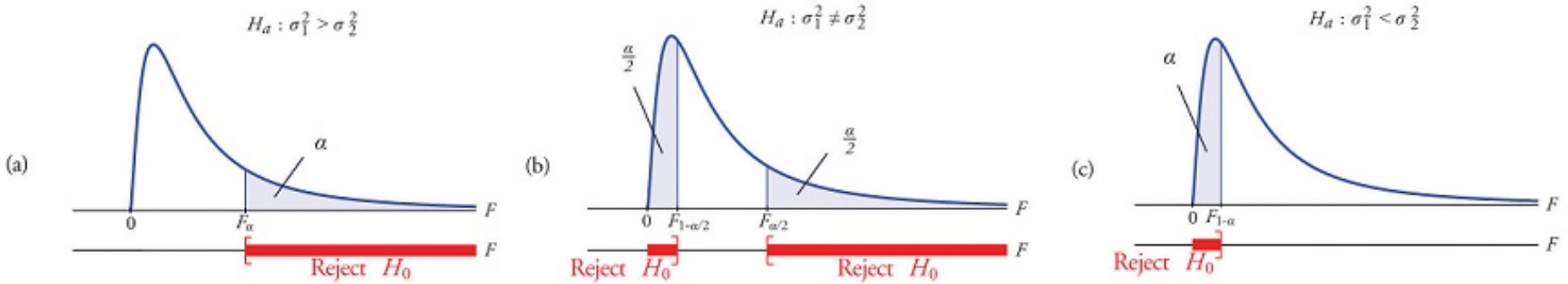
where s_1^2, s_2^2 are the sample variances, n, m are the sample sizes.

F test for the variance of two populations II

- The critical region:

$$
\begin{aligned}
H_a: \quad & \sigma_1^2 \neq \sigma_2^2 \rightarrow W = (0, F_{\alpha/2}) \cup (F_{1-\alpha/2}, \infty) \\
& \sigma_1^2 < \sigma_2^2 \rightarrow W = (0, F_{\alpha/2}) \\
& \sigma_1^2 > \sigma_2^2 \rightarrow W = (F_{1-\alpha/2}, \infty)
\end{aligned}
$$

where $F_{\alpha/2}$ and $F_{1-\alpha/2}$ are the quantiles of the distribution $F(n-1, m-1)$.



F test for the variance of two populations III

Example

The "ToothGrowth" data frame contains the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC). Test if the variance of the length of cells is the same in the two groups determined by the delivery method. Test if the standard deviation of the length of cells is the same depending on the dose given (Bartlett test).

Estimation and Hypothesis Testing II

Table of contents

One-sample and two-sample (paired and unpaired) non-parametric tests I

- when the populations **do not have a normal distribution** or the sample size is small $\rightarrow$ **Wilcoxon** (paired observations) and **Mann-Whitney** tests.
- in R: **wilcox.test(x, y, alternative="two.sided","less","greater", paired=T,F)**
- the test hypothesis are formulated more generally than for parametric tests (i.e. in regard to the distribution of the sample(s)):

$$H_0 : P(X < Y) = \frac{1}{2} \text{ (the values in the two samples are similar)}$$

$$H_a : P(X < Y) \neq \frac{1}{2}, P(X < Y) < \frac{1}{2}, P(X < Y) > \frac{1}{2}$$

One-sample and two-sample (paired and unpaired) non-parametric tests II

Example

In the built-in data set named *immer* (*MASS* package), the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame columns Y1 and Y2. Without assuming the data to have normal distribution, test at .05 significance level if the barley yields of 1931 and 1932 have identical data distributions (if there is a significant difference between them or not).

One-sample and two-sample (paired and unpaired) non-parametric tests III

Example

A study is designed in order to test the efficiency of vitamin E supplements for Alzeihmer disease prevention. 20 subjects aged over 65 are randomly distributed to two groups. The first group (10 pleople) receives 400UI/day of vitamin E, while the second group receives a placebo treatment. The initial vitamin E levels are measured:
Group 1 : 7.5; 12.6; 3.8; 20.2; 6.8; 403.3; 2.9; 7.2; 10.5; 205.4
Group 2 : 8.2; 13.3; 102.0; 12.7; 6.3; 4.8; 19.5; 8.3; 407.1; 10.2
Test if there are a significant difference between the two groups initially.

Tests for the proportions I

- Hypothesis:

$$H_0 : p = p_0$$

$$H_a : \quad p \neq p_0$$
$$p < p_0$$
$$p > p_0$$

- The test statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

has the standard normal distribution $N(0, 1)$, where $\hat{p} = \frac{x}{n}$, x is the number of cases that exhibit the studied characteristic, n is the sample size.

- Conditions: $n \times p \geq 5$ și $n \times (1 - p) \geq 5$.
- In R: **binom.test()**

Tests for the proportions II

Example

We study the effect of birth weight on the cognitive abilities of babies. In order to do this the IQ score of 33 randomly chosen children that were underweight at birth ($< 1500gr$) is measured and it is found that 8 of them have a score less than 70. In normal children, this proportion is 3.2%. Test if the low birth weight has a significant effect on the cognitive abilities of children.

Confidence intervals for proportions:
The $(1-\alpha)\%$ confidence interval for the proportion of a population is

$$(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}),$$

where $\hat{p} = x/n$, x number of succes in the sample, n is the sample size, $z_{\alpha/2}$ is the corresponding quantile of the standard normal distribution.

Tests for the proportions III

Example

Compute the 95% and 99% confidence intervals for the proportion of underweight newborns that have an IQ score less than 70.

χ^2 test for a proportion I

Hypothesis:

$$H_0 : p = p_0$$

$$H_a : \quad p \neq p_0$$
$$p < p_0$$
$$p > p_0$$

Test statistic:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

has a $\chi^2(1)$ distribution, where O_i are the observed values of type i and E_i are the expected values of type i, $i = 1, 2$.
In R: $prop.test(x, n, p, alternative = "two.sided" / "less" / "greater")$ where

- x is the number of individuals in the sample exhibiting the studied characteristic
- n is the sample size

χ^2 test for a proportion II

- p is the proportion assumed as true by the null hypothesis H_0

Example

The breast cancer incidence rate in the US female population in the 50-54 group age is aproximately 2%. We would like to test if the proportion of breast cancer patients whose mothers where disgnosed with the same condition is larger than that of the general population. A random sample of 10000 women is chosen in the age group 50-54 whose mothers had breast cancer at some point in their lives and we observe that 400 of the women also have this condition. At the 0.05 level of significance, what is your conclusion?

The χ^2 test for several proportions I

Hypothesis:

- H_0: $p_1 = p_2 = \ldots = p_k$
- H_a : at least one proportion is different

Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

has a $\chi^2(k-1)$ distribution, where O_i are the observed values of type i and E_i are the expected values of type i, $i = \overline{1,k}$..
In R: $prop.test(x, n, alternative = "two.sided" / "less" / "greater")$ where

- x is the vector containing the number of individuals in each sample exhibiting the studied characteristic
- n is the vector of sample sizes

The χ^2 test for several proportions II

Example

A study is conducted in order to analyse the effects of oral contraception (OC) on heart conditions of women aged 40-44. The researchers find that among 5000 women which participated in the study and took OC, 13 suffered a heart attack in the next 3 years, while among 10000 women that did not take OC, 7 suffered a heart attack in the next 3 years. What is the conclusion of the study ($\alpha = 0.05$)?

Comparison of two proportions I

- Ipotezele statistice:

$$\begin{aligned} H_0: &\quad p_1 = p_2 \\ H_a: &\quad p_1 \neq p_2 \\ &\quad p_1 < p_2 \\ &\quad p_1 > p_2 \end{aligned}$$

- The test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

  has approximately the $N(0,1)$ distribution, where $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$ (proportion of succes in the two samples), n_1, n_2 sample sizes, $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$.

- Conditions: $n_1 \cdot \hat{p}, n_1 \cdot (1-\hat{p}), n_2 \cdot \hat{p}, n_2 \cdot (1-\hat{p}) > 5$.

Comparison of two proportions II

Example

A study is conducted to find out the factors that help spread tuberculosis among drug users. Two samples of 97 drug users that admittedly shared needles and 161 that didn't are chosen and tested. The reports show that 34 individuals in the first group and 28 in the second have tuberculosis. Test whether the proportion of TBC infected people is larger among drug users that share needles, with a 0.05 level of significance.

Comparison of means of $k > 2$ populations

Let $\mu_1, \mu_2, \ldots, \mu_k$ be the means of k populations. We test the hypothesis:

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
- H_a: at least one mean is diferent

The following must hold:

i) the observations must be independent

ii) the groups (samples) have the same variance

iii) the errors (differences between the values and the group means) have a normal distribution

The ANOVA test is used.

ANOVA I

In R: **summary(aov(var∼group))**

Example

The data set "WeightLoss" in the "car" package contains data about the weight loss of 34 subjects included in a study, belonging to three groups - control, diet and diet+sport. We would like to compare the weight loss after two months ("wl2" variable) between these groups.

Notations:

- x_{ij} is the *jth* observation (value) in group i
- n_i the number of observations in group i
- $\bar{x}_i$ the mean for group i

The ANOVA model:

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where

ANOVA II

- μ is a constant
- α_i is a constant specific to the *ith* group; $\sum_{i=1}^{k} \alpha_i = 0$
- ϵ_{ij} is an error term with the $N(0, \sigma^2)$ distribution.

Equivalent hypethesis:

- H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
- H_a: there exists an $\alpha_i \neq 0$

We define:

- between groups variance:

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

ANOVA III

- within groups variance:

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

- total variance:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

We have that

$$SST = SSB + SSW$$

The principle behind ANOVA: if the means of the groups are significantly different, the SSB will be larger than SSW.
The test statistic:

$$F = \frac{SSB/(k-1)}{SSW/(n-k)}.$$

ANOVA IV

- Considering the null hypothesis as true, F has a F(k-1,n-k) distribution.

- We reject H_0 if the associated p-value to the calculated value of the test statistic is below the significance level of the test, α (or equivalently, when the computed value of F for the data at hand is large, $F_{obs} > F_{1-\alpha,k-1,n-k}$).
- If the ANOVA model shows that there are significant differences between the groups, we can then use t-tests for each pair of groups to determine which ones differ (**post-hoc ANOVA**).
- In R: **pairwise.t.test(quantitative.variable,group.criteria)**

The Kruskal-Wallis test

- If the assumptions ii) or iii) of the ANOVA test are not met, then we use the Kruskal-Wallis test instead.
- In R: **kruskal.test(var∼group)**
- If there are significant differences, we can then use Mann-Whitney tests to compare pairs of groups.

Testing normality

Hypothesis tested:

$$H_0 : \text{the population has a normal distribution}$$

$$H_a : \text{the population does not have a normal distribution}$$

Tests:

- Shapiro-Wilk: sample size ≤ 5000; shapiro.test(x)
- Anderson-Darling: sample size> 7; ad.test(x) in the *nortest* package

Example

We would like to test if the distribution of life expectancy of 50 batteries produced by factory A is normal. The data can be found in the "Battery" data frame in the "PASWR" package.

Correlation and Linear Regression

Table of contents

We're now going to shift our attention to one of the most exciting and practically useful topics in data analysis: **predictive analytics**.

Whereas in the last unit, we were using data to make inferences about the world, this unit is primarily about using data to make inferences (or **predictions**) about other data.

At the surface level, linear regression is a method that is used both to predict the values that continuous variables take on, and to make inferences about how certain variables are related to a continuous variable.

Although linear regression is, at a high level, conceptually *quite simple*, it is absolutely indispensable to modern applied statistics.

- **Linear regression is used to predict** the value of an outcome variable Y based on one or more input predictor variables X.
- The aim is to **establish a linear relationship** (a mathematical formula) between the predictor variable(s) and the response variable.
- We can use this formula to **estimate** the value of the response Y, when only the predictors (Xs) values are known.

Example

Consider the data frame "cars" in R which contains 50 observations(rows) and 2 variables (columns) – dist and speed.

- print the first 6 observations in the data set
- attach the data set.
- describe the 2 variables "dist" and "speed".

Can we say that there might be a natural relationship between the distance taken to stop and the speed of the car?

How can we find that relationship?

We want to find a mathematical model of the type

$$Y = f(X).$$

How can we find the function f? Hint: our method is called **linear** regression.

In order to find the function f or to find the type of relationship, we could plot the two variables against each other, thus constructing a **scatter plot** of the data.

- do this in R using the function **plot(x,y)**

- what can we see from the graphical image?

- other useful graphs: boxplot of the two variable (for detecting the outliers) and a density plot/histogram/QQ plot to see if the variables have a normal distribution.

- what test can we use to establish if the the 2 variables have a normal distribution?

Correlation I

- **Correlation** (called Pearson correlation coefficient) is a statistical measure that suggests the **level of linear dependence** between two continuous variables.
- **Correlation can take values between -1 to +1**.
  - if the speed increases, the distance also increases along with it, then the correlation between them will be closer to 1.
  - the opposite is true for an inverse relationship, in which case, the correlation between the variables will be close to -1.
- A value closer to 0 suggests a **weak linear relationship** between the variables. A low correlation ($-0.2 < x < 0.2$) probably suggests that much of variation of the response variable (Y) is unexplained by the predictor (X), in which case, we should probably look for better explanatory variables.

Correlation II

Examples of negative correlation:
- a student who has many absences has a decrease in grades
- as weather gets colder, heating bills will increase
- if a train increases speed, the length of time to get to the final point decreases
- the more one works out at the gym, the less body fat one may have.

Positive correlation examples:
- the more time you spend running on a treadmill, the more calories you will burn
- as the temperature goes up, ice cream sales also go up
- when an employee works more hours his paycheck increases proportionately.

Correlation III

The formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}}.$$

Example: Compute the Pearson correlation coefficient of the variables "dist" and "speed" using the function **cor()**. What is the conclusion, looking at the value we obtained?
cor.test()

Build the linear regression model

- The scatter plot of the data and the value of the correlation coefficient indicates that there is a linear relationship between the speed and the stopping distance.
- How do we construct the model in R?
- The function used for building linear models is **lm()**. The lm() function takes in two main arguments, namely: formula and data. The data is typically a data.frame and the formula is a object of class formula.
- Perform the linear regression for the data frame "cars". Explain the output of the lm() function. What is the mathematical form the model (the function f)?

The regression model I

Simple linear regression model has the form

$$y = a + b \cdot x + \varepsilon,$$

where y is the predicted continuous variable, x is the predictive continuous variable, ε is the error of the model and a and b are the unknown coefficients.

How do we find the coefficients a and b? Let's take a look at the **errors (residuals)** of the model.

The regression model II

An idea would be to **minimize the residuals**, but how?

The method of minimizing the residuals is called **least squares method**: it minimizes the sum of squares

$$SSE(\hat{a}, \hat{b}) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \hat{y}_i = a + b \cdot x_i.$$

We now have to find the minimum of function S by computing the partial derivatives with respect to a and b. This gives us

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{xy}\frac{s_x}{s_y}$$

Diagnostics

Now the linear model is built and we have a formula that we can use to predict the "dist" value if a corresponding "speed" is known.

Is this enough to actually use this model? NO! Before using a regression model, you have to ensure that it is **statistically significant**.

How do you ensure this? We have several clues if the model is a valid one or not.

Use the function **summary()** to extract detailed information about the model.

Explain the output of this function.

Tests I

The t-test for the b coefficient of the model:

H_0: $b = 0$
H_a: $b \neq 0$

The test statistic:

$$t = \frac{\hat{b}}{s_{\hat{b}}}, \quad s_{\hat{b}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

has a Student t distribution with parameter $n - 2$.

The standard deviation of $\hat{a}$ is $s_{\hat{a}} = s_{\hat{b}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i}$.

The t-statistic for $\hat{a}$ is $t = \frac{\hat{a}}{s_{\hat{a}}}$.

Tests II

The F-test for the model coincides with the t-test in the case of simple linear regression. Therefore associated p-value of the F-test has the same value as the one for the t-test for coefficient b.

The statistic for the F test:

$$F = \frac{\text{explained variance}}{\text{unexplained variance}} = \frac{MSM}{MSE},$$

where

$$MSM(\text{mean squares for model}) = SSM/DFM,$$

$$MSE(\text{mean squares for error}) = SSE/DFE,$$

$$DFM(\text{degrees freedom model}) = p - 1,$$

$$DFE(\text{degrees freedom error}) = n - p,$$

Tests III

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

n is the number of observations and p is the number of predictors.

The F statistic has a F distribution with parameters DFM, DFE.

The standard error of the residuals: $\sqrt{SSE/(n-p)}$.
The coefficient of determination, denoted R^2 or r^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable. It is used to analyze how **differences in one variable can be**

explained by a difference in a second variable and is the square of the correlation coefficient, thus has values in $[0, 1]$. Another formula

$$R^2 = 1 - \frac{SSM}{SST}, \quad SST = \sum_{i=1}^{n} (y_i - \bar{y})$$

The adjusted R^2 is preferred instead of the R^2 because the latter is influenced by the number of variables in the model. The formula for the adjusted R^2 is

$$\bar{R}^2 = 1 - \frac{MSM}{MST} = 1 - \frac{(1 - R^2)(n-1)}{n-p}, \quad MST = SST/(n-1).$$

Tests V

The adjusted R^2 will increase only if a variable improves the regression more than you would expect by chance. $\bar{R}^2$ doesn't include all data points, is always lower than R^2 and can be negative (although it's usually positive). Negative values will likely happen if R^2 is close to zero — after the adjustment, the value will dip below zero a little.

Observation. Even if there is a strong connection between the two variables, determination does not prove causality.

Assumptions

Assumptions of the regression model:

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

A note about sample size. In linear regression the sample size rule of thumb is that the regression analysis requires at least 20 cases per independent variable in the analysis.

Using the plot() function check the assumptions for the model in the example. Eliminate the outliers and redo the model. Are the assumptions satisfied in the case of the second model?

Add the regression line to the scatter plot of the data. Predict the distance for a speed of 6 and 26 mph.

Try to change the model into a polynomial one with degrees from 2 to 4. Which is the best one?

Multiple linear regression

- In the case of multiple linear regression, the are more than one predictive variable (independent variable).
- The model:

$$y = b_0 + b_1 \cdot x_1 + \cdots + b_n \cdot x_n + \varepsilon$$

- The coefficients of the model are found using the least squares method.
- The assumptions are the same as for simple linear regression.
- The t-tests are the same for each coefficient of the model.
- The F-test has different hypothesis: H_0: $b_1 = b_2 = \cdots = b_n = 0$, H_a: at least one of the coefficients b_i is different from 0.
- The function in R is lm().

Linearization of nonlinear models I

What if the relationship between two continuous variables is not linear?

Linearization of nonlinear models II

We then take a look at the scatter plot and try to figure out what type of model we should fit:

- parabolic model $y = a + b_1 x + b_2 x^2$
- exponential model $y = ae^{bx}$
- logarithmic model $y = a \ln(bx)$
- saturation-growth model $y = \frac{ax}{b+x}$
- power model $y = ax^b$

In each case we have to transform the data, either the dependent, the independent variable or both.

What are the transformations in each case? After you perform the suitable transformations, you can perform linear regression.

Linearization of nonlinear models III

If the model is a polynomial of the type

$$y = a + b_1 x + b_2 x^2 + \cdots + b_p x^p,$$

then this can be treated as a multiple linear model, thus perform multiple linear regression with the predictors x, x^2, ..., x^p.

Logistic Regression

Logistic Regression

- Logistic regression is a regression model used when the dependent variable (Y) is categorical, more specifically binary.
- The output takes the values "0" and "1" which represent: pass/fail, win/lose, alive/dead or healthy/sick.
- The binary logistic model is used to estimate the **probability** of a binary response based on one or more predictor (or independent) variables.
- Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression.

Examples:

- Politics: Predicting whether an American voter will vote Democratic or Republican, based on age, income, sex, race, state of residence, votes in previous elections, etc.
- Spam Detection : Predicting if an email is Spam or not
- Credit Card Fraud : Predicting if a given credit card transaction is fraud or not
- Health : Predicting if a given mass of tissue is benign or malignant or if a patient has a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.)
- Marketing : Predicting if a given user will buy an insurance product or not
- Banking : Predicting if a customer will default on a loan.

Assumptions

- Y is not normally distributed, but has a binomial distribution with parameters n and p.
- The logistic regression does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables $logit(p) = \beta_0 + \beta_1 x$.
- The homogeneity of variance does NOT need to be satisfied.
- Errors need to be independent but NOT normally distributed.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters.

The logistic model

- Predict the probability of success "p".
- Construct the "odds" which is a non negative value:

$$\frac{p}{1-p}.$$

- Apply the log function to the previous value to get the log-odds which is continuous on $\mathbb{R}$:

$$\log(\frac{p}{1-p}).$$

- Assume that there is a linear relationship between the log-odds value and the independent variables (predictors):

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

- Use Maximum Likelihood Estimation(MLE) to estimate the parameters of the model.
- Validate the model.

- We can see that

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

- To minimize the mis-classification rate, we should predict $Y = 1$ when $p \geq 0.5$ and $Y = 0$ when $p < 0.5$.
- This means

$$Y = \begin{cases} 1, & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k > 0 \\ 0, & \text{otherwise} \end{cases}$$

.
- "Odds": $e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}$
- "Odds Ratio": e^{β_i}, for all independent variables X_i.

Interpretation of the model:
- interpretation for β_i: The odds multiply by e^{β_i} for every 1-unit increase in x_i
- if $\beta_j > 0$, then $e^{\beta_j} > 1$, and the odds (of the characteristic being present) increase.
- if $\beta_j < 0$,then $e^{\beta_j} < 1$, and the odds decrease.
- e^{β_0} = the odds that the characteristic is present in an observation i when $X_i = 0$, i.e., at baseline.

Maximum Likelihood Estimation for the parameters of the model with one predictor:

- the likelihood function:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

- the log-likelihood function:

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))) =$$

$$= -\sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_i}) + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i)$$

- we maximize the log-likelihood function in order to find the estimates for the parameters β_0 and β_1
- we thus compute the partial derivatives of the log-likelihood function and then set them to 0
- we can only solve the equations numerically, using Newton-Raphson for example.

Evaluation of model I

- **AIC** (Akaike Information Criteria) = counterpart of adjusted r square in multiple regression
  - smaller the better
  - penalizes increasing number of coefficients in the model
  - useful in comparing models (the model with lowest AIC will be the best)
- **Null Deviance and Residual Deviance**
  - Null deviance is calculated from the model with no features, i.e.,only intercept
  - Residual deviance is calculated from the model having all the features
  - the larger the difference between null and residual deviance, better the model
  - used to compare models (lower Null or Residual deviance, the better the model)

Evaluation of model II

- **Confusion matrix**
  - Accuracy$=\frac{\text{True Positives+True Negatives}}{\text{True Positives+True Negatives+False Positives+False Negatives}}$
  - True Positive Rate (TPR/ Sensitivity/Recall) $=\frac{TP}{TP+FN}$
  - False Positive Rate (FPR)$=\frac{FP}{FP+TN}$
  - True Negative Rate (TNR/Specificity)$=\frac{TN}{TN+FP}$
  - False Negative Rate (FNR)$=\frac{FN}{FN+TP}$
  - Precision$=\frac{TP}{TP+FP}$
  - F score= harmonic mean of precision and recall (the higher the value, the better the model)

Example I

Example

The data set "mtcars" describes different models of a car with their various engine specifications. In "mtcars" data set, the transmission mode (automatic or manual) is described by the column am which is a binary value (0 or 1). We can create a logistic regression model between the columns "am" and 3 other columns - hp, wt and cyl.

In R we use the function **glm()** (generalized linear model).

glm(Y X1+X2+X3, family=binomial(link="logit"), data=mydata)

Example II

```
library(MASS)
attach(mtcars)
head(mtcars)

log_regr=glm(formula = am ~ cyl + hp + wt, data = mtcars,
+family = binomial(link=logit))
log_regr

summary(log_regr)

predict(log_regr,data.frame(cyl=6,hp=100,wt=3.5),type="response")
```

Redo the model so that you eliminate the variable with the highest p-value, "cyl". Discuss your findings.

Sample size

Let p be the smallest of the proportions of negative or positive cases in the population and k the number of covariates (the number of independent variables), then the minimum number of cases to include is:

$$N = 10k/p$$

For example: you have 3 covariates to include in the model and the proportion of positive cases in the population is 0.20. The minimum number of cases required is

$$N = 10 \times 3/0.20 = 150$$

If the resulting number is less than 100 you should increase it to 100 as suggested by Long (1997).

Exercises

1. Consider the data set logregr1.csv which has 200 observations and the outcome variable used will be "hon", indicating if a student is in an honors class or not. Import the data in R using the function "read.csv()" and perform logistic regression including the variables "female" and "math" (a math score). Discuss your findings and say with what percentage do the odds of being in honors class increase with we have an increase of one unit in each variable separately.
2. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable. Read the data into R from the source https://stats.idre.ucla.edu/stat/data/binary.csv and perform logistic regression. Discuss your findings.

References

- Hosmer DW Jr, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression. Third Edition. New Jersey: John Wiley & Sons.
- Long JS (1997) Regression Models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications.