

ALGORITHM	DESCRIPTION	R PACKAGE::FUNCTION	SAMPLE CODE
<b>NBC</b> Naïve Bayes classifier	A classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature	e1071::naiveBayes	naiveBayes(class ~., data = x)
<b>KNN</b> k-Nearest Neighbours	A non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression	class::knn	knn(train, test, cl, k = 1, l = 0, prob = FALSE, use.all = TRUE)
<b>REG</b> Linear Regression	Model the linear relationship between a scalar dependant variable Y and one or more explanatory variables (or independent variables) denoted X	stats::lm	lm(dist ~ speed, data=cars)
<b>LOG</b> Logistic Regression	Used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.	stats::glm	glm(Y ~., family = binomial (link = 'logit'), data = X)
<b>TRE</b> Tree-Based Models	The idea is to consecutively divide (branch) the training dataset based on the input features until an assignment criterion with respect to the target variable into a "data bucket" (leaf) is reached	rpart::rpart	rpart(Kyphosis ~ Age + Number + Start, data = kyphosis)
<b>ANN</b> Artificial Neural Network	Neural networks are built from units called perceptrons. Perceptrons have one or more inputs, an activation function and an output. An ANN model is built up by combining perceptrons in structured layers.	neuralnet::neuralnet	neuralnet(f,data=train_hidden=c(5,3),linear.output=T)
<b>SVM</b> Support Vector Machine	A data classification method that separates data using hyperplanes	e1071::svm	svm(formula, data = NULL, ..., subset, na.action = na.omit, scale = TRUE)
<b>PCA</b> Principal Component Analysis	A procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.	stats::prcomp stats::princomp FactoMineR::PCA ade4::dudi.pca amap::acp	<b>stats</b> : prcomp(formula, data = NULL, subset, na.action, ...) <b>stats</b> : princomp(formula, data = NULL, subset, na.action, ...) <b>FactoMineR</b> : PCA(decathlon, quanti.sup = 11:12, quali.sup=13) <b>ade4</b> : dudi.pca(deugStab, center = deugScen, scale = FALSE, scan = FALSE) <b>amap</b> : acp(lubisch)
<b>KMC</b> k-Mean Clustering	Aims at partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean	stats::kmeans	kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)
<b>HCL</b> Hierarchical Clustering	An approach which builds a hierarchy from the bottom-up, and doesn't require the number of clusters to be specified beforehand.	stats::hclust	hclust(d, method = "complete", members = NULL)

	ALGORITHM	DESCRIPTION	R PACKAGE::FUNCTION	SAMPLE CODE
META ALGORITHM	<b>REGU</b> Regularisation L1 (Lasso) L2 (Ridge)	Regularization adds a penalty on the different parameters of a model to reduce the freedom of the model. Hence, the model will be less likely to fit the noise of the training data and will improve the generalization abilities of the model	glmnet::glmnet	<b>L1</b> : glmnet(myMatrixA, myMatrixB, family = "gaussian", alpha = 1) <b>L2</b> : glmnet(myMatrixA, myMatrixB, family = "gaussian", alpha = 0)
	<b>BOO</b> Boosting	A process of iteratively refining, e.g. by reweighting, of estimated regression and classification functions (though it has primarily been applied to the latter), in order to improve predictive ability.	<b>Parametric model</b> - mboost::glmboost	glmboost(Yen ~ ., data = curr1[trnidxs,])
	<b>BAG</b> Bagging	Bagging is a way to increase the power of a predictive statistical model by taking multiple random samples (with replacement) of the training data set, and using each of them to construct a separate model and separate predictions for the original test set	<b>All models</b> : foreach <b>Tree models</b> : ipred::bagging	<b>foreach</b> : d <- data.frame(x=1:10, y=rnorm(10)) s <- foreach(d=iter(d, by='row'), .combine=rbind) %dopar% identical(s, d) <b>ipred</b> : bagging(formula, data, subset, na.action=na.rpart, \dots)
	<b>PRU</b> Pruning	Pruning is a technique that reduces the size of decision tree by removing sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier and hence improves predictive accuracy by reducing overfitting	rpart::prune	prune(x, cp = 0.1)
	<b>RFO</b> Random Forrest	An ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression)	randomForest::randomForest	randomForest(X ~ ., data = Y, subset = mySub)
TIME SERIES	<b>STS</b> Lead-lag analysis, Auto-correlation, Spectral analysis, Time series clustering, Seasonality, Trend....	Random sampling of observations for training and testing a model can be an issue when faced with a times dimension. Random sampling may either destroy serial correlation properties in the data which we would like to exploit	stats xts forecast spectral TTR ....	<b>Auto-correlation</b> : acf(x, lag.max = NULL, type = c("correlation", "covariance", "partial")) <b>Spectral Analysis</b> : spec.pgram(myTs, spans = NULL) <b>Seasonal Decomposition of Time Series</b> - stl(x, s.window = 7, t.window = 50, t.jump = 1) .....
MODEL VALIDATION	<b>PM</b> Performance metrics	Depends on the problem: • <b>Regression</b> : squared errors, outliers, error rate... • <b>Classification</b> : Accuracy, precision, recall, F-score...	<b>Regression</b> -stats::outlierTest, stats::qqPlot .... <b>Classification</b> -ROCR:: <b>Tree</b> : caret:: confusionMatrix	<b>Regresion</b> : fit <- lm(Y~X,data=myData) outlierTest(fit) qqPlot(fit, main="QQ Plot")
	<b>BVT</b> Biases-Variance Tradeoff	• Simple models with few parameters are easier to compute but may lead to poorer fits ( <b>high bias</b> ). • Complex models may provide more accurate fits but may over-fit the data ( <b>high variance</b> )	Tailored to the analysis	Tailored to the analysis
	<b>CV</b> Cross validation	Cross validation compares the test performances of different model realisations with different sets or values of parameters	caret::createDataPartition caret::createFolds	createDataPartition(classes, p = 0.8, list = FALSE)
	<b>LC</b> Learning Curves	Learning curves plot a model's training and test errors, or the chosen performance metric, depending on the training set size	caret:: learing_curve_dat	learing_curve_dat(dat, outcome = NULL, proportion = (1:10)/10, test_prop = 0, verbose = TRUE, ...)



# R Reference Card

by Tom Short, EPRI PEAC, tshort@epri-peac.com 2004-11-07

Granted to the public domain. See [www.Rpad.org](http://www.Rpad.org) for the source and latest version. Includes material from *R for Beginners* by Emmanuel Paradis (with permission).

## Getting help

Most R functions have online documentation.

**help(topic)** documentation on *topic*

**?topic** id.

**help.search("topic")** search the help system

**apropos("topic")** the names of all objects in the search list matching the regular expression "topic"

**help.start()** start the HTML version of help

**str(a)** display the internal \*str\*ucture of an R object

**summary(a)** gives a "summary" of *a*, usually a statistical summary but it is *generic* meaning it has different operations for different classes of *a*

**ls()** show objects in the search path; specify *pat*="pat" to search on a pattern

**ls.str()** str() for each variable in the search path

**dir()** show files in the current directory

**methods(a)** shows S3 methods of *a*

**methods(class=class(a))** lists all the methods to handle objects of class *a*

## Input and output

**load()** load the datasets written with *save*

**data(x)** loads specified data sets

**library(x)** load add-on packages

**read.table(file)** reads a file in table format and creates a data frame from it; the default separator *sep*=" " is any whitespace; use *header*=TRUE to read the first line as a header of column names; use *as.is*=TRUE to prevent character vectors from being converted to factors; use *comment.char*=" " to prevent "#" from being interpreted as a comment; use *skip*=*n* to skip *n* lines before reading data; see the help for options on row naming, NA treatment, and others

**read.csv("filename", header=TRUE)** id. but with defaults set for reading comma-delimited files

**read.delim("filename", header=TRUE)** id. but with defaults set for reading tab-delimited files

**read.fwf(file, widths, header=FALSE, sep=" ", as.is=FALSE)** read a table of fixed width formatted data into a 'data.frame'; *widths* is an integer vector, giving the widths of the fixed-width fields

**save(file, ...)** saves the specified objects (...) in the XDR platform-independent binary format

**save.image(file)** saves all objects

**cat(..., file="", sep=" ")** prints the arguments after coercing to character; *sep* is the character separator between arguments

**print(a, ...)** prints its arguments; *generic*, meaning it can have different methods for different objects

**format(x, ...)** format an R object for pretty printing

**write.table(x, file="", row.names=TRUE, col.names=TRUE, sep=" ")** prints *x* after converting to a data frame; if *quote* is TRUE,

character or factor columns are surrounded by quotes ("); *sep* is the field separator; *eol* is the end-of-line separator; *na* is the string for missing values; use *col.names*=NA to add a blank column header to get the column headers aligned correctly for spreadsheet input

**sink(file)** output to file, until sink()

Most of the I/O functions have a *file* argument. This can often be a character string naming a file or a connection. *file*="" means the standard input or output. Connections can include files, pipes, zipped files, and R variables.

On windows, the file connection can also be used with *description* = "clipboard". To read a table copied from Excel, use

```
x <- read.delim("clipboard")
```

To write a table to the clipboard for Excel, use

```
write.table(x, "clipboard", sep="\t", col.names=NA)
```

For database interaction, see packages RODBC, DBI, RMySQL, RPgSQL, and ROracle. See packages XML, hdf5, netCDF for reading other file formats.

## Data creation

**c(...)** generic function to combine arguments with the default forming a vector; with *recursive*=TRUE descends through lists combining all elements into one vector

**from:to** generates a sequence; ":" has operator priority; 1:4 + 1 is "2,3,4,5"

**seq(from, to)** generates a sequence by= specifies increment; length= specifies desired length

**seq(along=x)** generates 1, 2, ..., length(along); useful for for loops

**rep(x, times)** replicate *x* times; use *each*= to repeat "each" element of *x* each times; *rep*(c(1,2,3), 2) is 1 2 3 1 2 3; *rep*(c(1,2,3), *each*=2) is 1 1 2 2 3 3

**data.frame(...)** create a data frame of the named or unnamed arguments; *data.frame*(*v*=1:4, *ch*=c("a", "B", "c", "d"), *n*=10); shorter vectors are recycled to the length of the longest

**list(...)** create a list of the named or unnamed arguments; *list*(*a*=c(1,2), *b*="hi", *c*=3*i*);

**array(x, dim=)** array with data *x*; specify dimensions like *dim*=c(3,4,2); elements of *x* recycle if *x* is not long enough

**matrix(x, nrow=, ncol=)** matrix; elements of *x* recycle

**factor(x, levels=)** encodes a vector *x* as a factor

**gl(n, k, length=n\*k, labels=1:n)** generate levels (factors) by specifying the pattern of their levels; *k* is the number of levels, and *n* is the number of replications

**expand.grid()** a data frame from all combinations of the supplied vectors or factors

**rbind(...)** combine arguments by rows for matrices, data frames, and others

**cbind(...)** id. by columns

## Slicing and extracting data

Indexing vectors

<i>x</i> [ <i>n</i> ]	<i>n</i> <sup>th</sup> element
<i>x</i> [- <i>n</i> ]	all <i>but</i> the <i>n</i> <sup>th</sup> element
<i>x</i> [1: <i>n</i> ]	first <i>n</i> elements
<i>x</i> [-(1: <i>n</i> )]	elements from <i>n</i> +1 to the end
<i>x</i> [c(1,4,2)]	specific elements
<i>x</i> ["name"]	element named "name"
<i>x</i> [ <i>x</i> > 3]	all elements greater than 3
<i>x</i> [ <i>x</i> > 3 & <i>x</i> < 5]	all elements between 3 and 5
<i>x</i> [ <i>x</i> %in% c("a", "and", "the")]	elements in the given set

Indexing lists

<i>x</i> [ <i>n</i> ]	list with elements <i>n</i>
<i>x</i> [[ <i>n</i> ]]	<i>n</i> <sup>th</sup> element of the list
<i>x</i> [[ "name" ]]	element of the list named "name"
<i>x</i> \$ <i>name</i>	id.

Indexing matrices

<i>x</i> [ <i>i</i> , <i>j</i> ]	element at row <i>i</i> , column <i>j</i>
<i>x</i> [ <i>i</i> , ]	row <i>i</i>
<i>x</i> [, <i>j</i> ]	column <i>j</i>
<i>x</i> [, c(1,3)]	columns 1 and 3

*x*["name", ] row named "name"

Indexing data frames (matrix indexing plus the following)

<i>x</i> [[ "name" ]]	column named "name"
<i>x</i> \$ <i>name</i>	id.

## Variable conversion

**as.array(x), as.data.frame(x), as.numeric(x), as.logical(x), as.complex(x), as.character(x), ...** convert type; for a complete list, use *methods*(*as*)

## Variable information

**is.na(x), is.null(x), is.array(x), is.data.frame(x), is.numeric(x), is.complex(x), is.character(x), ...** test for type; for a complete list, use *methods*(*is*)

**length(x)** number of elements in *x*

**dim(x)** Retrieve or set the dimension of an object; *dim*(*x*) <- *c*(3,2)

**dimnames(x)** Retrieve or set the dimension names of an object

**nrow(x)** number of rows; *NROW*(*x*) is the same but treats a vector as a one-row matrix

**ncol(x)** and **NCOL(x)** id. for columns

**class(x)** get or set the class of *x*; *class*(*x*) <- "myclass"

**unclass(x)** remove the class attribute of *x*

**attr(x, which)** get or set the attribute *which* of *x*

**attributes(obj)** get or set the list of attributes of *obj*

## Data selection and manipulation

**which.max(x)** returns the index of the greatest element of *x*

**which.min(x)** returns the index of the smallest element of *x*

**rev(x)** reverses the elements of *x*

**sort(x)** sorts the elements of *x* in increasing order; to sort in decreasing order: *rev*(*sort*(*x*))

**cut(x, breaks)** divides *x* into intervals (factors); *breaks* is the number of cut intervals or a vector of cut points

**match(x, y)** returns a vector of the same length than *x* with the elements of *x* which are in *y* (NA otherwise)

**which(x == a)** returns a vector of the indices of *x* if the comparison operation is true (TRUE), in this example the values of *i* for which *x*[*i*] == *a* (the argument of this function must be a variable of mode logical)

**choose(n, k)** computes the combinations of *k* events among *n* repetitions = *n*! / [(*n* - *k*)! *k*!]

**na.omit(x)** suppresses the observations with missing data (NA) (suppresses the corresponding line if *x* is a matrix or a data frame)

**na.fail(x)** returns an error message if *x* contains at least one NA

**unique(x)** if x is a vector or a data frame, returns a similar object but with the duplicate elements suppressed

**table(x)** returns a table with the numbers of the differents values of x (typically for integers or factors)

**subset(x, ...)** returns a selection of x with respect to criteria (... , typically comparisons: x\$V1 < 10); if x is a data frame, the option select gives the variables to be kept or dropped using a minus sign

**sample(x, size)** resample randomly and without replacement size elements in the vector x, the option replace = TRUE allows to resample with replacement

**prop.table(x,margin=)** table entries as fraction of marginal table

Math

**sin,cos,tan,asin,acos,atan,atan2,log,log10,exp**

**max(x)** maximum of the elements of x

**min(x)** minimum of the elements of x

**range(x)** id. then c(min(x), max(x))

**sum(x)** sum of the elements of x

**diff(x)** lagged and iterated differences of vector x

**prod(x)** product of the elements of x

**mean(x)** mean of the elements of x

**median(x)** median of the elements of x

**quantile(x,probs=)** sample quantiles corresponding to the given probabilities (defaults to 0,25,5,75,1)

**weighted.mean(x, w)** mean of x with weights w

**rank(x)** ranks of the elements of x

**var(x)** or cov(x) variance of the elements of x (calculated on n – 1); if x is a matrix or a data frame, the variance-covariance matrix is calculated

**sd(x)** standard deviation of x

**cor(x)** correlation matrix of x if it is a matrix or a data frame (1 if x is a vector)

**var(x, y)** or cov(x, y) covariance between x and y, or between the columns of x and those of y if they are matrices or data frames

**cor(x, y)** linear correlation between x and y, or correlation matrix if they are matrices or data frames

**round(x, n)** rounds the elements of x to n decimals

**log(x, base)** computes the logarithm of x with base base

**scale(x)** if x is a matrix, centers and reduces the data; to center only use the option center=FALSE, to reduce only scale=FALSE (by default center=TRUE, scale=TRUE)

**pmin(x,y,...)** a vector which ith element is the minimum of x[i], y[i],...

**pmax(x,y,...)** id. for the maximum

**cumsum(x)** a vector which ith element is the sum from x[1] to x[i]

**cumprod(x)** id. for the product

**cummin(x)** id. for the minimum

**cummax(x)** id. for the maximum

**union(x,y),intersect(x,y),setdiff(x,y),setequal(x,y),is.element(e1,set)** “set” functions

**Re(x)** real part of a complex number

**Im(x)** imaginary part

**Mod(x)** modulus; abs(x) is the same

**Arg(x)** angle in radians of the complex number

**Conj(x)** complex conjugate

**convolve(x,y)** compute the several kinds of convolutions of two sequences

**fft(x)** Fast Fourier Transform of an array

**mvfft(x)** FFT of each column of a matrix

**filter(x,filter)** applies linear filtering to a univariate time series or to each series separately of a multivariate time series

Many math functions have a logical parameter na.rm=FALSE to specify missing data (NA) removal.

Matrices

**t(x)** transpose

**diag(x)** diagonal

**%%** matrix multiplication

**solve(a,b)** solves a %% x = b for x

**solve(a)** matrix inverse of a

**rowsum(x)** sum of rows for a matrix-like object; **rowSums(x)** is a faster version

**colsum(x),colSums(x)** id. for columns

**rowMeans(x)** fast version of row means

**colMeans(x)** id. for columns

Advanced data processing

**apply(X,INDEX,FUN=)** a vector or array or list of values obtained by applying a function FUN to margins (INDEX) of X

**lapply(X,FUN)** apply FUN to each element of the list X

**tapply(X,INDEX,FUN=)** apply FUN to each cell of a ragged array given by X with indexes INDEX

**by(data,INDEX,FUN)** apply FUN to data frame data subsetted by INDEX

**merge(a,b)** merge two data frames by common columns or row names

**xtabs(a b,data=x)** a contingency table from cross-classifying factors

**aggregate(x,by,FUN)** splits the data frame x into subsets, computes summary statistics for each, and returns the result in a convenient form; by is a list of grouping elements, each as long as the variables in x

**stack(x, ...)** transform data available as separate columns in a data frame or list into a single column

**unstack(x, ...)** inverse of stack()

**reshape(x, ...)** reshapes a data frame between ‘wide’ format with repeated measurements in separate columns of the same record and ‘long’ format with the repeated measurements in separate records; use (direction=“wide”) or (direction=“long”)

Strings

**paste(...)** concatenate vectors after converting to character; sep= is the string to separate terms (a single space is the default); collapse= is an optional string to separate “collapsed” results

**substr(x,start,stop)** substrings in a character vector; can also assign,as substr(x, start, stop) <- value

**strsplit(x,split)** split x according to the substring split

**grep(pattern,x)** searches for matches to pattern within x; see ?regex

**gsub(pattern,replacement,x)** replacement of matches determined by regular expression matching sub() is the same but only replaces the first occurrence.

**tolower(x)** convert to lowercase

**toupper(x)** convert to uppercase

**match(x,table)** a vector of the positions of first matches for the elements of x among table

**x %in% table** id. but returns a logical vector

**pmatch(x,table)** partial matches for the elements of x among table

**nchar(x)** number of characters

Dates and Times

The class Date has dates without times. POSIXct has dates and times, including time zones. Comparisons (e.g. >), seq(), and difftime() are useful. Date also allows + and -. ?DateTimeClasses gives more information. See also package chron.

**as.Date(s)** and **as.POSIXct(s)** convert to the respective class; format(dt) converts to a string representation. The default string format is “2001-02-21”. These accept a second argument to specify a format for conversion. Some common formats are:

- %a,%A Abbreviated and full weekday name.
- %b,%B Abbreviated and full month name.
- %d Day of the month (01–31).
- %H Hours (00–23).
- %I Hours (01–12).
- %j Day of year (001–366).
- %m Month (01–12).
- %M Minute (00–59).
- %p AM/PM indicator.
- %S Second as decimal number (00–61).
- %U Week (00–53); the first Sunday as day 1 of week 1.
- %w Weekday (0–6, Sunday is 0).
- %W Week (00–53); the first Monday as day 1 of week 1.
- %y Year without century (00–99). Don’t use.
- %Y Year with century.
- %z (output only.) Offset from Greenwich; -0800 is 8 hours west of.
- %Z (output only.) Time zone as a character string (empty if not available).

Where leading zeros are shown they will be used on output but are optional on input. See ?strftime.

Plotting

**plot(x)** plot of the values of x (on the y-axis) ordered on the x-axis

**plot(x, y)** bivariate plot of x (on the x-axis) and y (on the y-axis)

**hist(x)** histogram of the frequencies of x

**barplot(x)** histogram of the values of x; use horiz=FALSE for horizontal bars

**dotchart(x)** if x is a data frame, plots a Cleveland dot plot (stacked plots line-by-line and column-by-column)

**pie(x)** circular pie-chart

**boxplot(x)** “box-and-whiskers” plot

**sunflowerplot(x, y)** id. than plot() but the points with similar coordinates are drawn as flowers which petal number represents the number of points

**stripplot(x)** plot of the values of x on a line (an alternative to boxplot() for small sample sizes)

**coplot(x~y | z)** bivariate plot of x and y for each value or interval of values of z

**interaction.plot(f1, f2, y)** if f1 and f2 are factors, plots the means of y (on the y-axis) with respect to the values of f1 (on the x-axis) and of f2 (different curves); the option fun allows to choose the summary statistic of y (by default fun=mean)

**matplot(x,y)** bivariate plot of the first column of x vs. the first one of y, the second one of x vs. the second one of y, etc.

**fourfoldplot(x)** visualizes, with quarters of circles, the association between two dichotomous variables for different populations (x must be an array with dim=c(2, 2, k), or a matrix with dim=c(2, 2) if k = 1)

**assocplot(x)** Cohen–Friendly graph showing the deviations from independence of rows and columns in a two dimensional contingency table

**mosaicplot(x)** ‘mosaic’ graph of the residuals from a log-linear regression of a contingency table

**pairs(x)** if x is a matrix or a data frame, draws all possible bivariate plots between the columns of x

**plot.ts(x)** if x is an object of class "ts", plot of x with respect to time, x may be multivariate but the series must have the same frequency and dates

**ts.plot(x)** id. but if x is multivariate the series may have different dates and must have the same frequency

**qqnorm(x)** quantiles of x with respect to the values expected under a normal law

**qqplot(x, y)** quantiles of y with respect to the quantiles of x

**contour(x, y, z)** contour plot (data are interpolated to draw the curves), x and y must be vectors and z must be a matrix so that dim(z)=c(length(x), length(y)) (x and y may be omitted)

**filled.contour(x, y, z)** id. but the areas between the contours are coloured, and a legend of the colours is drawn as well

**image(x, y, z)** id. but with colours (actual data are plotted)

**persp(x, y, z)** id. but in perspective (actual data are plotted)

**stars(x)** if x is a matrix or a data frame, draws a graph with segments or a star where each row of x is represented by a star and the columns are the lengths of the segments

**symbols(x, y, ...)** draws, at the coordinates given by x and y, symbols (circles, squares, rectangles, stars, thermometres or “boxplots”) which sizes, colours ... are specified by supplementary arguments

**termpLOT(mod.obj)** plot of the (partial) effects of a regression model (mod.obj)

The following parameters are common to many plotting functions:

**add=FALSE** if TRUE superposes the plot on the previous one (if it exists)

**axes=TRUE** if FALSE does not draw the axes and the box

**type="p"** specifies the type of plot, "p": points, "l": lines, "b": points connected by lines, "o": id. but the lines are over the points, "h": vertical lines, "s": steps, the data are represented by the top of the vertical lines, "S": id. but the data are represented by the bottom of the vertical lines

**xlim=, ylim=** specifies the lower and upper limits of the axes, for example plot(xlim=c(1, 10) or xlim=range(x))

**xlab=, ylab=** annotates the axes, must be variables of mode character

**main=** main title, must be a variable of mode character

**sub=** sub-title (written in a smaller font)

## Low-level plotting commands

**points(x, y)** adds points (the option type= can be used)

**lines(x, y)** id. but with lines

**text(x, y, labels, ...)** adds text given by labels at coordinates (x,y); a typical use is: plot(x, y, type="n"); text(x, y, names)

**mtext(text, side=3, line=0, ...)** adds text given by text in the margin specified by side (see axis() below); line specifies the line from the plotting area

**segments(x0, y0, x1, y1)** draws lines from points (x0,y0) to points (x1,y1)

**arrows(x0, y0, x1, y1, angle= 30, code=2)** id. with arrows at points (x0,y0) if code=2, at points (x1,y1) if code=1, or both if code=3; angle controls the angle from the shaft of the arrow to the edge of the arrow head

**abline(a,b)** draws a line of slope b and intercept a

**abline(h=y)** draws a horizontal line at ordinate y

**abline(v=x)** draws a vertical line at abscissa x

**abline(lm.obj)** draws the regression line given by lm.obj

**rect(x1, y1, x2, y2)** draws a rectangle which left, right, bottom, and top limits are x1, x2, y1, and y2, respectively

**polygon(x, y)** draws a polygon linking the points with coordinates given by x and y

**legend(x, y, legend)** adds the legend at the point (x,y) with the symbols given by legend

**title()** adds a title and optionally a sub-title

**axis(side, vect)** adds an axis at the bottom (side=1), on the left (2), at the top (3), or on the right (4); vect (optional) gives the abscissa (or ordinates) where tick-marks are drawn

**rug(x)** draws the data x on the x-axis as small vertical lines

**locator(n, type="n", ...)** returns the coordinates (x,y) after the user has clicked n times on the plot with the mouse; also draws symbols (type="p") or lines (type="l") with respect to optional graphic parameters (...); by default nothing is drawn (type="n")

## Graphical parameters

These can be set globally with **par(...)**; many can be passed as parameters to plotting commands.

**adj** controls text justification (0 left-justified, 0.5 centred, 1 right-justified)

**bg** specifies the colour of the background (ex. : bg="red", bg="blue", ... the list of the 657 available colours is displayed with colors())

**bty** controls the type of box drawn around the plot, allowed values are: "o", "l", "7", "c", "u" ou "j" (the box looks like the corresponding character); if bty="n" the box is not drawn

**cex** a value controlling the size of texts and symbols with respect to the default; the following parameters have the same control for numbers on the axes, cex.axis, the axis labels, cex.lab, the title, cex.main, and the sub-title, cex.sub

**col** controls the color of symbols and lines; use color names: "red", "blue" see colors() or as "#RRGGBB"; see rgb(), hsv(), gray(), and rainbow(); as for cex there are: col.axis, col.lab, col.main, col.sub

**font** an integer which controls the style of text (1: normal, 2: italics, 3: bold, 4: bold italics); as for cex there are: font.axis, font.lab, font.main, font.sub

**las** an integer which controls the orientation of the axis labels (0: parallel to the axes, 1: horizontal, 2: perpendicular to the axes, 3: vertical)

**lty** controls the type of lines, can be an integer or string (1: "solid", 2: "dashed", 3: "dotted", 4: "dotdash", 5: "longdash", 6: "twodash", or a string of up to eight characters (between "0" and "9") which specifies alternatively the length, in points or pixels, of the drawn elements and the blanks, for example lty="44" will have the same effect than lty=2

**lwd** a numeric which controls the width of lines, default 1

**mar** a vector of 4 numeric values which control the space between the axes and the border of the graph of the form c(bottom, left, top, right), the default values are c(5.1, 4.1, 4.1, 2.1)

**mfcOL** a vector of the form c(nr,nc) which partitions the graphic window as a matrix of nr lines and nc columns, the plots are then drawn in columns

**mfrow** id. but the plots are drawn by row

**pch** controls the type of symbol, either an integer between 1 and 25, or any single character within ""

1 ○ 2 △ 3 + 4 × 5 ◇ 6 ▽ 7 ⊗ 8 ✱ 9 ⊕ 10 ⊕ 11 ✎ 12 ⊞ 13 ⊞ 14 ⊞ 15 ■  
16 ● 17 ▲ 18 ◆ 19 ● 20 ● 21 ○ 22 □ 23 ◇ 24 △ 25 ▽ \* · · · X X a a ? ?

**ps** an integer which controls the size in points of texts and symbols

**pty** a character which specifies the type of the plotting region, "s": square, "m": maximal

**tck** a value which specifies the length of tick-marks on the axes as a fraction of the smallest of the width or height of the plot; if tck=1 a grid is drawn

**tcl** a value which specifies the length of tick-marks on the axes as a fraction of the height of a line of text (by default tcl=-0.5)

**xaxt** if xaxt="n" the x-axis is set but not drawn (useful in conjunction with axis(side=1, ...))

**yaxt** if yaxt="n" the y-axis is set but not drawn (useful in conjunction with axis(side=2, ...))

## Lattice (Trellis) graphics

**xyplot(y~x)** bivariate plots (with many functionalities)

**barchart(y~x)** histogram of the values of y with respect to those of x

**dotplot(y~x)** Cleveland dot plot (stacked plots line-by-line and column-by-column)

**densityplot(~x)** density functions plot

**histogram(~x)** histogram of the frequencies of x

**bwplot(y~x)** “box-and-whiskers” plot

**qqmath(~x)** quantiles of x with respect to the values expected under a theoretical distribution

**stripplot(y~x)** single dimension plot, x must be numeric, y may be a factor

**qq(y~x)** quantiles to compare two distributions, x must be numeric, y may be numeric, character, or factor but must have two ‘levels’

**spLOm(~x)** matrix of bivariate plots

**parallel(~x)** parallel coordinates plot

**levelplot(z~x\*y|g1\*g2)** coloured plot of the values of z at the coordinates given by x and y (x, y and z are all of the same length)

**wireframe(z~x\*y|g1\*g2)** 3d surface plot

**cloud(z~x\*y|g1\*g2)** 3d scatter plot

# 1 Sannolikhetslära

## 1.1 Några diskreta fördelningar

- Binomialfördelning

$$X \sim \text{Bin}(n, p)$$

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n$$

$$E[X] = np, \quad \text{var}(X) = np(1-p), \quad G_X(s) = (1 + p(s-1))^n$$

- Poissonfördelning

$$X \sim \text{Po}(\mu)$$

$$p_X(k) = \frac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, 2, \dots$$

$$E[X] = \mu, \quad \text{var}(X) = \mu, \quad G_X(s) = e^{(s-1)\mu}$$

- Hypergeometriskfördelning

$$X \sim \text{Hyp}(N, n, p)$$

$$p_X(k) = \frac{\binom{Np}{k} \binom{N(1-p)}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, \dots, n$$

$$E[X] = np, \quad \text{var}(X) = \frac{N-n}{N-1} np(1-p)$$

- Geometrisk fördelning

$$X \sim \text{Ge}(p)$$

$$p_X(k) = (1-p)^k p, \quad k = 0, 1, 2, \dots$$

$$E[X] = \frac{1-p}{p}, \quad \text{var}(X) = \frac{1-p}{p^2}, \quad G_X(s) = \frac{p}{1 - (1-p)s}$$

- För första-gången-fördelning

$$X \sim \text{Ffg}(p)$$

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \dots$$

$$E[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1-p}{p^2}, \quad G_X(s) = \frac{sp}{1 - (1-p)s}$$

- Negativ Binomialfördelning

$$X \sim \text{NB}(r, p)$$

$$p_X(k) = \binom{k+r-1}{r-1} p^r (1-p)^{k-r}, \quad k = 0, 1, 2, \dots$$

$$E[X] = \frac{r(1-p)}{p}, \quad \text{var}(X) = \frac{r(1-p)}{p^2}, \quad G_X(s) = \left( \frac{p}{1 - (1-p)s} \right)^r$$

## 1.2 Några kontinuerliga fördelningar

- Likformig (rektangulär) fördelning på intervallet (a,b)

$$X \sim U(a, b)$$

$$f_X(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

$$E[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12}, \quad M_X(p) = \frac{e^{pb} - e^{pa}}{p(b-a)}$$

- Exponentialfördelning

$$X \sim \text{Exp}(\lambda),$$

där  $\lambda$  betecknar intensiteten. Ibland används väntevärdet  $\mu = \frac{1}{\lambda}$  som parameter.

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

$$E[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}, \quad M_X(p) = \frac{\lambda}{\lambda - p}, \quad \lambda > p.$$

- Normalfördelning

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

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad -\infty < x < +\infty$$

$$E[X] = \mu, \quad \text{var}(X) = \sigma^2, \quad M_X(p) = \exp \left\{ p\mu + \frac{p^2}{2}\sigma^2 \right\}$$

- $\chi^2$ -fördelning

$$Y \sim \chi^2(n)$$

Uppkomst: Om  $X_1, \dots, X_n$  är oberoende, var och en  $N(0, 1)$ , gäller att  $Y = X_1^2 + \dots + X_n^2$  får en  $\chi^2$  fördelning med  $n$  frihetsgrader.

$$f_Y(x) = \frac{x^{(n/2)-1} e^{-x/2}}{2^{(n/2)} \Gamma(n/2)}, \quad x \geq 0,$$

där  $\Gamma(\cdot)$  är gammafunktionen

$$\Gamma(c) = \int_0^\infty x^{c-1} e^{-x} dx, \quad \text{där } c > 0.$$

$$E[Y] = n, \quad \text{var}(Y) = 2n, \quad M_Y(p) = \frac{1}{(1-p)^{n/2}}, \quad p < 1$$