Exploratory Data Analysis with R[1]

Luís Paquete

paquete@dei.uc.pt

CISUC, Department of Informatics Engineering,
Faculty of Science and Technology,
University of Coimbra

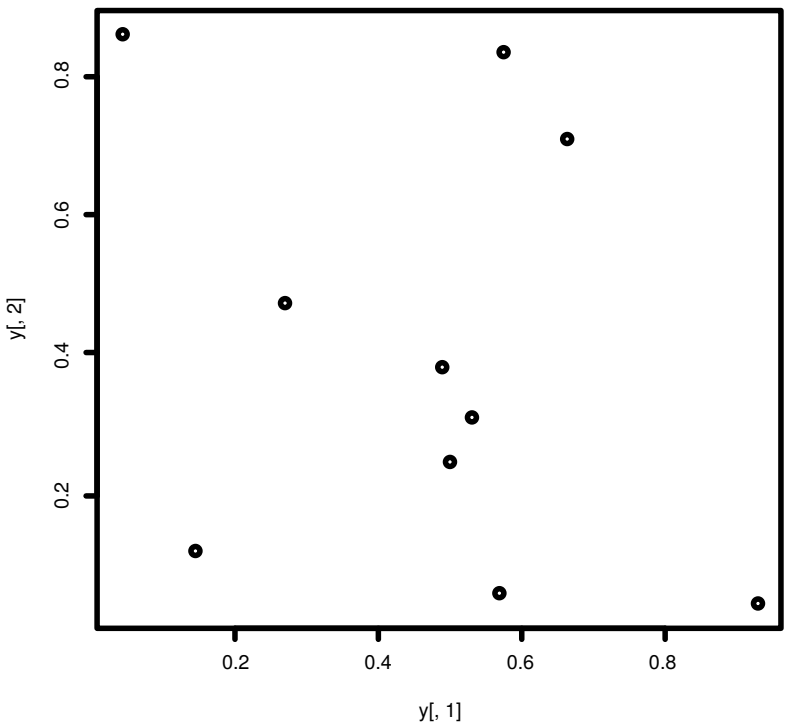
[1] Based on the material of Thomas Girke, U California, 2013

High-level plotting functions

- `plot`: generic x-y plotting
- `barplot`: bar plots
- `boxplot`: box-and-whisker plot
- `hist`: histograms
- `pie`: pie charts
- `qqnorm,qqline,qqplot`: distribution comparison plots
- `pairs,coplot`: display of multivariant data

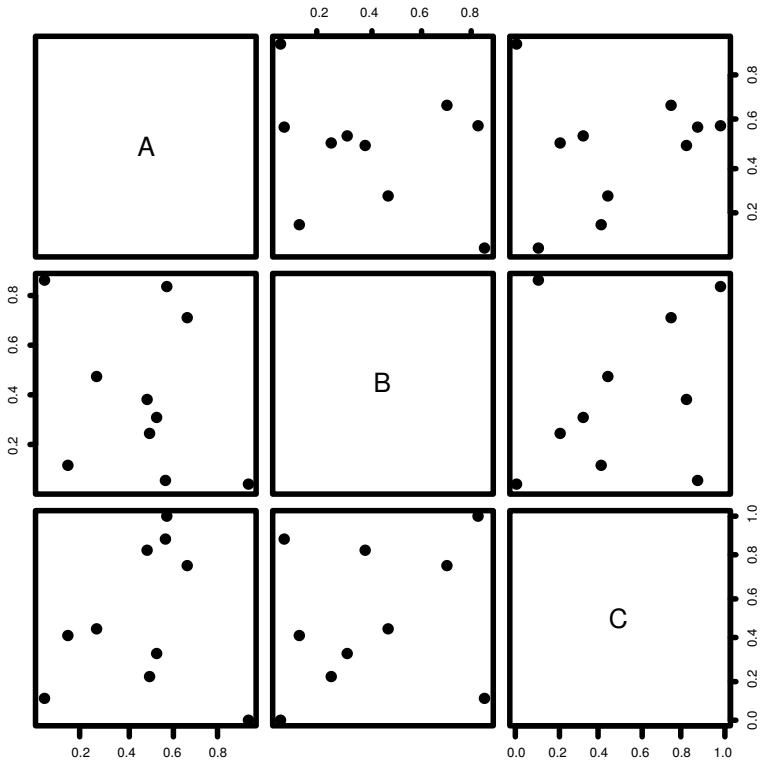
Sample data set

```
set.seed(1410)
y <- matrix(runif(30),ncol=3,dimnames=list(letters[1:10],LETTERS[1:3]))
plot(y[,1],y[,2])
```



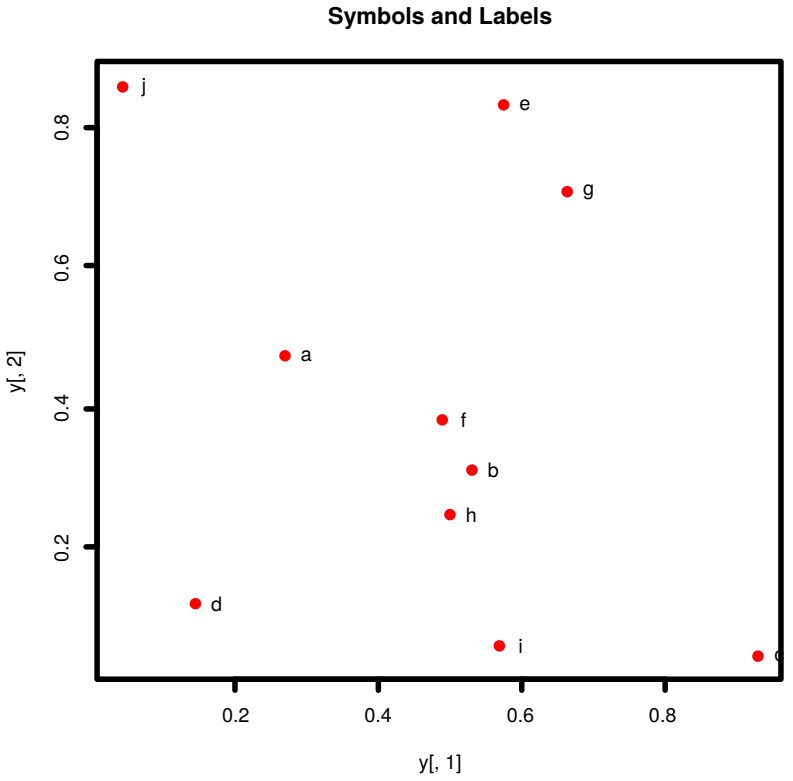
Scatter plots: all pairs

```
pairs(y)
```



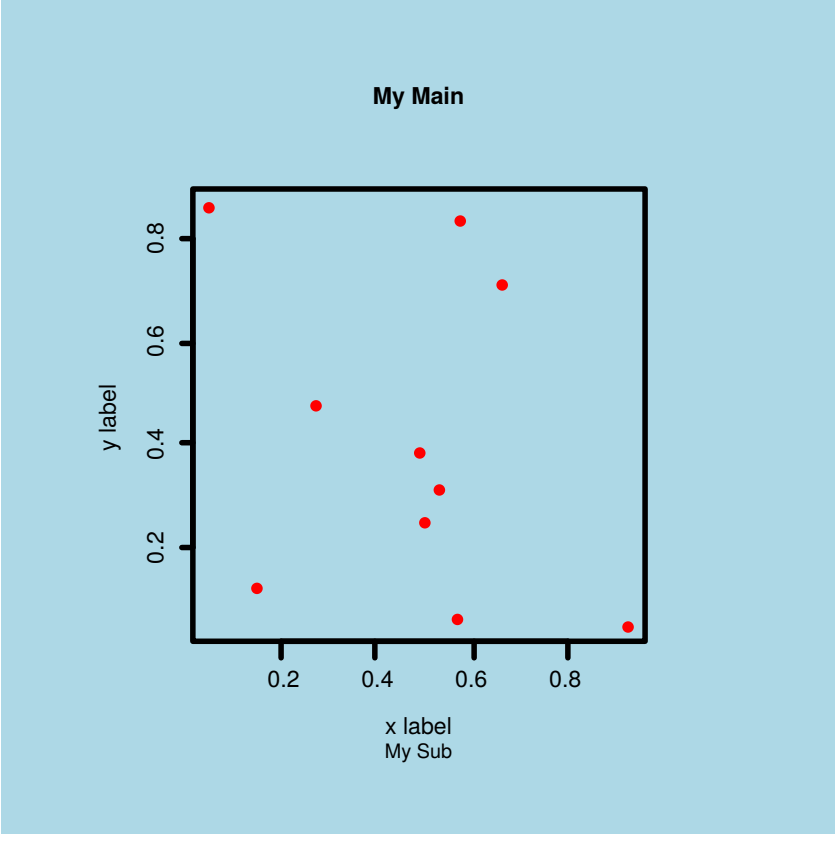
Scatter plots: with labels

```
plot(y[,1], y[,2], pch=20, col="red", main="Symbols and Labels")
text(y[,1]+0.03, y[,2], rownames(y))
```



Scatter plots: more examples

```
op <- par(mar=c(8,8,8,8), bg="lightblue")
plot(y[,1], y[,2], type="p", col="red", cex.lab=1.2, cex.axis=1.2,
        cex.main=1.2,cex.sub=1, lwd=4, pch=20, xlab="x label",
        ylab="y label", main="My Main", sub="My Sub")
par(op)
```

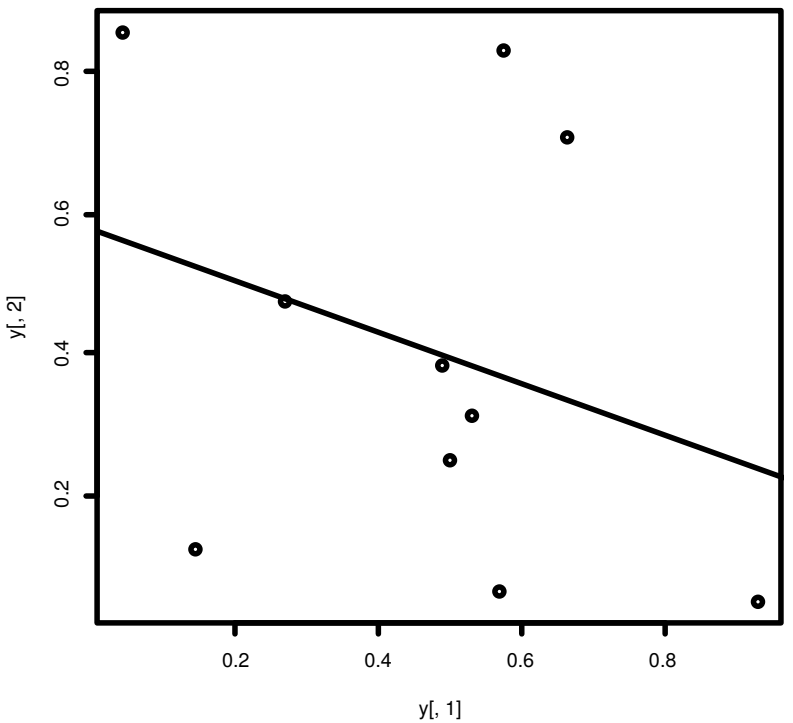


Important arguments

- `mar`: specifies the margin sizes around the plotting area in order: `c(bottom,left, top, right)`
- `col`: color of symbols
- `pch`: type of symbols, samples: `example(points)`
- `lwd`: size of symbols
- `cex.*`: control font sizes
- For details see `?par`

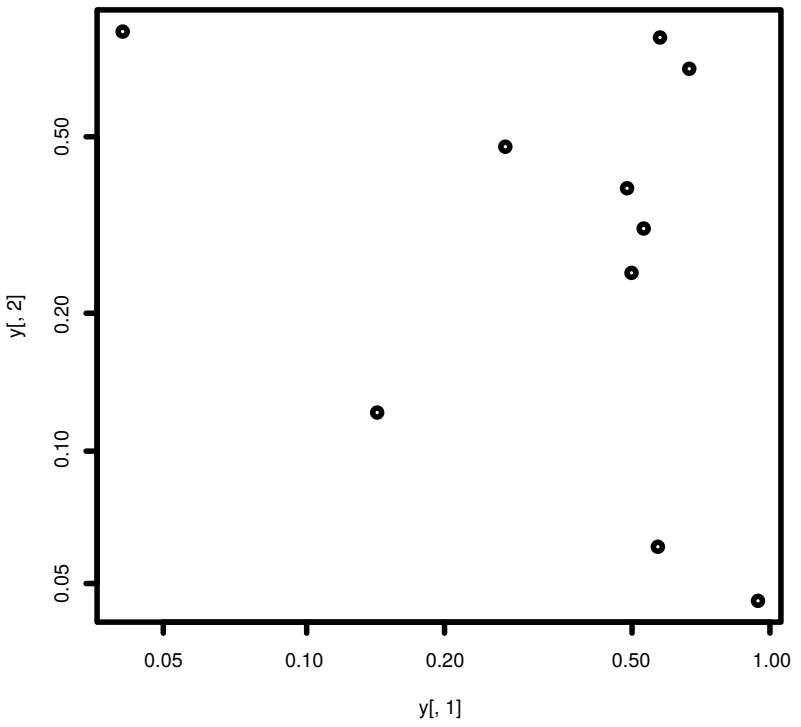
Regression line to a plot

```
plot(y[,1], y[,2])
myline <- lm(y[,2]~y[,1])
abline(myline, lwd=2)
summary(myline)
```



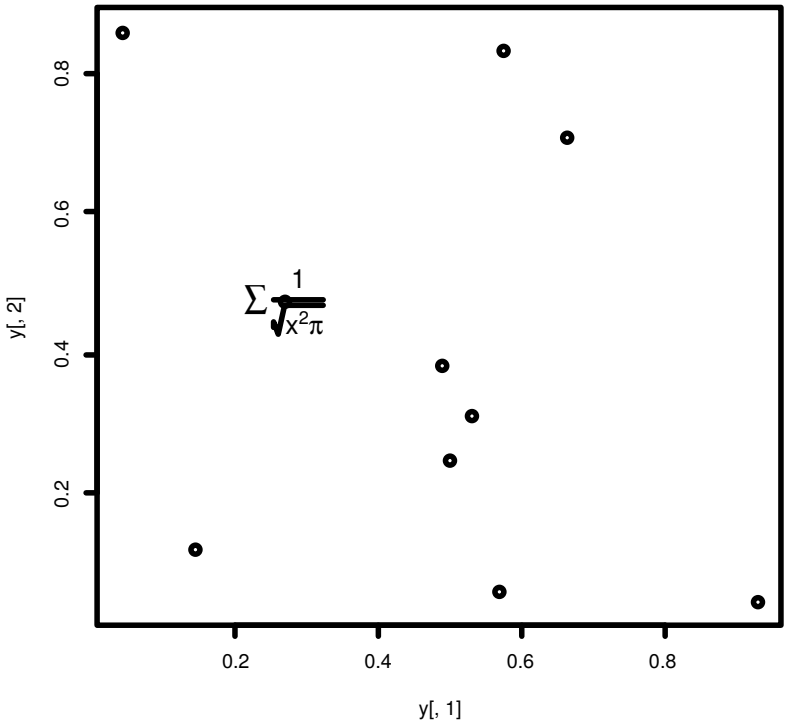
Log scale

```
plot(y[,1], y[,2], log="xy")
```



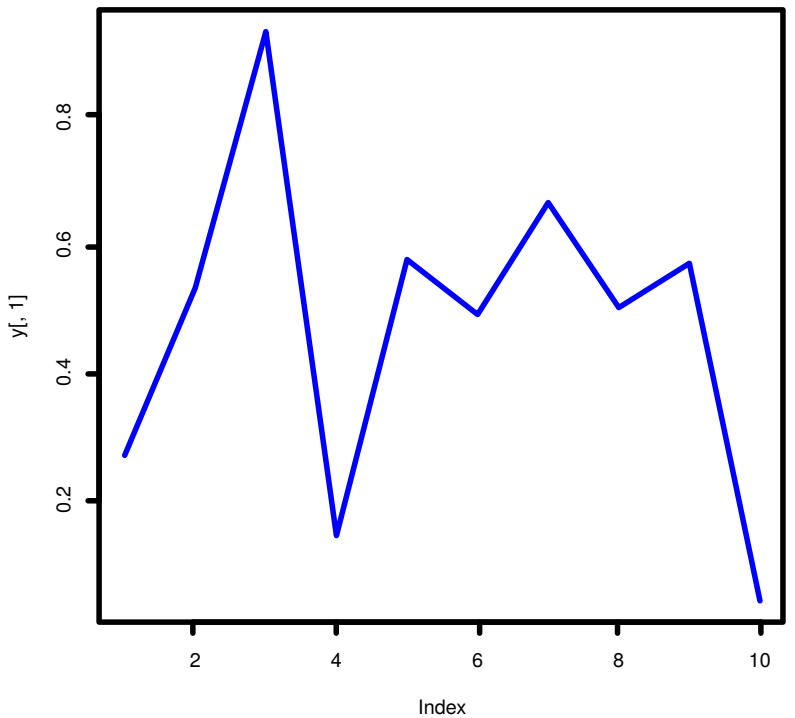
Add mathematical expression to the plot

```
plot(y[,1], y[,2])
text(y[1,1], y[1,2],expression(sum(frac(1,sqrt(x^2*pi)))), cex=1.3)
```



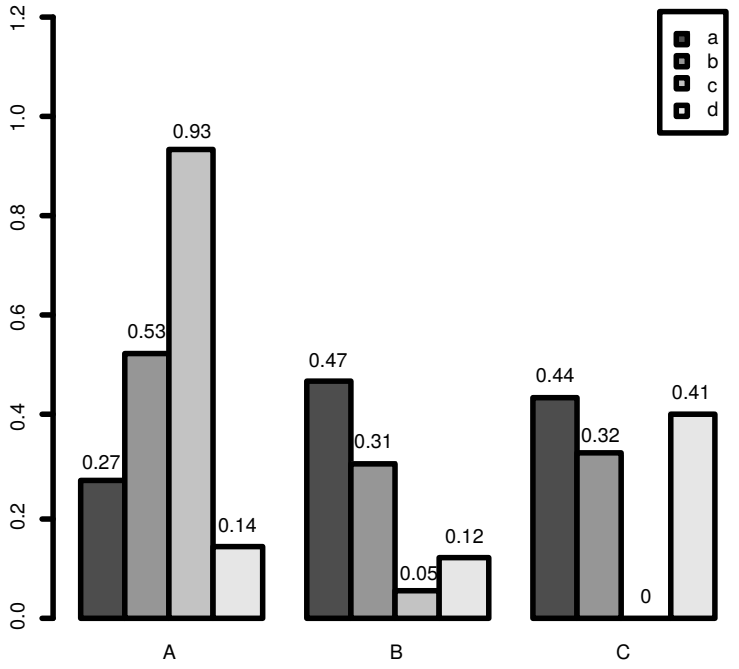
Line plot: Single data set

```
plot(y[,1], type="l", lwd=2, col="blue")
```



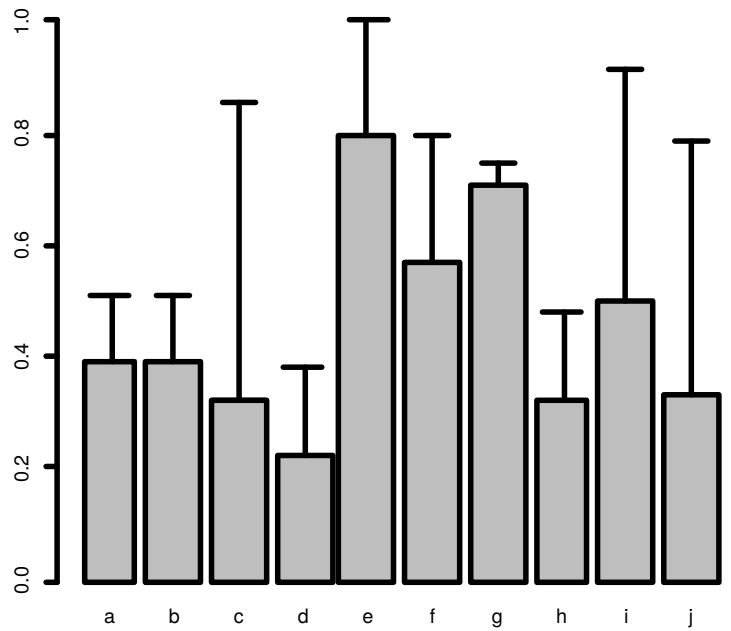
Bar plot

```
barplot(y[1:4,], ylim=c(0, max(y[1:4,])+0.3),beside=TRUE,legend=letters[1:4])
text(labels=round(as.vector(y[1:4,]),2), x=seq(1.5, 13, by=1)
     +sort(rep(c(0,1,2), 4)), y=as.vector(y[1:4,])+0.04)
```



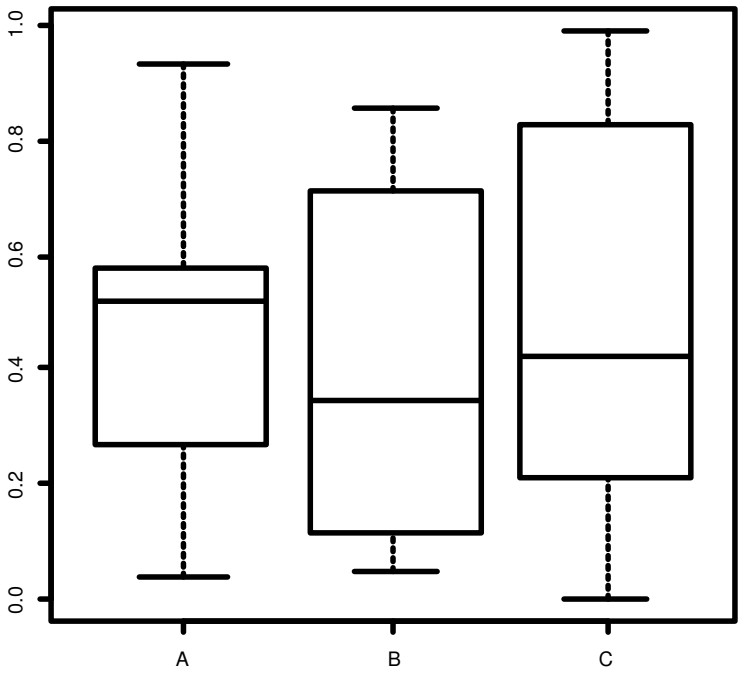
Bar plot with Error Bars

```
bar <- barplot(m <- rowMeans(y), ylim=c(0, 1.1))
stdev <- apply(y,1,sd)
arrows(bar, m, bar, m + stdev, length=0.15, angle = 90)
```



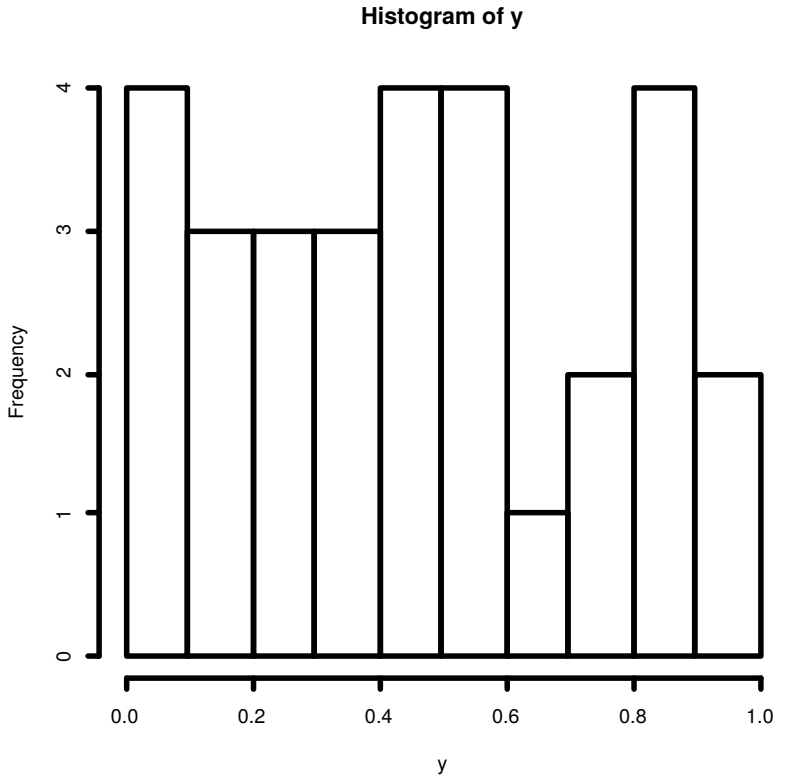
Boxplot

```
boxplot(y)
```



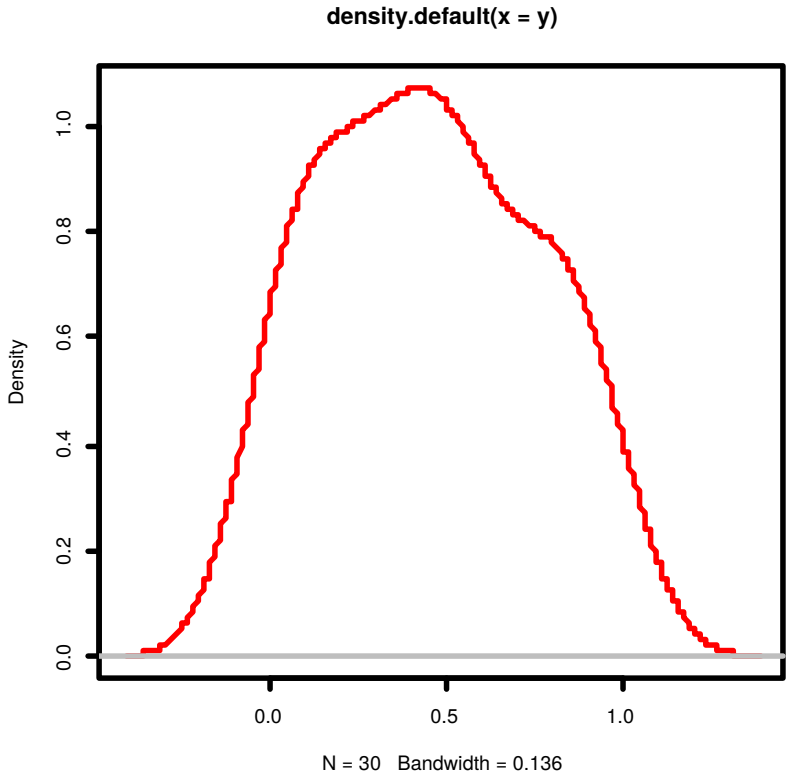
Histogram

```
hist(y, freq=TRUE, breaks=10)
```



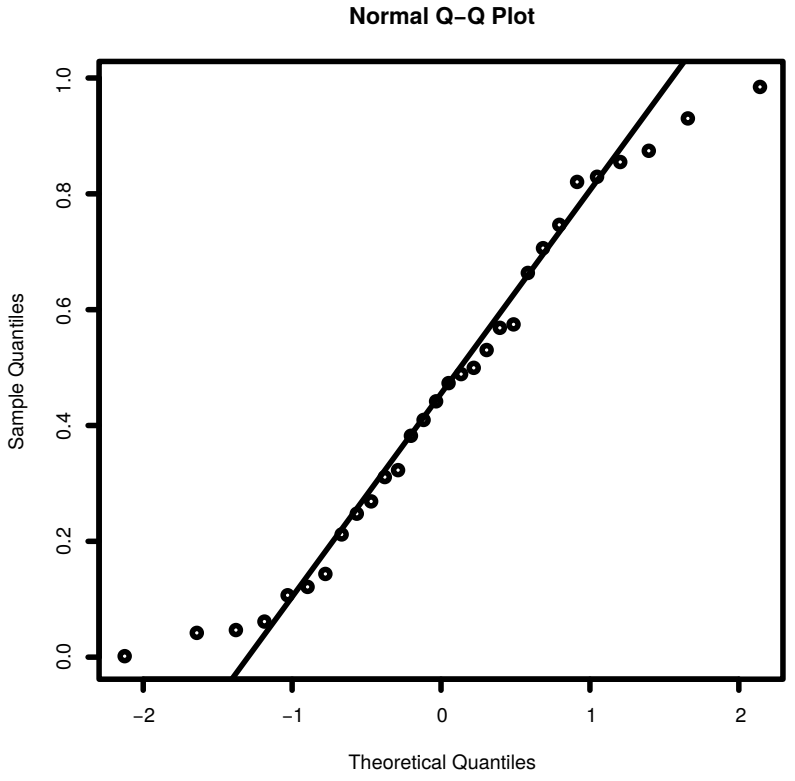
Density Plots

```
plot(density(y), col="red")
```



QQ-plots for Normal Distribution

```
qqnorm(as.vector(y))
qqline(as.vector(y))
```



Distributions

- Binomial: `binom`
- Cauchy: `cauchy`
- Chisquare: `chisq`
- Exponential: `exp`
- Gamma: `gamma`
- Geometric: `geom`
- Lognormal: `lnorm`
- Normal: `norm`
- Poisson: `pois`
- Uniform: `unif`
- Weibull: `weibull`

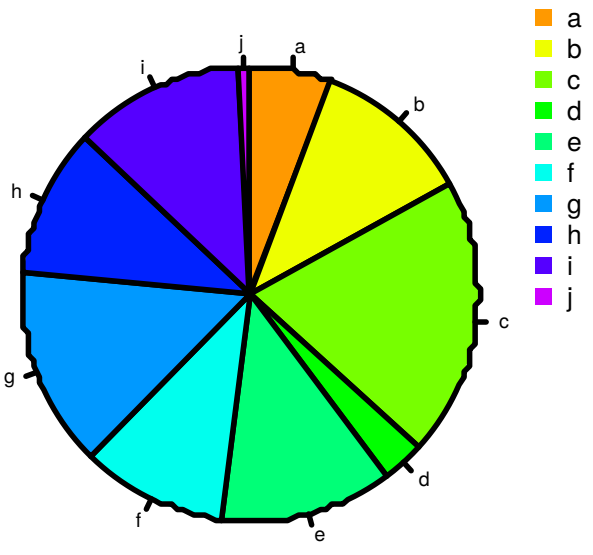
- prefix `r`: random variable generator (`runif`)
- prefix `d`: probability density function (`dunif`)
- prefix `p`: cumulative density function (`punif`)
- prefix `q`: quantile function (`qunif`)

See more in:

`https://en.wikibooks.org/wiki/R_Programming/Probability_Distributions`

Pie charts

```
pie(y[,1], col=rainbow(length(y[,1]), start=0.1, end=0.8), clockwise=TRUE)
legend("topright", legend=row.names(y), cex=1.3, bty="n", pch=15, pt.cex=1.8,
       col=rainbow(length(y[,1]), start=0.1, end=0.8), ncol=1)
```



Color Selection

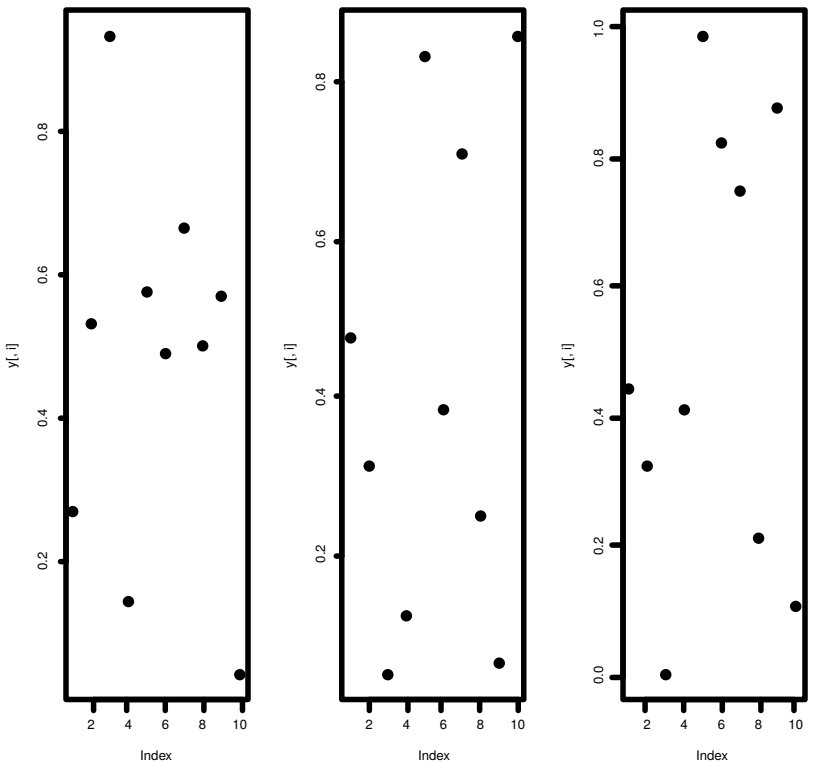
- Default color palette:
  `palette()`
- Change color palette:
  `palette(rainbow(5, start=0.1, end=0.2))`
- Change palette to default:
  `palette(default)`
- Gray shades: `gray(seq(0.1, 1, by= 0.2))`

See color chart in R :

`http://research.stowers-institute.org/efg/R/Color/Chart/`

Several plots

```
par(mfrow=c(1,3))
for(i in 1:3) {
    plot(y[,i]
}
```



Saving graphics to files

- After the `pdf()`, all graphics are redirected to file `test.pdf`:
  `pdf("test.pdf"); plot(1:10, 1:10); dev.off()`
- Works similarly to `bmp`, `jpeg`, `png`, `tiff` and `svg`.

Other graphical packages

- `grid` package:

  `https://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html`

- `lattice` package:

  `http://lmdvr.r-forge.r-project.org/figures/figures.html`

- `ggplot2` package:

  `http://ggplot2.org/`