This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2020.

The >eR-Biostat initiative

Making R based education materials in statistics accessible for all

An introduction to R: Short Version (2017)

Part 1: a quick start

Developed by

Dan Lin (Hasselt University) and Ziv Shkedy (Hasselt University)

LAST UPDATE: 15/10/2017

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

Email: erbiostat@gmail.com

twitter @erbiostat

Overview

A (very) quick start:

- Sampling from a normal distribution.
- Working with data: the cars data.
- Two sample t-test.
- Basic plots.

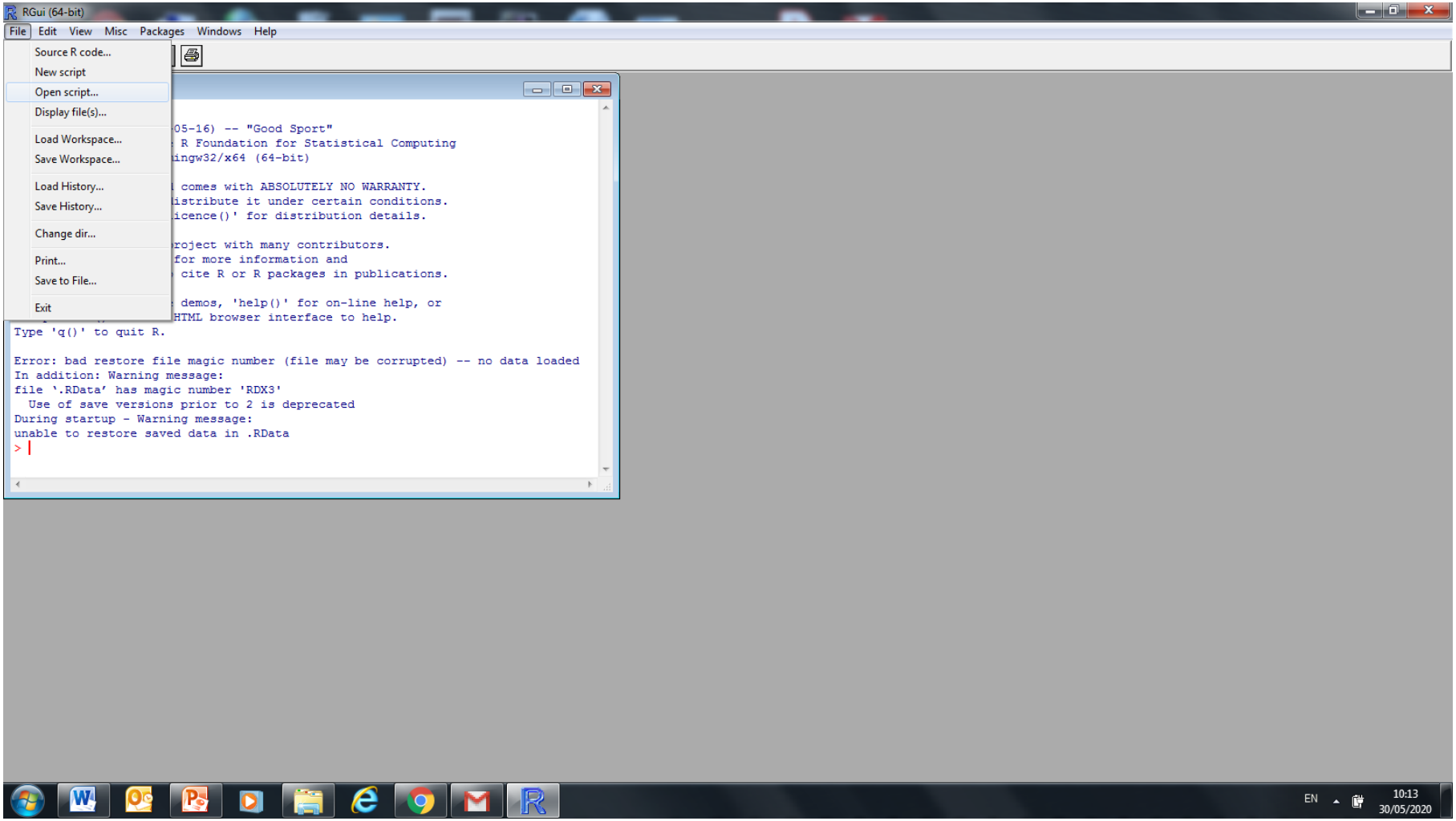
A (very) quick Start

…the first step…

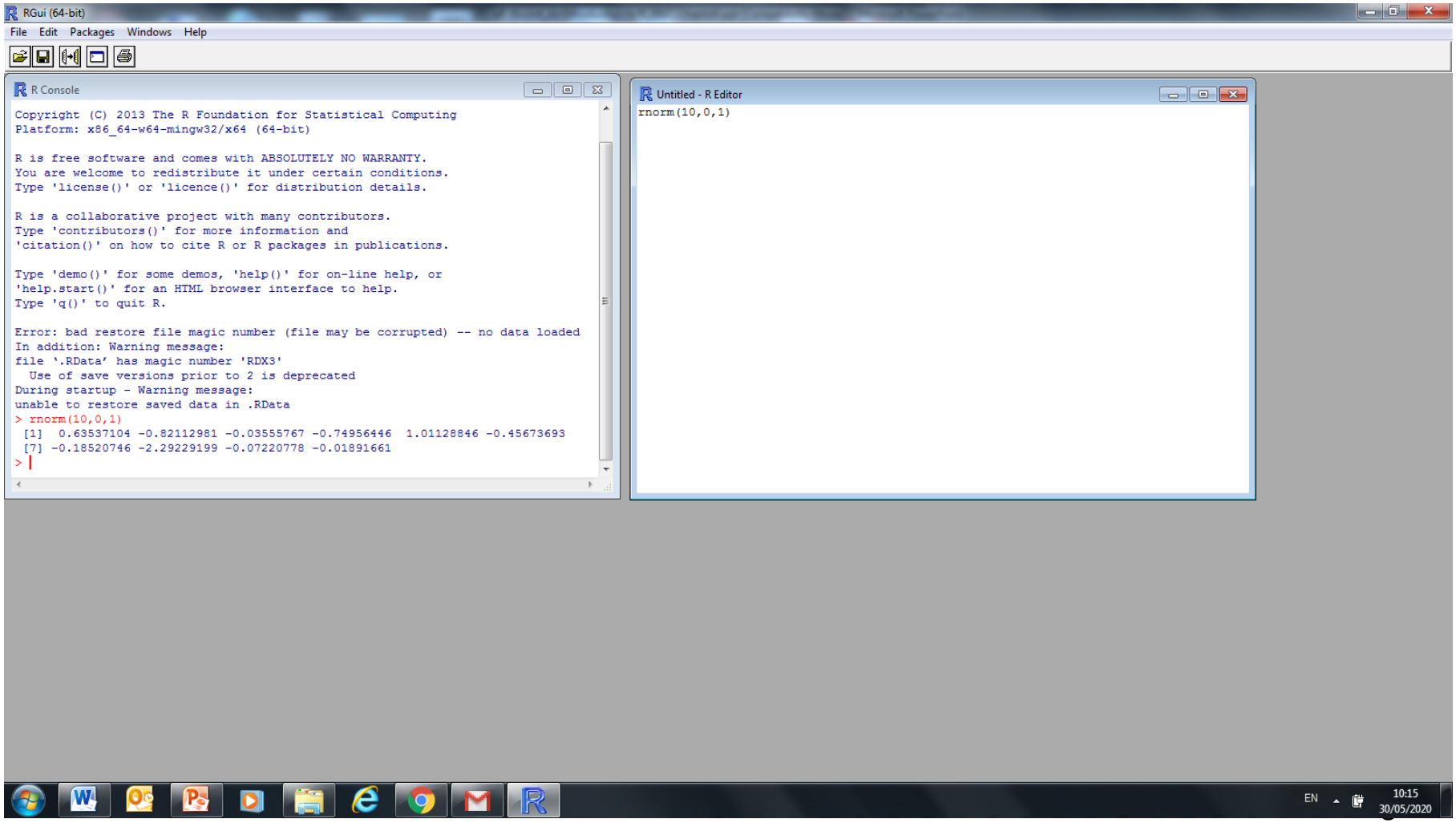
The R environment

- Open R.
- Open a new script window.

Open a script window



The script & the output

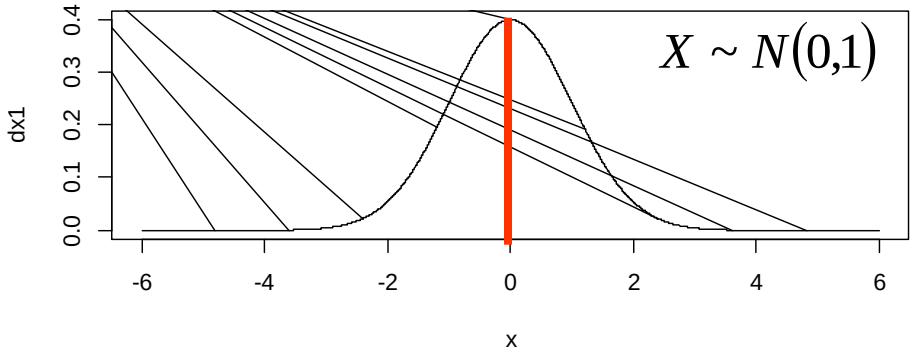


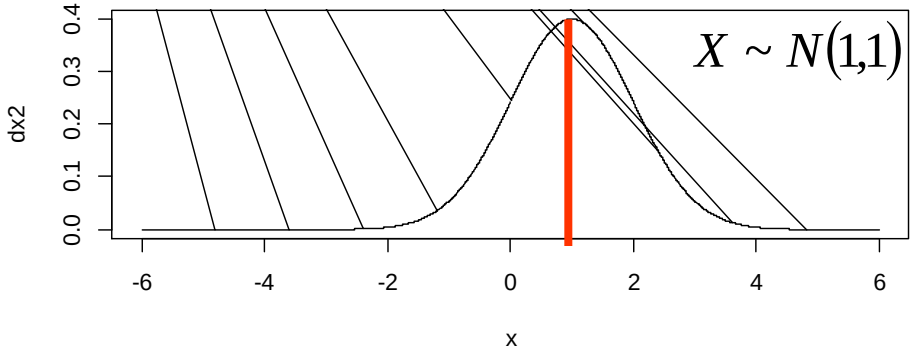
Example: a normal distribution

The normal distribution: location

Density function of a normal distribution

$$X \sim N\left(\mu, \sigma^2\right)$$

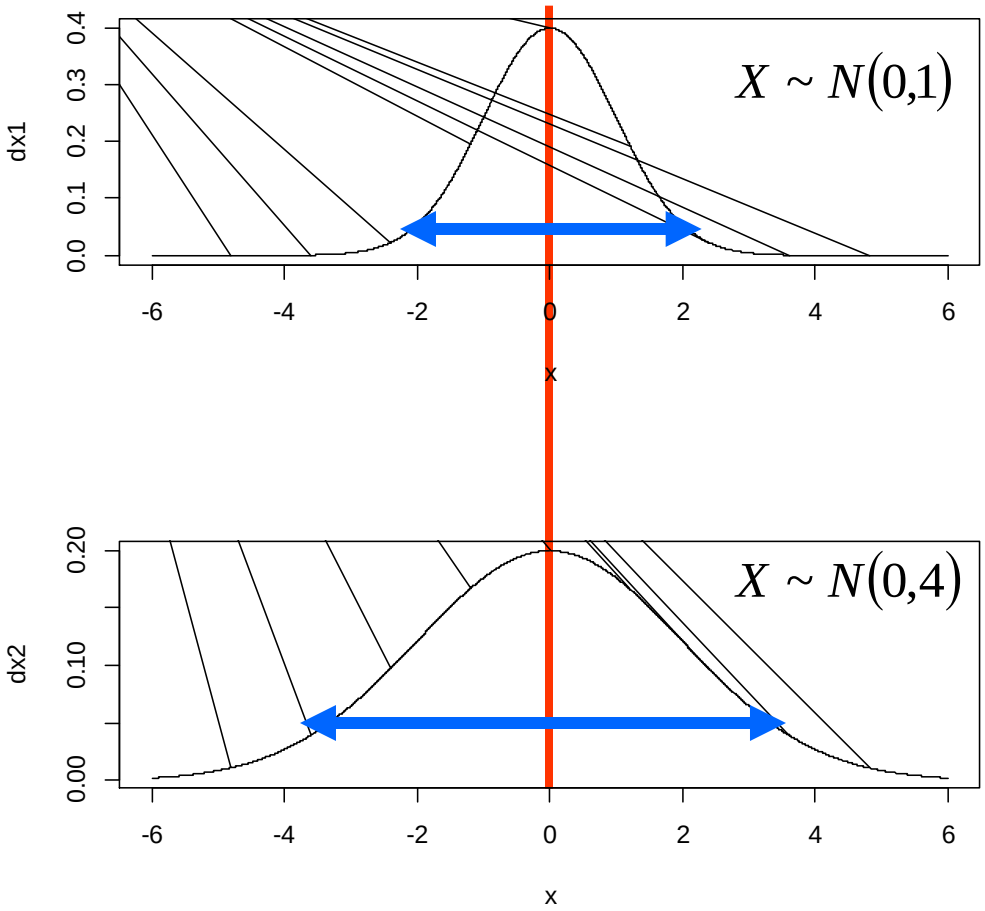




The normal distribution: variability

Density function of a normal distribution

$$X \sim N(\mu, \sigma^2)$$



R functions

```
function(data)
```

A procedure that was programed in R that uses data to produce output.

```
> var(x)
```

The r function

data

Calculate the sample variance.

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

Random sample from a normal distribution in R

Draw a random sample of size 100 from a normal distribution with mean – and variance 1

$$X \sim N(\mu, \sigma^2)$$

In R

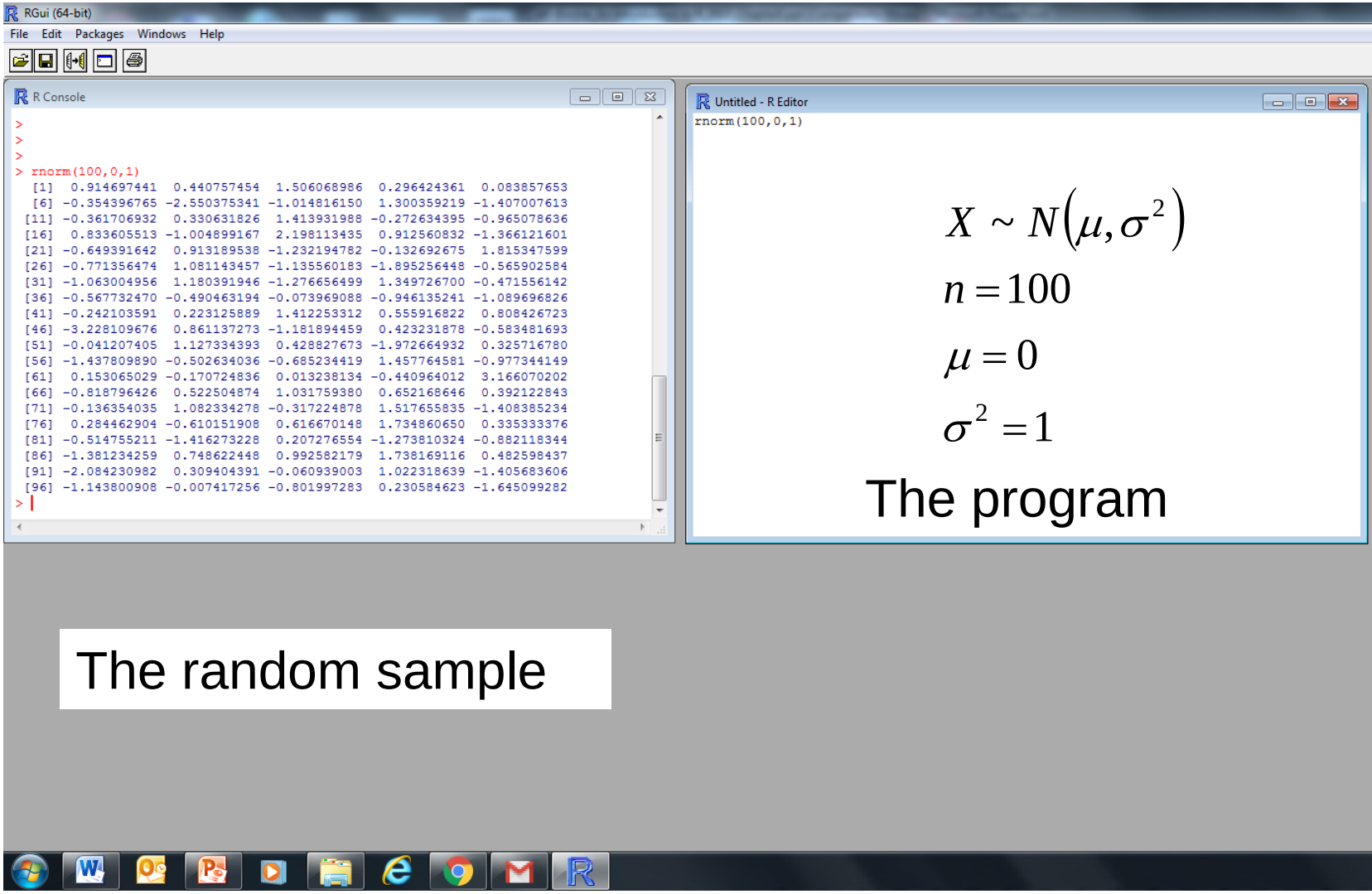
```
rnorm(sample size, mean,standard deviation)
```

$$X \sim N(0,1)$$

```
rnorm(100,0,1)
```

A function in R that draw a sample from $N(\mu, \sigma^2)$

The script & the output



The random sample

Random sample from a normal distribution in R

Draw a random sample of size **100** from a normal distribution with mean 0 and variance 1

$$X \sim N\left(\mu, \sigma^2\right) \Rightarrow \; X \sim N(0,1)$$

```
> rnorm(100,0,1)
  [1] -0.173911348 -0.463196096 -1.084838332  2.373958677 -1.685884982
  [6] -1.952672126 -0.055601310 -0.241913096 -0.999586206  0.308335895
 [11]  0.556993818  2.337451275  0.778734465 -0.501354458  0.004525392
 [16] -1.468709822  0.109901143  0.109103689  0.662434110 -0.177097648
 [21] -1.442033566  0.615239368  0.254080126  1.152977602 -0.089559002
 [26]  0.065022482  0.300405204 -0.190196930 -0.244365328  0.886735849
 [31] -0.667671228 -1.009209277  0.388362272 -0.041883373  0.750480061
 [36] -2.103109677 -1.515839684 -0.477250540 -0.344581482  0.072570862
 [41] -0.364485234 -0.920898769  1.148778190  1.092225688 -0.832389361
 [46] -1.914844153 -0.384265110  0.528078353  1.319149374  0.226817654
 [51] -0.605867376 -0.658048328  0.086126314  0.711404951  1.190303122
 [56]  2.499314086  2.201924724  0.591527333 -0.733622099 -0.656031690
 [61] -0.194759316  0.864421699  0.813854743 -0.628803589  0.362077258
 [66]  0.312250497  1.451227963  1.107136623  0.680487861  1.585879056
 [71] -0.249983835 -1.436293634 -0.470710524 -2.330088808  0.265551343
 [76] -0.847238216 -1.199413581 -1.866542460  0.826973063 -0.592073631
 [81] -1.751735134  0.077115620 -0.306869702  0.120083596 -0.303521155
 [86] -0.644268518  0.295067198  2.004409939  0.310290927  0.221898330
 [91] -1.450606907 -1.264043444 -0.257282348  0.078120141 -0.902925645
 [96]  0.499980835 -0.596173525 -1.085097601 -0.773094391  0.693319162
```

100 observations

Creating an R object

```
> x<-rnorm(100,0,1)
```

An R object contains the results

```
> x
 [1] -1.91083203  1.04955497 -2.40884482  0.33493954  1.45434660 -2.42198672
 [7]  0.44232862 -0.73804911 -0.36354587  0.39064194 -0.31993512 -1.30809569
[13]  0.11409195  0.43549125 -0.29501115  0.29197212  0.50983934 -0.80452037
[19] -0.61008244  1.80780477  1.31535974 -1.33155401  0.29044725 -0.63380504

[85]  1.03861350  0.89381884  0.86323215 -0.24199953  1.64380126  0.45445204
[91]  1.90708641  0.34088349 -0.25727644 -0.26498359  0.80095645  1.42711451
[97]  1.27998167 -0.54106317 -1.29443674  0.36046722
```

Print the R object

Summary statistics

```
 > mean(x)
[1] 0.02149641
```

Sample mean

```
> var(x)
[1] 1.061159
```

Sample variance

A **function in R** that calculate the mean:

```
mean(my sample)
```

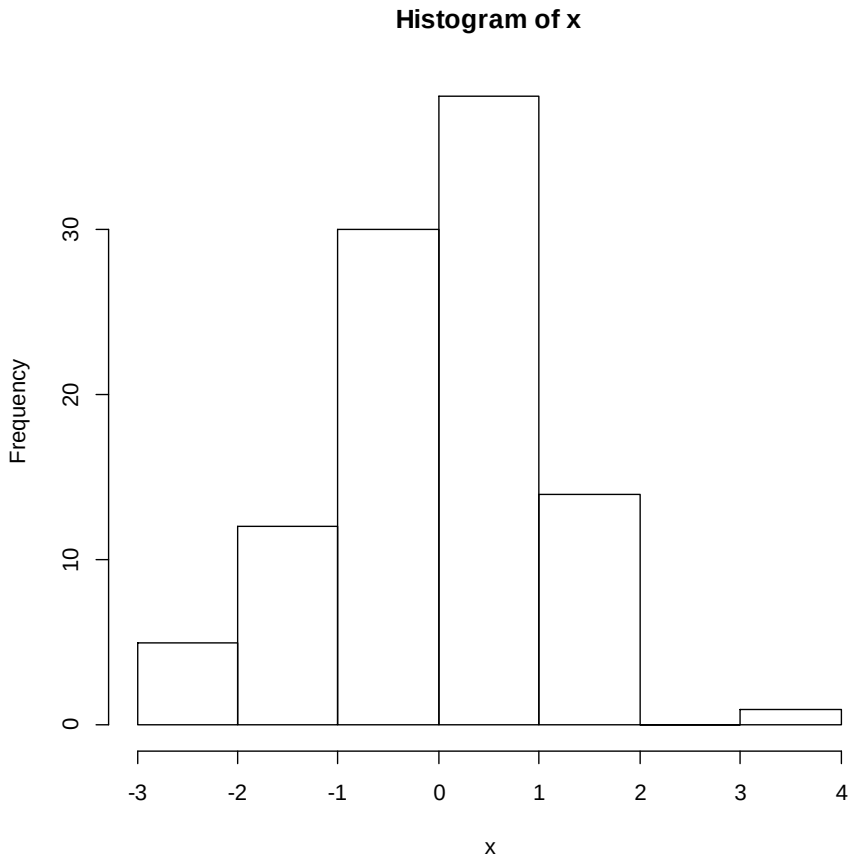
A **function in R** that calculate the variance:

```
var(my sample)
```

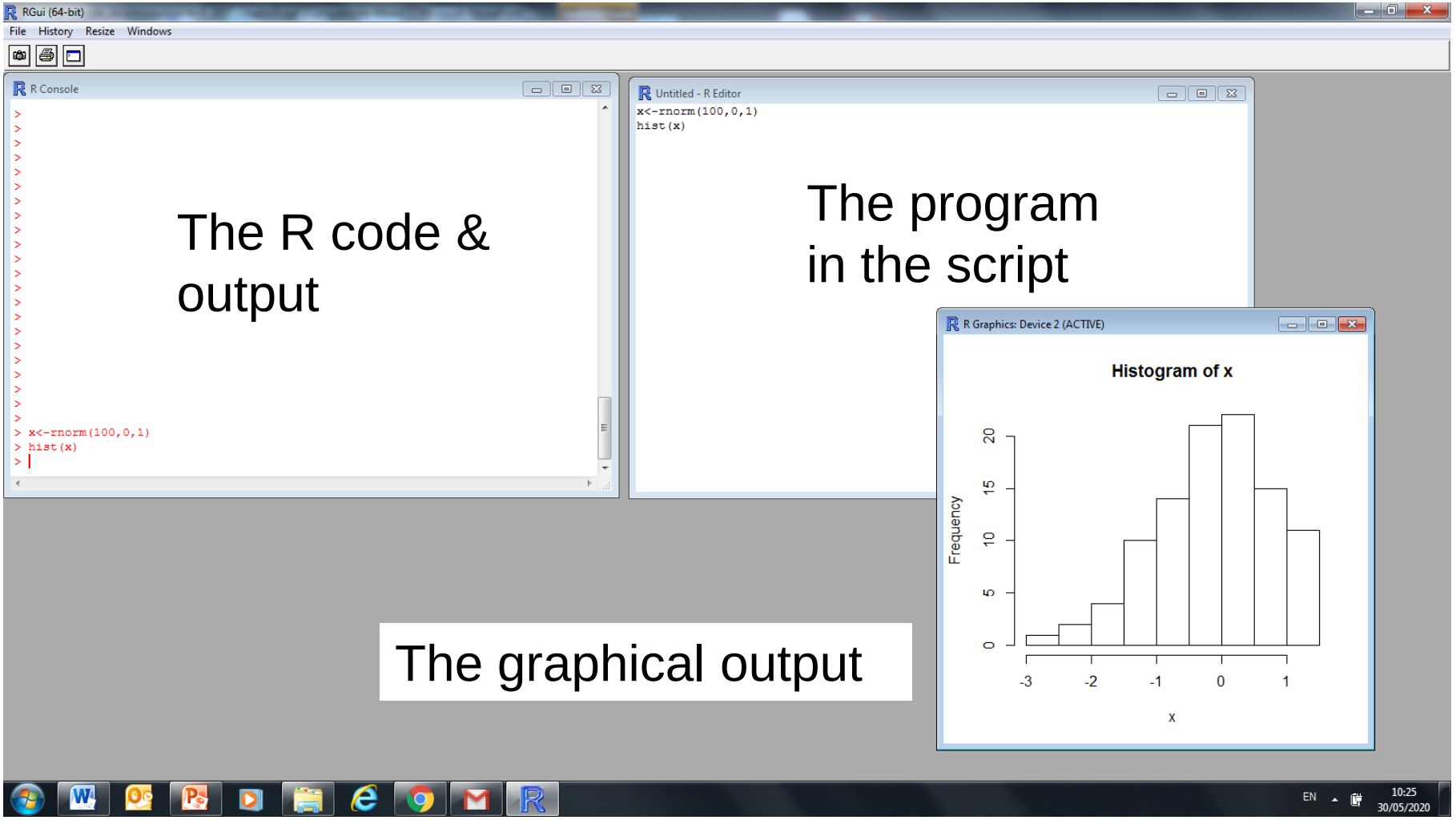
Histogram of the sample

```
> hist(x)
```

A function in R that produces a histogram



Histogram of the sample



The R code & output

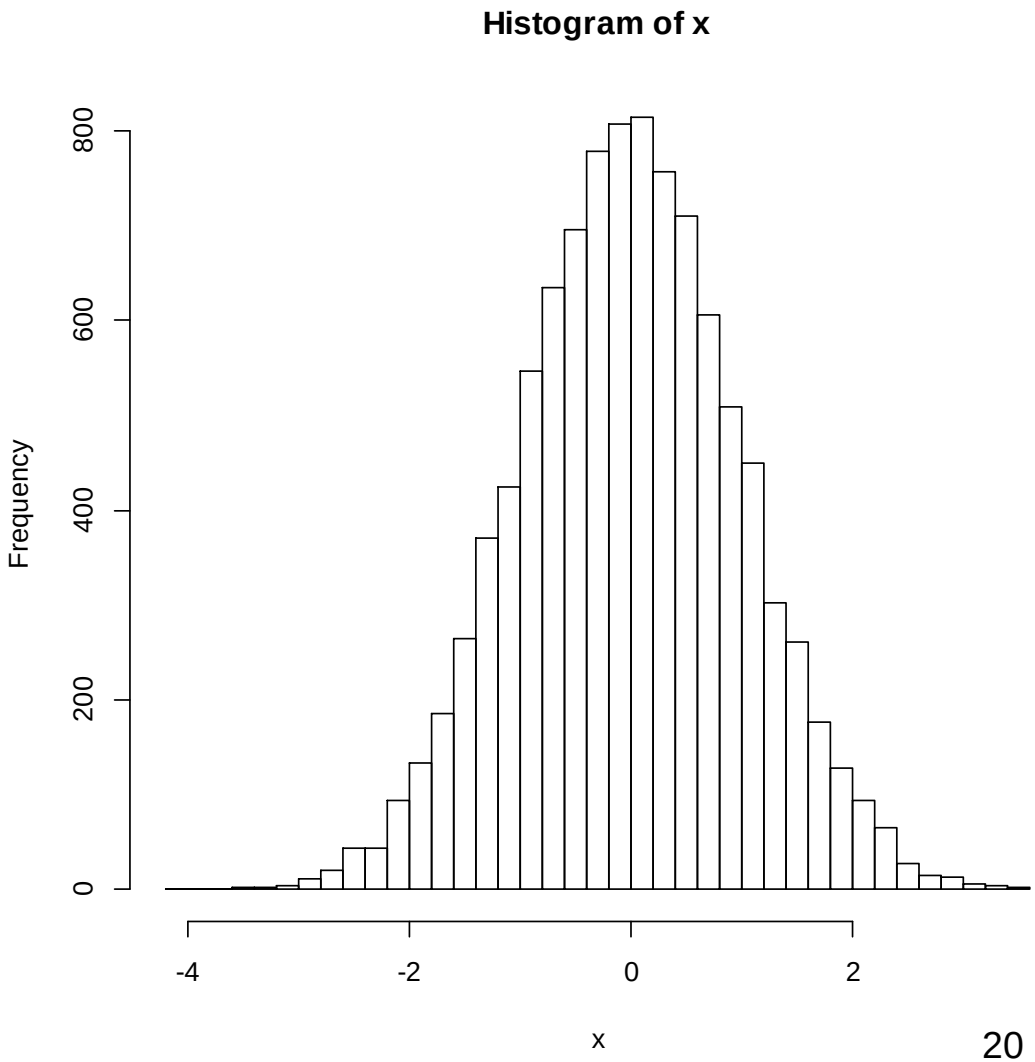
The program in the script

The graphical output

Histogram of the sample

```
> x<-rnorm(10000,0,1)
> mean(x)
[1] -0.01259969
> var(x)
[1] 0.9871957
> hist(x,nclass=50)
```

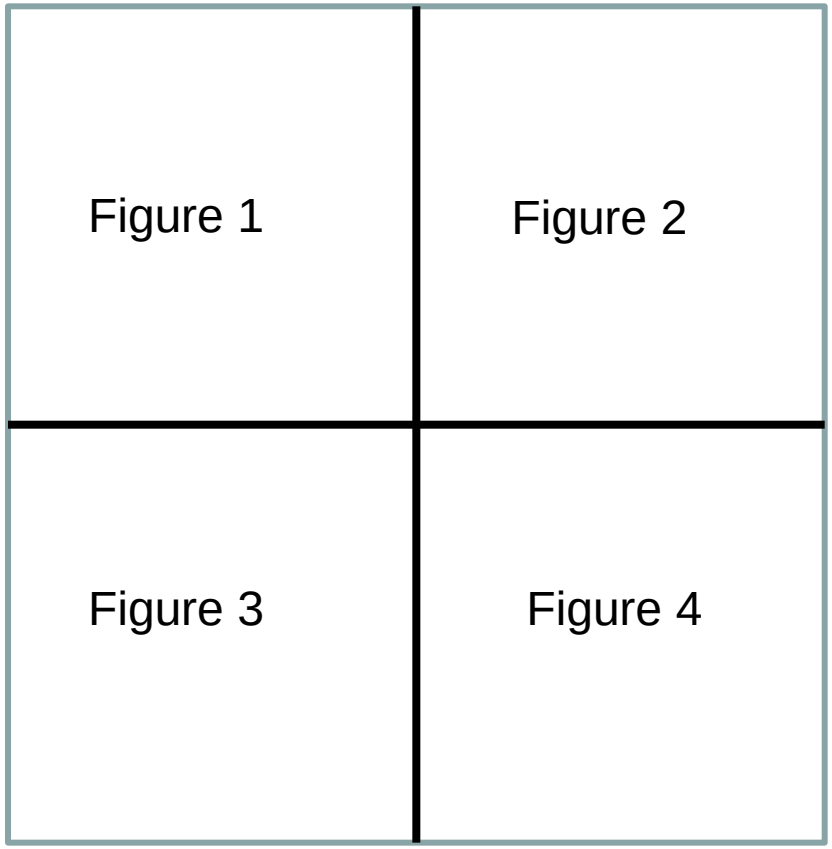
A function in R that produces a histogram



Controlling the graphical output

```
> par(mfrow=c(2,2))
```

Split the graphical window to a panel with 2 rows and 2 columns.



In general:

```
> par(mfrow=c(n,m))
```

Number of rows

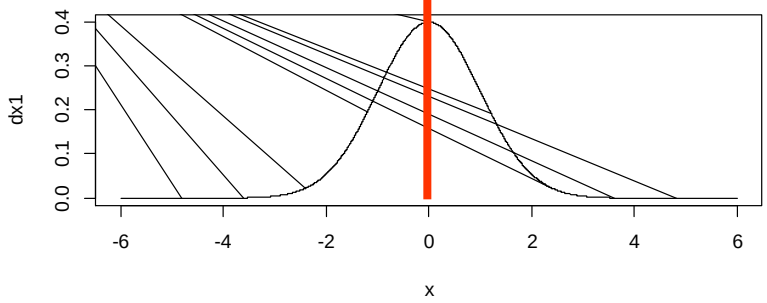
Number of columns

The normal distribution: location

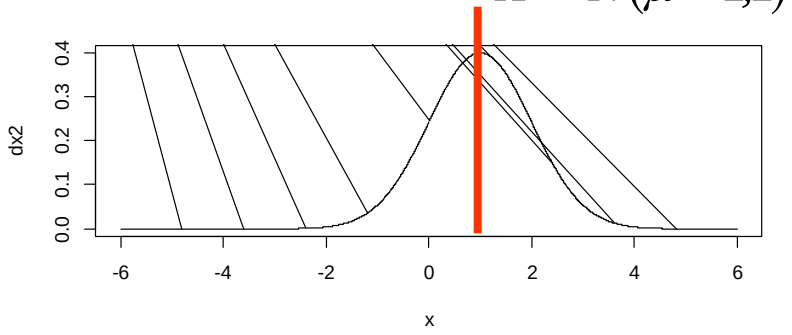
$$X \sim N(\mu, \sigma^2)$$

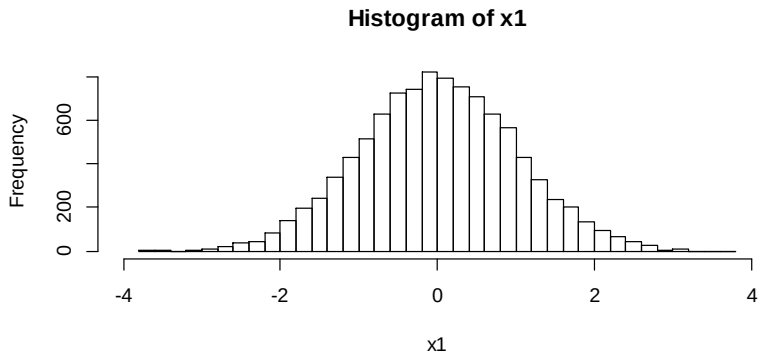
```
> x1<-rnorm(10000,0,1)
> x2<-rnorm(10000,1,1)
> par(mfrow=c(2,1))
> hist(x1,nclass=50,xlim=c(-4,4))
> hist(x2,nclass=50,xlim=c(-4,4))
```

$X \sim N(\mu = 0,1)$



$X \sim N(\mu = 1,1)$

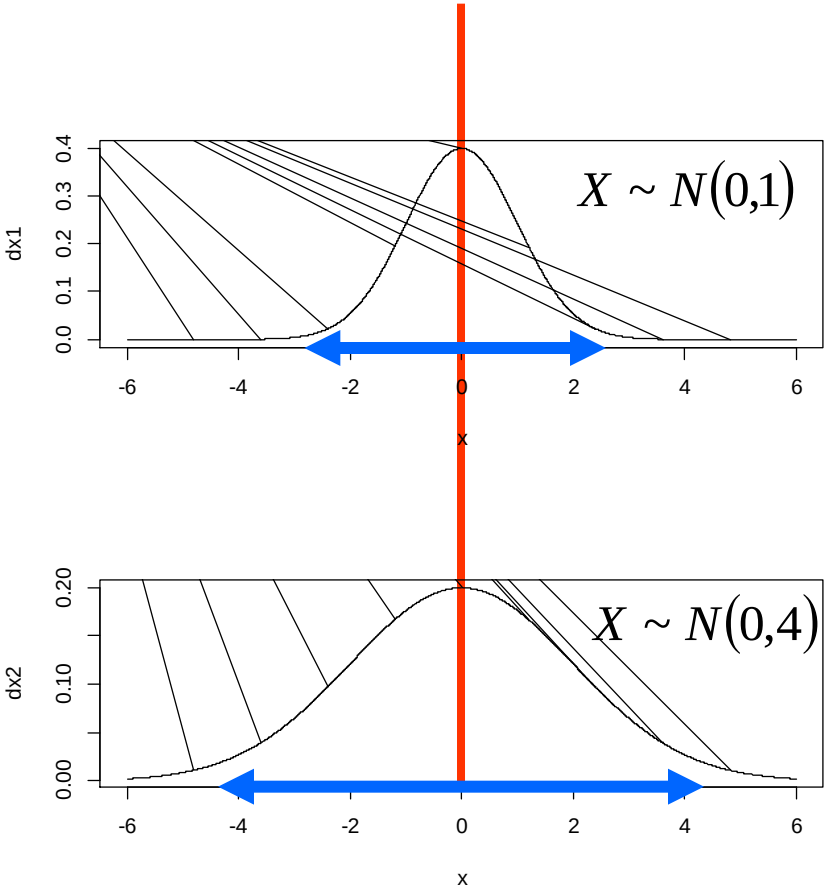


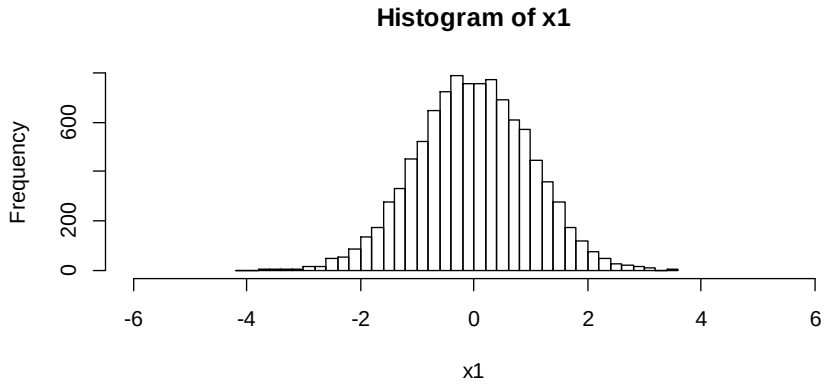


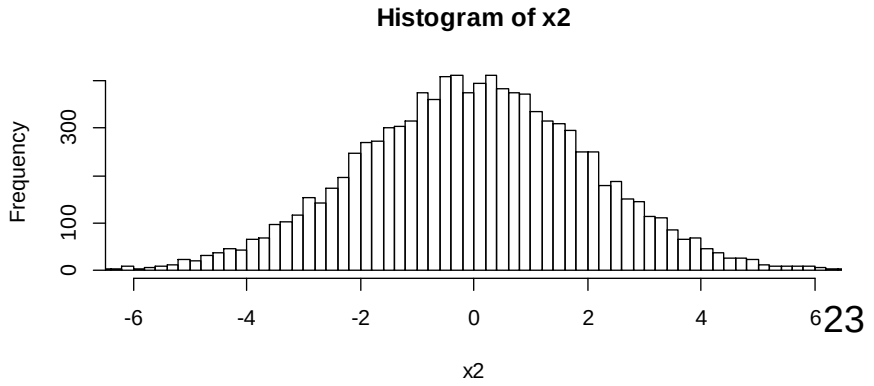
The normal distribution: variability

$$X \sim N(\mu, \sigma^2)$$

```
> x1<-rnorm(10000,0,1)
> x2<-rnorm(10000,0,2)
> par(mfrow=c(2,1))
> hist(x1,nclass=50,xlim=c(-6,6))
> hist(x2,nclass=100,xlim=c(-6,6))
```







Example: working with data

The cars Data set in R

1. Write **cars** in the script window.
2. Submit

```
> cars
```

```
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
.      .    .
.      .    .
48    24   93
49    24  120
50    25   85
```

```
> help(cars)
>
```

Speed and Stopping Distances of Cars
Description

The data give the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s.

[,1]	speed	numeric	Speed (mph)
[,2]	dist	numeric	Stopping distance (ft)

The cars Data set in R: the $ sign

```
> speed
Error: object 'speed' not found
>
> cars$speed
 [1]  4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15
[26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 20 22 23 24 24 24 24
25
>
```

cars$speed: the variable speed in the object cars

The cars Data set in R: creating a new object

```
> cars[,1]
 [1]  4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15
[26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 20 22 23 24 24 24 24
25

> x=cars[,1]

> print(x)
 [1]  4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15
[26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 20 22 23 24 24 24 24
25
>
```

Basic plot and descriptive statistics

- What is the average speed of the cars ?
- What is the variance of the cars' speed ?
- What is the min. (max.) speed ?

- What is the association between speed and stopping distance ?

Descriptive statistics

```
> mean(cars$speed)
[1] 15.4

> max(cars$speed)
[1] 25

> min(cars$speed)
[1] 4
```

The variable speed in the dataset cars

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

attach(data)

```
> attach(cars)
```

Tells R to work with the dataset cars.

```
> mean(speed)
[1] 15.4
> max(speed)
[1] 25
> min(speed)
[1] 4
```

We can work with the variables by using thier names.

```
> detach(cars)
```

Stop using the dataset cars.

Descriptive statistics

Correlation between the variables speed and stopping distance.

```
> cor(cars)
          speed      dist
speed 1.0000000 0.8068949
dist  0.8068949 1.0000000
```

R functions

```
function(data)
```

A procedure that was programed in R that uses data to produce output.

```
> var(cars$speed)
[1] 27.95918
```

Calculate the variance.

```
> var(cars$speed)
```

function data

R functions

```
> print(cars)
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
45    23   54
46    24   70
47    24   92
48    24   93
49    24  120
50    25   85
```

```
> cor(cars)
           speed      dist
speed 1.0000000 0.8068949
dist  0.8068949 1.0000000
```

R functions

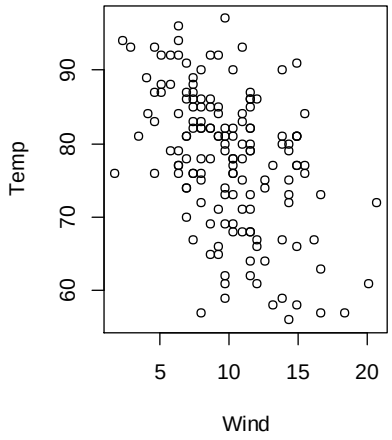
- Can be used for
  - Data analysis: descriptive statistics, testing, modeling, etc.
  - Data manipulation: selection of cases, variables…
  - Data management: reading and writing datasets into/from R.
  - Visualization: plots for the data.
  - …..

Discussion

- R Objects: data frame.
- R functions.
- $.

Practical session

- The airquality is a dataset available in R.
- How many variables there are in the data ?
- Define an R object which contain the information about the wind speed.
- Calculate the mean, and variance for the wind speed.



Working with R function: Two-sample t-test

The sleep data in R

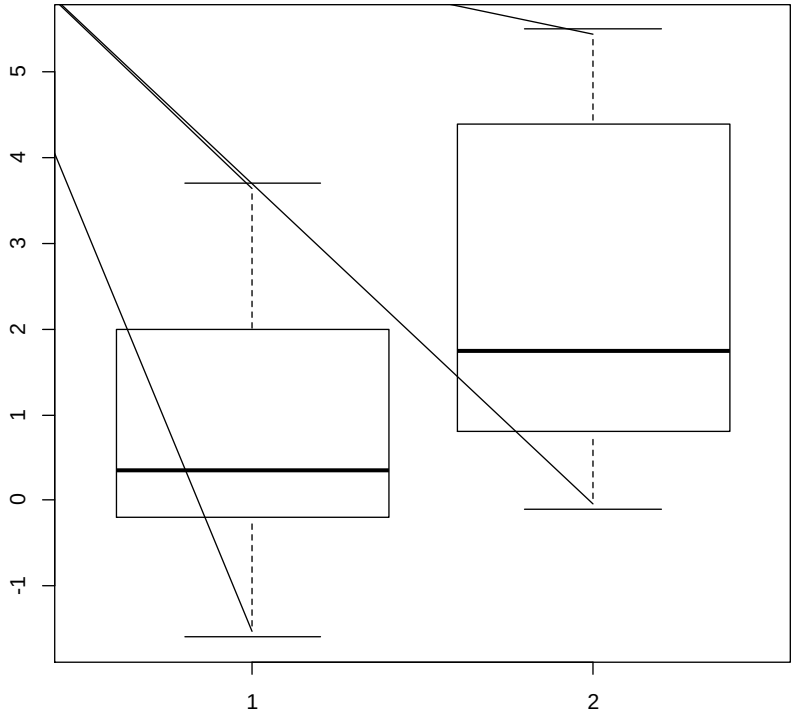
```
> help(sleep)
```

Data which show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.

extra	numeric	increase in hours of sleep
group	factor	drug given
ID	factor	patient ID

```
> sleep
   extra group ID
1    0.7     1  1
2   -1.6     1  2
3   -0.2     1  3
4   -1.2     1  4
5   -0.1     1  5
.      .     .  .
14   0.1     2  4
15  -0.1     2  5
16   4.4     2  6
17   5.5     2  7
18   1.6     2  8
19   4.6     2  9
20   3.4     2 10
```

Two samples t-test




```
> extra=sleep$extra
> group=sleep$group
> boxplot(split(extra,group))
```

The aim of the analysis:

Test for a difference between the two soporific drugs

Two samples t-test

```
> t.test(extra~group,var.equal=TRUE)

        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

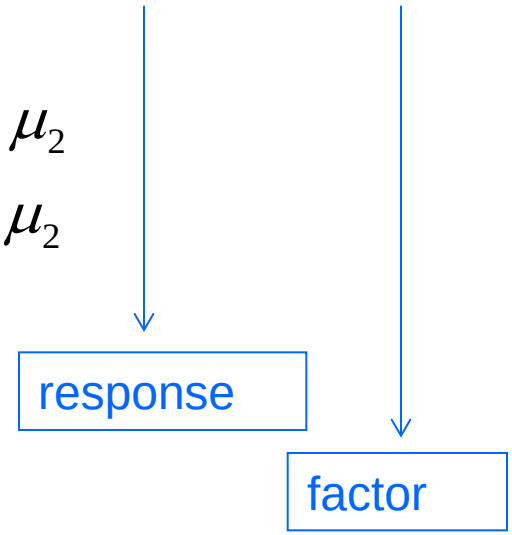
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Two samples t-test

```
> t.test(extra~group,var.equal=TRUE)
```

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$



R object for the output

```
> t.obj=t.test(extra~group,var.equal=TRUE)
> summary(t.obj)
```

```
             Length Class  Mode
statistic    1      -none- numeric
parameter    1      -none- numeric
p.value      1      -none- numeric
conf.int     2      -none- numeric
estimate     2      -none- numeric
null.value   1      -none- numeric
alternative  1      -none- character
method       1      -none- character
data.name    1      -none- character
```

t.obj:

R object contains the output of the analysis

R object for the output

```
> print(t.obj)

        Two Sample t-test

data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -3.363874  0.203874
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

```
> t.obj$p.value
[1] 0.07918671
> t.obj$statistic
        t
-1.860813
>
```

Objects in the “output”

Discussion

- R Objects: output of the analysis.
- R functions: t.test
- $.

Practical session

- The ToothGrowth is a dataset available in R.
- Use `help(ToothGrowth)` for more details.
- The response variable is the Tooth length.
- Test if the Supplement type has an effect on the tooth length.

t.test(response ~ group, data = …)

Basic plots

The faithful data in R

```
> help(faithful)
```

Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

```
> Faithful
```

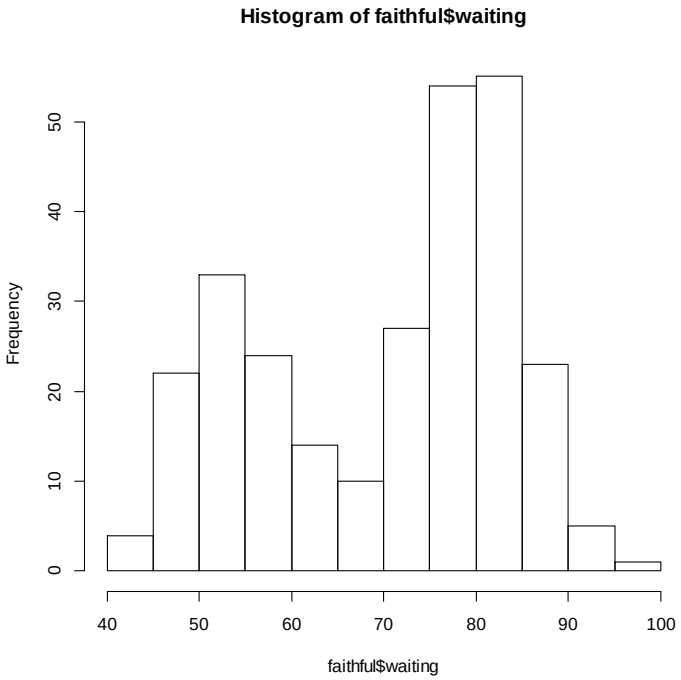
	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51

The faithful data in R

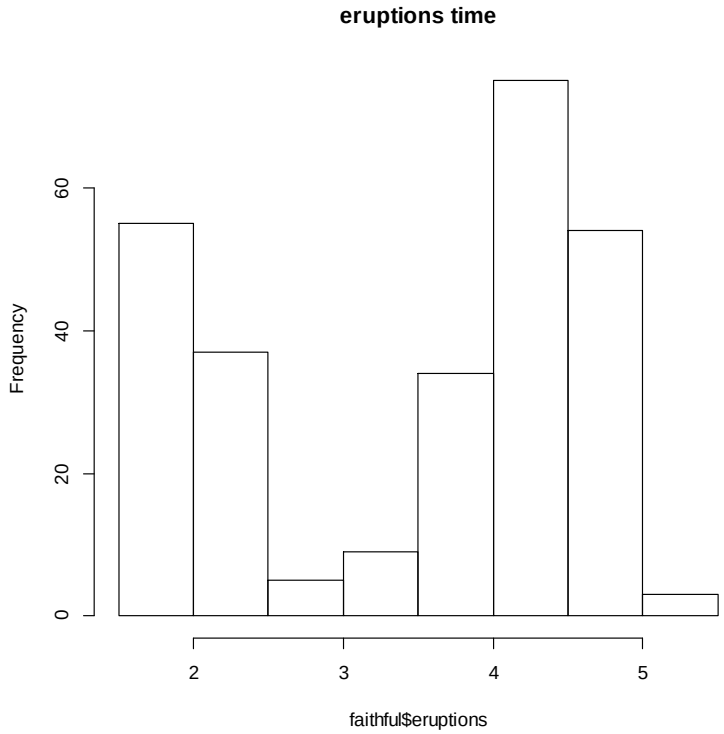
```
> faithful$eruption
  [1] 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950
4.350 1.833 3.917 ,...., .4.750 4.117 2.150 4.417 1.817 4.467

> mean(faithful$eruption)
[1] 3.487783

aithful$eruption
> mean(x)
[1] 3.487783
> median(x)
[1] 4
> range(x)
[1] 1.6 5.1
> min(x)
[1] 1.6
> max(x)
[1] 5.1
```

Basic plot

```
> hist(faithful$waiting)
```

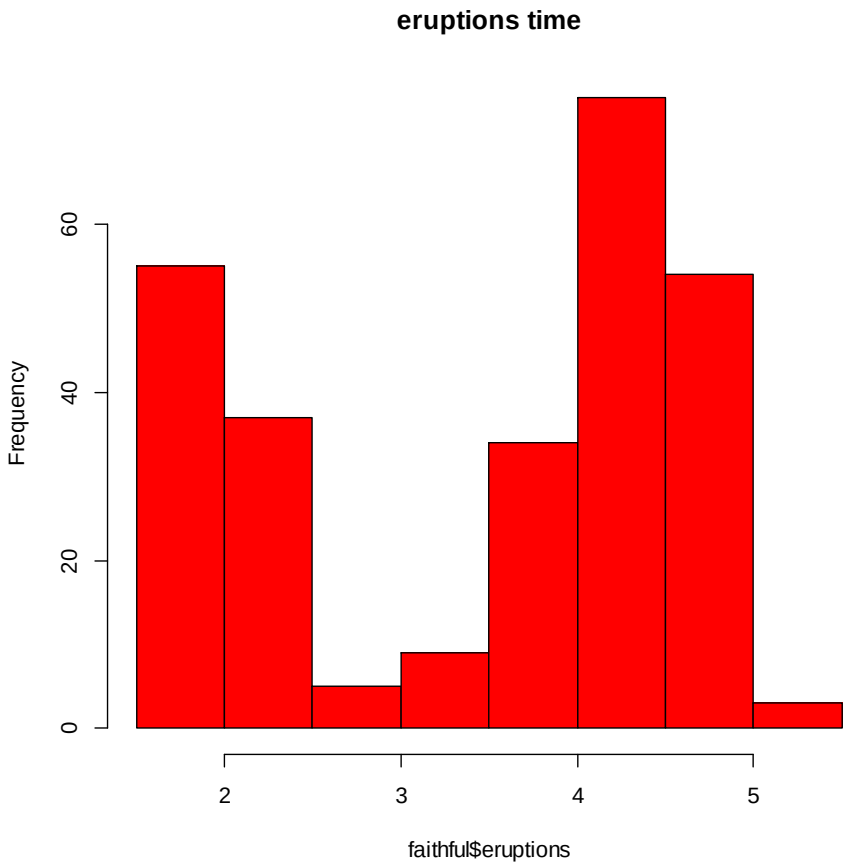



```
>hist(faithful$eruptions,
      main="eruptions time")
```

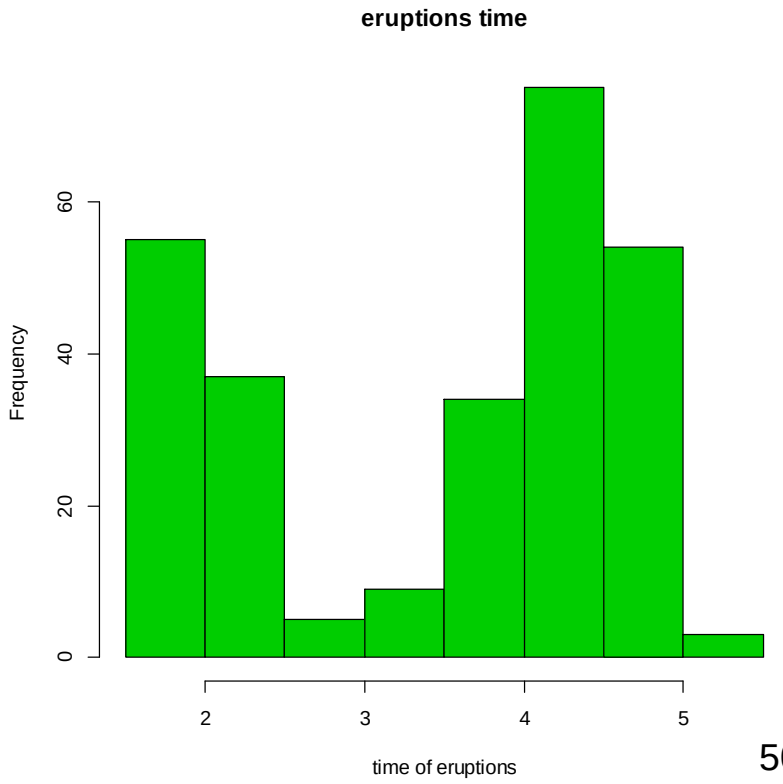


Basic plot

```
>hist(faithful$eruptions,
 main="eruptions time",
 col=2)
```

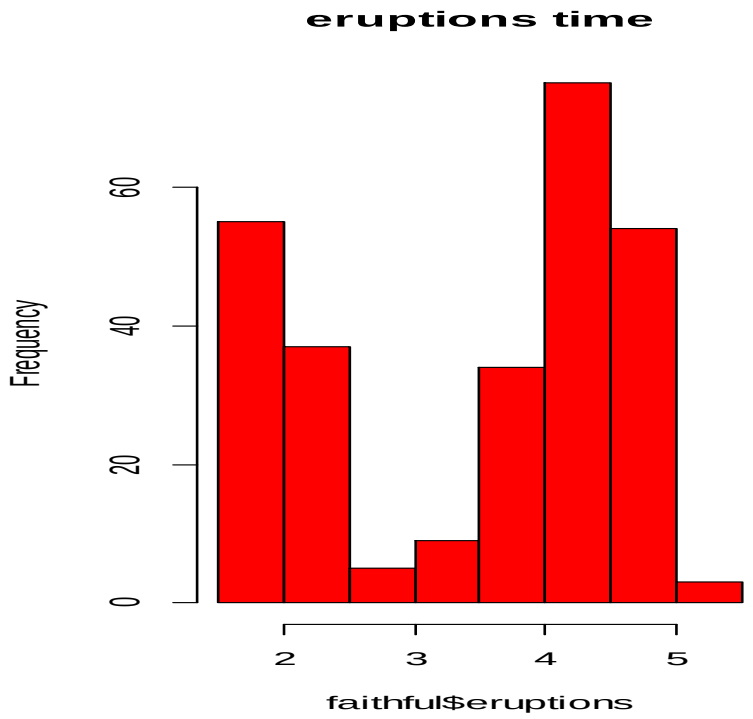


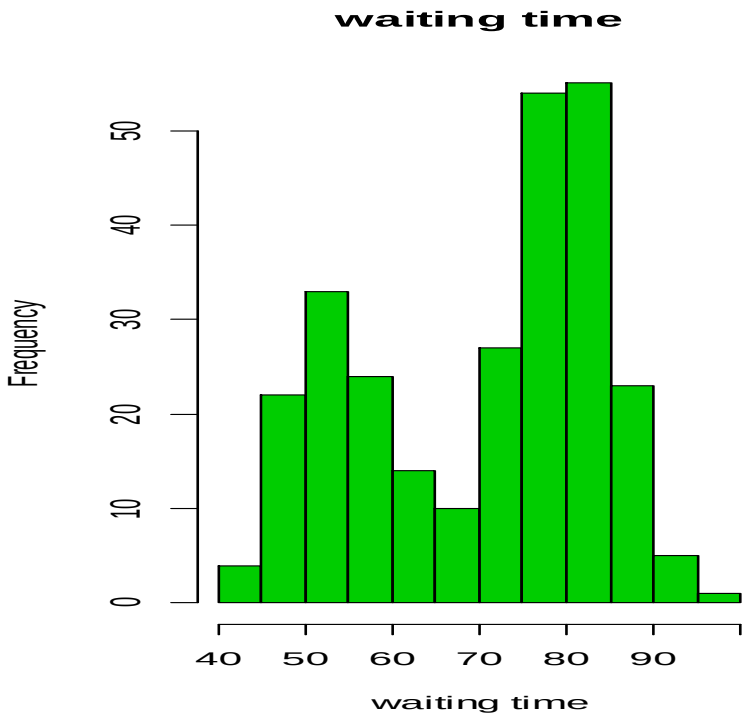

```
>hist(faithful$eruptions,
 main="eruptions time",
 col=3,
 xlab="time of eruptions")
```



Basic plot

```
> mfrow=c(1,2))
> hist(faithful$eruptions,main="eruptions time",col=2)
> hist(faithful$waiting,main="waiting
  time",col=3,xlab="waiting time")
```





Practical session

- Use the ToothGrowth data.
- Produce an histogram for the tooth length with the following structure.