STARTING WITH R

Karine Chalvet-Monfray

Faculty of Veterinary Science – Mahidol University – 22-24 July 2015

Pedagogical aims

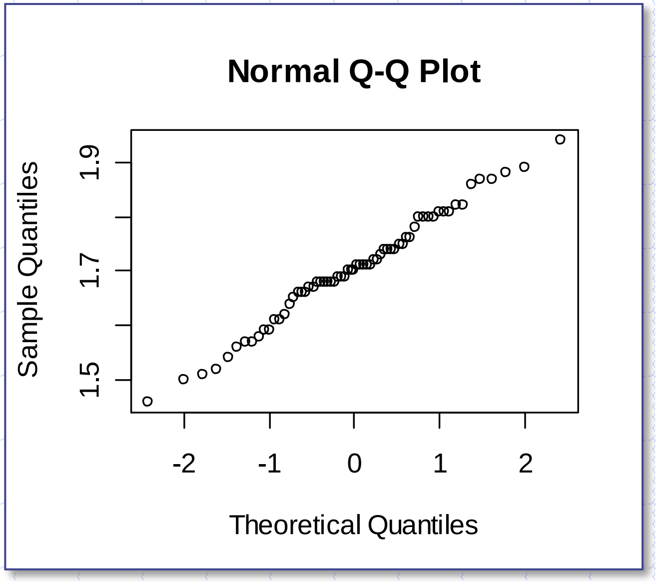
- Know how import data in
- Know how do basic representation
- Know how do simple statistic tests
- Know how do calculus of epidemiology
- Know how learn more on

Color code concerning R code

function()

d1 is a data frame
d1 has 2 column=2 variables

```
2+2
[1]  4
```



Don't forget to close before asking a new code

Big sample (N>50)

Color code concerning variables

V1
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

Quantitative variable

V1

V1
+
+
-
+
±
-
-
+

Qualitative variable

V1	V2
	+
	+
	-
	+
	±
	-
	-
	+

2 qualitative variables

V1	V2
	9.0
	0.8
	5.9
	1.6
	7.6
	9.9
	0.1
	3.1

1 qualitative variable
+ 1 quantitative variable

V1	V2
7.4	9.0
1.8	0.8
5.4	5.9
2.0	1.6
0.6	7.6
5.3	9.9
0.7	0.1
4.8	3.1

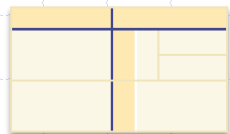
2 quantitative variables

Date
22/06/1965
20/03/1968
30/08/1965
27/09/1965
22/07/1966
05/05/1964
16/01/1964
11/02/1968

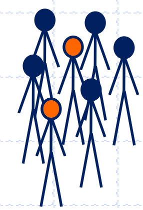
Qualitative variable
-> text

Color code concerning plan

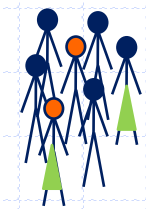
Chap. 1



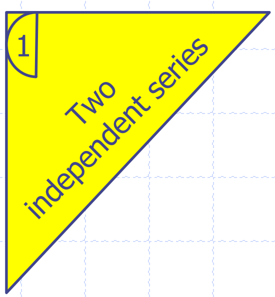
Chap. 4



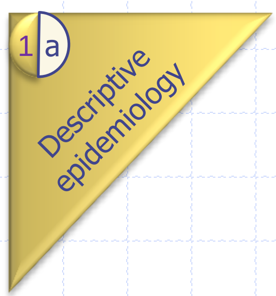
Part 5.1



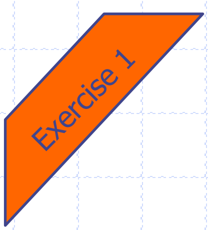
Part 5.2



Basic card



Perfecting card



Basic exercise

Perfecting exercise

What is R ?

- Origin

  -> R = Language and Software created by Ross Ihaka and Robert Gentleman from the language S (S+: Software) for statistics

- Interests

  -> worldly used with a lot of help

  -> free and open source

  -> build by good statisticians and numericians

  -> flexibility because using interpreted language

  -> increasing tools and field (mathematics, GIS,…)

- disadvantages

  -> less convivial because using language, need help a the beginning

Plan

1. Principle of R language
2. Importing data
3. Simple descriptive statistics -> *Graphics*
4. Simple analytical statistics -> *tests of comparison mean and frequencies*
5. Specific tools for epidemiology
6. Specific tools for clinical study

Plan

1. Principle of R language
2. Importing data
3. Simple descriptive statistics -> *Graphics*
4. Simple analytical statistics -> *tests of comparison mean and frequencies*
5. Specific tools for epidemiology
6. Specific tools for clinical study

Principle of R language

1. R can calculate as a calculator

```
2+2
[1] 4
```

2. R use functions with arguments [to be specified or optional (set by default)]

function()

```
read.table("cohort.txt",header=TRUE)
```

optional argument

argument to be specified

3. The result can be assigned

```
d<-read.table("cohort.txt",header=TRUE)
```

```
read.table("cohort.txt",header=TRUE)->d
```

Exercise 1

Practical on principle of R language

- Calculate:

$\sqrt{4}$

```
4^(0.5)
[1] 2
```

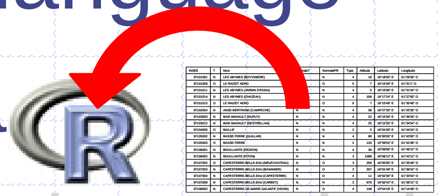
$\ln(\pi \times (4 \times 10^2 + 543))$

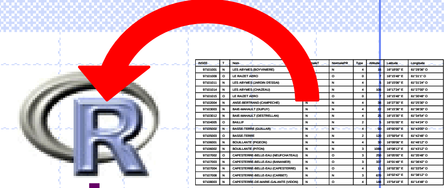
```
log(pi*(4e2+543))
[1] 7.993796
```

$e^{2.36-1.96*0.49}$

```
exp(2.36-1.96*0.49)
[1] 4.053578
```

Plan

1. Principle of R language
2. Importing data 
3. Simple descriptive statistics -> *Graphics*
4. Simple analytical statistics -> *tests of comparison mean and frequencies*
5. Specific tools for epidemiology
6. Specific tools for clinical study



Importing data: one method

1. In Excel save as txt format (tabulation as a separator) the data frame

   If there is empty cell write NA for Not Available

2. Read the table in R and assign the result

```
d<-read.table("cohort.txt",header=T)
```

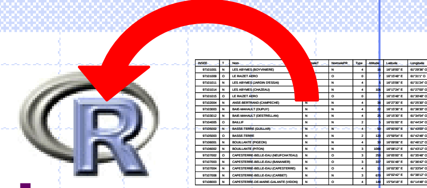
name of results = name of data frame

function which reads a table

“name of txt file”

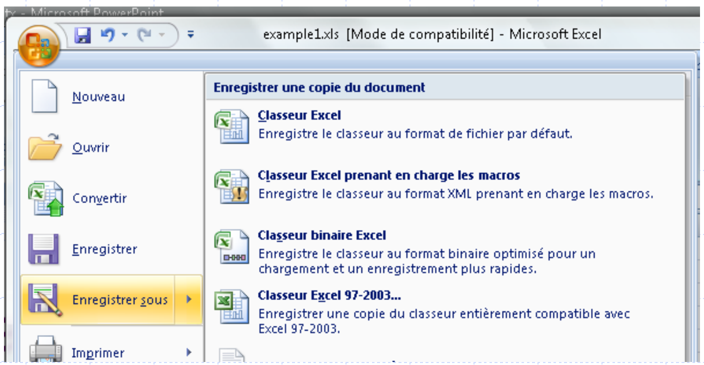
the table has header

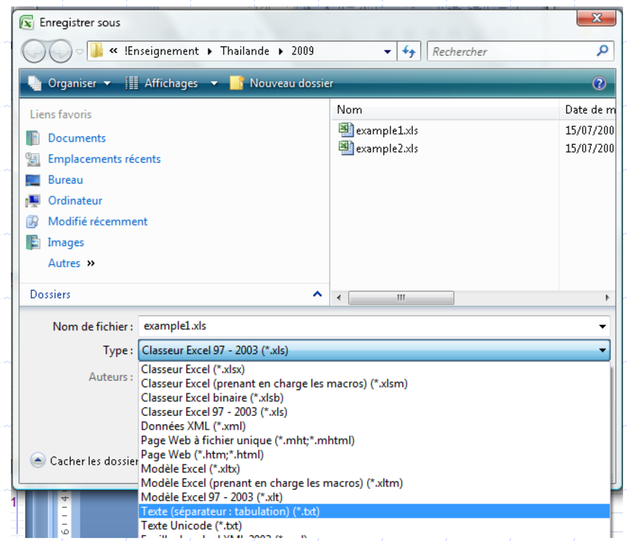
Exercise 2a



Practical on importing data

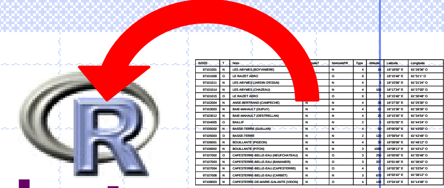
Using the file example1.xls, import the data in R with assigning the result to d1.






```
d1<-read.table("example1.txt",header=T)
```

Exercise 2b



Practical on importing data

Do it the same with example2.xls.

It doesn't't work. See data in Excel.

	A	B	C
1	origin	size body	dosage
2	A	1.5	295.7
3	A	1.64	50.5
4	A	1.83	136.6
5	A	1.57	107.1
6	A	1.73	329.9
7	A		418.7
8	A	1.66	209.9
9	A	1.73	140.3

	A	B	C
1	origin	size_body	dosage
2	A	1.5	295.7
3	A	1.64	50.5
4	A	1.83	136.6
5	A	1.57	107.1
6	A	1.73	329.9
7	A	NA	418.7
8	A	1.66	209.9
9	A	1.73	140.3
10	A	1.75	36.5

Importing data: quick visualization 1

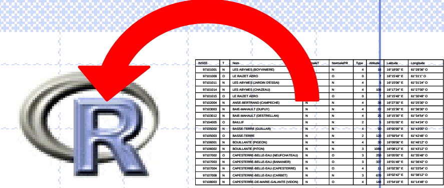
- Some R functions are very useful to visualize the data frame: `view()`

`view(d1)`

function which edits an object

the data frame

	origin	size
1	A	1.64
2	A	1.61
3	A	1.68
4	A	1.87
5	A	1.52
6	A	1.76
7	A	1.69
8	A	1.81
9	A	1.71
10	A	1.80
11	A	1.70



Importing data: quick visualization 2

- Some R functions are very useful to visualize the data frame: `str()`

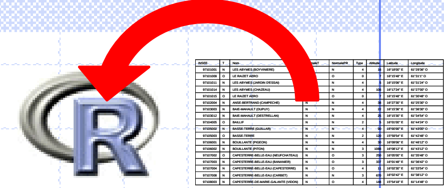
function which displays compactly the structure of an object

d1 is a data frame
d1 has 65 rows=65 observations
d1 has 2 column=2 variables

```
str(d1)
'data.frame':   65 obs. of  2 variables:
 $ origin: Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ size  : num  1.64 1.61 1.68 1.87 1.52 1.76 1.69 1.81 1.71 1.8 ...
```

The first column is named origin. It is a factor variable (qualitative) with 5 levels "A","B",... The first values are A,A,A... (1 is for the first level =A)

The second column is named size. It is a numerical variable (quantitative data). The first values are 1.64,1.61,1.68 ...



Importing data: quick visualization 3

Some R functions are very useful to visualize the data frame: `summary()`

function which produces result summaries

```
summary(d1)
origin       size
 A:15    Min.   :1.460
 B:14    1st Qu.:1.660
 C:13    Median:1.700
 D:12    Mean   :1.704
 E:11    3rd Qu.:1.760
         Max.   :1.940
 ...
```

The first variable named origin is qualitative variable. There is 15 observations with level A, 14 obs…

The second variable named size is quantitative variable. The minimum value is 1.46, The first quartile is 1.66…

Exercise 3

Practical on importing data

Make a quick visualization with the data of the example 2.

```
view(d2)
```

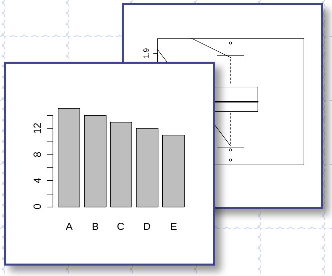
	origin	size_body	dosage	v
1	A	1.5	295.7	
2	A	1.64	50.5	
3	A	1.83	136.6	
4	A	1.57	107.1	
5	A	1.73	329.9	
6	A	NA	418.7	

```
summary(d2)
 origin    size_body          dosage
 A:15     Min.   :1.470   Min.   :   6.5
 B:14     1st Qu.:1.647   1st Qu.:  60.2
 C:13     Median :1.715   Median : 139.6
 D:13     Mean   :1.711   Mean   : 224.4
 E:10     3rd Qu.:1.765   3rd Qu.: 326.7
          Max.   :1.980   Max.   :1023.6
          NA's   :1.000
```

```
str(d2)
'data.frame':   65 obs. of  3 variables:
 $ origin   : Factor w/ 5 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ size_body: num  1.5 1.64 1.83 1.57 1.73 NA 1.66 1.73 1.75 1.71 ...
 $ dosage   : num  295.7  50.5 136.6 107.1 329.9 ...
```

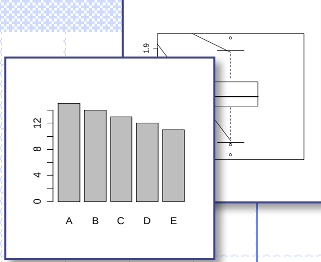
Plan

1. Principle of R language
2. Importing data
3. Simple descriptive statistics -> *Graphics*



4. Simple analytical statistics -> *tests of comparison mean and frequencies*
5. Specific tools for epidemiology
6. Specific tools for clinical study

Introduction for simple descriptive statistics



V1

Qualitative

V1

or

V1
+
+
-
+
±
-
-
+

…

1 a b

Quantitative

V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

or

V2
1
5
3
4
4
5
2
3

…

2 a b c

Normal distribution ?

3 a b

1 a

Simple descriptive statistics for qualitative variable:

`table()` and `barplot()`

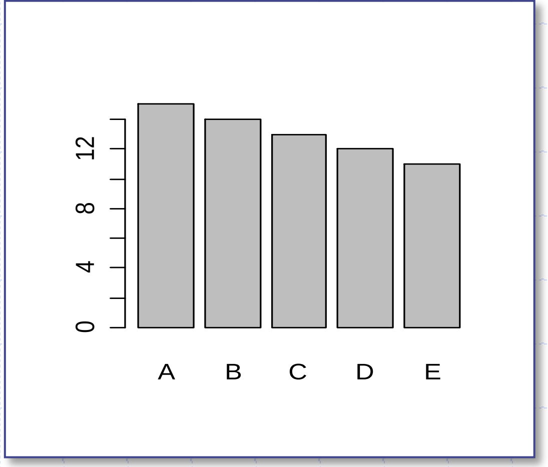
function which builds a contingency table

a qualitative variable of the data frame

function which creates a bar plot

```
table(d1$origin)

 A  B  C  D  E
15 14 13 12 11
```

```
barplot(table(d1$origin))
```



V1

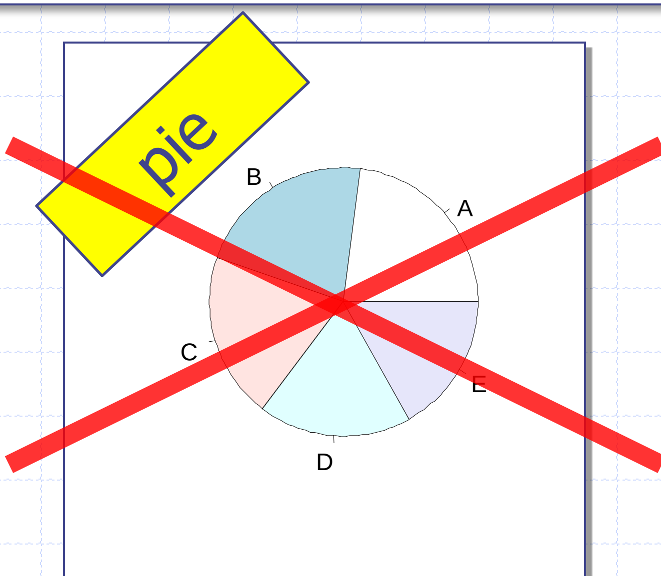
1b

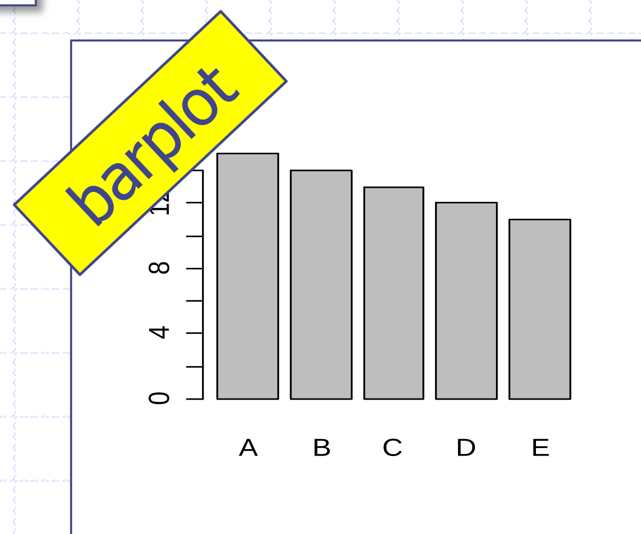
Simple descriptive statistics for qualitative variable:

~~pie()~~

function which draws a pie chart

```
pie(table(d1$origin))
```

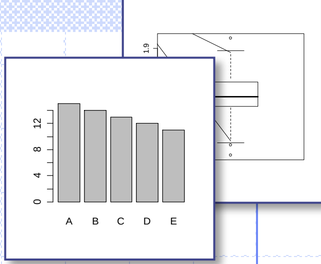




V1

Exercise 4a

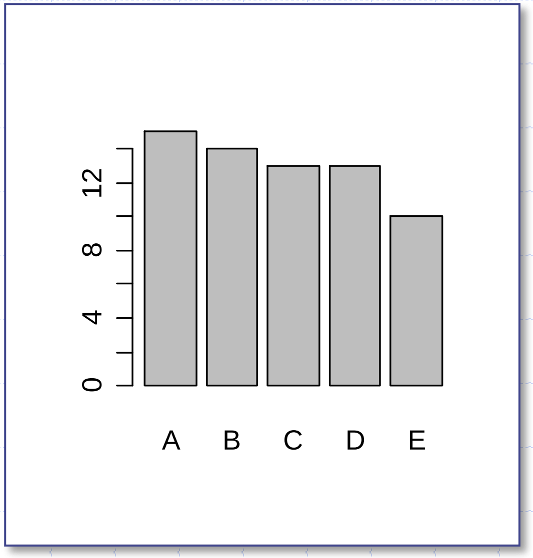
Practical on simple descriptive statistics



Make a contingency table and a bar plot for the variable “origin” of the example 2.

```
table(d2$origin)

 A  B  C  D  E
15 14 13 13 10
```

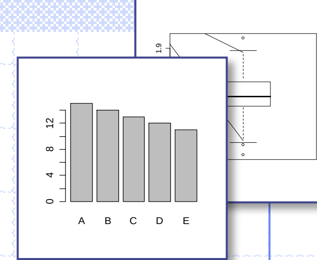
```
barplot(table(d2$origin))
```



V1

2a

Simple descriptive statistics for quantitative variable:



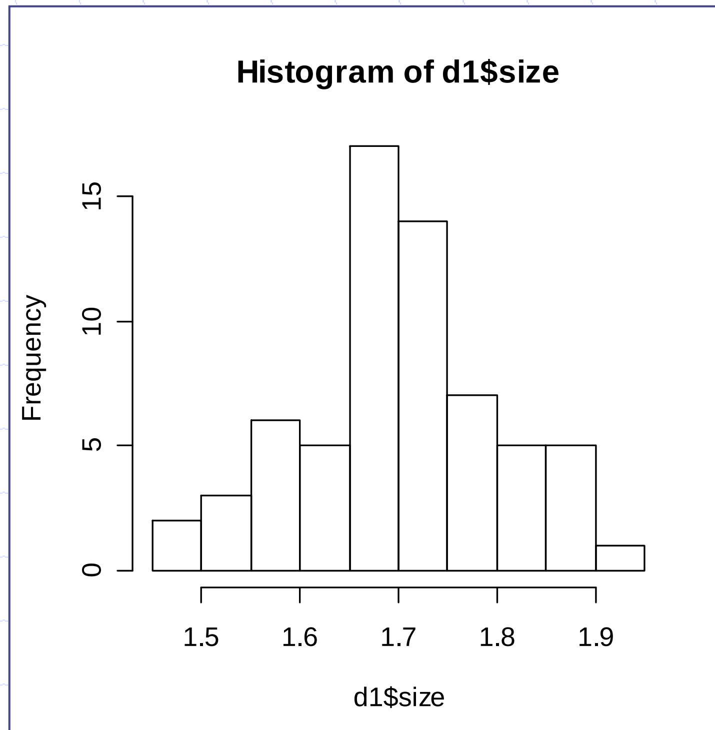

```
hist()
```

function which computes a histogram

a quantitative variable of the data frame

```
hist(d1$size)
```

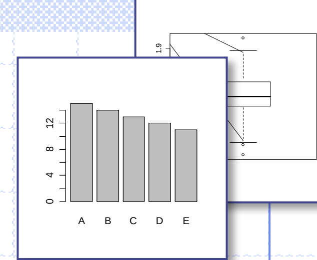
Big sample (N>50)



V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

2b

Simple descriptive statistics for quantitative variable:

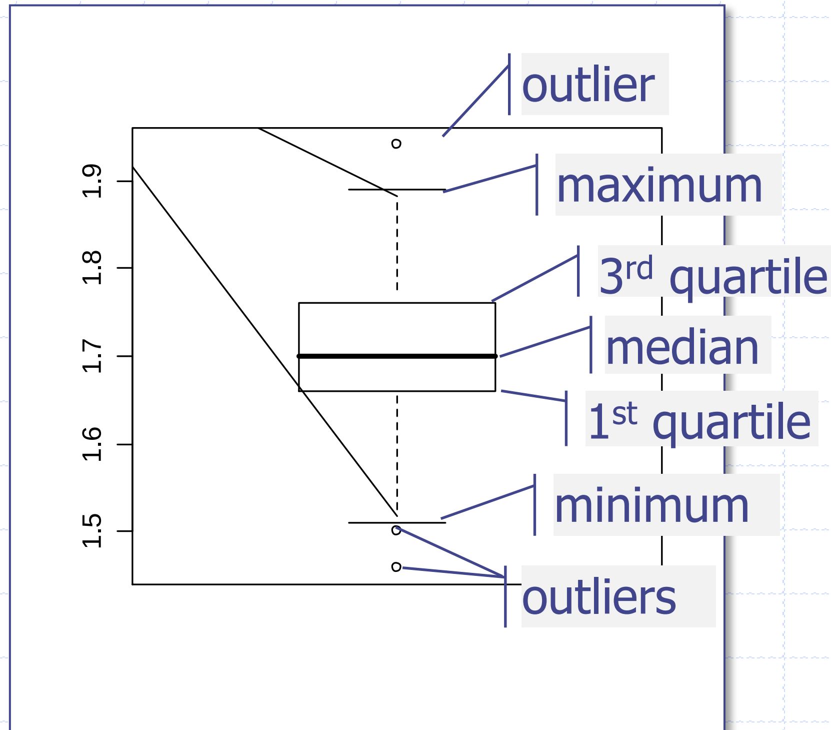


boxplot()

function which produces a box-and-whisker plot

```
boxplot(d1$size)
```

Not too small sample (N>15)

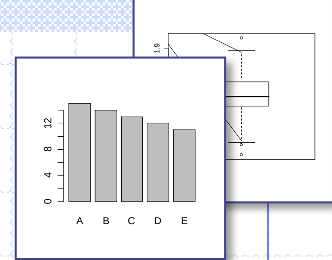


Maximum and minimum are calculated as the most extreme data points which are no more than 1.5 times the interquartile (3rd quartile-1st quartile) from the box.

V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

2 c

Simple descriptive statistics for quantitative variable:

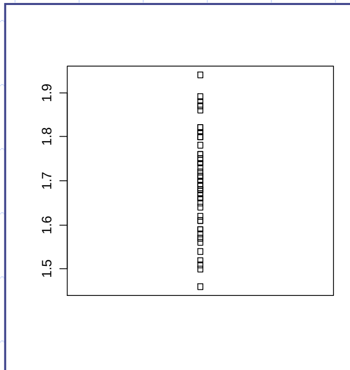


stripchart()

function which produces a dot plot (1D scatter plot)

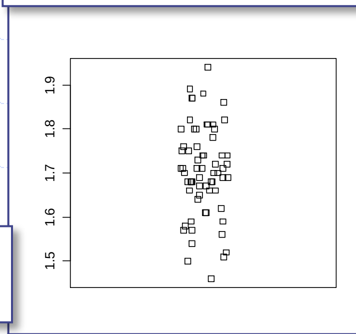
argument for which plot is drawn vertically

```
stripchart(d1$size,vertical=TRUE)
```



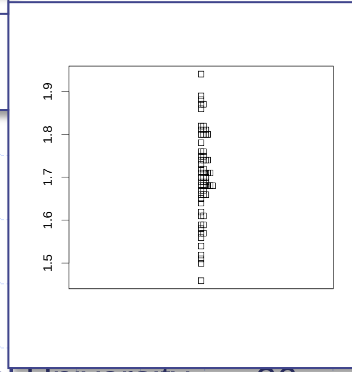
argument for which dots are not overplotted but jittered

```
stripchart(d1$size,vertical=TRUE,method="jitter")
```




```
stripchart(d1$size,vertical=TRUE,method="stack")
```

argument for which dots are not overplotted but stacked

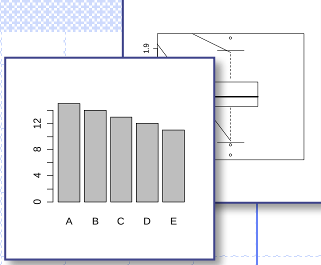


Even for small sample

V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

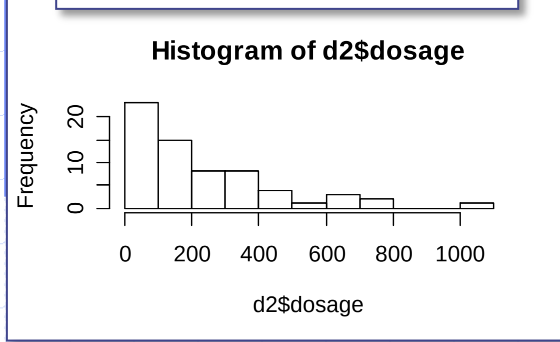
Exercise 4b

Practical on simple descriptive statistics

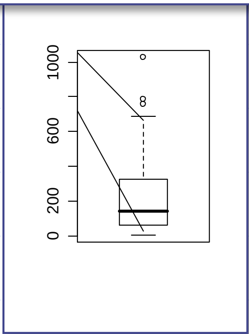


Make a histogram, a box plot and dot plot of the variable "dosage" of the example 2

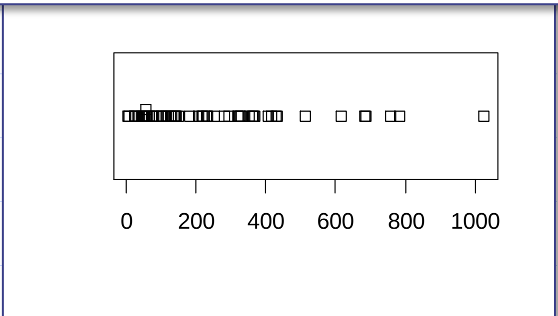
```
hist(d2$dosage)
```




```
boxplot(d2$dosage)
```




```
stripchart(d2$dosage,method="stack")
```



V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

3a

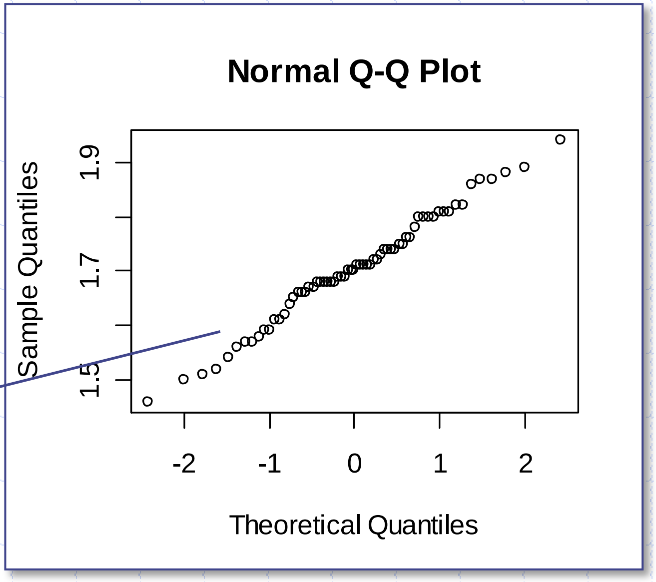
Simple descriptive statistics: looking for normality:

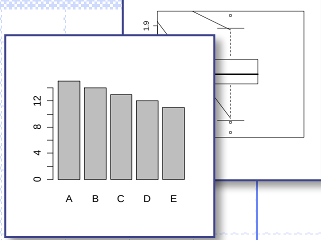
```
qqnorm()
```

function which produces a normal QQ plot

```
qqnorm(d1$size)
```

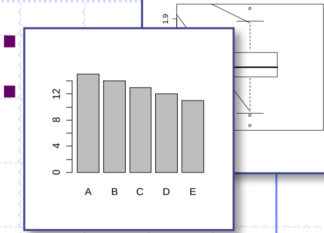
As the plot looks like a straight line, the distribution seems normal





V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

Simple descriptive statistics: looking for normality:




```
shapiro.test()
```

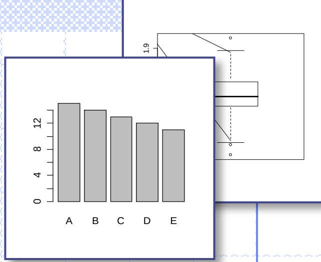
function which performs the Shapiro-Wilk test of normality

```
shapiro.test(d1$size)

Shapiro-Wilk normality test

data:  d1$size
W = 0.9873, p-value = 0.7448
```

p>0.05, the normality hypothesis test is not rejected;
the distribution seems normal.

V2
9.0
0.8
5.9
1.6
7.6
9.9
0.1
3.1

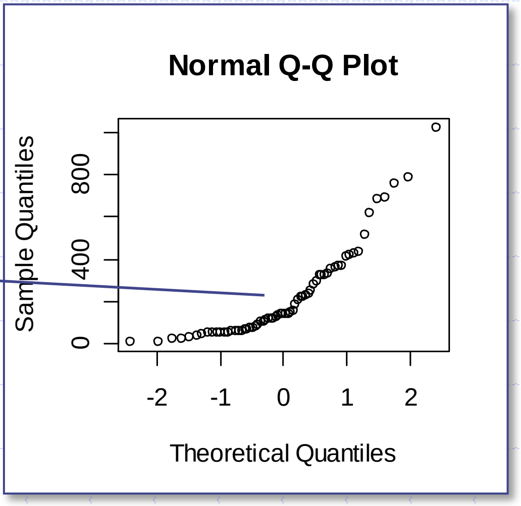
Exercise 4c



Simple descriptive statistics

Looking for normality of the variable dosage of the example 2.

```
qqnorm(d2$dosage)
```



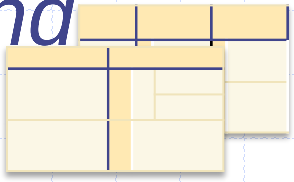
As the plot doesn't look like a straight line, the distribution is not normal.

```
shapiro.test(d2$dosage)

Shapiro-Wilk normality test

data:  d2$dosage
W = 0.8218, p-value = 2.135e-07
```

p<0.001, the normality hypothesis is rejected; the distribution is not normal.

Plan

1. Principle of R language
2. Importing data
3. Simple descriptive statistics -> *Graphics*
4. Simple analytical statistics -> *tests of comparison 2 observed mean and frequencies*



5. Specific tools for epidemiology
6. Specific tools for clinical study

Introduction for simple analytical statistics

In order to simplify, we envisaged only to test two series of data.

The series can be:

either independent or paired.

The variable can be:

either qualitative or quantitative.

The test for quantitative(s) variable(s) can be:

either parametric or non parametric.

We envisaged R code mainly for raw data.

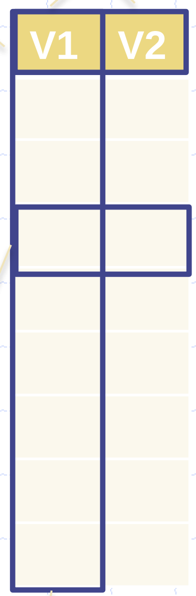
Introduction for data organization for two series

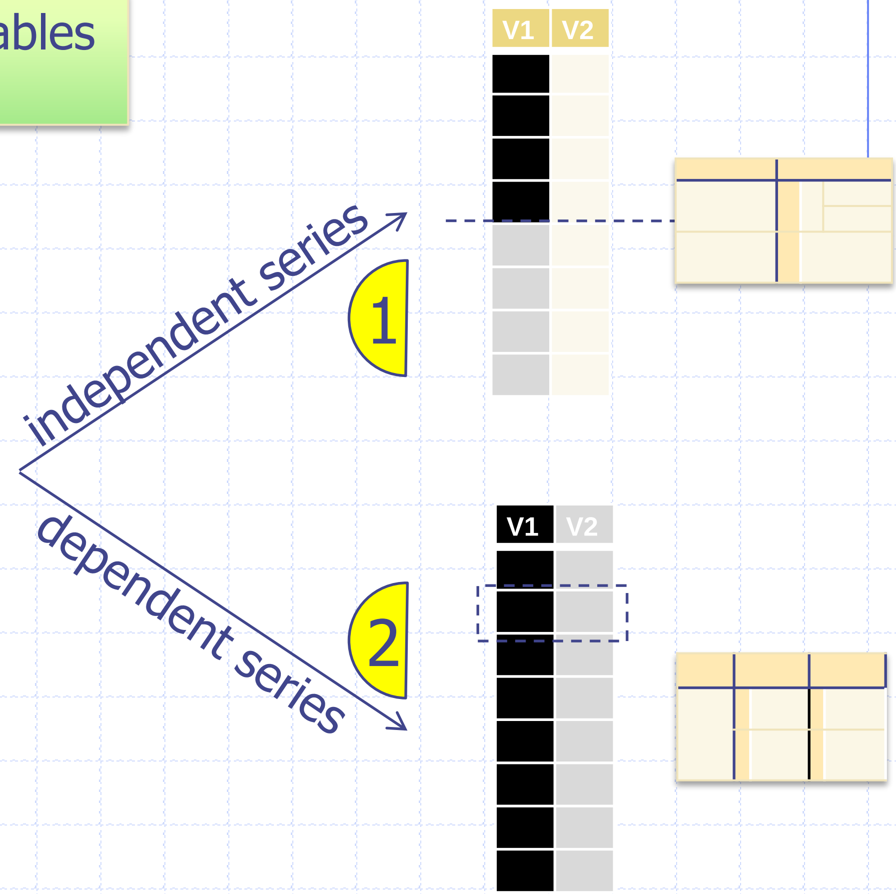
two series: only 2 variables per observation

first row for header of variable

one row per observation subject

one column per variable





1

Two independent series

Different cases for two independent series

Series variable, is dichotomic because there is two series

V1	V2

The other variable can be:

qualitative (dichotomic or not)

quantitative (normal distribution or not)

V1	V2
	+
	+
	-
	+
	±
	-
	-
	+

a

b

V1	V2
	9.0
	0.8
	5.9
	1.6
	7.6
	9.9
	0.1
	3.1

c

d

e

f

1

Two independent series

Different tests

One dichotomic variable(series) + one qualitative (dichotomic or not) variable		One dichotomic variable (series) + one quantitative (normal or not) variable	
Pearson's Chi-squared Test (a)	parametric tests	F test for two variances (c)	Welch two sample t test (d) Student two samples t test (e)
Particularly adapted for small samples Fisher's exact test only for 2 dichotomic variables (b)	non parametric tests	Particularly adapted for small samples Mann-Whitney-Wilcoxon rank sum test (f)	

V1	V2
■	+
■	+
■	-
■	+
□	±
□	-
□	-
□	+

V1	V2
■	9.0
■	0.8
■	5.9
■	1.6
□	7.6
□	9.9
□	0.1
□	3.1

The choice between tests for quantitative variable is supposed known

1 a Two independent series

Pearson's Chi-squared test:

chisq.test()

two qualitative variables

```
chisq.test(d3$exposition,d3$disease)

Pearson's Chi-squared test with Yates'
continuity correction

data:  d3$exposition and d3$disease
X-squared = 5.1212, df = 1, p-value = 0.02363
```

p<0.05, the independence hypothesis test is rejected; the difference is significant.

```
Warning message:
```

Sample too small: expected value<5

V1	V2
	+
	+
	-
	+
	±
	-
	-
	+

1|b Two independent series

Fisher's exact test:

fisher.test()

2 qualitative variables

```
fisher.test(d3$exposition,d3$disease)

Fisher's Exact Test for Count Data

data:  d3$exposition and d3$disease
p-value = 0.01773
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.103106 4.395369
sample estimates:
odds ratio
  2.178838
```

Particularly adapted for small samples (only for 2 dichotomic variables)

$p<0.05$, the independence hypothesis test is rejected; the difference is significant.

V1	V2
■	+
■	+
■	-
■	+
□	-
□	-
□	-
□	+

1 c

Two independent series

F test for 2 variances:

var.test()

a quantitative variable ~ a qualitative dichotomic variable

```
var.test(d3$response~d3$disease)

        F test to compare two variances

data:  d3$response by d3$disease
F = 1.1702, num df = 143, denom df = 54, p-value = 0.5147
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
 0.7320177 1.7876995
sample estimates:
ratio of variances
          1.170242
```

p>0.05, the hypothesis test H0 is not rejected;
the difference is not significant.

V1	V2
	9.0
	0.8
	5.9
	1.6
	7.6
	9.9
	0.1
	3.1

1d

Two independent series

Welch two samples t test:

t.test()

a quantitative variable ~ a qualitative dichotomic variable

```
t.test(d3$response~d3$disease)

        Welch Two Sample t-test

data:  d3$response by d3$disease
t = -3.4249, df = 105.129, p-value = 0.0008781
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.7937418 -0.2116779
sample estimates:
 mean in group no mean in group yes
         5.110023    <    5.612732
```

p<0.001, the hypothesis test H0 is rejected; the difference is highly significant.

V1	V2
	9.0
	0.8
	5.9
	1.6
	7.6
	9.9
	0.1
	3.1

1e Two independent series

Student two samples t test:

t.test()

a quantitative variable ~ a qualitative dichotomic variable

var.equal=TRUE

```
t.test(d3$response~d3$disease, var.equal=TRUE)

        Two Sample t-test

data:  d3$response by d3$disease
t = -3.3062, df = 197, p-value = 0.001124
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.8025613 -0.2028584
sample estimates:
 mean in group no mean in group yes
        5.110023    <    5.612732
```

p<0.01, the hypothesis test H0 is rejected; the difference is highly significant.

V1	V2
	9.0
	0.8
	5.9
	1.6
	7.6
	9.9
	0.1
	3.1

1 | f

Two independent series

Mann-Whitney-Wilcoxon rank sum test:

```
wilcox.test()
```

a quantitative variable ~ a qualitative dichotomic variable

```
wilcox.test(d3$response~d3$disease)

        Wilcoxon rank sum test with continuity
correction

data:  d3$response by d3$disease
W = 2774, p-value = 0.001102
alternative hypothesis: true location shift is not
equal to 0
```

P<0.01, the hypothesis test H0 is rejected;
the difference is highly significant.

Particularly adapted for small samples

V1	V2
	9.0
	0.8
	5.9
	1.6
	7.6
	9.9
	0.1
	3.1

1

Exercise 5a

Practical on simple analytical statistics

Test if the two variables "exposition" and "disease" of the example 3 are independent with a Chi^2 test.

```
chisq.test(d6$exposition,d6$disease)

        Pearson's Chi-squared test with Yates'
continuity correction

data:  d6$exposition and d6$disease
X-squared = 4.5608, df = 1, p-value = 0.03271
```

S+

p<0.05, the independence hypothesis test is rejected; the difference is significant.

1

Exercise 5b

Practical on simple analytical statistics

With the same data, test if the variable "dosage" is different according the "exposition" with a student two sample t test. We need to test the homogeneity of the both variances before.

```
var.test(d6$dosage~d6$exposition)

        F test to compare two variances

data:  d6$dosage by d6$exposition
F = 0.9378, num df = 100, denom df = 98, p-value = 0.7493
...
```

NS

```
t.test(d6$dosage~d6$exposition,var.equal=T)

Two Sample t-test
data:  d6$dosage by d6$exposition
t = 2.2108, df = 198, p-value = 0.02820
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03456140 0.60549861
sample estimates:
 mean in group no mean in group yes
        5.458515          5.138485
```

S+

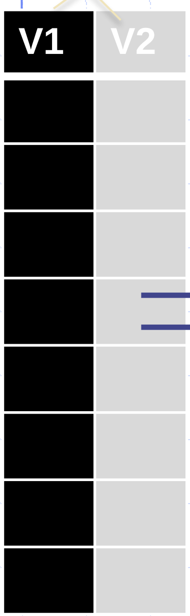
5.458515 > 5.138485

2

Two dependent series

Different cases for two dependent series

We will consider only the case where both variables have the same type.



The both variables can be:

qualitative (dichotomic only)

quantitative (normal distribution or not)

a

V1	V2
+	+
+	+
+	-
-	-
-	+
-	+
+	+
-	-

V1	V2
6.0	4.4
1.7	8.6
7.1	9.0
3.0	5.0
4.9	4.3
5.2	4.3
0.2	10.0
9.7	6.1

b

c

V1-V2
1.6
-6.9
-1.9
-2.0
0.6
0.9
-9.8
3.6

d

e

V1	V2
6.0	4.4
1.7	8.6
7.1	9.0
3.0	5.0
4.9	4.3
5.2	4.3
0.2	10.0
9.7	6.1

2

Two dependent series

Different tests

dichotomic variable	One quantitative variable (normal or not)		Two quantitative variables (normal or not)	
McNemar's Chi-squared Test (a) — Particularly adapted for small samples	parametric tests	Paired t test (b)	parametric tests	Pearson's correlation coefficient test (d)
	non parametric tests	Particularly adapted for small samples — Wilcoxon signed rank test (c)	non parametric tests	Particularly adapted for small samples — Spearman's rank correlation coefficient test (e)

V1	V2
+	+
+	+
+	-
-	-
-	+
-	+
+	+
-	-

V1-V2
1.6
-6.9
-1.9
-2.0
0.6
0.9
-9.8
3.6

V1	V2
6.0	4.4
1.7	8.6
7.1	9.0
3.0	5.0
4.9	4.3
5.2	4.3
0.2	10.0
9.7	6.1

The linear regression is not envisaged here.

2a Two dependent series

Mcnemar's Chi-squared test:

mcnemar.test()

two dichotomic variables

```
mcnemar.test(d4$test1,d4$test2)

McNemar's Chi-squared test with continuity correction

data:  d4$test1 and d4$test2
McNemar's chi-squared = 32.0727, df = 1, p-value = 1.485e-08
```

p<0.001, the hypothesis test is rejected;
the difference is highly significant.

Particularly adapted for small samples

V1	V2
+	+
+	+
+	-
-	-
-	+
-	+
+	+
-	-

2 b

Two dependent series

Paired t test:

t.test()

two quantitative variables

paired=TRUE

```
t.test(d4$femurRsize,d4$femurLsize,paired=TRUE)

        Paired t-test

data:  d5$femurRsize and d5$femurLsize
t = 0.5755, df = 199, p-value = 0.5656
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04974689  0.09074689
sample estimates:
mean of the differences
                 0.0205
```

p>0.05, the hypothesis test is not rejected; the difference is not significant.

V1-V2
1.6
-6.9
-1.9
-2.0
0.6
0.9
-9.8
3.6

2 | c

Two dependent series

Wilcoxon signed rank test:

wilcox.test()

two quantitative variables

paired=TRUE

```
wilcox.test(d4$femurRsize,d4$femurLsize,paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  d5$femurRsize and d5$femurLsize
V = 8238, p-value = 0.5966
alternative hypothesis: true location shift is not equal to 0
```

p>0.05, the hypothesis test is not rejected; the difference is not significant.

Particularly adapted for small samples

V1-V2
1.6
-6.9
-1.9
-2.0
0.6
0.9
-9.8
3.6

2d

Two dependent series

Pearson's correlation coefficient test:

cor.test()

two quantitative variables

```
cor.test(d4$size,d4$weightbefore)

        Pearson's product-moment correlation

data:  d5$size and d5$weightbefore
t = 9.5205, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4571813 0.6486851
sample estimates:
      cor
0.5603776
```

p<0.001, the hypothesis test is rejected; the correlation is highly significant.

95% confidence interval of correlation coefficient r

Correlation coefficient r

2e

Two dependent series

Spearman's rank correlation coefficient test:

cor.test()

two quantitative variables

method="spearman"

```
cor.test(d4$size,d4$weightbefore,method="spearman")
 Spearman's rank correlation rho

data:  d5$size and d5$weightbefore
S = 644607, p-value = 4.907e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5165327

Warning message:
In cor.test.default(d5$size, d5$weightbefore, method =
"spearman") :
  Impossible de calculer les p-values exactes avec des ex-
aequos
```

p<0.001, the hypothesis test is rejected;
the correlation is highly significant.

Correlation coefficient Rho

Particularly adapted for small samples

V1	V2
6.0	4.4
1.7	8.6
7.1	9.0
3.0	5.0
4.9	4.3
5.2	4.3
0.2	10.0
9.7	6.1

2

Exercise 5c

Practical on simple analytical statistics

Using data of example 4, test if the two variables “testA” and “testB” realized on the same subject are different with a Mcnemar Chi^2 test.

```
mcnemar.test(d7$testA,d7$testB)

 McNemar's Chi-squared test with continuity correction

data:  d7$testA and d7$testB
McNemar's chi-squared = 4.3478, df = 1, p-value = 0.03706
```

S+

```
table(d7$testA,d7$testB)

      no yes
  no   97   6
  yes  17  80
```

2

Exercise 5d

Practical on simple analytical statistics

Using same data, test if there is a significant difference between the "weigh tafter" and the "weight before" with a paired t test.

```
t.test(d7$weightafter,d7$weightbefore,paired=TRUE)

        Paired t-test

data:  d7$weightafter and d7$weightbefore
t = 5.8572, df = 199, p-value = 1.922e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.711051 3.447949
sample estimates:
mean of the differences
              2.5795
```

S+++

Increasing of weight

2

Exercise 5e

Practical on simple analytical statistics

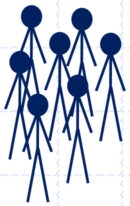
Using same data, test if there is a significant correlation between the “weightbefore” and the “size” with the Spearman rank correlation coefficient.

```
cor.test(d7$size,d7$weightbefore,method="spearman")

        Spearman's rank correlation rho

data:  d7$size and d7$weightbefore
S = 644607, p-value = 4.907e-15
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5165327
```

S+++

Positive correlation

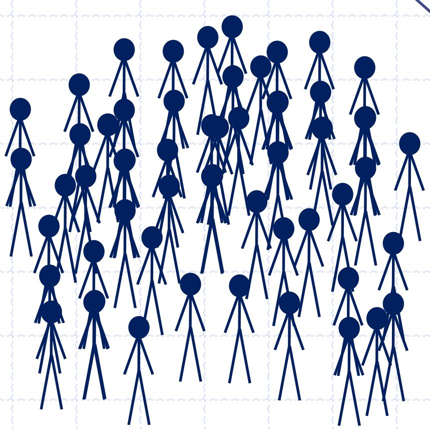
Plan

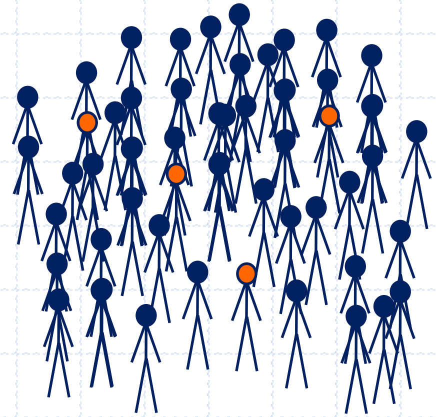
1. Principle of R language
2. Importing data
3. Simple descriptive statistics -> *Graphics*
4. Simple analytical statistics -> *tests of comparison mean and frequencies*
5. Specific tools for epidemiology 
6. Specific tools for clinical study

Introduction for specific tools for epidemiology

Health parameter or disease

+

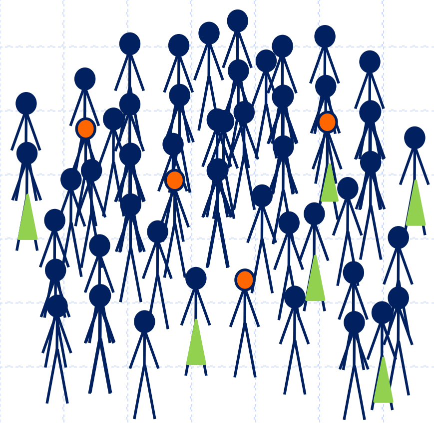




only status “disease”

Descriptive epidemiology

1

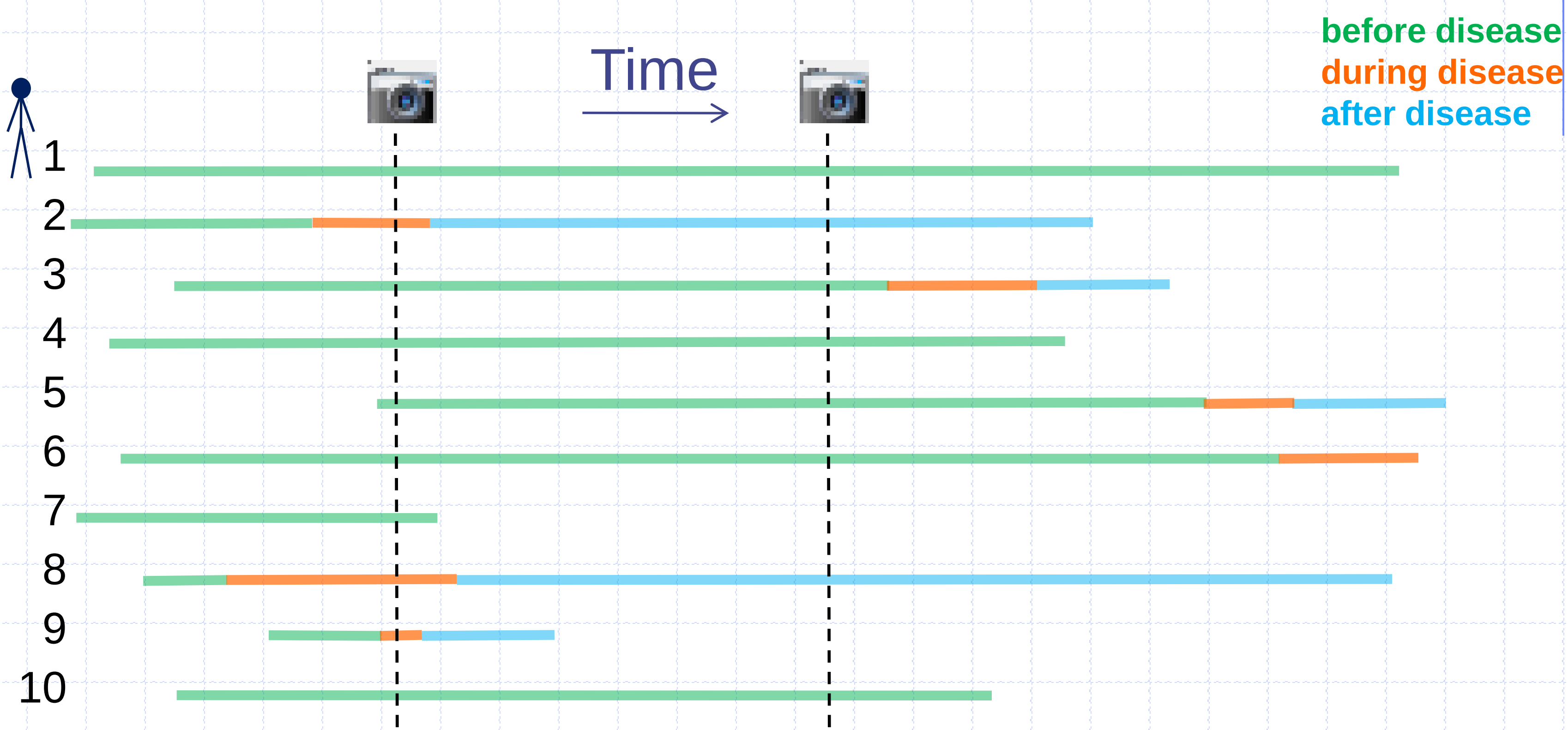


status “disease” + other status

Analytical epidemiology

2

Introduction to descriptive epidemiology



a

Prevalence rate (P)= frequency of ill people

P=3/10

P=0/8

Prevalence rate (P) and its Confidence Interval

binom.test()

Number of disease case — Total number of person

```
binom.test(6,122)

        Exact binomial test

data:  6 and 122
number of successes = 6, number of trials = 122, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.01825960 0.10397262
sample estimates:
probability of success
            0.04918033
```

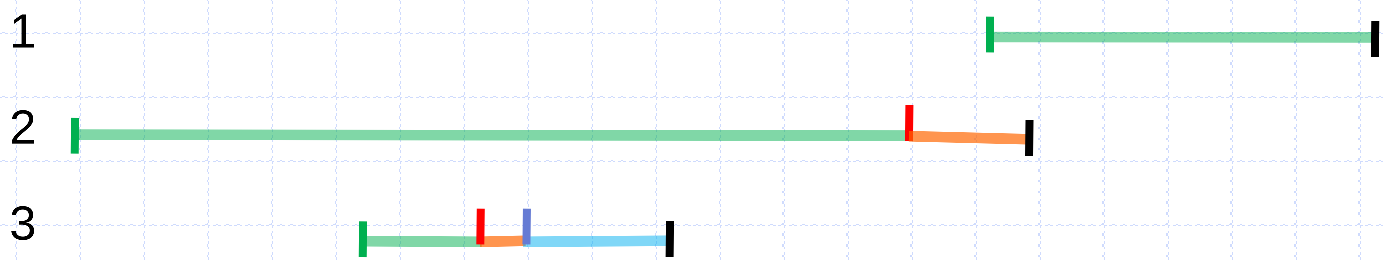
Confidence Interval of Prevalence rate P

Prevalence rate P

Particularly adapted for small samples

Prevalence rate (P) in cohort data -> What is it ?

```
d8[1:3,]
          DO          DLN           DI          DEI
1 27/10/2007 31/01/2008         <NA>         <NA>
2 23/10/2006 15/11/2007 07/10/2007 15/11/2007
3 12/01/2007 15/06/2007 24/02/2007 28/02/2007
```



Prevalence rate (P) in cohort data : the problem of the date

Date
22/06/1965
20/03/1968
30/08/1965
27/09/1965
22/07/1966
05/05/1964
16/01/1964
11/02/1968

→

Date n
1965.472
1968.215
1965.661
1965.737
1966.553
1964.341
1964.040
1968.111

The dates are text for R so this a qualitative data. It need to transform to a quantitative value

Prevalence rate (P) in cohort data: transform dates to fractional numbers (quantitative variables)

```
As.Date("03/07/2007", "%d/%m/%Y")
[1] "2007-07-03"
```

1a Descriptive epidemiology

Prevalence rate (P) in cohort data : adapt the data frame define the date of P (DPn)

```
d8$DLNn<-as.Date(d8$DLN,"%d/%m/%Y")
d8$DIn<-as.Date(d8$DI,"%d/%m/%Y")
d8$DEIn<-as.Date(d8$DEI,"%d/%m/%Y")
d8$DOn<-as.Date(d8$DO,"%d/%m/%Y")
```

```
d8[1:3,]
          DO        DLN         DI        DEI      DLNn        DIn       DEIn        DOn
1 27/10/2007 31/01/2008       <NA>       <NA> 2008-01-31       <NA>       <NA> 2007-10-27
2 23/10/2006 15/11/2007 07/10/2007 15/11/2007 2007-11-15 2007-10-07 2007-11-15 2006-10-23
3 12/01/2007 15/06/2007 24/02/2007 28/02/2007 2007-06-15 2007-02-24 2007-02-28 2007-01-12
```

```
DPn<-as.Date("01/04/2007","%d/%m/%Y")
```

Prevalence rate (P) in cohort data : count the total number of persons present at the date of P (DPn)

```
pres<-subset(d, DOn<=DPn & DLNn>=DPn)
npres<-nrow(pres)
npres
[1] 7
```

`subset()` return subsets of data frames which meet conditions.
The conditions here are the date of origin (DOn) is before the date for which Prevalence is calculated (DPn) and the date of the last news (DLNn) is after the DPn.
⇒Present at the date DPn
`nrow()` return the number of rows.

1 a Descriptive epidemiology

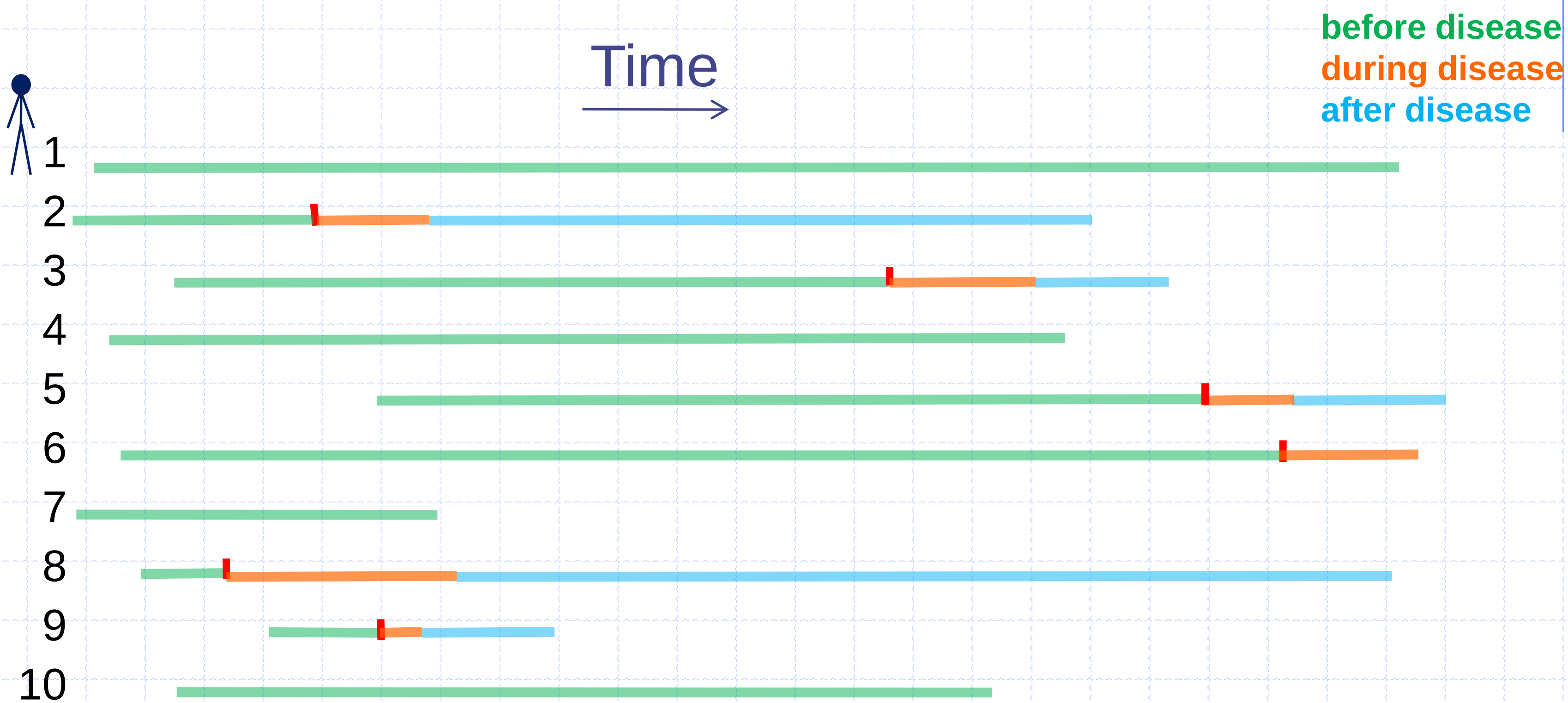
Prevalence rate (P) in cohort data : count the total number of cases

```
cases<-subset(d, DIn<=DPn & DEIn>=DPn)
ncases<-nrow(cases)
ncases
[1] 2
```

The conditions here are the date of illness (DIn) is before the date for which Prevalence is calculated (DPn) and the date of the end of ilness (DEIn) is after the DPn.

⇒Cases at the date DPn

Introduction to descriptive epidemiology



b Incidence rate (I)= number of new case by person-time = number of ( ) / sum of ( )

1b Descriptive epidemiology

Incidence rate (I) and its Confidence Interval

`pois.exact()` of the package epitools

The package epitools need to be previously download on the computer and declare it.

Number of new cases

Total number of person-time

```
library(epitools)
pois.exact(3,100)
  x  pt rate       lower      upper conf.level
1 3 100 0.03 0.006186712 0.08767277       0.95
```

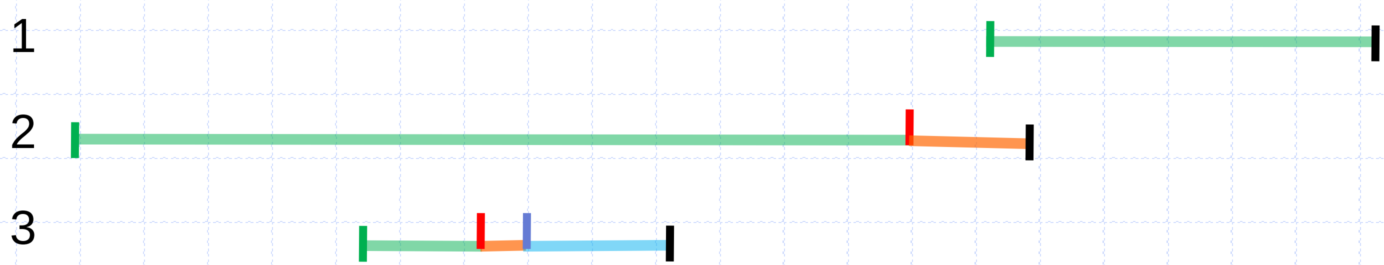
95% confidence interval of Incidence rate I

Particularly adapted for small samples

Incidence rate I

Incidence rate (I) in cohort data

```
d8[1:3,]
          DO        DLN         DI        DEI
1 27/10/2007 31/01/2008       <NA>       <NA>
2 23/10/2006 15/11/2007 07/10/2007 15/11/2007
3 12/01/2007 15/06/2007 24/02/2007 28/02/2007
```



Participation Period (PP)

Sum of every participation time = number of person-time (nPT)

Incidence rate (I) in cohort data : adapt the data frame

```
d8$DLNn<-as.Date(d8$DLN,"%d/%m/%Y")
d8$DIn<-as.Date(d8$DI,"%d/%m/%Y")
d8$DOn<-as.Date(d8$DO,"%d/%m/%Y")
```

```
d[1:3,]
          DO        DLN         DI       DLNn        DIn        DOn
1 27/10/2007 31/01/2008       <NA> 2008-01-31       <NA> 2007-10-27
2 23/10/2006 15/11/2007 07/10/2007 2007-11-15 2007-10-07 2006-10-23
3 12/01/2007 15/06/2007 24/02/2007 2007-06-15 2007-02-24 2007-01-12
```

Incidence rate (I) in cohort data : count the total number of person-time

```
dd<-subset(d,(DIn>DOn|is.na(DIn)))
DEn<-pmin(dd$DLNn,dd$DIn,na.rm=T)
DBn<-dd$DOn
PP<-DEn-DBn
nPT<-sum(PP)
nPT
[1] 6.551677
```

Na.rm=T removes all the NA values

The conditions here are the date of illness(DIn) is after the date of origin (DOn) or the date of illness is not available (person still healthy).

The date of the end of time of participation (DEn) is minimum between DLNn et DIn.

The date of the beginning of time of participation (DBn) is Don

The participation periods (PP) is the difference between DEn and DBn

The number of Person-Time (nPT) is the sum of every participation period (PP)

Incidence rate (I) in cohort data : count the total number of cases

```
newcases<-subset(dd,DIn>DOn)
nnewcases<-nrow(nnewcases)
nnewcases
[1] 4
```

The conditions here are the date of illness (DIn) is after the date of Origin (DOn).
⇒New cases during the study

1

Exercise 6a

Practical on simple descriptive epidemiology

In a survey on 1538 person, 19 have antibodies against a disease. What are the seroprevalence rate of the disease and its 95% confidence interval?

```
binom.test(19,1538)

        Exact binomial test

data:  19 and 1538
number of successes = 19, number of trials = 1538, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.007453676 0.019224878
sample estimates:
probability of success
           0.01235371
```

P=1.24% [0.75%,1.92%]

1

Exercise 6b

Practical on simple descriptive epidemiology

In the same survey, what is the 90% confidence interval of the sero-prevalence rate? Use the argument conf.level=0.9.

```
binom.test(19,1538,conf.level=0.9)

        Exact binomial test

data:  19 and 1538
number of successes = 19, number of trials = 1538, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
90 percent confidence interval:
 0.008104379 0.018074569
sample estimates:
probability of success
            0.01235371
```

More narrow confident interval

P=1.24% [0.81%,1.81%]

1

Exercise 6c

Practical on simple descriptive epidemiology

In a cohort study, what are the prevalence rate and its 95% confidence interval at the 1st April 2007? (“cohort.txt”)

```
d8$DLNn<-as.Date(d8$DLN,"%d/%m/%Y")
d8$DIn<-as.Date(d8$DI,"%d/%m/%Y")
d8$DEIn<-as.Date(d8$DEI,"%d/%m/%Y")
d8$DOn<-as.Date(d8$DO,"%d/%m/%Y")
DPn<-as.Date("01/04/2007","%d/%m/%Y")
pres<-subset(d8, DOn<=DPn & DLNn>=DPn)
(npres<-nrow(pres))
case<-subset(d8, DIn<=DPn & DEIn>=DPn)
(ncase<-nrow(case))
binom.test(ncase,npres)
```

1

Exercise 6c

Practical on simple descriptive epidemiology

In a cohort study, what are the prevalence rate and its 95% confidence interval at the 1st April 2007? ("cohort.txt")

```
binom.test(ncase,npres)
     Exact binomial test

data:  ncase and npres
number of successes = 2, number of trials = 7, p-value = 0.4531
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.03669257 0.70957914
sample estimates:
probability of success
            0.2857143
```

P = 29% [3.7%,71%]

1

Exercise 6d

Practical on simple descriptive epidemiology

In a cohort study, the total number of person-time at risk is 34590 women-month, a specific pathology is observed 15. What are the incidence rate and its 95% confidence interval?

```
library(epitools)
pois.exact(15,34590)
   x    pt         rate        lower       upper conf.level
1  15 34590 0.0004336513 0.0002427113 0.000715242      0.95
```

I=433 [243,715] cases for 100,000 women-month

1

Exercise 6e

Practical on simple descriptive epidemiology

In the cohort study of exercise 6e, what are the incidence rate and its 95% CI during all the study ? (“cohort.txt”)

```
dd<-subset(d,DIn>DOn|is.na(DIn))
DEn<-pmin(dd$DLNn,dd$DIn,na.rm=T)
DBn<-dd$DOn
PP<-DEn-DBn
nPT<-sum(PP)
newcases<-subset(dd,DIn>DOn)
nnewcases<-nrow(newcases)
pois.exact(nnewcases,nPT)
```

1 Exercise 6e

Practical on simple descriptive epidemiology

In the cohort study of exercise 6e, what are the incidence rate in person-year and its 95% CI during all the study? ("cohort.txt")

```
pois.exact(nnewcases,nPT)
  x       pt       rate       lower      upper conf.level
 8 1315 0.00608365 0.002626488 0.01198722        0.95

pois.exact(nnewcases,nPT/365.25)
  x       pt       rate       lower      upper conf.level
1 8 3.600274 2.222053 0.9593247 4.378331        0.95
```

I=2.2 [0.96,4.4] new cases for 1 person-years

⚠ The unit of number of Person-time is the person-days because the unit of the time is the day (Cf. as.Date)

2

Analytical epidemiology

Introduction to analytical epidemiology

-> association between exposition and disease

Ex. : Lung cancer

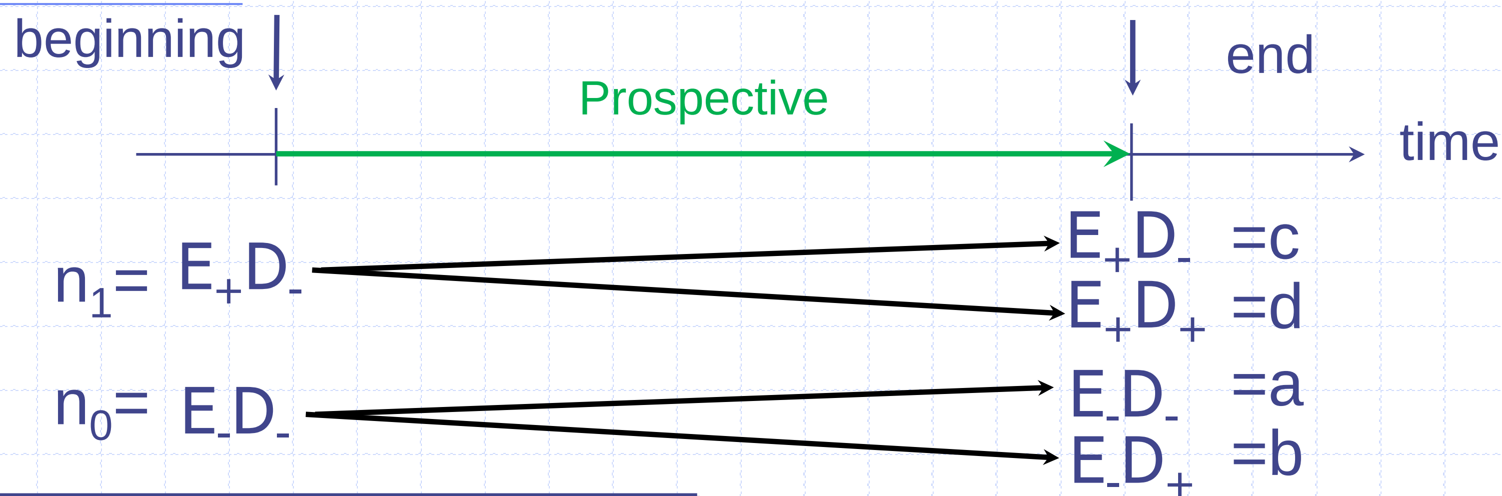
Ex. : tabacco

		Disease		
		-	+	
Exposition	-	a	b	n_0
	+	c	d	n_1
		m_0	m_1	

Cohort survey a b

Case-control study c

Cohort survey and Risk Ratio (RR)



		Disease		
		-	+	
Exposition	-	a	b	n_0
	+	c	d	n_1
		m_0	m_1	

Risk to get ill in exposed = $R_1 = \frac{d}{n_1}$

Risk to be ill in not-exposed = $R_0 = \frac{b}{n_0}$

Risk Ratio=RR= $\frac{R_1}{R_0}$

Risk Ratio (RR) and its Confidence Interval

`riskratio()` of the package epitools

The package epitools need to be previously download on the computer and declare it.

$a=E_-D_-$

$b=E_-D_+$

$c=E_+D_-$

$d=E_+D_+$

c()

```
library(epitools)
riskratio(c(29, 35, 64, 12))
```

Or

exposition variable

disease variable

```
library(epitools)
riskratio(d9$exposition,d9$disease)
```

Risk Ratio (RR) and its Confidence Interval

```
library(epitools)
riskratio(c(29, 35, 64, 12))
 $data
          Outcome
Predictor  Disease1 Disease2 Total
  Exposed1       29       35    64
  Exposed2       64       12    76
  Total          93       47   140

$measure
          risk ratio with 95% C.I.
Predictor   estimate     lower     upper
  Exposed1 1.0000000        NA        NA
  Exposed2 0.2887218 0.1640857 0.5080288

$p.value
          two-sided
Predictor    midp.exact fisher.exact   chi.square
  Exposed1           NA           NA           NA
  Exposed2 1.282338e-06 1.876171e-06 1.203275e-06

$correction
[1] FALSE

attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
```

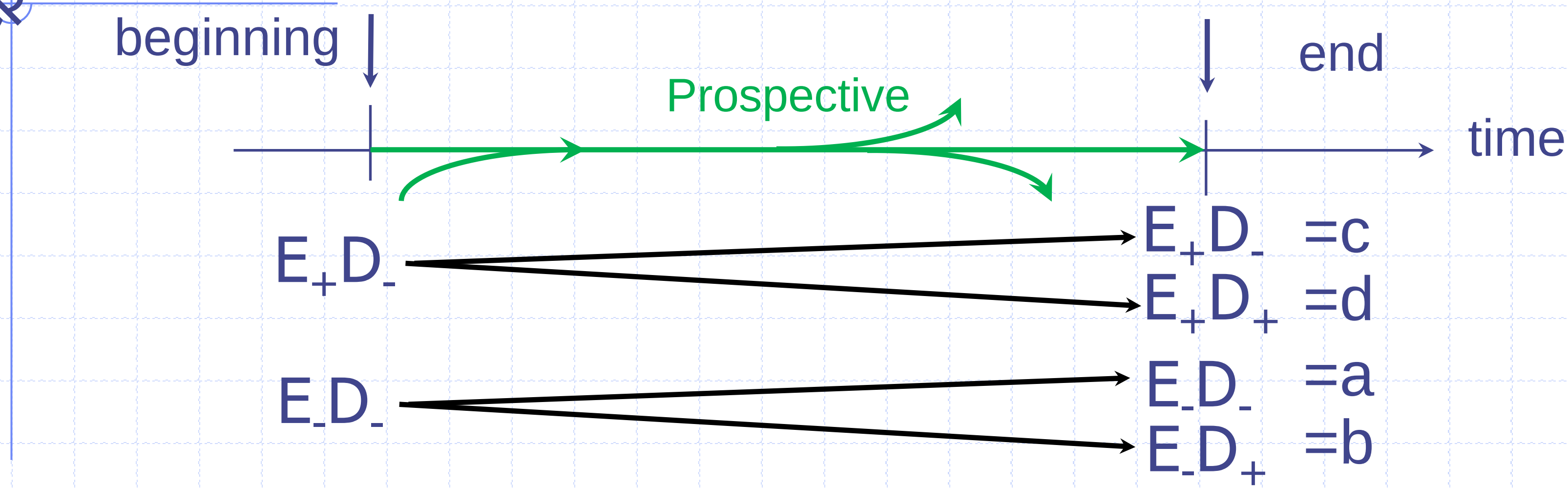
Risk Ratio RR

95% confidence interval of Risk Ratio RR

p value of fisher exact test -> RR≠1?

If small sample

Cohort survey and Incidence Rate Ratio (IRR)



N_{pt1} = number of person-time in exposed

N_{pt0} = number of person-time in not-exposed

I_1 = Incidence rate in exposed = a/N_{pt1}

I_0 = Incidence rate in not-exposed = b/N_{pt0}

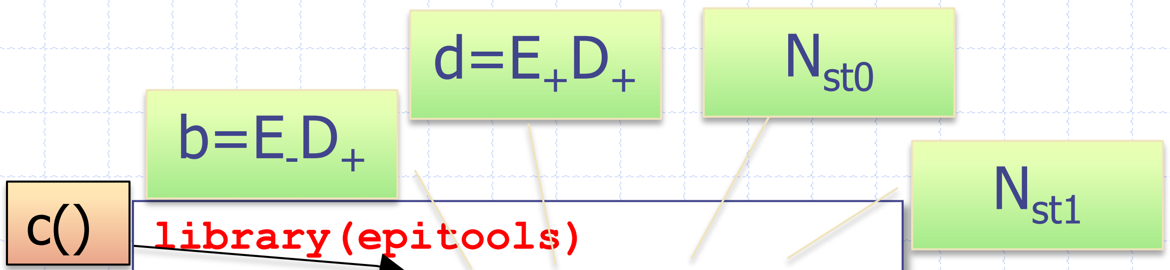
$$RR = \frac{R_1}{R_0}$$

$$\text{Incidence Rate Ratio(IRR)} = \frac{I_1}{I_0} = \frac{\frac{a}{N_{pt1}}}{\frac{b}{N_{pt0}}}$$

Incidence Rate Ratio (IRR) and its Confidence Interval

`rateratio()` of the package epitools

The package epitools need to be previously download on the computer and declare it.




```
library(epitools)
rateratio(c(41,35,1201,632))
```

2b Analytical epidemiology

Incidence Rate Ratio (IRR) and its Confidence Interval

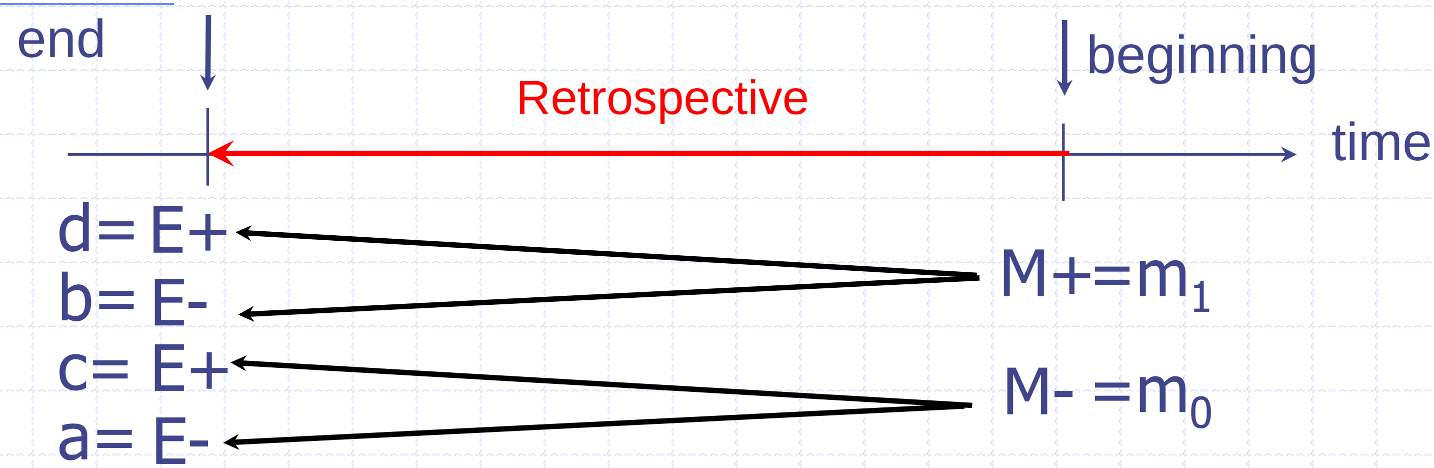
```
library(epitools)
rateratio(c(41,35,1201,632))
 $data
          Outcome
Predictor  Cases Person-time
  Exposed1    41        1201
  Exposed2    35         632
  Total       76        1833

$measure
          rate ratio with 95% C.I.
Predictor  estimate    lower    upper
  Exposed1 1.000000       NA       NA
  Exposed2 1.623489 1.027673 2.550396

$p.value
          two-sided
Predictor  midp.exact       wald
  Exposed1         NA         NA
  Exposed2 0.03797212 0.03377064

attr(,"method")
[1] "Median unbiased estimate & mid-p exact CI"
```

Incidence Rate Ratio IRR

95% confidence interval of Incidence Rate Ratio RR

p value of midp exact test -> IRR≠1?

Case-control Study and Odds Ratio (OR)



		Disease	
		-	+
Exposition	-	a	b	n_0
	+	c	d	n_1
		m_0	m_1	

~~$RR = \frac{R_1}{R_0}$~~

$$\text{Odds Ratio} = OR = \frac{ad}{bc}$$

Relation between Odds Ratio (OR) and Risk Ratio (RR)

$$RR = \frac{OR}{1 + R_0(OR - 1)}$$

with R_0 risk to get the disease in not-exposed

$$\text{if } R_0 << 1 \text{ then } RR \approx OR$$

$$\text{if } OR > 1 \text{ then } OR > RR > 1$$

Exposition linked to the disease

$$\text{si } OR < 1 \text{ alors } OR < RR < 1$$

Exposition linked to the protection

OddsRatio (OR) and its Confidence Interval

`oddsratio()` of the package epitools

The package epitools need to be previously download on the computer and declare it.

$b=E_-D_+$ | $c=E_+D_-$

$a=E_-D_-$ | $d=E_+D_+$

c()

```
library(epitools)
oddsratio(c(29, 35, 64, 12))
```

Or

exposition variable | disease variable

```
library(epitools)
oddsratio(d10$exposition,d10$disease)
```

Odds Ratio (OR) and its Confidence Interval

```
library(epitools)
oddsratio(c(29, 35, 64, 12))
 $data
          Outcome
Predictor  Disease1 Disease2 Total
  Exposed1       29       35    64
  Exposed2       64       12    76
  Total          93       47   140

$measure
          odds ratio with 95% C.I.
Predictor   estimate      lower     upper
  Exposed1 1.0000000         NA        NA
  Exposed2 0.1587889 0.06939932 0.3427073

$p.value
          two-sided
Predictor    midp.exact fisher.exact   chi.square
  Exposed1           NA           NA           NA
  Exposed2 1.282338e-06 1.876171e-06 1.203275e-06

$correction
[1] FALSE

attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
```

Odds Ratio OR

95% confidence interval of Odds Ratio OR

p value of fisher exact test -> OR≠1?

2

Exercise 7a

Practical on simple analytical epidemiology

In a case-control study, among 123 cases, 37 are smokers and among 239 controls, 25 are smokers. What are the Odds Ratio and its 95% Confidence Interval?

$b=E_-D_+$

$c=E_+D_-$

$a=E_-D_-$

$d=E_+D_+$

c()

```
library(epitools)
oddsratio(c(?, ?, ?, ?))
```

```
oddsratio(c(239-25,123-37,25,37))
```

2

Exercise 7a

Practical on simple analytical epidemiology

```
      $data
          Outcome
Predictor  Disease1 Disease2 Total
  Exposed1      214       86   300
  Exposed2       25       37    62
  Total         239      123   362

$measure
          odds ratio with 95% C.I.
Predictor  estimate    lower    upper
  Exposed1 1.000000       NA       NA
  Exposed2 3.660792 2.085971 6.525364

$p.value
          two-sided
Predictor    midp.exact fisher.exact   chi.square
  Exposed1           NA           NA           NA
  Exposed2 5.765939e-06 6.427031e-06 2.689288e-06

$correction
[1] FALSE

attr(,"method")
[1] "median-unbiased estimate & mid-p exact CI"
```

OR=3.66 [2.09,6.52]

OR≠1 S+++ (p<0.001)

Exposition is linked with disease (OR>1)

1

Exercise 7b

Practical on simple analytical epidemiology

The results of a cohort study are summarized in the file “cohort2.txt”. The column exposition and disease correspond respectively to the information about the people which smoke or not and the information about the people which get ill or not. What are the Relative Risk and its 95% Confidence Interval?

exposition variable

disease variable

```
library(epitools)
riskratio(   ?    ,    ?   )
```

```
riskratio(d11$exposition,d11$disease)
```

2

Exercise 7b

Practical on simple analytical epidemiology

```
$data
         Outcome
Predictor  no yes Total
    no     375  25   400
    yes    351  49   400
    Total  726  74   800

$measure
         risk ratio with 95% C.I.
Predictor estimate    lower    upper
      no      1.00       NA       NA
      yes     1.96 1.235641 3.108994

$p.value
         two-sided
Predictor  midp.exact fisher.exact  chi.square
      no           NA           NA          NA
      yes 0.003414235  0.004725127 0.003404036

$correction
[1] FALSE

attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
```

RR=1.96 [1.23,3.11]

RR≠1 S++ (p<0.01)

Exposition is linked with disease (RR>1)

2

Exercise 7c

Practical on simple analytical epidemiology

The results of a cohort study, all the people don't stay the same period in the cohort. In the exposed population (smoker), the total number of person-time at risk is 6546 people-years and 423 get ill. In the not-exposed population (no-smoker), the total number of person-time at risk is 9330 people-years and 57 get ill. What are the Incidence rate ratio and its 95% Confidence Interval?

$b=E_-D_+$ $d=E_+D_+$ N_{st0} N_{st1}

c()

```
library(epitools)
rateratio(c( ? , ? , ? , ?))
```

```
rateratio(c(57,423,9330,6546))
```

2

Exercise 7c

Practical on simple analytical epidemiology

```
$data
           Outcome
Predictor  Cases Person-time
  Exposed1    57        9330
  Exposed2   423        6546
  Total      480       15876

$measure
           rate ratio with 95% C.I.
Predictor  estimate    lower    upper
  Exposed1  1.00000       NA       NA
  Exposed2 10.55062 8.075443 14.05593

$p.value
           two-sided
Predictor  midp.exact wald
  Exposed1         NA   NA
  Exposed2          0    0

attr(,"method")
[1] "Median unbiased estimate & mid-p exact CI"
```

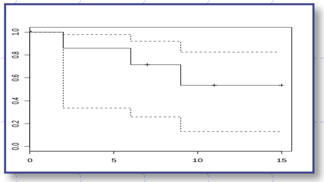
IRR=10.1 [8.08,14.1]

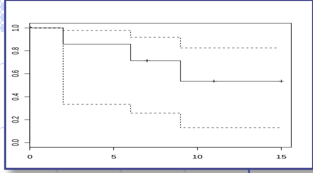
IRR≠1 S+++ (p<0.001)

Exposition is linked with disease (IRR>1)

Plan

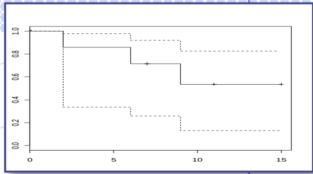
1. Principle of R language
2. Importing data
3. Simple descriptive statistics -> *Graphics*
4. Simple analytical statistics -> *tests of comparison mean and frequencies*
5. Specific tools for epidemiology
6. Specific tools for clinical study





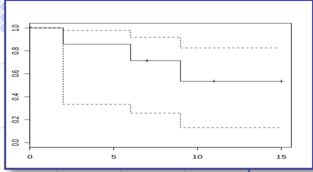
Clinical study -> survival curves

- Understand the principle of the survival curves and their confidence interval
- Know how realize survival curves from data frame with R.



Interest of survival curves

- Describe the survival rate in a sample;
- Estimate the survival rate in the population;
- Compare survival rates according groups (therapeutic and epidemiologic research).

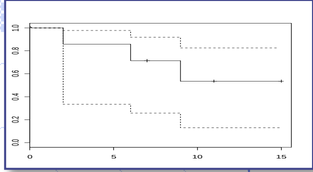


Survival data

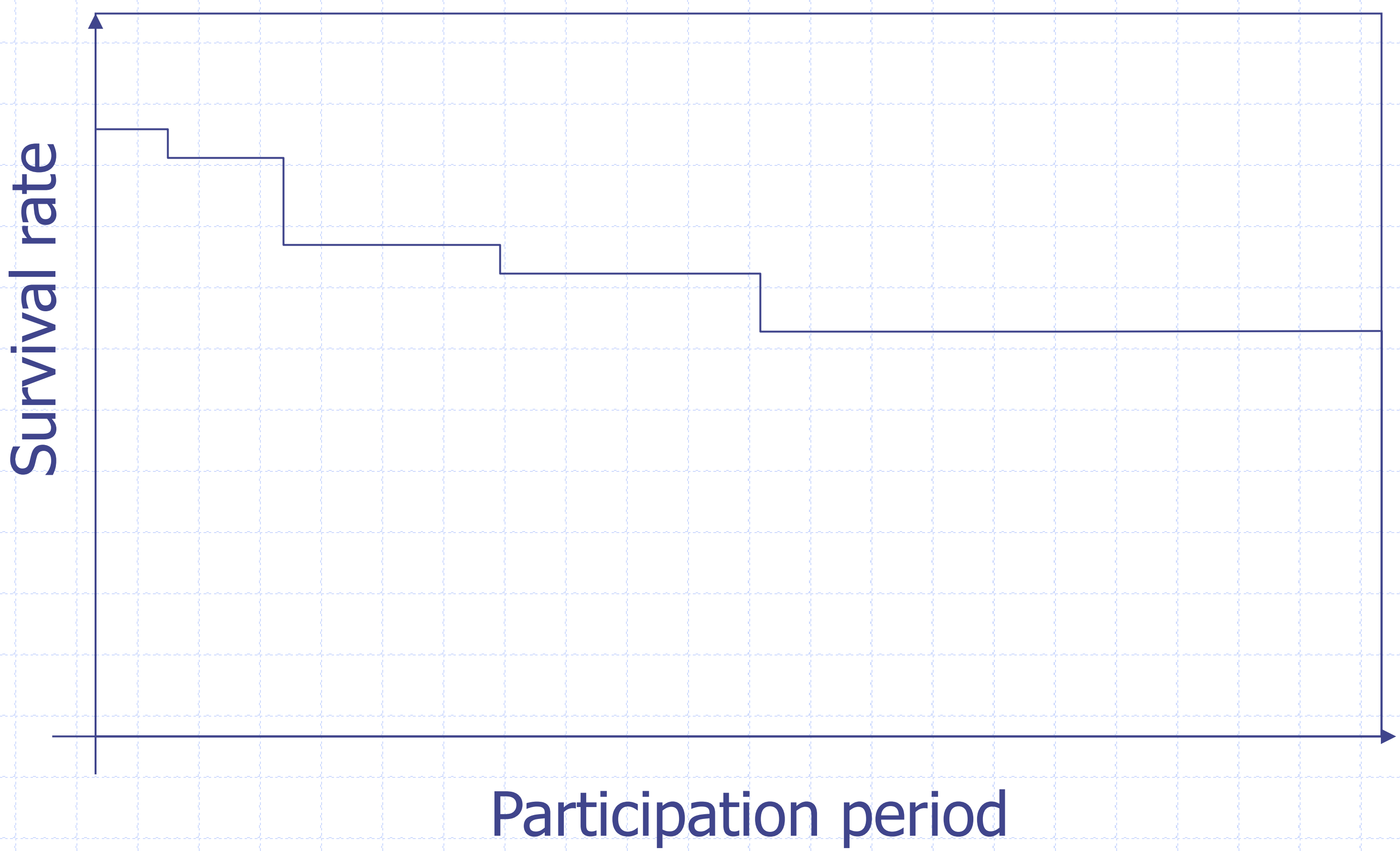
- Survival : time between two dates. The final date is not necessarily the date of death e.g., date of first relapse, of the first complication,...
- Particularity of survival data: possibility de censured data
- Survival rate : probability to be alive (or healthy)

  ≠1-mortality rate

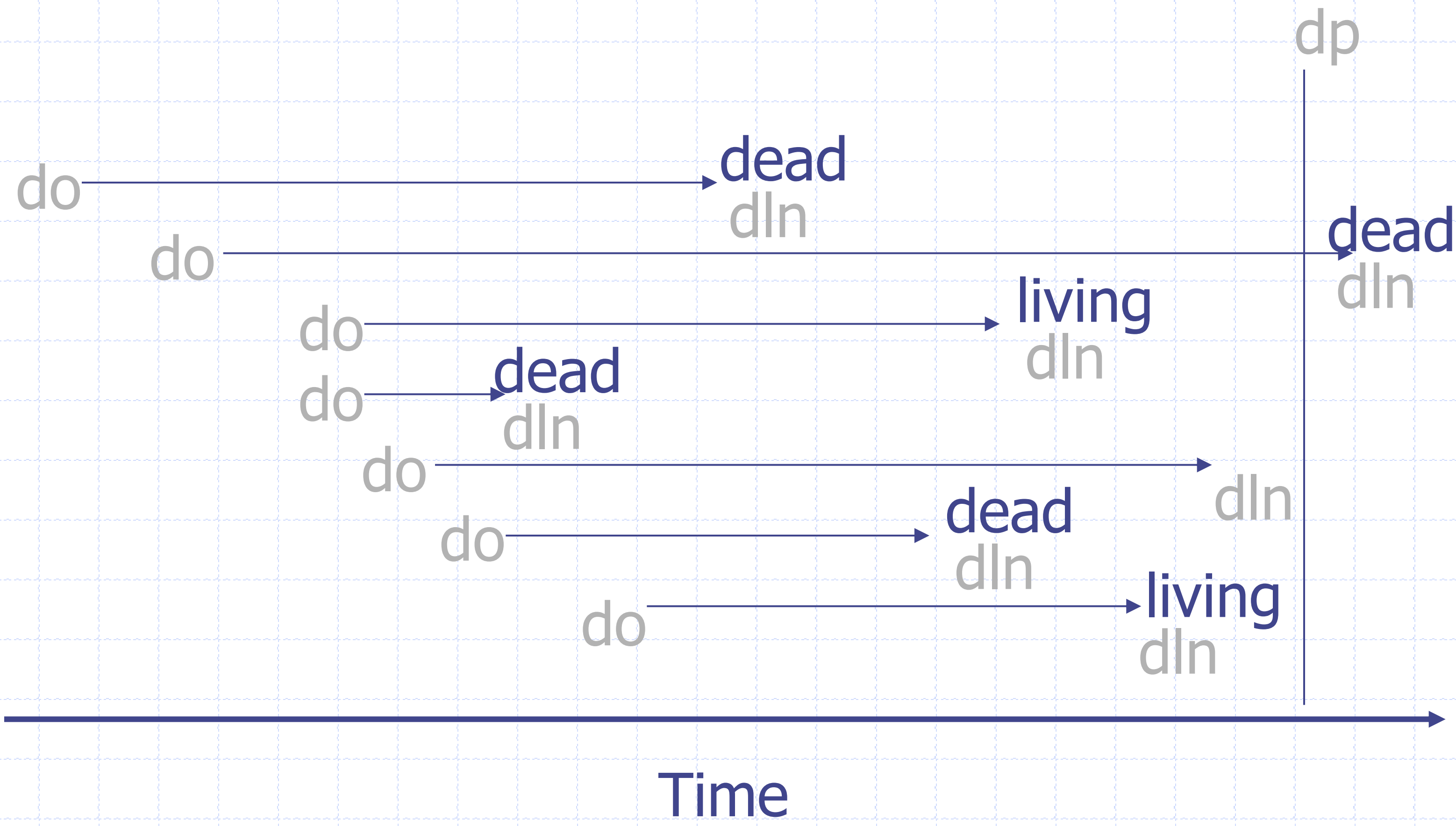
  mortality rate is an incidence rate
- Survival curve : Representation of the survival rate according the time of participation



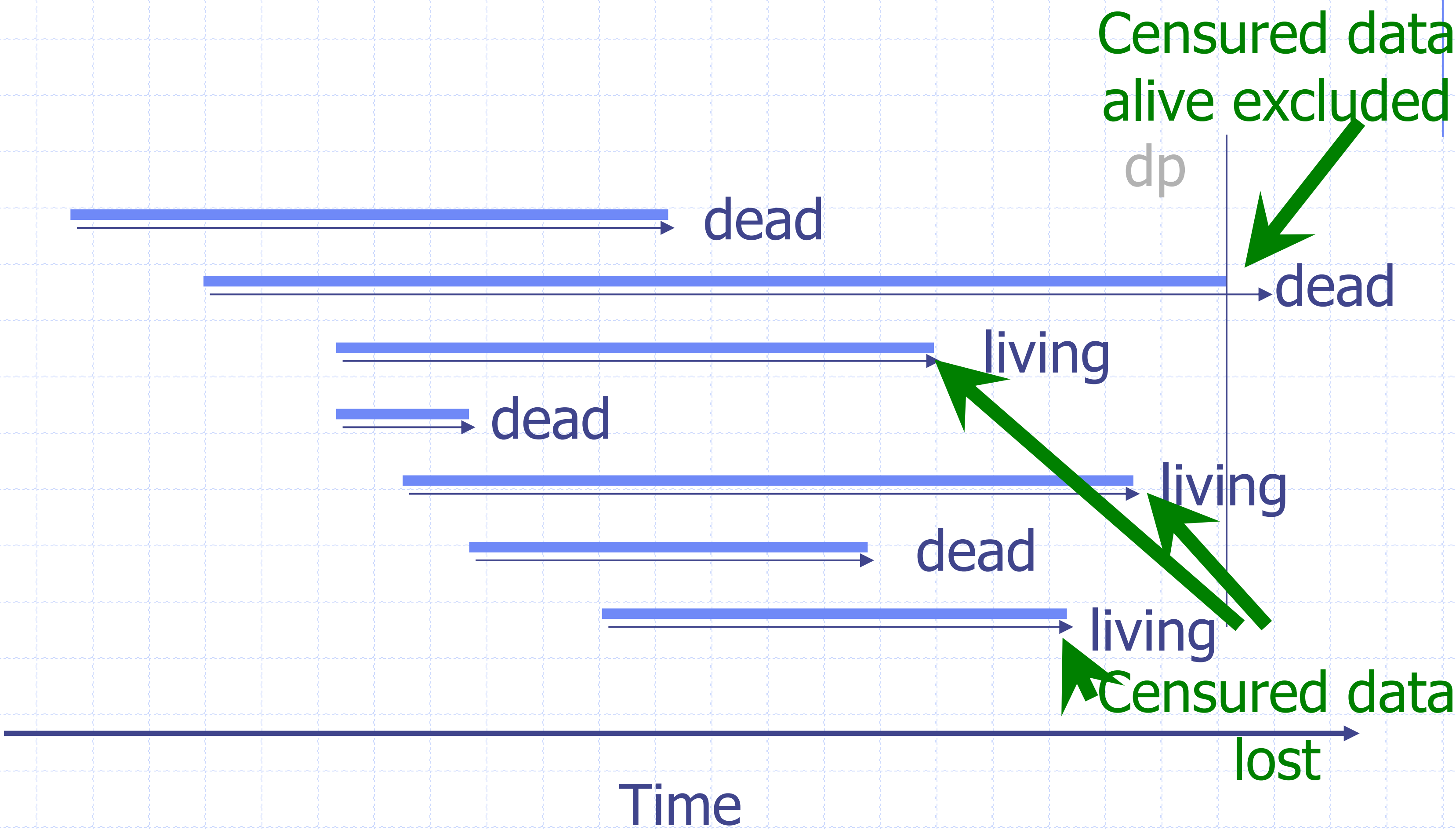
Survival curve



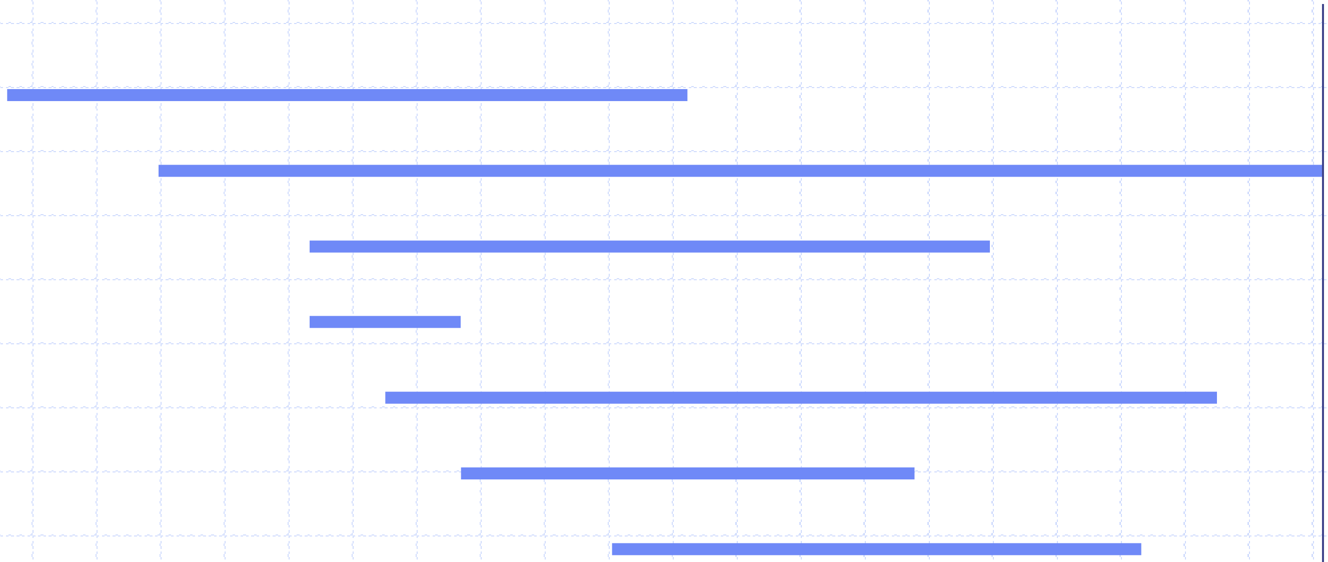
Participation period



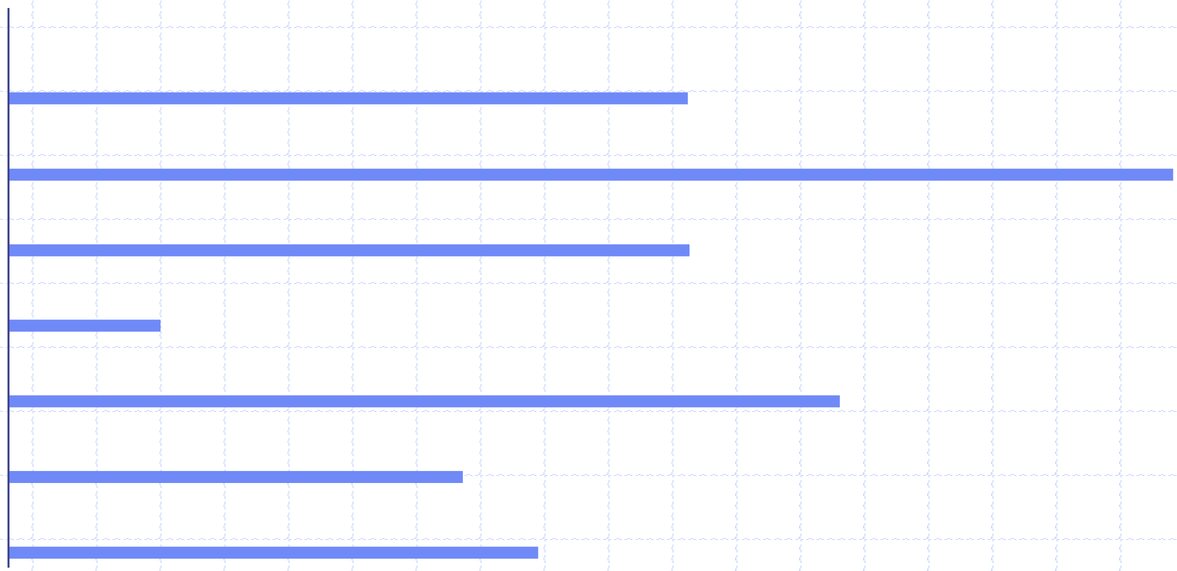
Participation period and censure

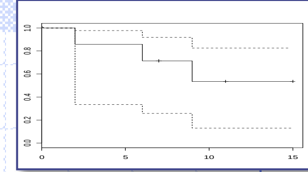


Participation period from 0

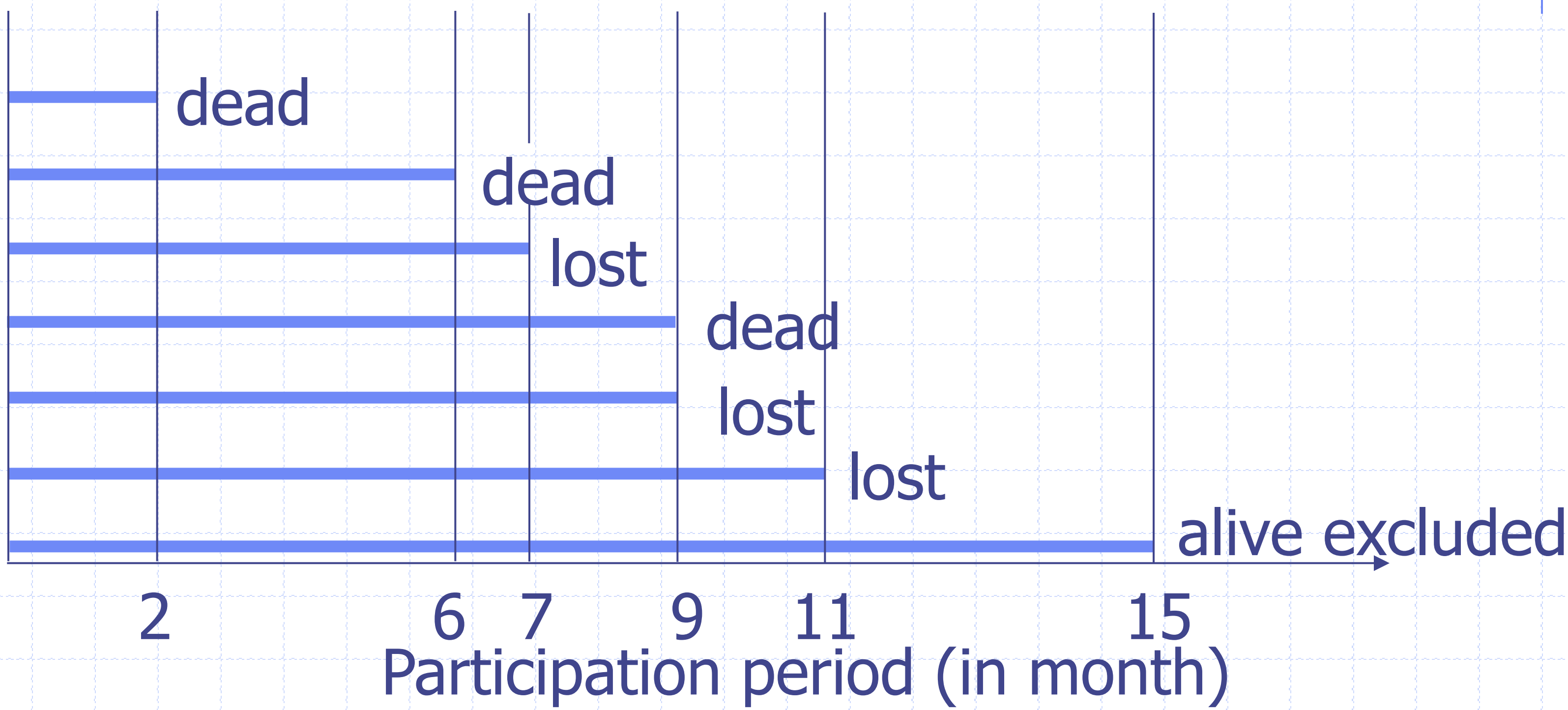


Ordered participation period

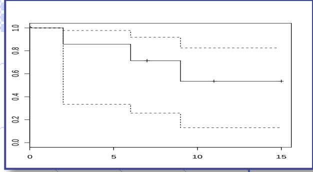




Summary of participation period



➡ Survival curve



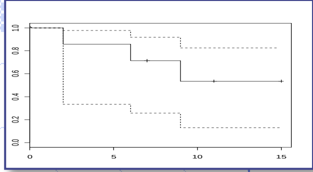
Survival rate

The survival rate represent the probability to be still alive a the time t. There are different non parametric methods to estimate the survival rate:

- Kaplan-Meier method
- Actuarial method

  These methods are based on the principle of the conditional probabilities.

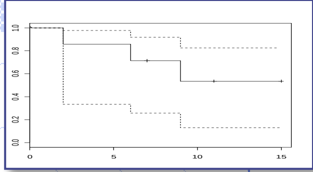
- Method of direct calculus



Kaplan-Meier method

In this method, the survival rate is calculated only at t_i when a time of participation is finishing.

- If the time of participation is finishing by a dead then the survival decrease.
- Otherwise, the survival rate is constant.



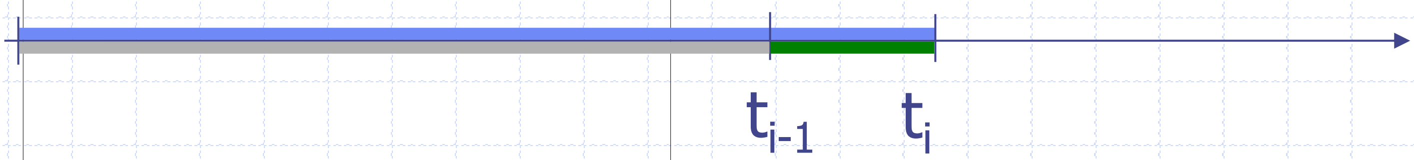
Principle of Kaplan-Meier method

So, the probability to be alive at time t_i is equal to:

- the probability to be alive at t_{i-1} multiplied by
- the probability to be alive at t_i conditionally to be alive at t_{i-1}.

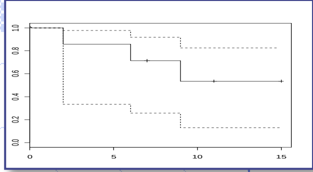
For the time between t_i and t_{i+1} excluded, the probability to be alive is constant.

Conditionnal probabilities



$$P(S_{t_i}) = P(S_{t_{i-1}}) \times P(S_{t_i | t_{i-1}})$$

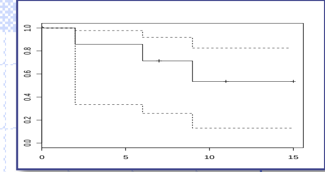
$$S_i = S_{i-1} \times S_{i|i-1}$$



Calculus of survival rate with the Kaplan-Meier method

$$S_{i|i-1} = \frac{N_i - D_i}{N_i}$$

$$S_i = \frac{N_1 - D_1}{N_1} \times \frac{N_2 - D_2}{N_2} \times \ldots \times \frac{N_i - D_i}{N_i}$$



Calculus of CI of survival rates with the Kaplan-Meier method

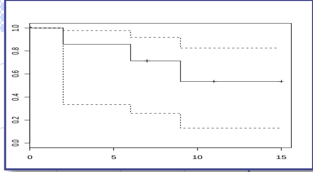
$$\text{IC de } S_i = S_i \pm u_{1-\alpha/2}\sqrt{\operatorname{var} S_i}$$

Variance of survival rate or Greenwood variance

$$\operatorname{var} S_i = S_i^2 \left[\frac{D_1}{N_1 (N_1 - D_1)} + \ldots + \frac{D_i}{N_i (N_i - D_i)} \right]$$

Be careful to the approximation by normal law

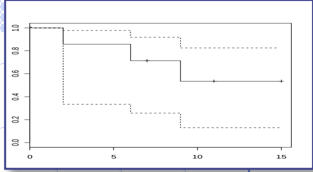
Calculus of CI of survival rates with the Kaplan-Meier method



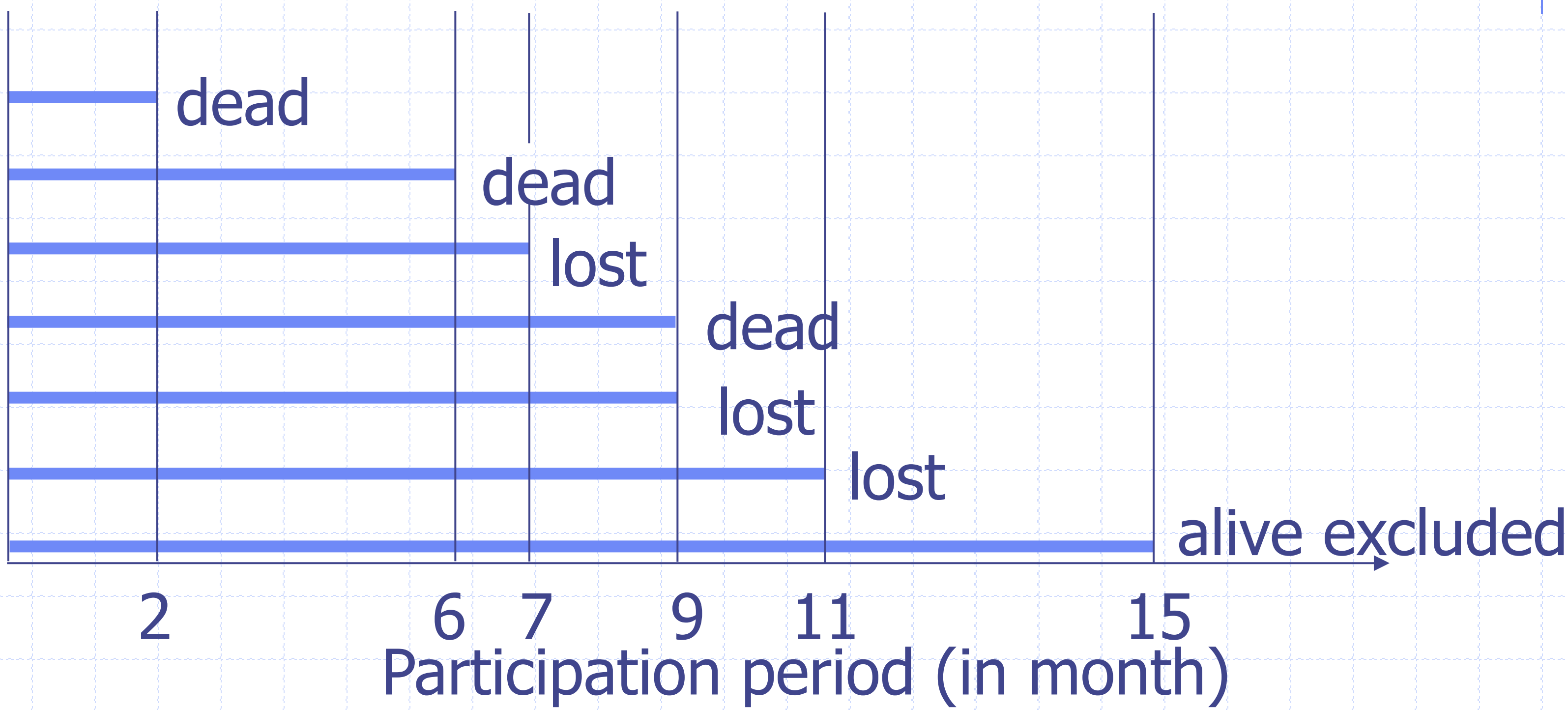
If the approximation by normal low is not proved-> correction of Rothman or surfit() in R

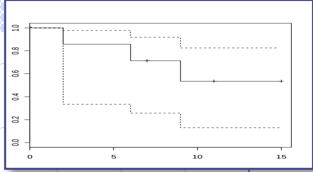
$$\text{IC de } S_i = \frac{M}{M + u_{1-\alpha/2}^{2}} \left[S_i + \frac{u_{1-\alpha/2}^{2}}{2M} \pm u_{1-\alpha/2} \sqrt{\operatorname{var} S_i + \frac{u_{1-\alpha/2}^{2}}{4M^2}} \right]$$

$$M = \frac{S_i(1 - S_i)}{\operatorname{var} S_i}$$



Summary of participation period

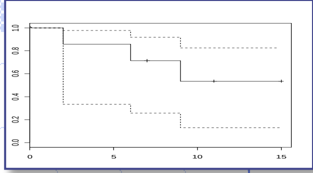


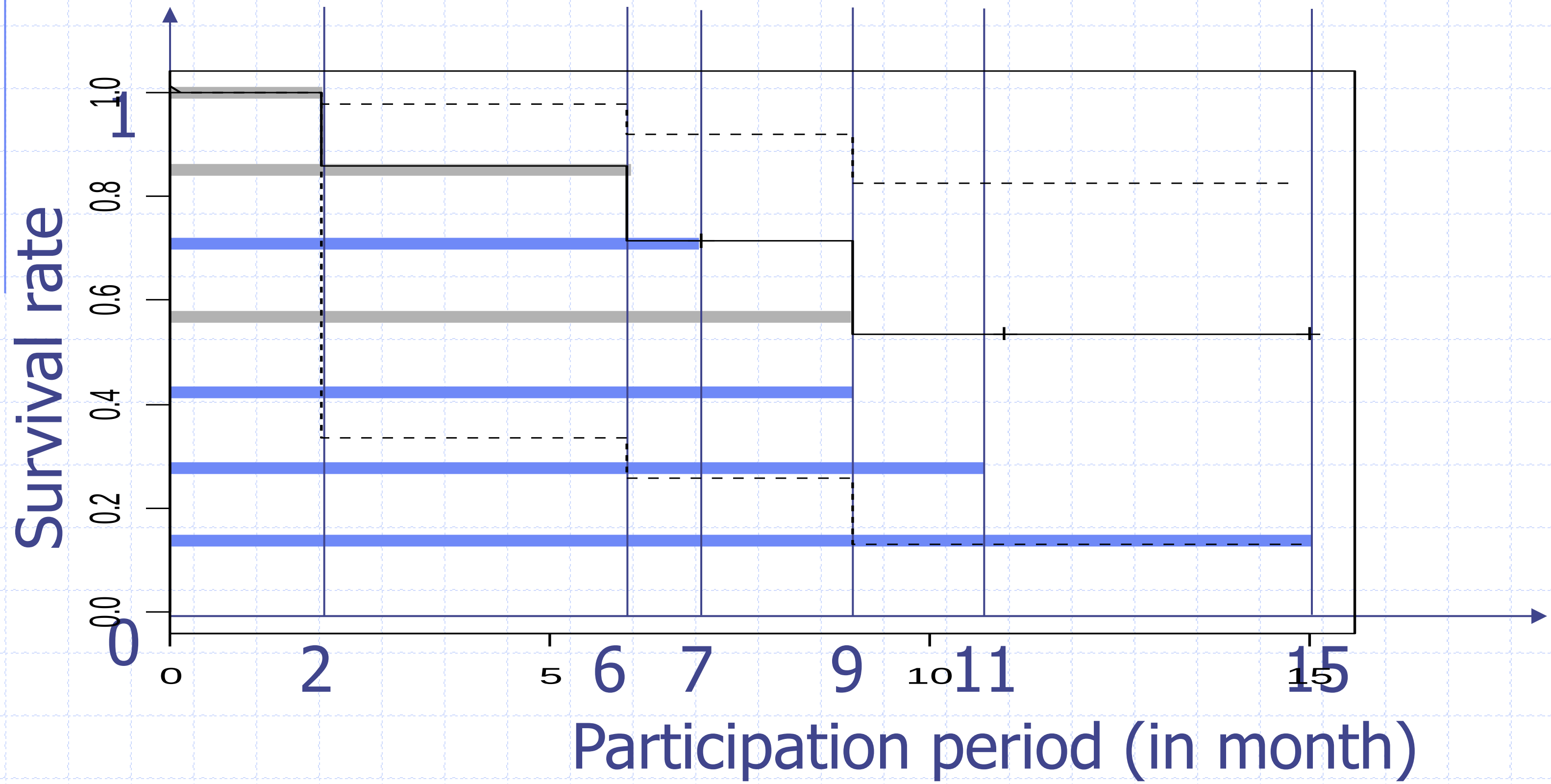


Kaplan-Meier method

t_i	$[t_i,t_{i+1}[$	N_i	D_i	C_i	$S_{i\|\ i-1}$	S_i	$varS_i$
0	[0,2[	7	0	0	1	1	-
2	[2,6[	7	1	0	6/7=0.857	0.857	0.017
6	[6,7[	6	1	0	5/6=0.833	0.833x 0.857=0.714	0.029
7	[7,9[	5	0	1	1	0.714	0.029
9	[9,11[	4	1	1	3/4=0.75	0.714x 0.75=0.536	0.040
11	[11,15]	2	0	1	1	0.536	0.040

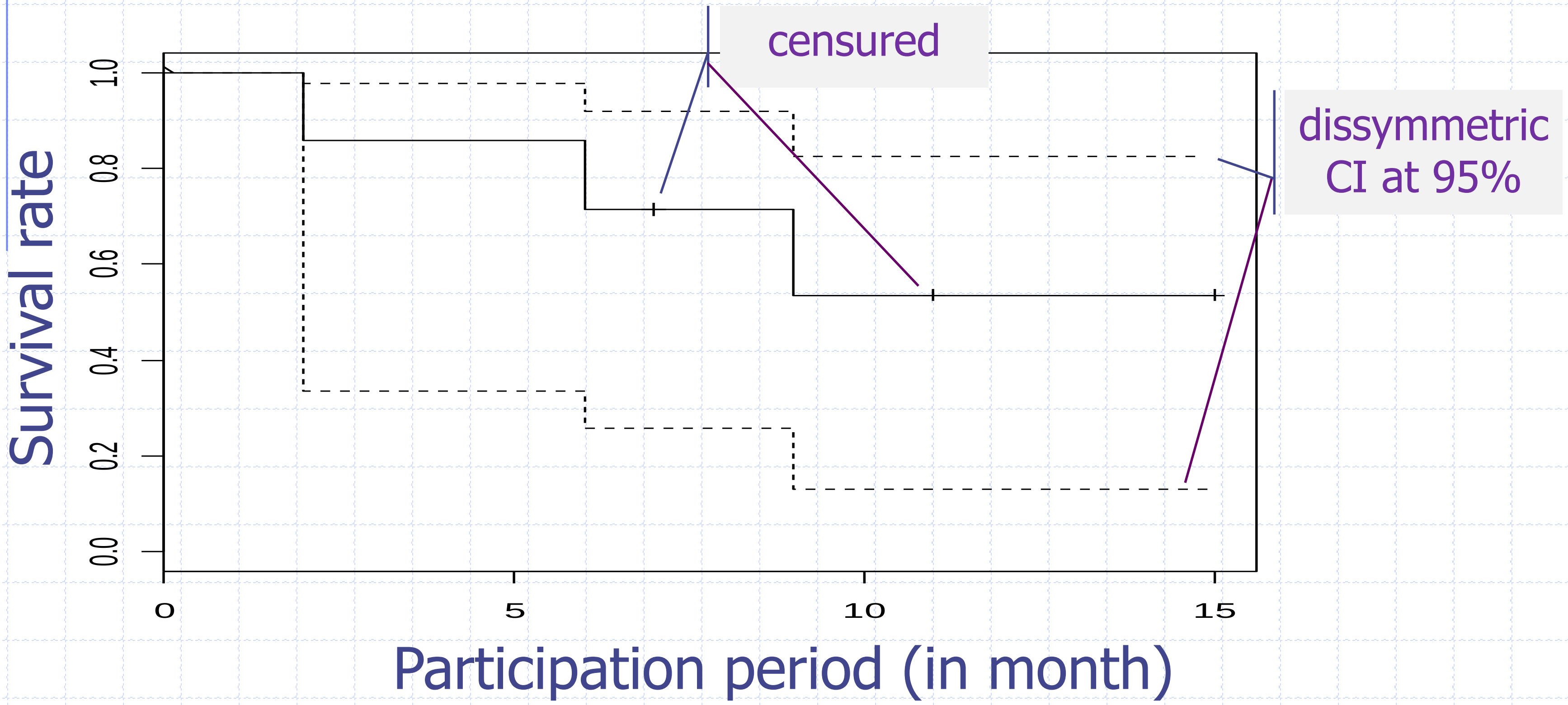
Graphic representation of the survival curve (Kaplan-Meier)

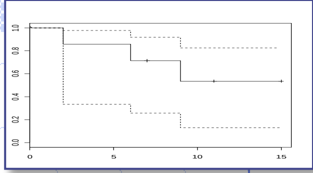




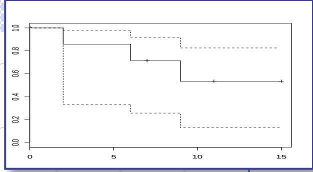
Graphic representation of the survival curve (Kaplan-Meier)

```
library(survival)
plot(survfit(Surv(PP,dead)~1,conf.type="log-log"))
```





Graphic representation of the survival curve (Kaplan-Meier)



```
library(survival)
plot(survfit(Surv(PPn,d11$dead)~1,conf.type="log-log"))
```

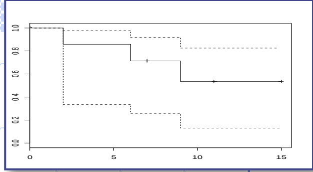
Participation period

Dichotomic variable: dead/alive status atthe end of the par

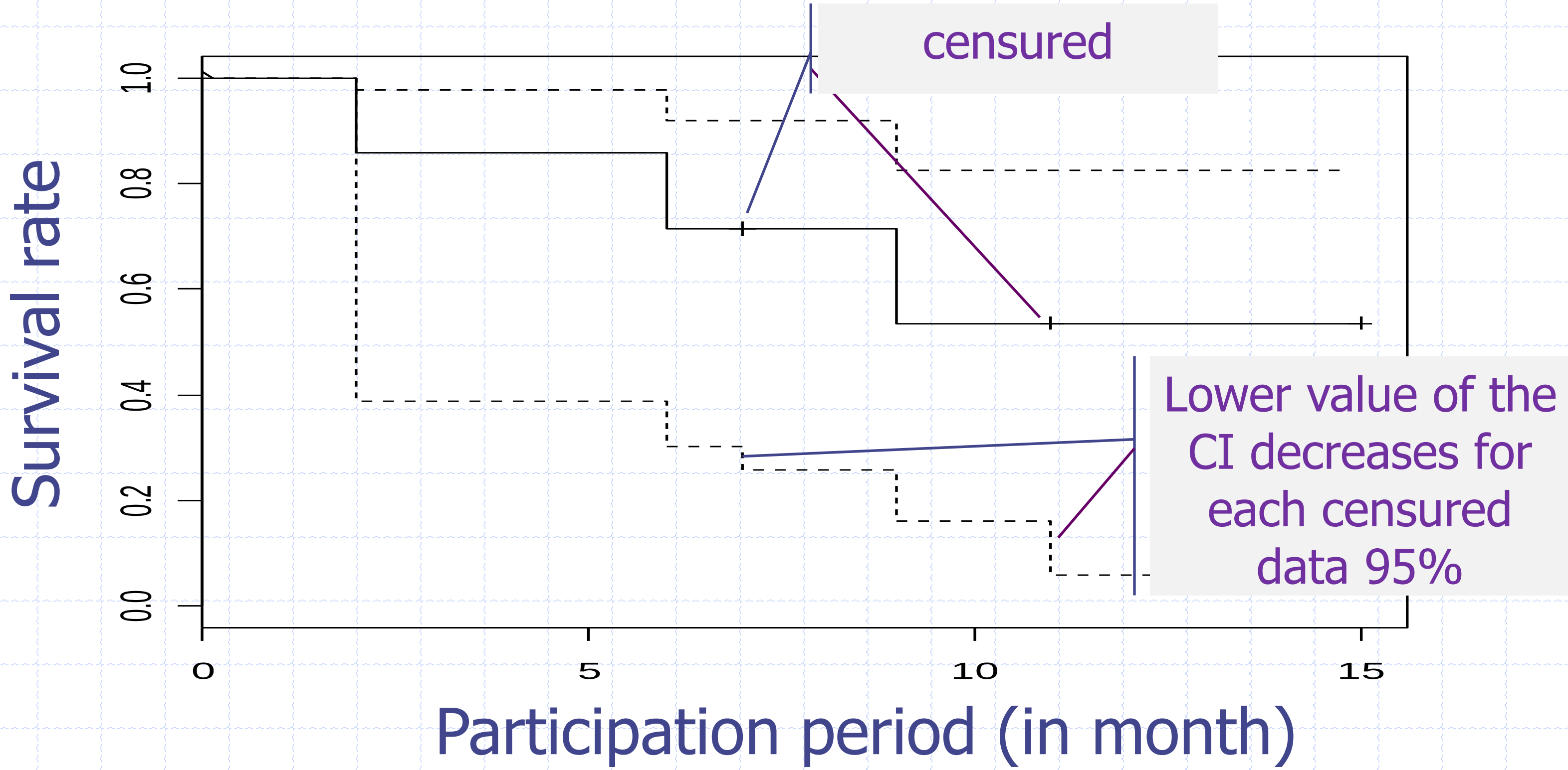
`surfit()` of the package survival computes all the value of the survival curve

The package survival need to be previously download on the computer and declare it.

Graphic representation of the survival curve (Kaplan-Meier)

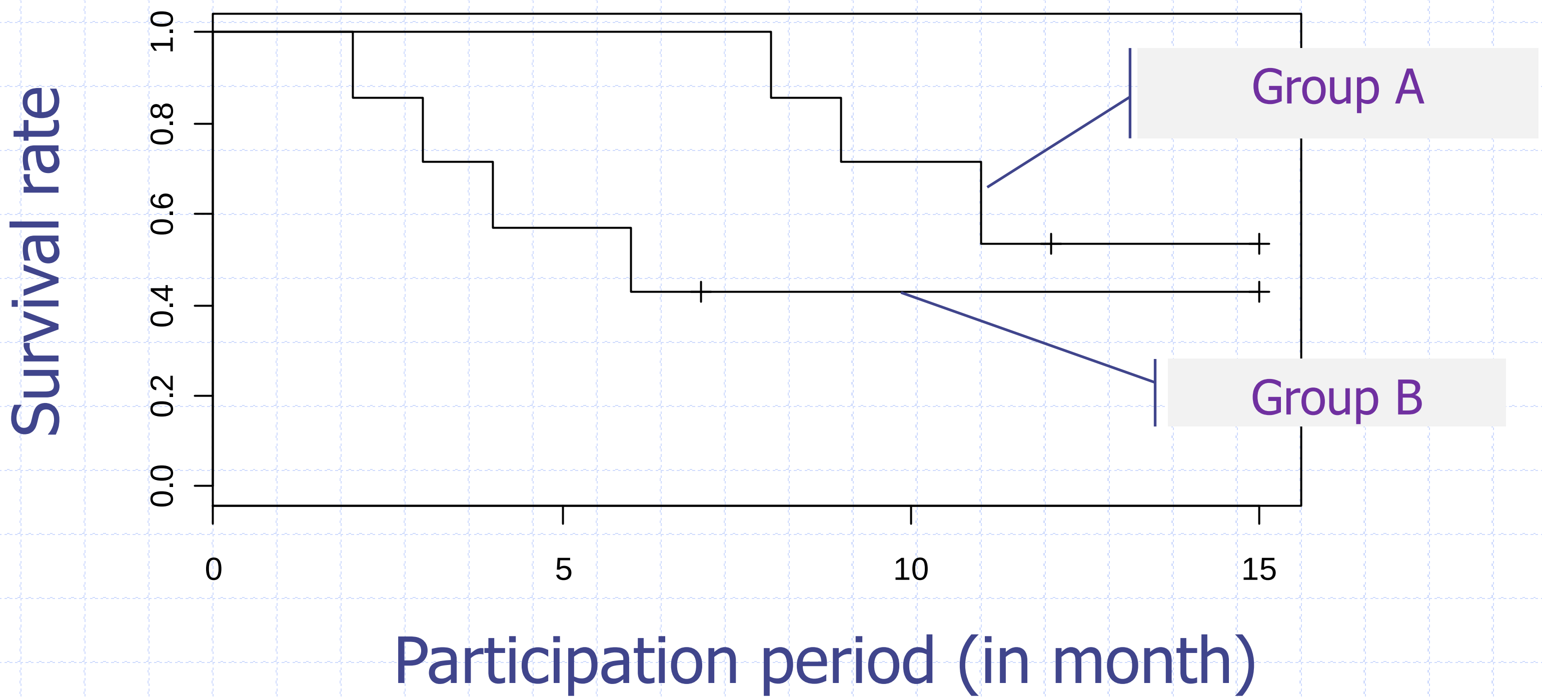


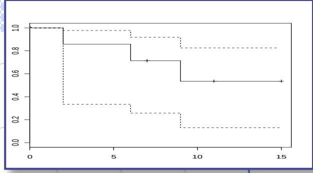
```
plot(survfit(Surv(PP,dead)~1,conf.type="log-log",conf.lower="modified"))
```



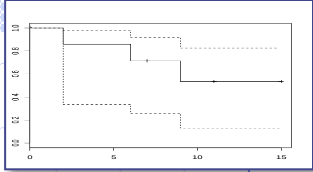
Graphic representation of survival curves for different groups

```
plot(survfit(Surv(PP,dead)~group,conf.type="log-log"))
```





Exercise 8a

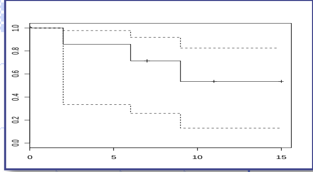


Practical on survival curve

A cohort clinical study follows more than 600 HIV sero-positive people, it is noted all the information about the date of origin (DO), the date of death (DCD).

a) Calculate the participation period for every people.

```
d10$DLNn<-as.Date(d10$DLN,"%d/%m/%Y")
d10$DIn<-as.Date(d10$DCD,"%d/%m/%Y")
d10$DOn<-as.Date(d10$DO,"%d/%m/%Y")
d10$DEn<-pmin(d10$DLNn,d10$DIn,na.rm=T)
PP<-d10$DEn-d10$DOn
PPn<-as.numeric(PP)
```

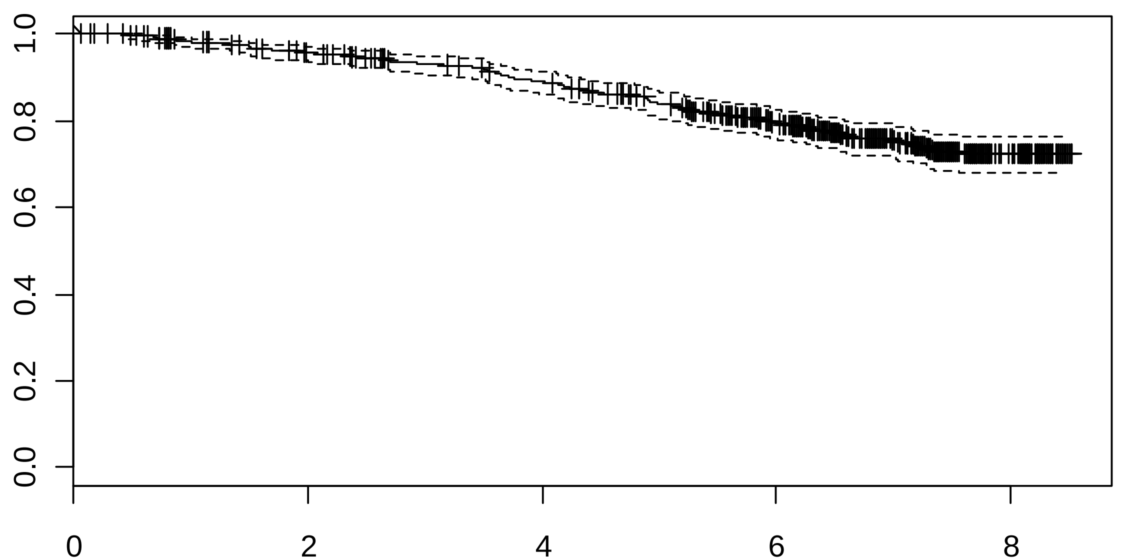


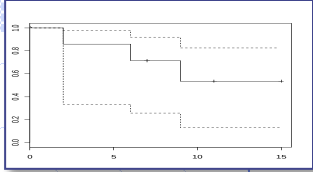
Exercise 8b

Practical on survival curve

b) Draw the survival curve.

```
plot(survfit(Surv(PPn,d10$dead)~1,conf.type="log-log"))
```



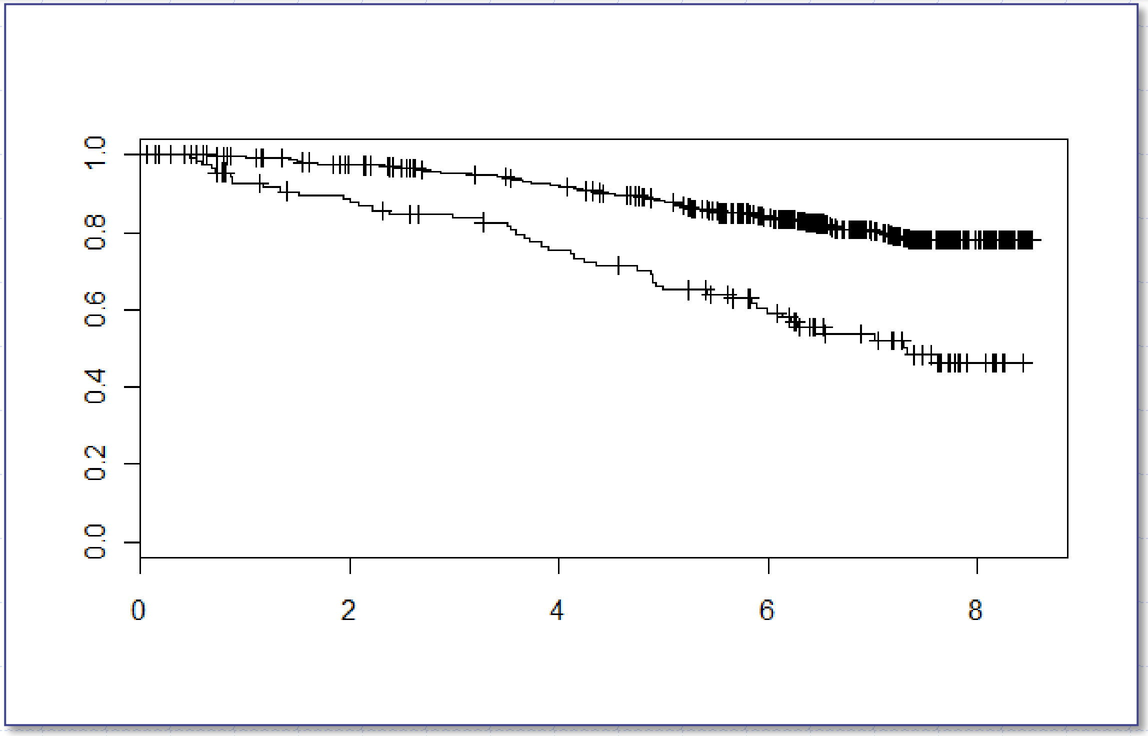


Exercise 8c

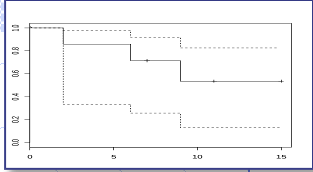
Practical on survival curve

c) Draw the survival curve according the categorie “AGE_CLASS”. AGE_CLASS=0 means younger than 40; AGE_CLASS=1 means equal or older than 40.

```
plot(survfit(Surv(PPn,d10$dead)~d10$AGE_CLASS,conf.type="log-log"))
```



Exercise 8d



Practical on survival curve

d) Read all the information about the survival curves.

```
summary(survfit(Surv(PPn,d10$dead)~d10$AGE_CLASS,conf.type="log-log"))

Call: survfit(formula = Surv(PP, d10$dead) ~ d10$AGE_CLASS, conf.type =
"log-log")

                d10$AGE_CLASS=0
  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.698    486       1    0.998 0.00206        0.985        1.000
 0.997    479       1    0.996 0.00292        0.984        0.999
 1.005    478       1    0.994 0.00358        0.981        0.998
 1.333    474       1    0.992 0.00414        0.978        0.997
 1.418    472       1    0.990 0.00464        0.975        0.996
 1.473    471       1    0.987 0.00508        0.972        0.994
...
```

To Finish….

…. Learn more

Learn more

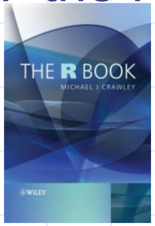
- In R software:

  `?name of the function`

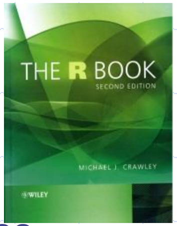
  Ex. `? cor.test`

- In web:
  - R cran: http://www.r-project.org/
  - Statistics with R of Vincent Zoonekynd: google + statistics with R
- With pdf documents:
  - R for Beginners of E. Paradis: google + R for beginners
- With Books:
  - There is now a lot of books on R and …
  - … there is the "Bible": the R book of Michael Crawley

First version
950 pages pdf available for free



Second version
1076 pages pdf available for 81 €



Exercise a

Learn more and get help

Look at the help on chi-squared test on R program.

chisq.test {stats} R Documentation

Pearson's Chi-squared Test for Count Data

Description

`chisq.test` performs chi-squared contingency table tests and goodness-of-fit tests.

Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Arguments

`x`	a numeric vector or matrix. `x` and `y` can also both be factors.
`y`	a numeric vector; ignored if `x` is a matrix. If `x` is a factor, `y` should be a factor of the same length.
`correct`	a logical indicating whether to apply continuity correction when computing the test statistic for 2 by 2 tables: one half is subtracted from all |*O* - *E*| differences; however, the correction will not be bigger than the differences themselves. No correction is done if `simulate.p.value = TRUE`.
`p`	a vector of probabilities of the same length of `x`. An error is given if any entry of `p` is negative.
`rescale.p`	a logical scalar; if TRUE then `p` is rescaled (if necessary) to sum to 1. If `rescale.p` is FALSE, and `p` does not sum to 1, an error is given.
`simulate.p.value`	a logical indicating whether to compute p-values by Monte Carlo simulation.
`B`	an integer specifying the number of replicates used in the Monte Carlo test.

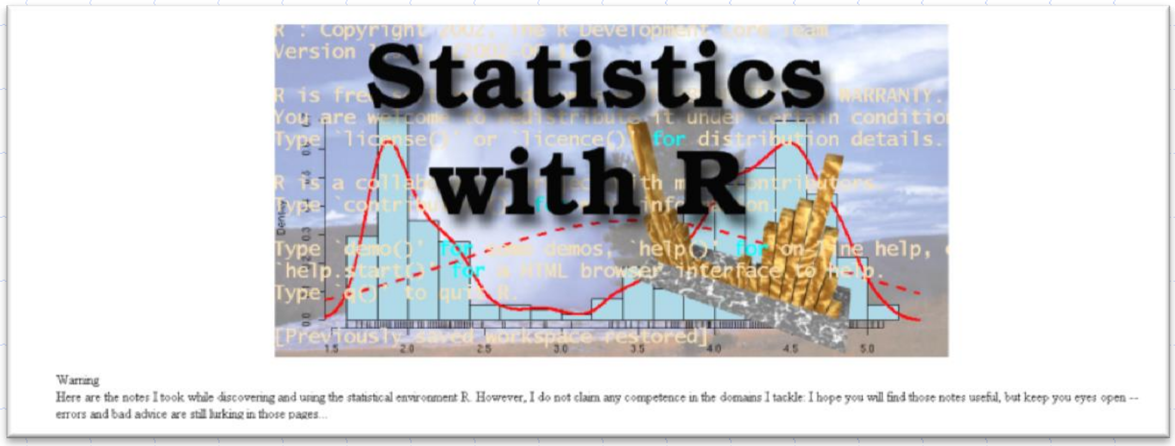
Details

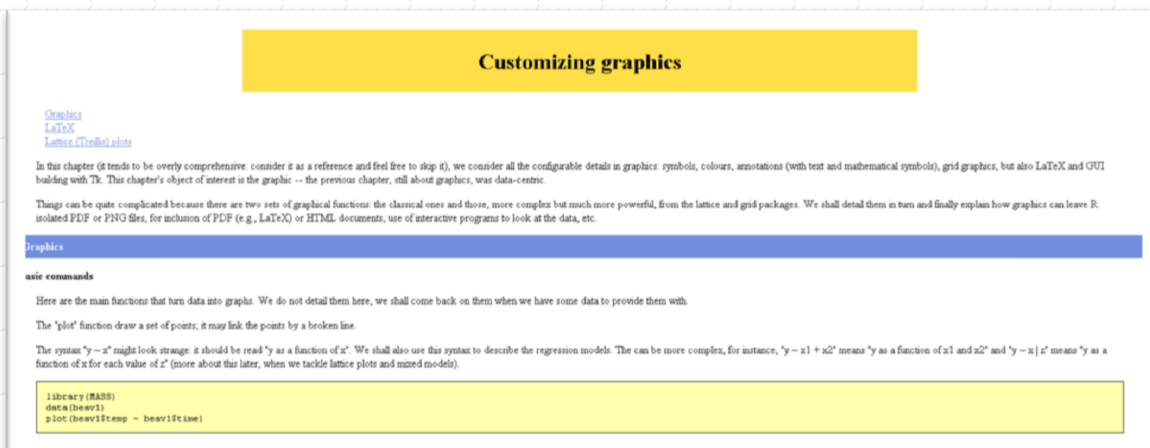
If `x` is a matrix with one row or column, or if `x` is a vector and `y` is not given, then a *goodness-of-fit test* is performed (`x` is treated as a one-dimensional contingency table). The entries of `x` must be non-negative integers. In this case, the hypothesis tested is whether the population

Exercise b

Learn more and get help

Look for the customizing graphic section in the website statistics with R

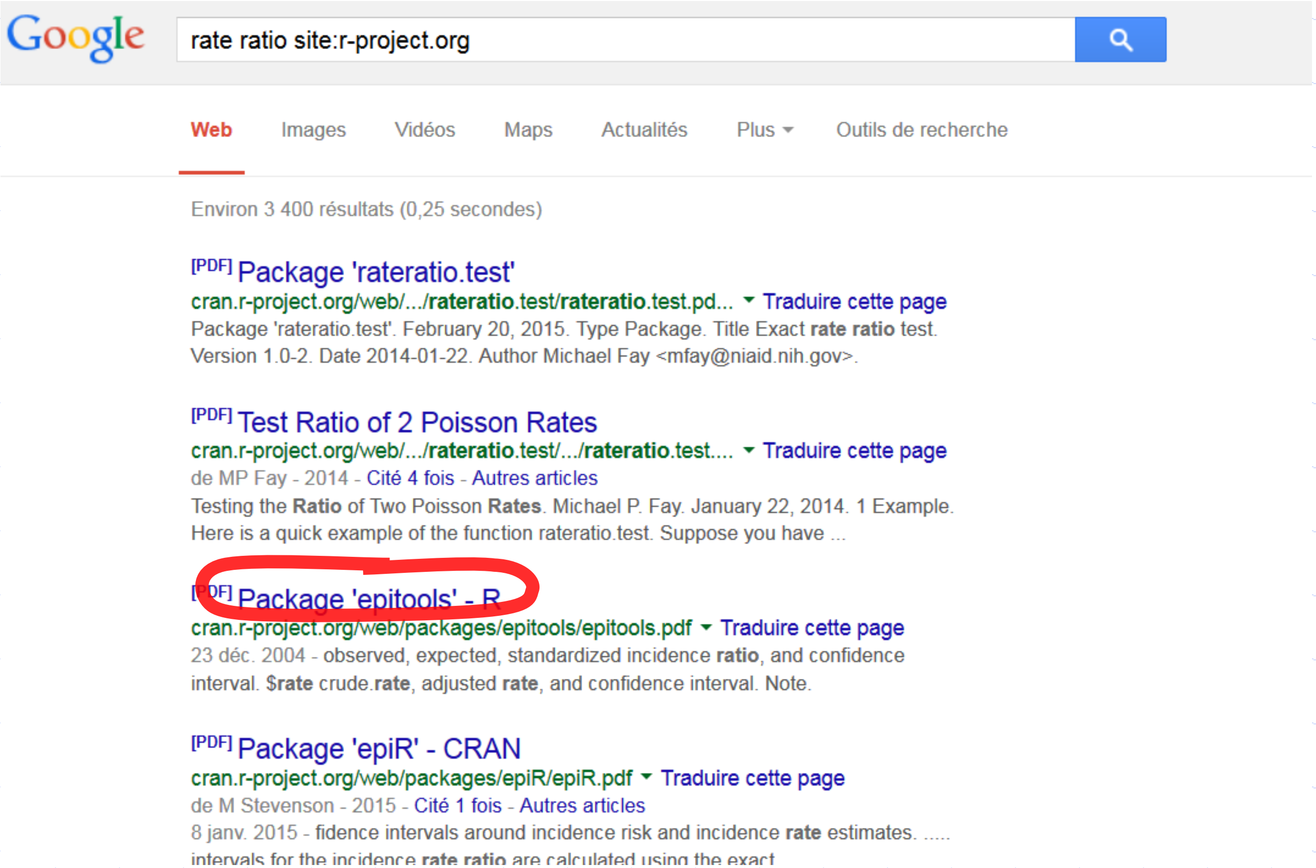




Exercise C

Learn more and get help

Make a research on risk ratio in the website of R cran project.



Thank you for your attention

ขอขอบคุณที่ท่านสนใจ