université PARIS DIDEROT PARIS 7

R programming

Club Bioinfo - Institut Jacques Monod

Leslie REGAD and Gaëlle LELANDAIS

Mails :
leslie.regad@univ-paris-diderot.fr ;
gaelle.lelandais@univ-paris-diderot.fr

Document version: 28/09/12

Statistics with R

Section 1

Random Sampling

- Principle
  - Getting a set of numbers according to a predefined statistical distribution (uniform distribution, normal distribution, etc.)
- Numbers are chosen according to a random process (repetitions give different results)

`sample ( )`
"takes a sample of the specified size from the elements of x using either with or without replacement"

`rnorm ( )`
"random generation for the normal distribution with mean equal to mean and standard deviation equal to sd"

Function « sample() »

- Example 1
  - Picking 4 numbers in a set of values comprised between 1 and 40

```
> sample(1:40,4)
[1] 26  6 25 34
> sample(1:40,4, replace=TRUE) # sampling with replacement
[1]  7 33 27  27
```

- Example 2
  - Simulate results of 10 tosses of a fair coin (« heads » and « Tails »)

```
> sample(c( "H", "T"), 10, replace=TRUE, prob=c(0.4,0.6))
 [1] "T" "T" "H" "T" "T" "H" "H" "H" "T" "H"
```

Function « rnorm() »

- Example 3
  - Generation of an artificial data vector of 10 normally distributed observations

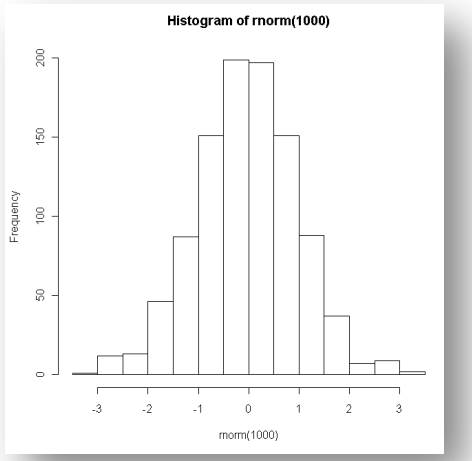
```
> rnorm(10)
 [1]  1.1451044 -1.1740811  2.1600010  0.8289392 -1.2881410
1.1022482 1.0495700 -0.4675296  0.3934182  1.0663837
```

  - Note : histogram of the values follow the probability density function when the size of the sample increases.

```
> hist(rnorm(100))
```

→

Default parameters,
mean = 0 and sd = 1



Summary Statistics

- Easy to calculate simple summary statistics
  - Mean, median, standard deviation, variance, empirical quantiles, etc.

```
mean ( )
"Generic function for the
(trimmed) arithmetic mean"
```

```
median ( )
"Compute the sample median"
```

```
var ( )
"Compute the variance of x "
```

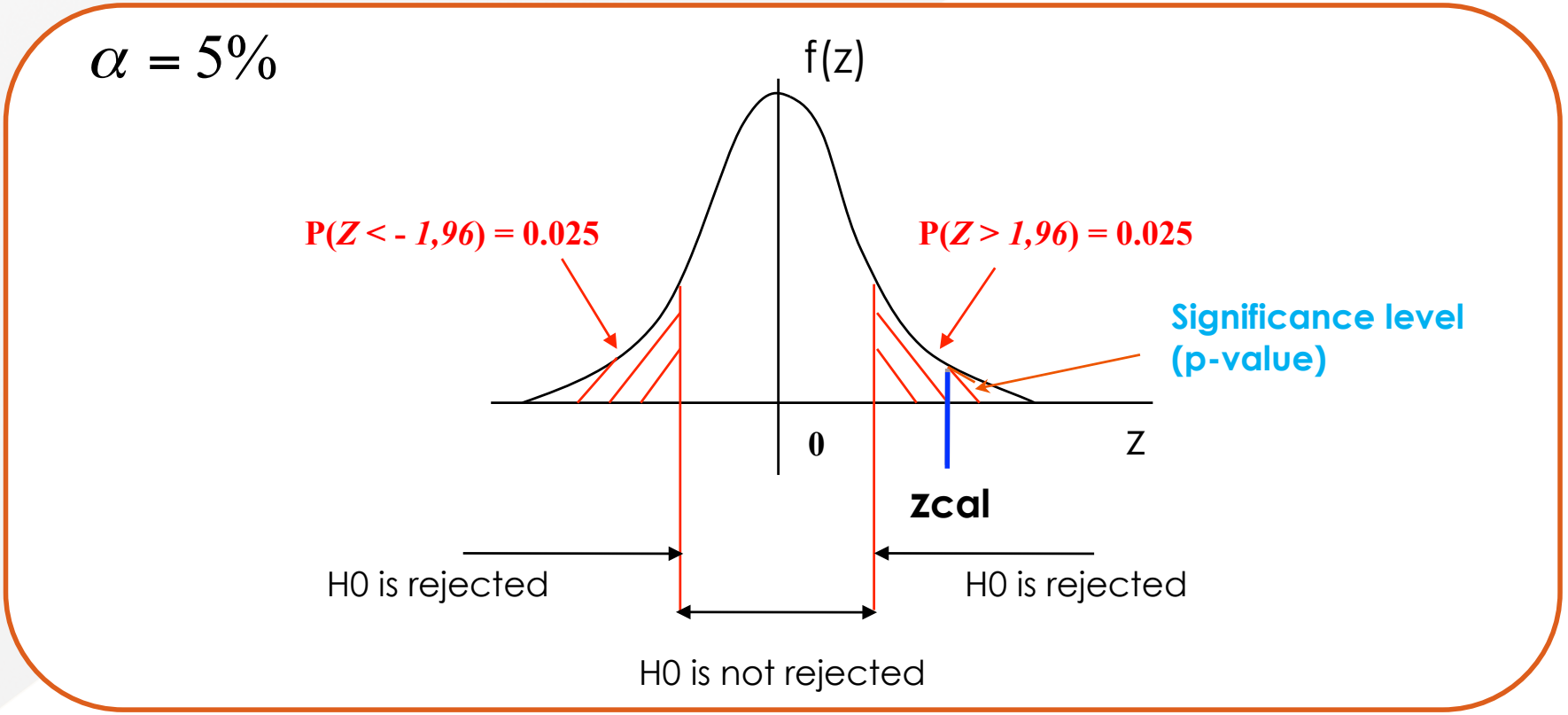
```
sd ( )
"This function computes the
standard deviation of the
values in x."
```

```
> data = rnorm(10)
> mean(data)
[1] 0.4275374
> median(data)
[1] 0.5847181
```

```
> var(data)
[1] 0.5645334
> sd(data)
[1] 0.7513544
```

Statistical Hypothesis Testing

- A statistical hypothesis test is a method of making decisions using data
  - The critical region of a hypothesis test is the set of all outcomes which cause the null hypothesis to be rejected in favor of the alternative hypothesis.



Testing for Differences

- Statistical procedure that consists in testing the hypothesis that two samples may be assumed to come from distributions with the same mean.

Hypotheses:

Bilateral Test

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Unilateral Test

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \text{ ou } \mu_1 > \mu_2 \end{cases}$$

```
t.test ( )
"Performs one and two sample
t-tests on vectors of data"
```

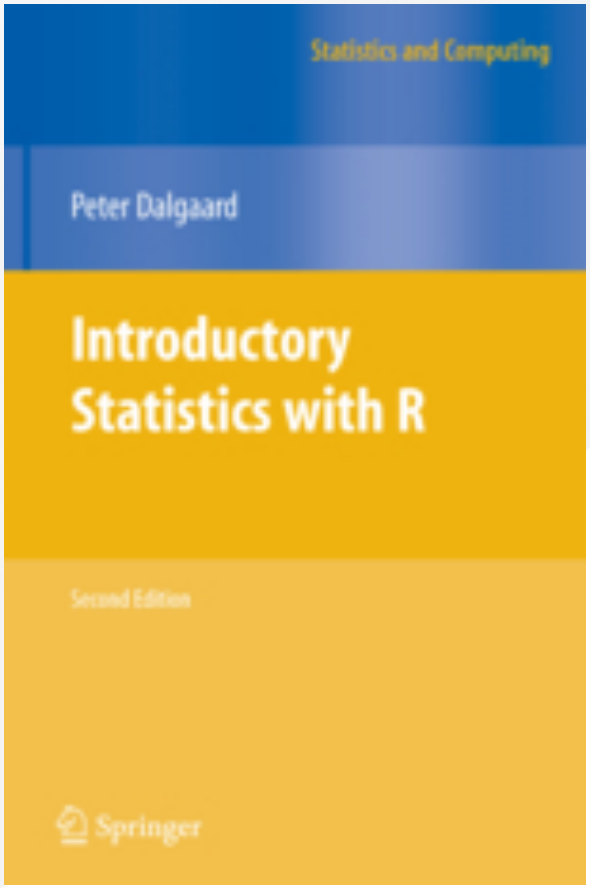
Function « t.test( ) »

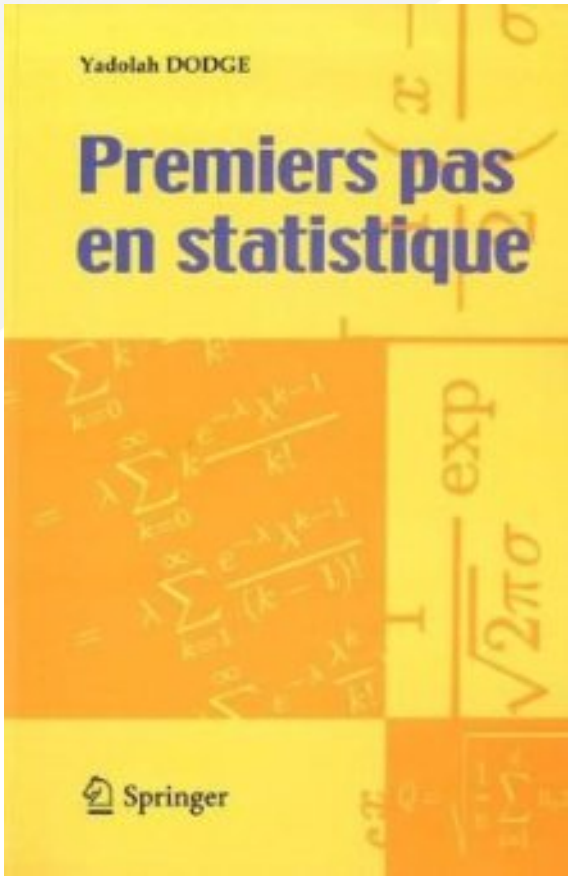

```
> data1 = rnorm(50)
> data2 = rnorm(40, mean = 4, sd = 1)
> t.test(data1, data2)

        Welch Two Sample t-test

data:  data1 and data2
t = -19.3996, df = 80.963, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -4.561019 -3.712457
sample estimates:
  mean of x   mean of y
-0.06069488  4.07604333
```

To go further…





Practical session

(applications for different statistical tests)

Leslie REGAD and Gaëlle LELANDAIS