Basic Concept of Statistics

Paolo Girardi and Livio Finos

22/10/2020

Contents

Hypothesis testing **2**
- Before we start (in R) . . . 2
- Diving reflex test - Dataset . . . 2

Measures of Dependence and the Simple linear model **3**
- Measuring the dependence . . . 3
- Covariance and Variance . . . 3
- Correlation . . . 3
- Linear Trend, the least squares method . . . 4
- Interpretation of the coefficients . . . 5

Permutation approach to Hypothesis Testing **5**
- Some remarks . . . 5
- Permutation tests - in a nutshell . . . 6
- All potential datasets . . . 7
- Random permutations . . . 7
- How likely $\hat{\beta}_1^{obs}$ was? . . . 9
- Calculation of the p-value . . . 9
- Interpretation . . . 9
- Composite alternatives (bilateral) . . . 10
- Some remarks . . . 10

Parametric Linear Model **10**
- From permutation tests (nonparametric) to parametric tests . . . 10
- The (simple) linear model . . . 11
- Hypothesis testing . . . 11
- Power is nothing without control . . . 12
- Terminology . . . 12
- Properties . . . 12

Descriptive Analysis and Test for Longitudinal Data **12**
- Local Field Potentials DATASET . . . 12
- Univariate Analisys . . . 13
  - by time . . . 14
  - plot all the timeseries . . . 15
  - correlation between sites . . . 16
  - cross-correlation at different lag . . . 16
- A simple test . . . 18

Ph.D. Course in Neuroscience
Calendar of the Basic Courses – Academic Year 2020-2021
Basic Concept of Statistics

Lesson 4 - Introduction to testing

Hypothesis testing

Before we start (in R)

```
#clean the memory
rm (list=ls ())
ECHO=FALSE
# customize the output of knitr
knitr :: opts_chunk$set (fig.align="center")#, fig.width=6, fig.height=6)
```

Diving reflex test - Dataset

The diving reflex, also known as the diving response and mammalian diving reflex, is a set of physiological responses to immersion. One of the main outcomes is a heart rate reduction. The effect is enhanced by the water temperature.

We measure the reduction in the heart rate of 10 subjects (children) for different temperature of the water (in Fahrenheit degree).

The values of `Temperature` are in Fahrenheit degree. The values of `Reduction` are in beat per second.

(data are fictitious)

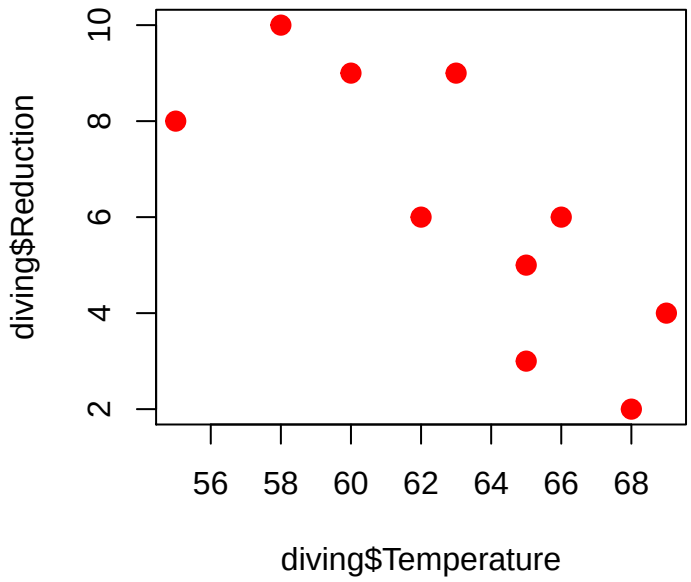
To read the data

```
setwd("/Users/Paolo/Dropbox/Dottorato_Neurosciences/2020_2021")
diving<-read.csv("diving_reflex.csv")
str(diving)
```

```
## 'data.frame':    10 obs. of  3 variables:
##  $ Children   : int  1 2 3 4 5 6 7 8 9 10
##  $ Temperature: int  68 65 66 62 60 55 58 65 69 63
##  $ Reduction  : num  2 5 6 6 9 8 10 3 4 9
```

We plot the data

```
plot(x=diving$Temperature,y=diving$Reduction,pch=20,col=2,cex=2)
```



Measures of Dependence and the Simple linear model

Measuring the dependence

we define:

- $X = Temperature$
- $Y = Reduction$

We review some famous index to measure the (linear) dependence among two variables

Covariance and Variance

Covariance between X and Y:

$\sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$

- values between $-\infty$ and ∞
- $\sigma_{xy} \approx 0$: there is no dependency between X and Y
- $\sigma_{xy} >> (<<) 0$: there is a strong positive (negative) dependency between X and Y

Correlation

With the Covariance it is difficult to understand when the relationship between X and Y is strong / weak. We note that

$-\sigma_x\sigma_y \leq \sigma_{xy} \leq \sigma_x\sigma_y$ is quivalent to $-1 \leq \frac{\sigma_{xy}}{\sigma_x\sigma_y} \leq 1$

Correlation between X and Y:

$\rho_{xy} = \frac{\sigma xy}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

- values between -1 and 1
- $\rho_{xy} \approx 0$: there is no dependency between X and Y
- $\rho_{xy} \approx 1 (-1)$: there is a strong positive (negative) dependency between X and Y

Linear Trend, the least squares method

We describe the relationship between
`Reduction` and `Temperature` with a straight line.

$Reduction \approx \beta_0 + \beta_1 Temperature$
$Y = \beta_0 + \beta_1 X$

Let's draw a line 'in the middle' of the data.

The **least-squares estimator**

We look for the one that passes more 'in the middle', the one that minimizes the sum of the squares of the residues:

$\hat{\beta}_0$ and $\hat{\beta}_1$ such that
$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ is minimum.

Estimates:

- Angular coefficient: $\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_{xx}} = \rho_{xy}\frac{\sigma_y}{\sigma_x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ = -0.481628
- Intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ = 36.5907292
- Response (estimated y): $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residuals (from the estimated response): $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$

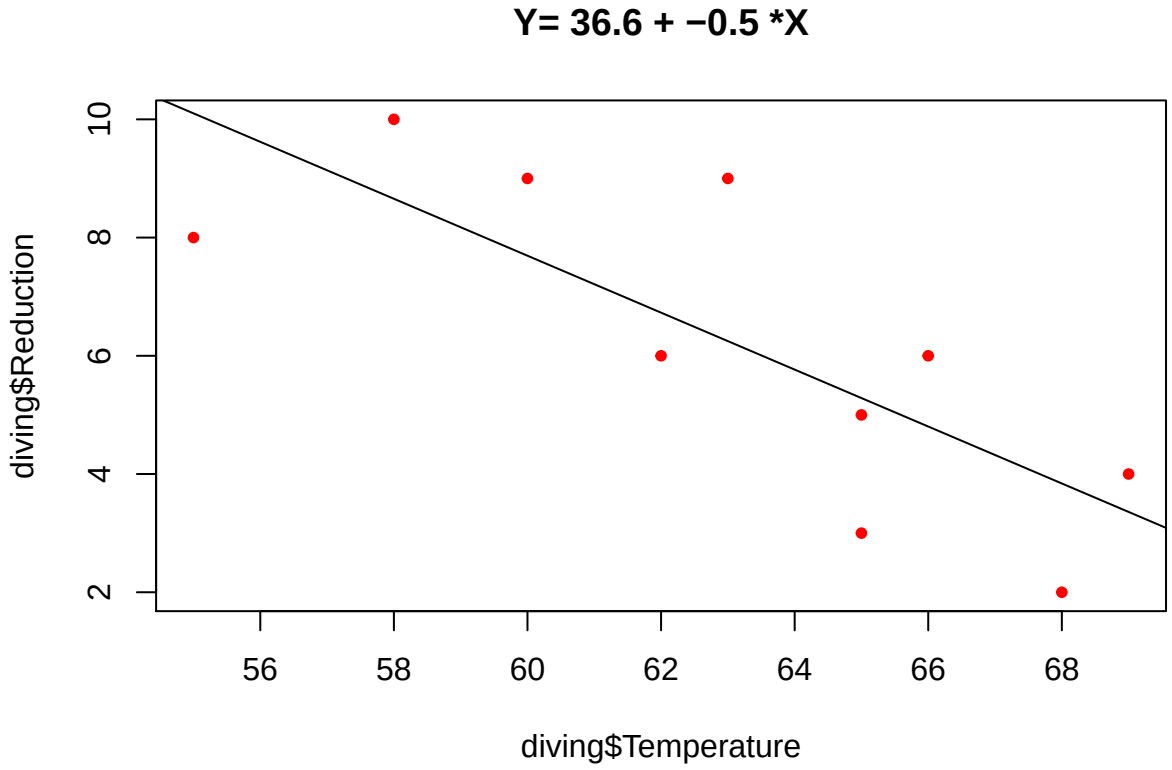
and therefore the least squares are the sum of the squared residuals: $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

A graphical representation:

```
# lm() is the function to estimate a classical linear model in R
model=lm(Reduction~Temperature,data=diving)
coefficients(model)
```

```
## (Intercept) Temperature
##   36.590729   -0.481628
```

```
plot(diving$Temperature,diving$Reduction,pch=20,col=2,cex=1)
coeff=round(coefficients(model),1)
title(paste("Y=",coeff[1],"+",coeff[2],"*X"))
abline(model,col=1)
```



Interpretation of the coefficients

- β_0 indicates the value of y when $x = 0$ (where the line intersects the ordinate axis).
- β_1 indicates how much y grows or decreases as a unit of x grows
  - If $\beta_1 = 0$ there is no relation between x and y.$Y$is constant (horizontal), knowing$x$does not change the estimate of y
  - If $\beta_1 > (<)0$ the relation between x and y is positive (negative). When X passes from x a $x + 1$ the estimate of Y changes from $\hat{y}$ to $\hat{y} + \hat{\beta}_1$

Permutation approach to Hypothesis Testing

Some remarks

Let's note that all the measures above does not make any assumptions on the random process that generate them.

Let's assume that Y - and possibly X - is not fix, while it is generated by a random variable.

The question: **Is there a relationship between Y and X?**

We estimated $\hat{\beta}_1$ = -0.481628

but the **true value** β_1 is really different from 0 (i.e. no relationship)?
Otherwise, is the distance to 0 is due to the random sampling?

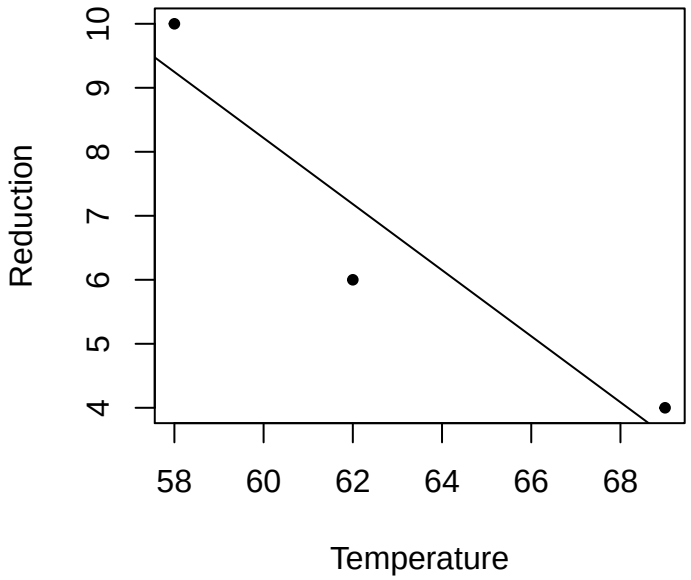
- **Null Hypothesis** H_0 : $\beta_1 = 0$ (the **true** β_1, not its estimate $\hat{\beta}_1$!). There is no relationship between X and Y.
- **Alternative Hypothesis** H_1 : $\beta_1 < 0$ The relationship is negative

Other possible specifications of H_1 : $\beta_1 > 0$ and, more commonly, H_1 : $\beta_1 \neq 0$.

Permutation tests - in a nutshell

As a toy example, let use a sub-set of the data of 3 subjects:

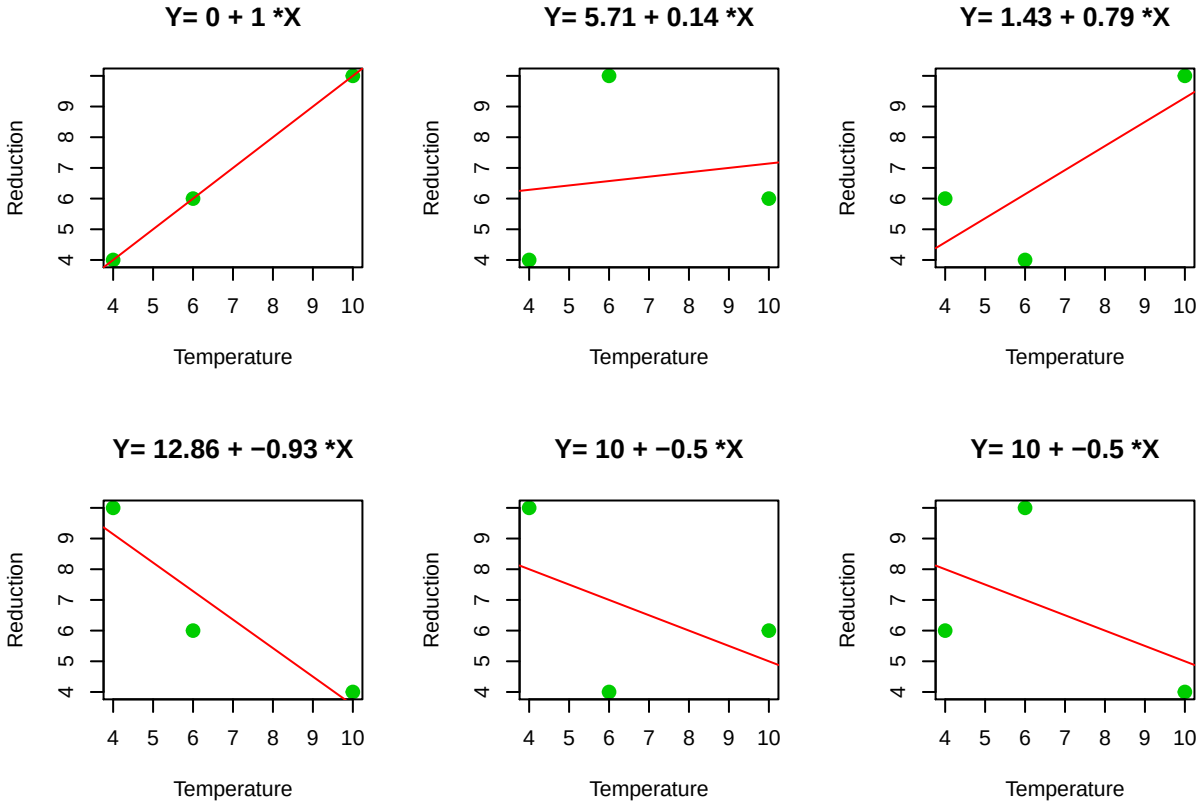
```
##   Children Temperature Reduction
## 9        9          69         4
## 4        4          62         6
## 7        7          58        10
```



- *If* H_0 is true: there is no linear relationship between X and Y
- Therefore, the trend observed on the data is due to chance.
- Any other match of x_i and y_i was equally likely to occur
- I can generate the datasets of other hypothetical experiments by exchanging the order of the observations in Y.
- How many equally likely datasets could I get with$X$ and Yobserved? $3 * 2 * 1 = 3! = 6$ possible datasets.

Remark: Here we only assume that y is a random variable. The only assumption here is the exchangeability of the observations: the joint density $f(y_1, \ldots, y_n)$ does not change when the ordering of $y_1, \ldots, y_n$ is changed.

All potential datasets



Random permutations

In our data set, if we apply the same principle...

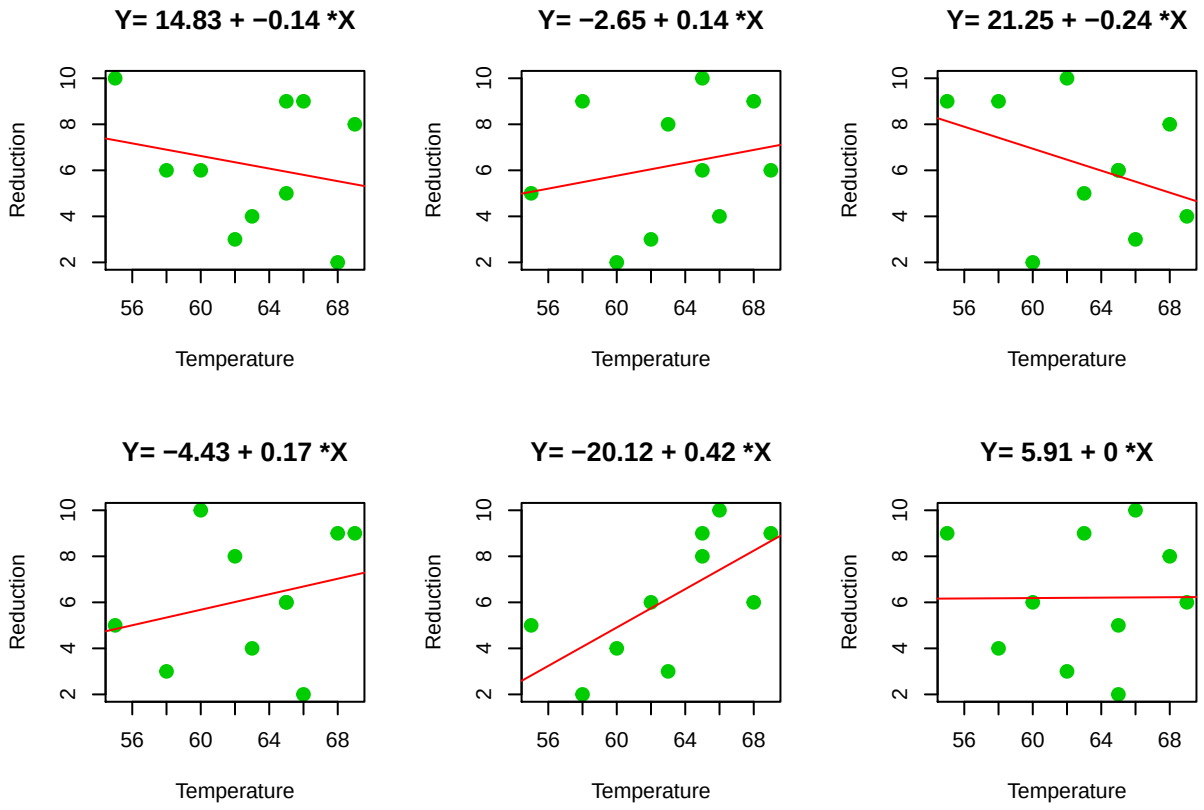
How many permutations of the vector $y_1, \ldots, y_n$ are possible? $n! = 10! = 3628800$.

big, perhaps not too big ... but what happen with, for example, $n = 20$? We got $20! = 2.432902e + 18$. This is too big, definitely!

We calculate a smaller (but sufficiently large) B of random permutations.

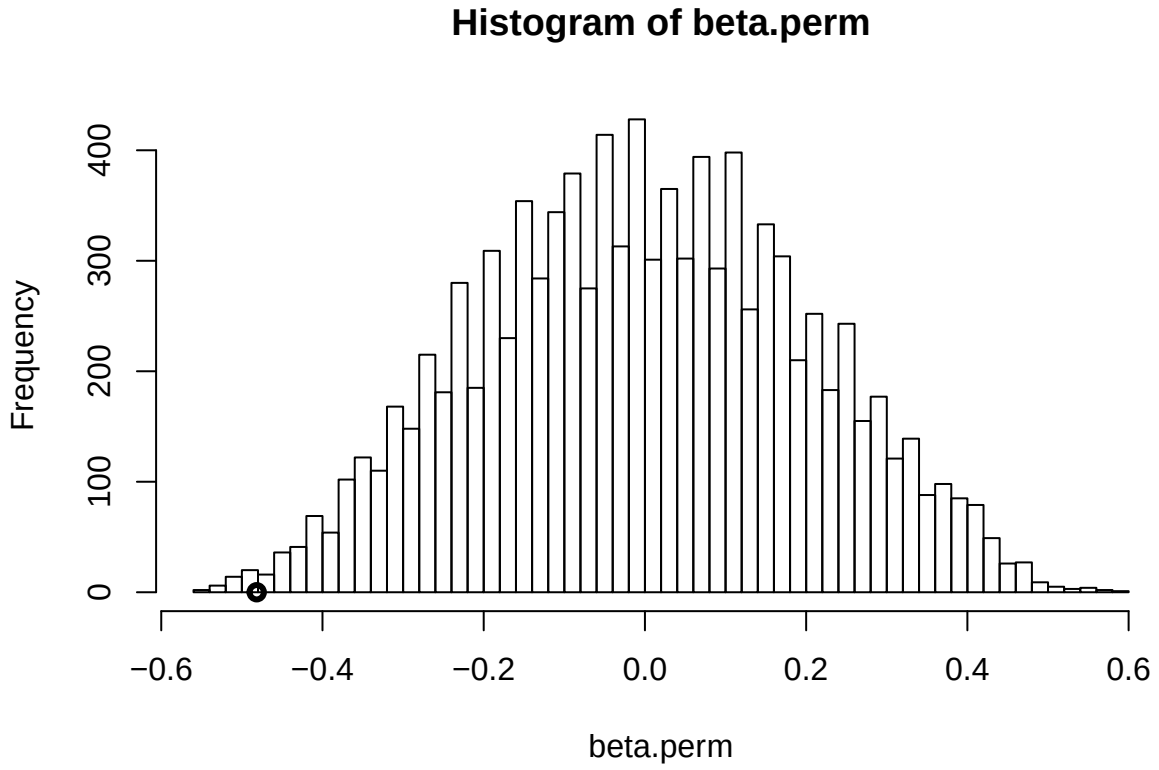
here some example

`Temperature` vs a permutations of `Reduction`



We repeat 10^4 times and we look at the histogram of the $\hat{\beta}_1$

```
# beta_1 estimated on the observed data:
beta1=coefficients(lm(Reduction~Temperature,data=diving))[2]
# function that permutes the y values and calculates the coeff beta_1
my.beta.perm <- function(Y,X){
  model=lm(sample(Y)~X)
  coefficients(model)[2]
}
#replicate it B-1 times
beta.perm= replicate(B,my.beta.perm(diving$Reduction, diving$Temperature ))
```



How likely $\hat{\beta}_1^{obs}$ was?

(before the experiment!)

How likely was it to get a $\leq \hat{\beta}_1^{obs}$ value among the many possible values of $\hat{\beta}_1^{*b}$ (obtained by permuting data)?

Remarks:

- $\hat{\beta}_1^{*b} < \hat{\beta}_1^{obs}$ (closer to 0): less evidence against H_1 than $\hat{\beta}_1^{obs}$
- $\hat{\beta}_1^{*b} \geq \hat{\beta}_1^{obs}$: equal or more evidence towards H_1 than $\hat{\beta}_1^{obs}$

Calculation of the p-value

Over B=10^4 permutations we got 41 times a $\hat{\beta}_1^{*b} \leq \hat{\beta}_1^{obs}$.

The p-value (significance) is $p = \frac{\#(\hat{\beta}_1^{*b} \geq \hat{\beta}_1^{obs})}{B+1} = 0.0041$ In our example is

```
mean((beta.perm<=beta1))
```

```
## [1] 0.00409959
```

Interpretation

The probability of $p = P(\hat{\beta}_1^* \leq \hat{\beta}_1 = \text{-}0.482 \mid H_0)$ is equal to $p = 0.0041$, i.e. very small.
So, it was unlikely to get a value like this **IF H_0 is true**.

Neyman-Pearson's approach has made common the use of a significance threshold for example $\alpha = .05$ (or $= .01$). When $p \leq \alpha$ rejects the hypothesis that there is no relationship between X and Y (H_0). If so, we are inclined to think that H_1 is true (there is a positive relationship).

- Type I error: False Positive
  the true hypo is H_0 (null correlation), BUT we accept H_1 (correlation is positive)

- Type II error: False Negative
  the true hypo is H_1 (positive correlation), BUT we do not reject H_0 (null correlation)

Type I error control

We want to guarantee not to get false relationships (a few false positives), better to be conservative. To make this, we want to bound the probability to make a false discovery:

$P(p - value \leq \alpha | H_0) \leq \alpha$

We built a machinery that in the long run (many replicates of the experiment) finds false correlations with probability α (e.g. $0.05 = 5\%$).

Composite alternatives (bilateral)

The hypothesis $H_1 : \beta_1 < 0$ (the relation is negative) must be justified with a priori knowledge.

More frequently, the Alternative hypothesis is appropriate: $H_1 : \beta_1 \neq 0$ (there is a relationship, I do not assume the direction)

I consider anomalous coefficients estimated as very small but also very large ('far from 0'). The p-value is $p = \frac{\#(|\hat{\beta}_1^{*b}| \geq |\hat{\beta}_1^{obs}|)}{B+1} = 0.0062994$

In our example

```
mean(beta.perm<beta1)+mean(beta.perm>-beta1)
```

```
## [1] 0.00629937
```

Some remarks

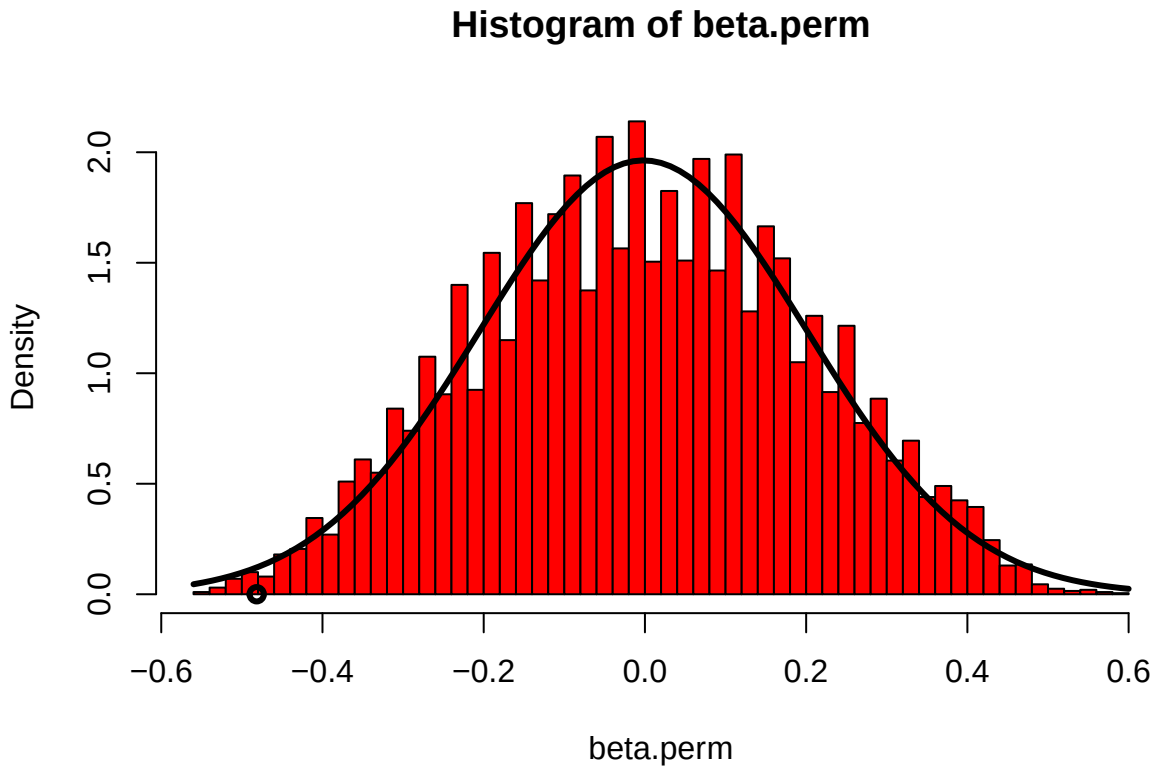
- Do not be confused with bootstrap methods. The former are extractions without reintegration, the latter with. The former have almost optimal properties and have (almost always) an exact control of the first type errors.
- A general approach and are applicable in many contexts. Very few assumptions.
- Some dedicated R packages:
- flip (the development version is on github)
- coin
- permuco
- They are of limited applicability when there are many variables involved.

Parametric Linear Model

From permutation tests (nonparametric) to parametric tests

We can see that the histogram of the statistical tests (calculated on the permuted data) is well described by a **Gaussian** (normal) curve.

```
hist(beta.perm,50,probability=TRUE,col=2)
curve(dnorm(x,mean(beta.perm),sd(beta.perm)),add=TRUE,col=1,lwd=3)
points(beta1,0,lwd=3,col=1)
```



The (simple) linear model

We assume that the observed values are distributed around true values $\beta_0 + \beta_1 X$ according to a Gaussian law:

$Y =$ linear part + normal error

$Y = \beta_0 + \beta_1 X + \varepsilon$

Assumptions of the linear model

- the $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$ the relationship between X and the true (mean) Y is linear.
- the **observations** are **independent** each others (knowing the value of the y_i observation does not help me to predict the value of y_{i+1}). The random part is ε_i, these are the independent terms.
- $\boldsymbol{\varepsilon_i \sim N(0, \sigma^2),\ \forall i = 1, \ldots, n}$ errors have normal distribution with zero mean and common variance (homoschedasticity: same variance).

Hypothesis testing

If these assumptions are true,

$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum (x_i - \bar{x})^2)$

We calculate the test statistic:

$t = \frac{\hat{\beta}_1}{std.dev\ \hat{\beta}_1} = \frac{\hat{\beta}_1}{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum (x_i - \bar{x})^2 / (n-2)}}$

If $H_0 : \beta_1 = 0$, $t \sim t(n-2)$ is true

On `reaction` data and $H_1 : \beta_1 \neq 0$ (bilateral alternative)

```
model=lm (Reduction ~ Temperature, data=diving)
summary(model) $ coefficients
```

```
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 36.590729  8.6644375  4.223093 0.002903553
## Temperature -0.481628  0.1370088 -3.515308 0.007899958
```

Similar result, but much more assumptions!

Power is nothing without control

We don't know if the data are generated under H_0 or under H_1.

But we have a tool (the test) that

- if the data are generated **under** H_0: it suggests (wrong) H_1 (i.e. $p \leq \alpha$, type I error, false positive) with probability α. E.g. $\alpha = .05$, low probability.
- if the data are generated **under** H_1: it suggests (correct) H_1 (i.e. true positive) with probability larger than α.
  This is the Power of a test. The Power is unknown, but we hope it is as high as possible.

Terminology

- Probability of **Type I error** (Probability of **False Positive**, α): the probability to find a relationship when it does not exist (true H_0, the test judges H_1).
- Probability of **Type II error** (Probability of **False Negative**): the probability NOT to find a relationship when it does exist (true H_1, the test judges H_0).
- **Specificity**: the probability NOT to find a relationship when it does NOT exist (true H_0, the test judges H_0). it is equal to 1 - Type I error.
- **Power** (**Sensitivity**): the probability to find a relationship when it does exist (true H_1, the test judges H_1). it is equal to 1 - Type II error.

Properties

If the parametric assumptions are valid, the test guarantes

- the control of the type I error at the α level,
- the maximum power (minimum error of type II β) among all the possible tests,
- asymptotic consistency (if they are under H_1 rejection always for sufficiently large n).

The permutation tests usually have slightly less power and converge to the corresponding parametric tests, IF they exist.

Descriptive Analysis and Test for Longitudinal Data

Local Field Potentials DATASET

The data are recordings of Local Field Potentials from an array with 220 channels inserted in the rat barrel cortex. The intracerebral local field potential (LFP) is a measure of brain activity that reflects the highly dynamic flow of information across neural networks.

The dataset contains 200 rows (# of channels) and 1000 columns (first 1000 time steps).

```
setwd("/Users/Paolo/Dropbox/Dottorato_Neurosciences/dataset")
lcp<-read.csv("lfpreduced_ex.csv",header=TRUE)
lcp<-as.matrix(lcp)
```

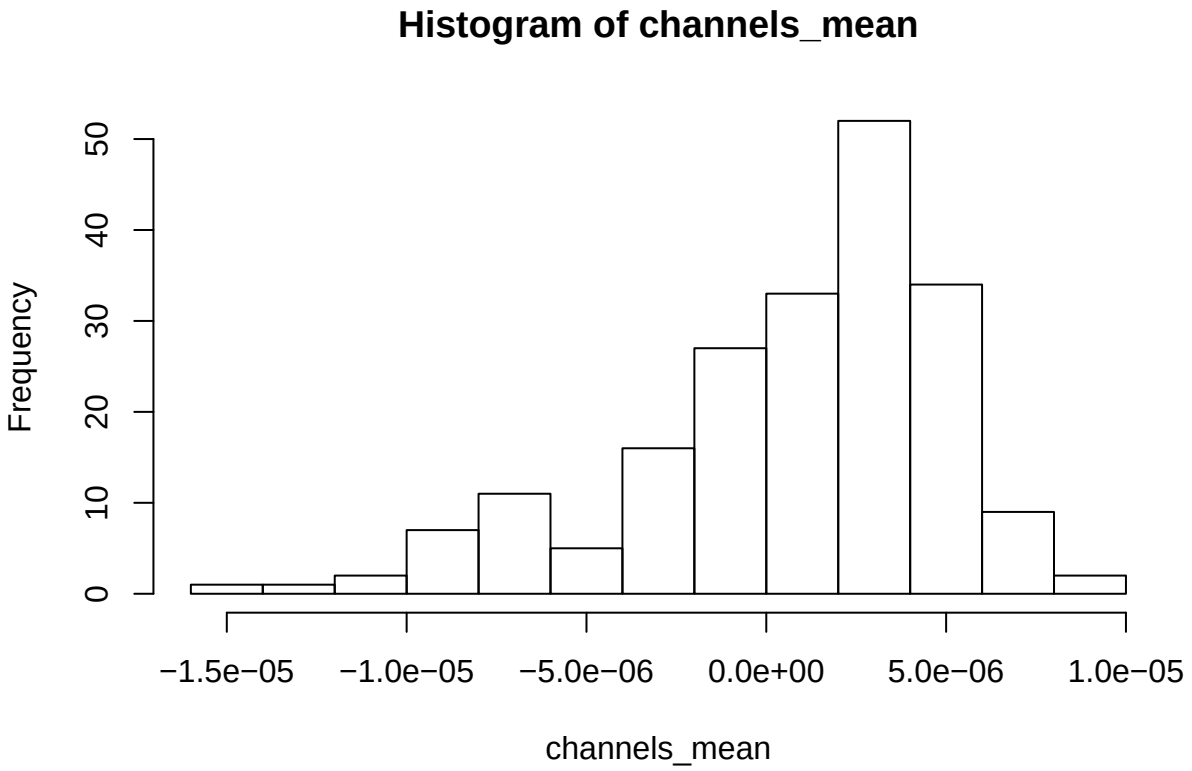
Univariate Analisys

There are 2 dimensions, channels and time ### by channels

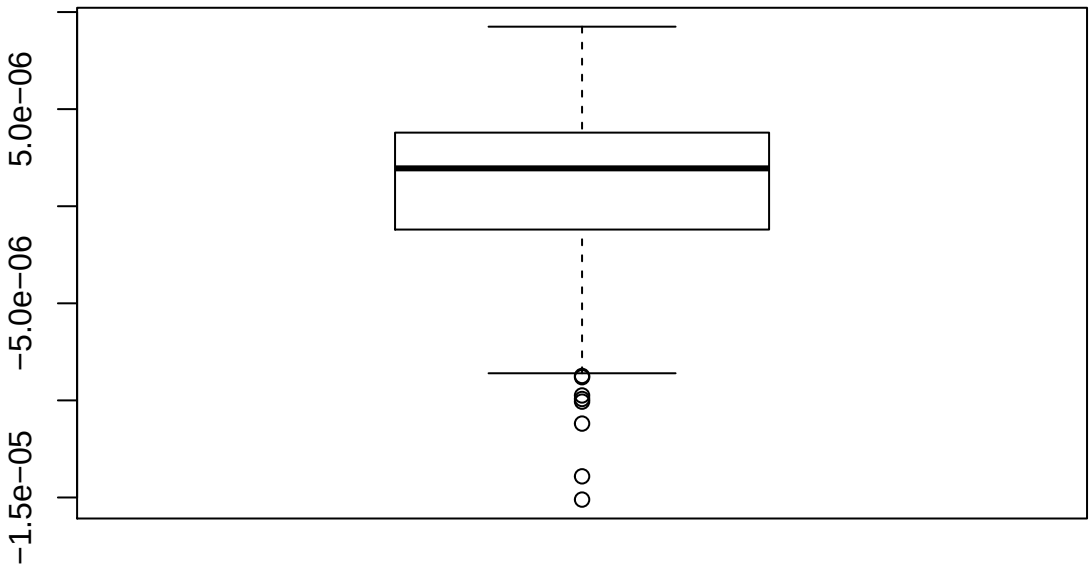
```
channels_mean<-apply(lcp,1,mean,na.m=T)
channels_var<-apply(lcp,1,mean,na.m=T)
summary(channels_mean)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -1.511e-05 -1.200e-06  1.950e-06  7.342e-07  3.789e-06  9.246e-06
```

```
hist(channels_mean)
```




```
boxplot(channels_mean)
```

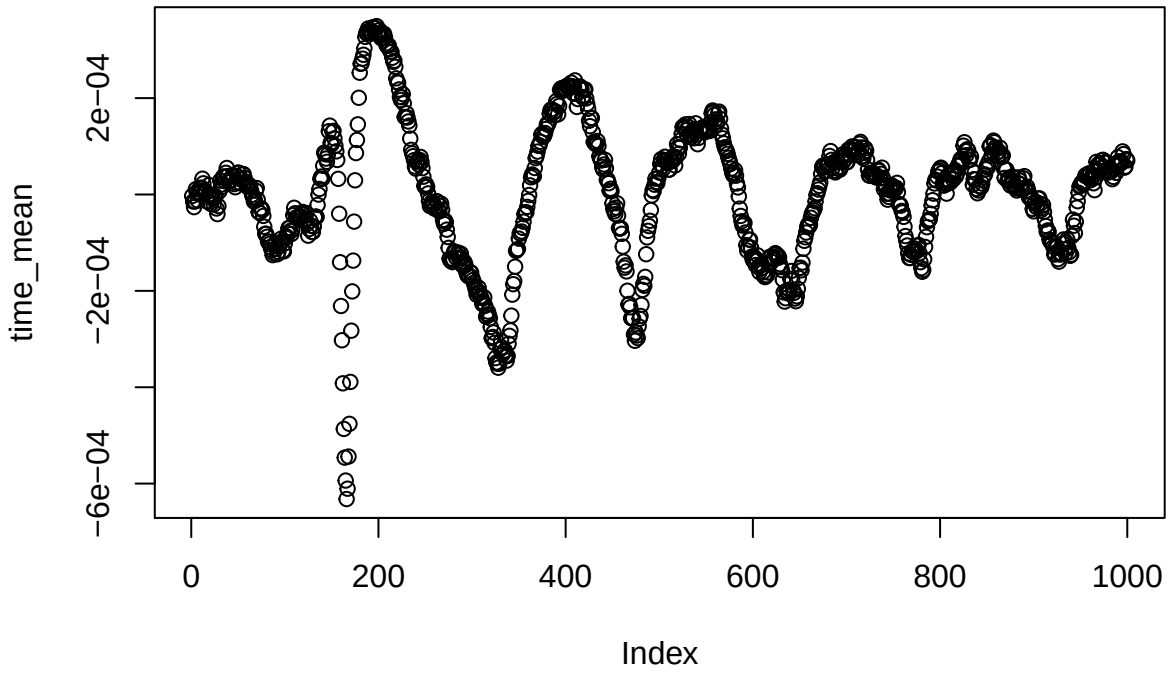


by time

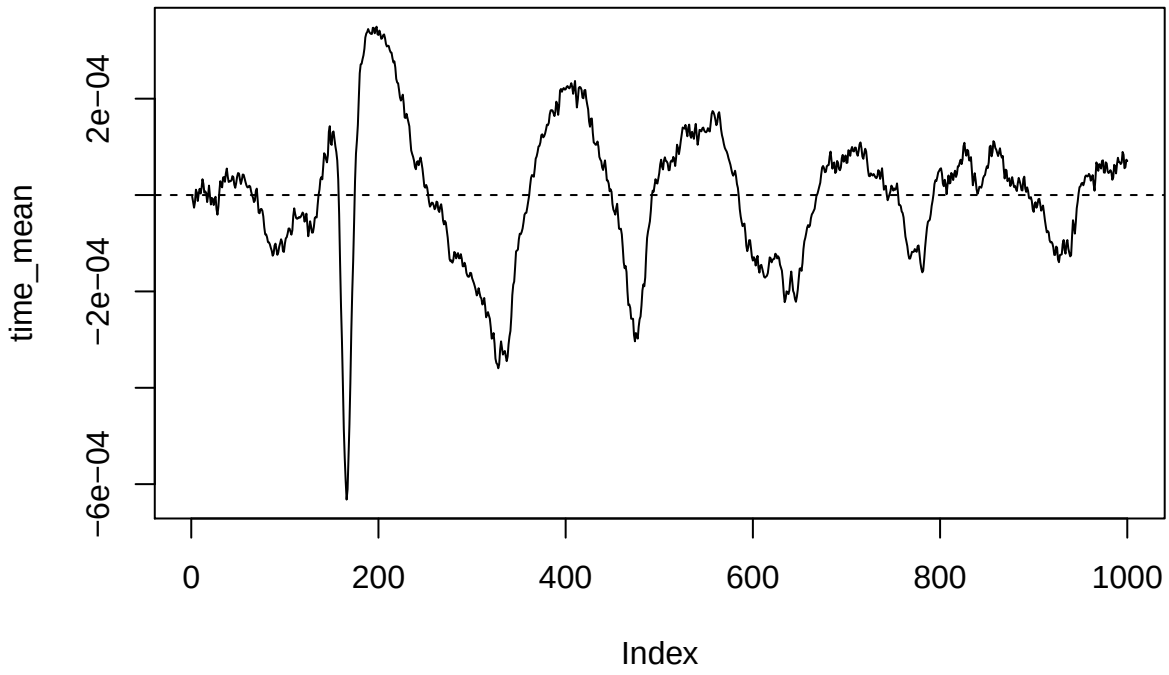
```
time_mean<-apply(lcp,2,mean,na.m=T)
time_var<-apply(lcp,2,var,na.rm=T)
summary(time_mean)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -6.320e-04 -8.762e-05  1.926e-05  7.342e-07  7.620e-05  3.494e-04
```

```
plot(time_mean)
```

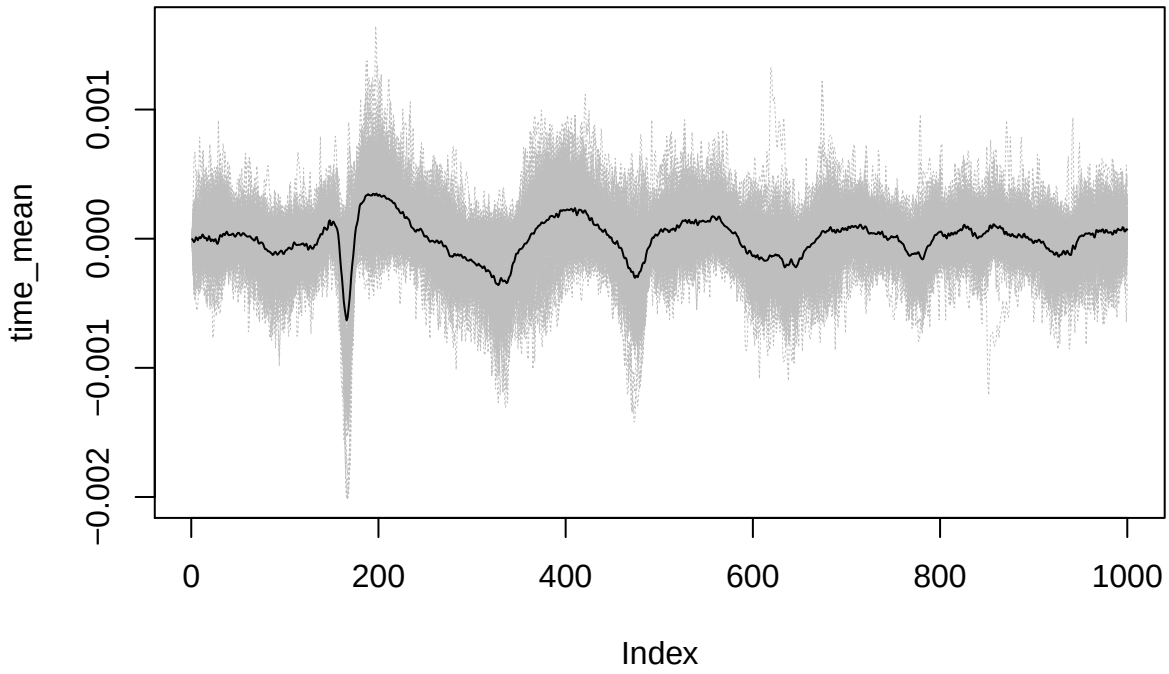



```
plot(time_mean,type="l")
abline(h=0,lty=2)
```



plot all the timeseries

```
plot(time_mean,ylim=range(lcp),type="n")
for(i in 1:200){
lines(lcp[i,],lwd=0.3,col="grey",lty=2)
}
lines(time_mean)
```



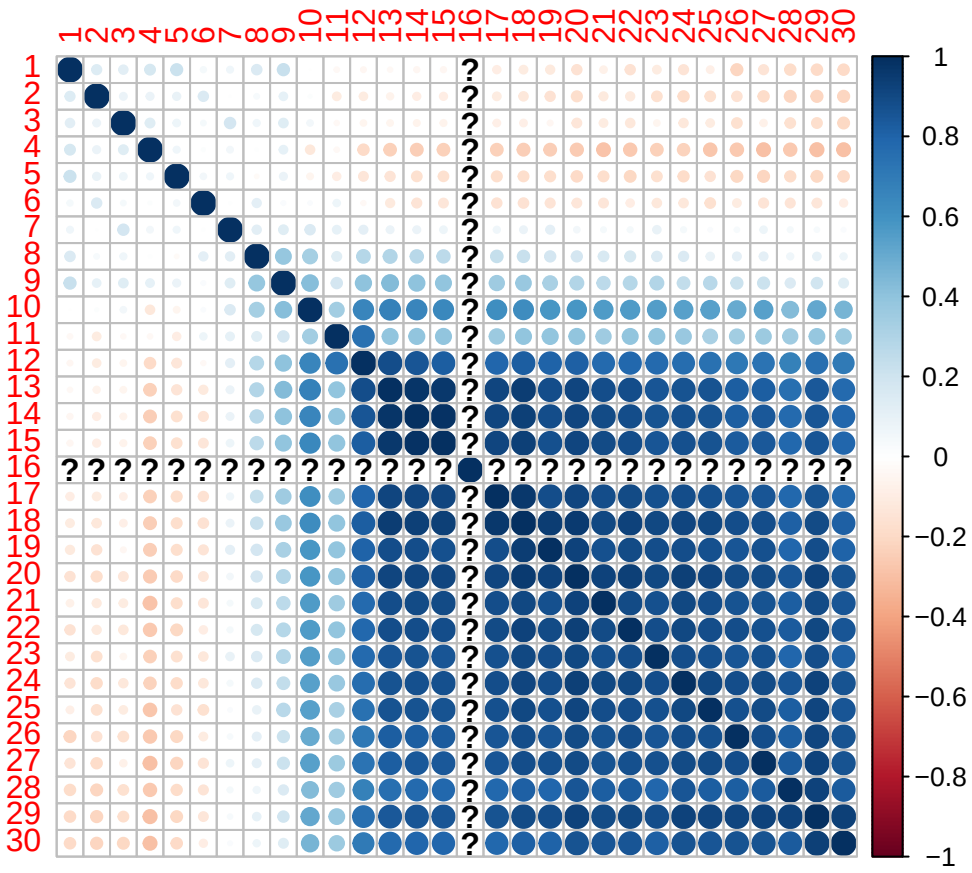
correlation between sites

```
library(corrplot) # install.packages("corrplot")
```

```
## corrplot 0.84 loaded
```

```
# we take the first 30 sites
corrplot(cor(t(lcp[1:30,])))
```

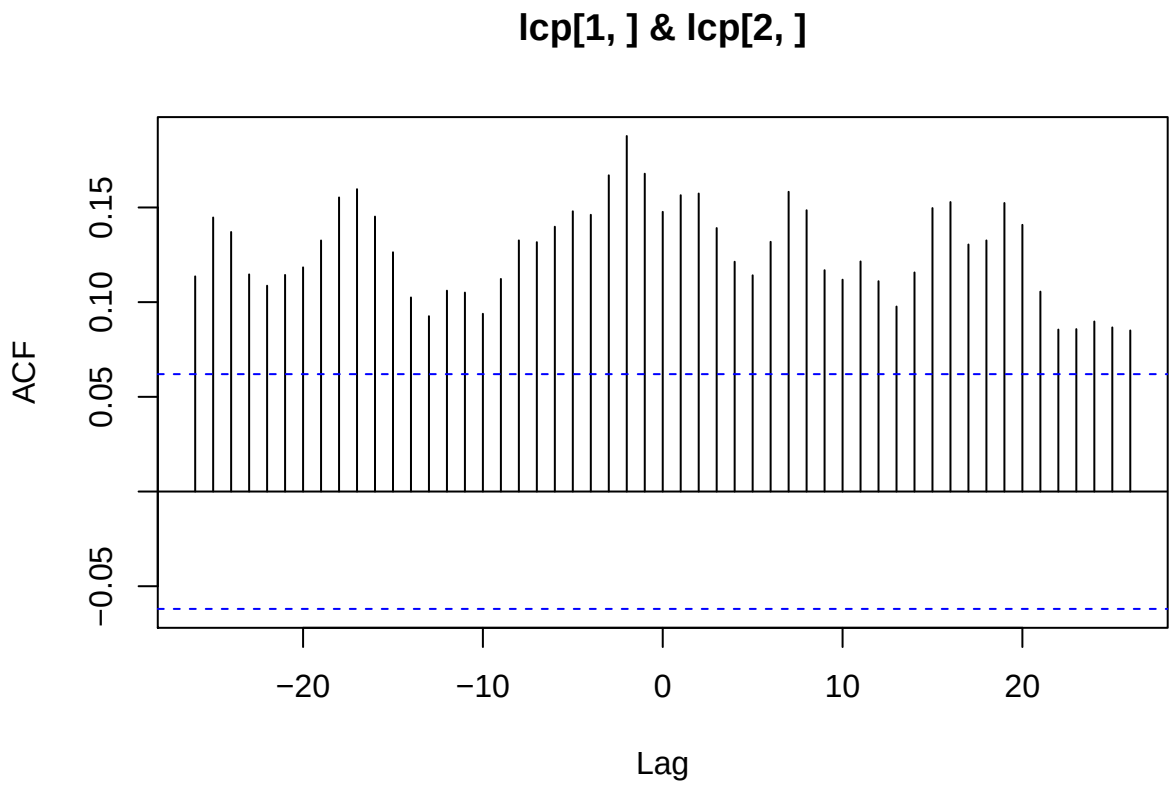
```
## Warning in cor(t(lcp[1:30, ])): la deviazione standard è zero
```



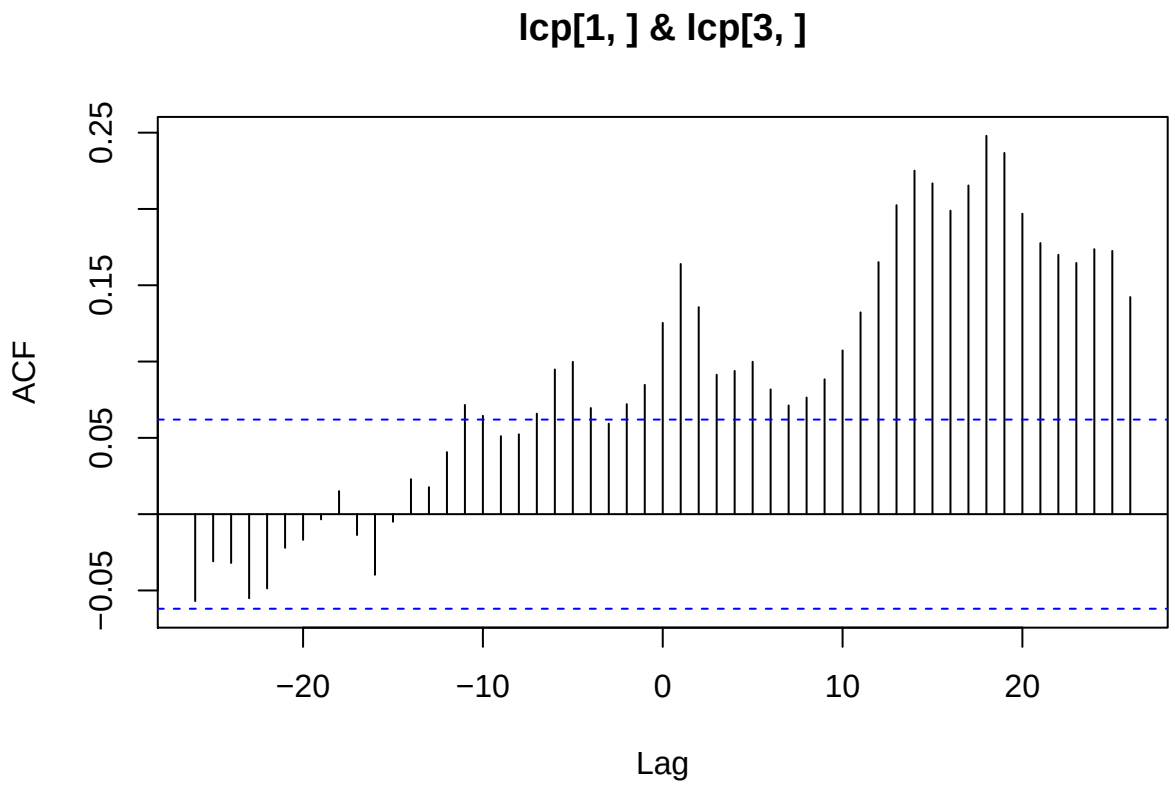

```
## time series 16 contains missing data
```

cross-correlation at different lag

```
ccf(lcp[1,],lcp[2,])
```




```
ccf(lcp[1,],lcp[3,])
```



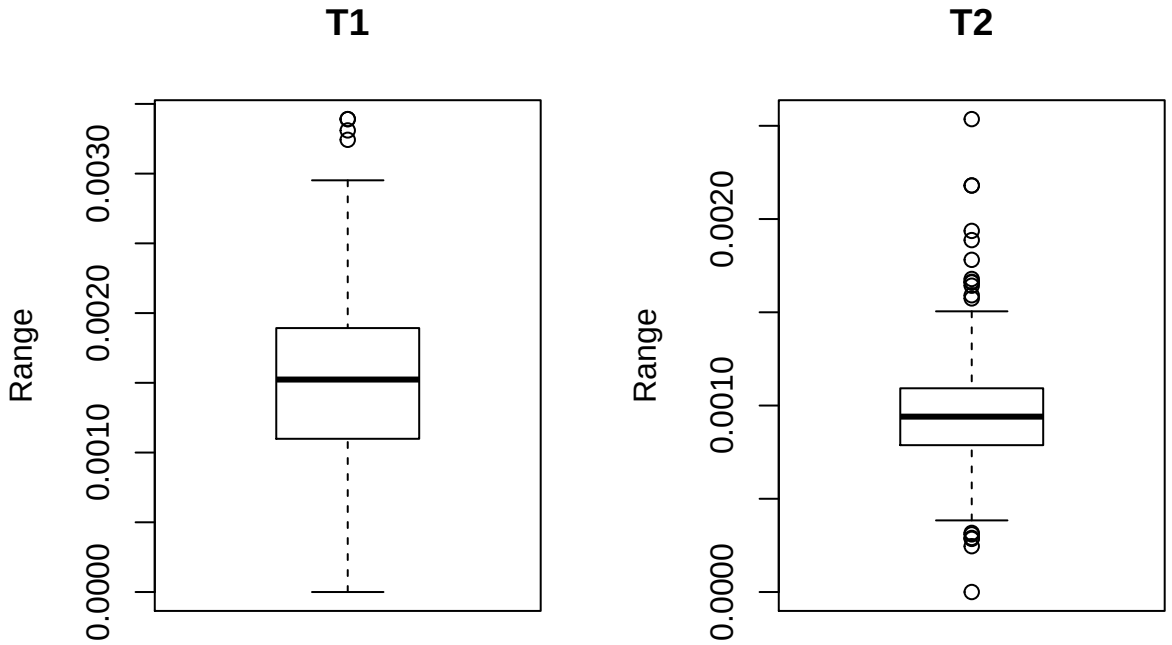
A simple test

There is some difference between the first 500 time points and the second 500 time points?
A measure of brain activity is the range of the measured LFP.
We calculate the range (max-min) for each channel.

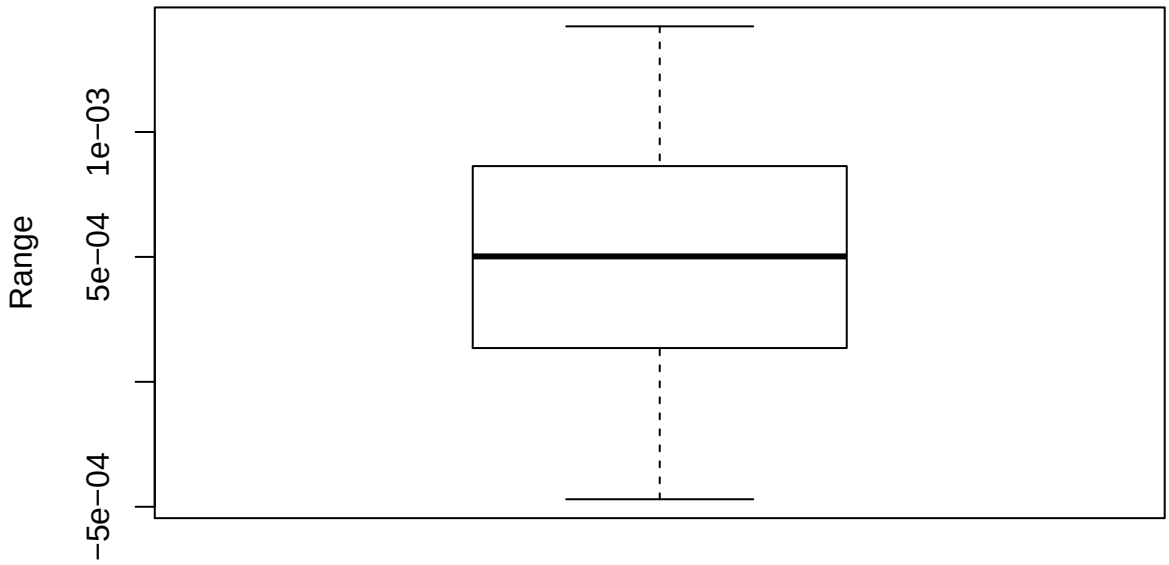
```
t1<-as.vector(diff(apply(lcp[,1:500],1,range)))
t2<-as.vector(diff(apply(lcp[,501:1000],1,range)))
```

The distribution of the measured range in t1 (1-500 timepoints) and in t2 (501-1000 timepoints) is reported below

```
par(mfrow=c(1,2))
boxplot(t1,main="T1",ylab="Range")
boxplot(t2,main="T2",ylab="Range")
```




```
par(mfrow=c(1,1))
boxplot(t1-t2,main="T1-T2",ylab="Range")
```




```
shapiro.test(t1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t1
## W = 0.97332, p-value = 0.0007405
```

```
shapiro.test(t2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t2
## W = 0.93696, p-value = 1.256e-07
```

Shapiro-Wilk normality test reported a non normal distribution for the range of the values of the channels. We assume a spatial independence between sites (is not true... but it is necessary to perform the test); a non parametric test (Wilcoxon test) is performed below

```
wilcox.test(t1,t2,paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  t1 and t2
## V = 19530, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

There is a strong difference ($p \approx 0$) in the range of the first 500 timepoints respect to the second 500 timepoints.