Exercise 1

Group C

16/1/2022

Prepare the data

```
source('exercise1.r')
```

First of all we load the genetic data about snp variants

```
mydata<-read.delim('GeneticData.txt', row.names = 1)
```

We initialize the vector valori, in which we will store the pvalues of the chi-square test, and the vector indici, in which we will store the indices of chromosome positions for which subjects have all the same variant, since for these the chi-square is not informative

```
indici<-c()
valori<-c()
```

Then we apply the chi-square test to all the chromosome locations, after having created the corresponding contingency matrix. To do this we use the funnction we created tabellariga(num,dati) that given in input the number of the line and the dataframe, gives in output the contingency matrix, as we can see in the following example

```
tabellariga(1,mydata)
```

```
##        REF C T
## group1   5   0
## group2  15   1
```

The column REF represents the number of counts for the reference, the other column the number of counts for the specific variant. It can happen that the same snp has different variants, in this case we will have the number of counts of each variant, like in this example

```
tabellariga(18,mydata)
```

```
##        A A G A REF
## group1   1   4   0
## group2   7   8   1
```

Apply chi-square test

Now that we have the contincey matrices, we can apply the chi-aquare test and extract the p-value

```
lung<-length(mydata[,1])
for (i in (1:lung)){
  pvalue<-as.double(chisq.test(tabellariga(i,mydata),simulate.p.value = TRUE)[3])
  valori<-append(valori,pvalue)
  if (length(tabellariga(i,mydata)[1,])==1) {
    indici<-append(indici,i)
  }
}
```

We construct a list with chromosome locations and corresponding p-values, we set to 1 the value of pvalues of the genes corresponding to the indices in indici and we visualize the first rows

```
pvalues<-matrix(valori)
colnames(pvalues)<-'p-value'
rownames(pvalues)<-rownames(mydata)
#setting to 1 the p-value for the genes corresponding to the indici vector, since
#these genes are not informative(all genotypes are equal)
pvalues[indici,1]<-1
head(pvalues,15)
```

```
##                 p-value
## chr1_11169676 1.0000000
## chr1_11174331 1.0000000
## chr1_11181327 0.4827586
## chr1_11181418 1.0000000
## chr1_11181457 1.0000000
## chr1_11181985 1.0000000
## chr1_11187893 1.0000000
## chr1_11188556 1.0000000
## chr1_11190646 0.4807596
## chr1_11193167 1.0000000
## chr1_11194591 1.0000000
## chr1_11199513 1.0000000
## chr1_11205058 0.2533733
## chr1_11206690 1.0000000
## chr1_11206852 0.6036982
```