BSG - Homework 2

HENRY QIU LO & MEYSAM ZAMANI

November 30, 2019

1. The file TSIChr22v4.raw contains genotype information of individuals from Tuscany in Italy, taken from the 1,000 Genomes project. The datafile contains all single nucleotide polymorphisms on chromosome 22 for which complete information is available. Load this data into the R environment. Use the fread instruction of the package data.table, which is more efficient for reading large datafiles. This data is in (0,1,2) format, where 0 and 2 represent the homozygotes AA and BB, and 1 represents the heterozygote AB. The first six leading columns of the data matrix can be ignored, as they do not contain any genetic information.

First we read the data and remove the first six columns since we are not using them. Since the data is too large we are saving in some steps this in R data formats.

```
#install.packages("data.table") # Install package data.table for having fread function ready to use.
#library("data.table")
#data <- fread(file= "TSIChr22v4.raw", header = TRUE)
#data <- data[,7:ncol(data)] #remove the first 6 columns
#saveRDS(data, file="TSIChr22v4_preprocess")
chr22 <- readRDS("TSIChr22v4_preprocess")
```

2. How many individuals does the database contain, and how many variants? What percentage of the variants is monomorphic? Remove all monomorphic SNPs from the database. How many variants remain in the database?

```
number.of.variants <- ncol(chr22)
number.of.variants
```

```
## [1] 1102156
```

Number of variants are: 1102156

```
number.of.individuals <- nrow(chr22)
number.of.individuals
```

```
## [1] 107
```

Number of individuals are: 107

```
#signfunction = s()
#for (a in 1:ncol(chr22)){
# if( var( chr22[ ,..a] ) == 0){
# signfunction <- s(signfunction, a)
# }
#}
#saveRDS(signfunction, file="monomorphicVar")
signfunction <- readRDS("monomorphicVar")
number.of.monomorphic <-length(signfunction)#number of monomorphic variants

print(100*length(signfunction)/(number.of.variants))
```

```
## [1] 81.03045
```

The percentage of monomorphic variants is: 81.0304530393157

```
print(number.of.variants-length(signfunction))
```

```
## [1] 209074
```

We have 209074 remaining variants (polymorphic variants)

Then we remove the monomorphic variants from the data, we also changed the values 0,1,2 for AA, AB, BB respectively.

```
#poly.data <- subset(chr22, select = -signfunction) #remove the monomorphic variants from data
#poly.data[poly.data==0] <- "AA"
#poly.data[poly.data==1] <- "AB"
#poly.data[poly.data==2] <- "BB"
#saveRDS(poly.data, file= "poly.data")
poly.data <- readRDS("poly.data")
```

3. Extract polymorphism rs587756191 T from the datamatrix, and determine its genotype counts. Apply a chi-square test for Hardy-Weinberg equilibrium, with and without continuity correction. Also try an exact test, and a permutation test. You can use function HWChisq, HWExact and HWPerm for this purpose. Do you think this variant is in equilibrium? Argue your answer.

As we can see in the following code, we can say it is in equilibrium since most of the test validate this except chi-square. Although chi-square test is not valid when the frequency is less than five, this case if for A/B.

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

```
rs <- poly.data$rs587756191_T
results <- genotype(rs,sep="")
out <- summary(results)
out$genotype.freq #genotype counts
```

```
##     Count  Proportion
## A/A   106 0.990654206
## A/B     1 0.009345794
```

```
#install.packages("'HardyWeinberg")
library(HardyWeinberg)
```

```
#data <- c(summary(results)$genotype[1,1],summary(results)$genotype[2,1], 0)
#names(data) <- c("AA", "AB", "BB")
#results.chi.cc <- HWChisq(data) #with continuity correction
```

Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal) Chi2 = 106.2512 DF = 1 p-value = 6.495738e-25 D = 0.002336449 f = -0.004694836

```
#results.chi <- HWChisq(data, cc= 0) #without continuity correction
```

Chi-square test for Hardy-Weinberg equilibrium (autosomal) Chi2 = 0.002358439 DF = 1 p-value = 0.961267 D = 0.002336449 f = -0.004694836

```
#results.exact <- HWExact(data)
```

Haldane Exact test for Hardy-Weinberg equilibrium (autosomal) using SELOME p-value sample counts: nAA = 106 nAB = 1 nBB = 0 H0: HWE (D==0), H1: D <> 0 D = 0.002336449 p-value = 1

```
#results.perm <- HWPerm(data)
```

Permutation test for Hardy-Weinberg equilibrium Observed statistic: 0.002358439 17000 permutations. p-value: 1

```
#results.all <- HWAlltests(data,include.permutation.test=TRUE)
```

```
                                          Statistic         p-value
```

Chi-square test: 2.358439e-03 9.612670e-01 Chi-square test with continuity correction: 1.062512e+02 6.495738e-25 Likelihood-ratio test: 4.694853e-03 9.453725e-01 Exact test with selome p-value: NA 1.000000e+00 Exact test with dost p-value: NA 1.000000e+00 Exact test with mid p-value: NA 5.000000e-01 Permutation test: 2.358439e-03 1.000000e+00

4. Determine the genotype counts for all these variants, and store them in a p x 3 matrix.

```
getGenotype<- function(x) {
 test <- summary(genotype(x,sep=""))
 set <- (test$genotype.freq[,1])
 name <- rownames(test$genotype.freq)
 name <- gsub("/","",name)
 if (length(name) == 3){
 return(c(set[1],set[2],set[3]))
 }else {
 if (name[1] == "AA" && name[2] == "AB"){
 return(c(set[1],set[2],0))
 } else if (name[1] == "AB" && name[2] == "BB"){
 return(c(0,set[1],set[2]))
 }else{
 return(c(set[1],0,set[2]))
 }
 }
}
#matrax <- apply(poly.data, 2, getGenotype)
#matrax <- t(matrax) #px3
#colnames(matrax) <- c("AA", "AB", "BB")
#saveRDS(matrax, file="matrax")
matrax <- readRDS("matrax")
```

5. Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. You can use HWChisqStats for this purpose. How many SNPs are significant (use alpha = 0.05)?

```
chisq.stats <- HWChisqStats(matrax, pvalues=TRUE)
print(length(chisq.stats[chisq.stats <= 0.05]) )
```

```
## [1] 8162
```

We have 8162 significant SNPs with alpha =0.05

6. How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone?

Considering an alpha of 0.05, we would have 200912 markers out of equilibrium.

```
print(length(chisq.stats[chisq.stats > 0.05]))
```

```
## [1] 200912
```

We have 200912 markers out of equilibrium

7. Which SNP is most significant according to the chi-square test results? Give it genotype counts. In which sense is this genotypic composition unusual?

We have 144 most significant SNPs, we are displaying the unique genotype counts since most of them are duplicated.

```
print(length(chisq.stats[chisq.stats ==min(chisq.stats)]) )
```

```
## [1] 144
```

We have 144 most significant SNPs

```
unique(matrax[chisq.stats ==min(chisq.stats),] )
```

```
##                 AA AB BB
## rs573187031_T 106  0  1
## rs62238771_C  105  0  2
## rs2629366_C    56  0 51
## rs2930745_C    80  0 27
## rs374366570_A  70  0 37
```

8. Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. You can use function HWExactStats for fast computation. How many SNPs are significant (use alpha = 0.05). Is the result consistent with the chi-square test?

We can say that it is consistent since 70.97% of the significants found by Exact test exist in the chi-square test

```
exact.stats <- HWExactStats(matrax)
print(length(exact.stats[exact.stats <= 0.05]))
```

```
## [1] 5793
```

We have 5793 significant SNPs

9. Which SNP is most significant according to the exact test results? Give its genotype counts. In which sense is this genotypic composition unusual?

Absence of any heterozygote AB is unusual.

```
matrax[exact.stats ==min(exact.stats),,drop=F]
```

```
##             AA AB BB
## rs2629366_C 56  0 51
```

10. Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP, using the HWLratio function. How many SNPs are significant (use alpha = 0:05). Is the result consistent with the chi-square test?

We can say it is consistent since 79% of the significant SNPs appears in the chi-square.

```
#ratio.test <- apply(matrax,1,function(x) HWLratio(x)$p )
#saveRDS(ratio.test, file="ratioTestHW")
#ratio.test <- readRDS("ratioTestHW")
#print(length(ratio.test[ratio.test <= 0.05]))
```

We have 7955 significant SNPs with alpha 0.05

11. Apply a permutation test for Hardy-Weinberg equilibrium to the first 10 SNPs, using the classical chi-square test (without continuity correction) as a test statistic. List the 10 p-values, together with the 10 p-values of the exact tests. Are the result consistent?

Yes, they are consistent but there are some values differents.

```
#perm.test <- apply(matrax[1:10,],1,function(x) HWPerm(x, verbose = FALSE)$p)
#saveRDS(perm.test, file="permTest11")
#perm.test <- readRDS("permTest11")
#perm.test #Perm test
```

[1] 1.000000000 1.000000000 1.000000000 1.000000000 0.645764706 [6] 1.000000000 1.000000000 1.000000000 0.123235294 0.008647059

```
#exact.stats[1:10] # Exact test
```

[1] 1.000000000 1.000000000 1.000000000 1.000000000 1.000000000 [6] 1.000000000 1.000000000 1.000000000 0.214715301 0.008643867

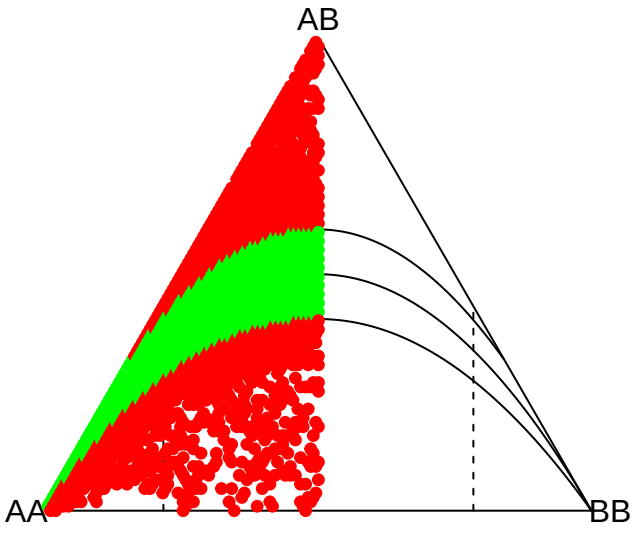
12. Depict all SNPs simultaeneously in a ternary plot with function HWTernaryPlot and comment on your result (because many genotype counts repeat, you may use UniqueGenotypeCounts to speed up the computations)

as you can see bellow, the right side of plot is empty and in red color the significant markers and in green color the non-significant ones which seems to be the most dense of the plot since the bottom there are not enough markers.

```
unique.matrax <- UniqueGenotypeCounts(matrax)
```

```
## 209074 rows in X
## 1900 unique rows in X
```

```
HWTernaryPlot(unique.matrax[,1:3])
```

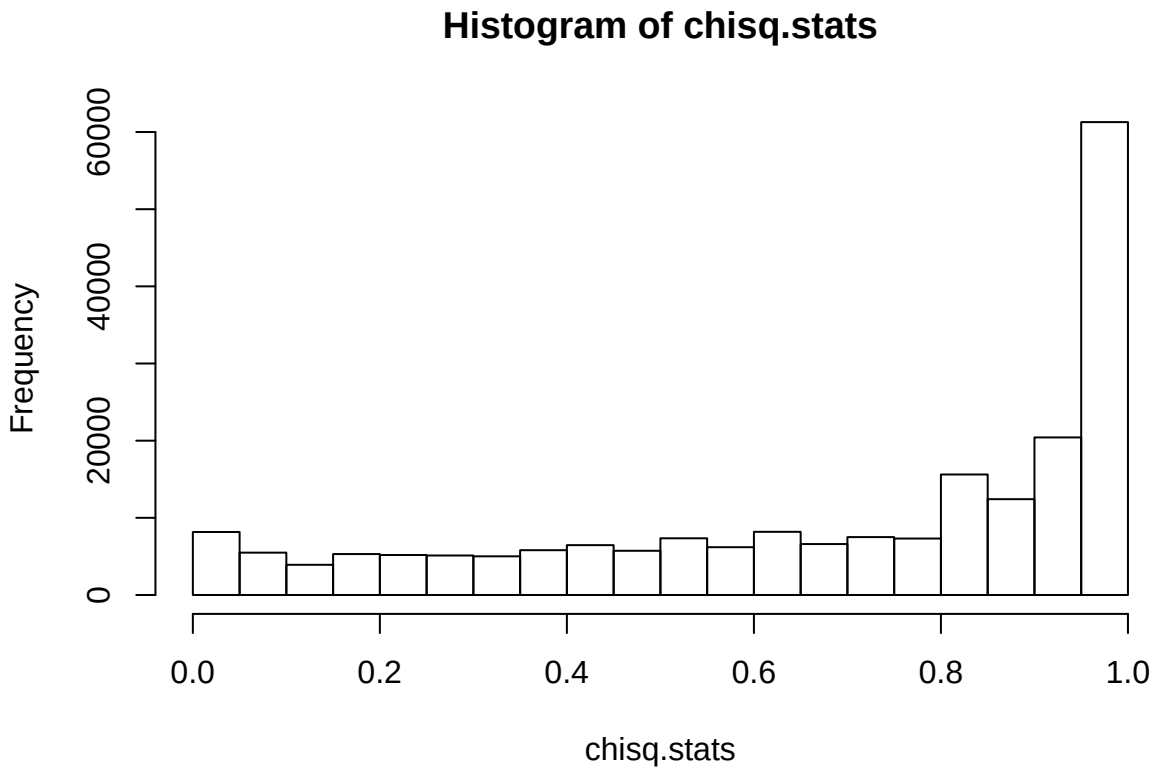


13. Can you explain why half of the ternary diagram is empty?

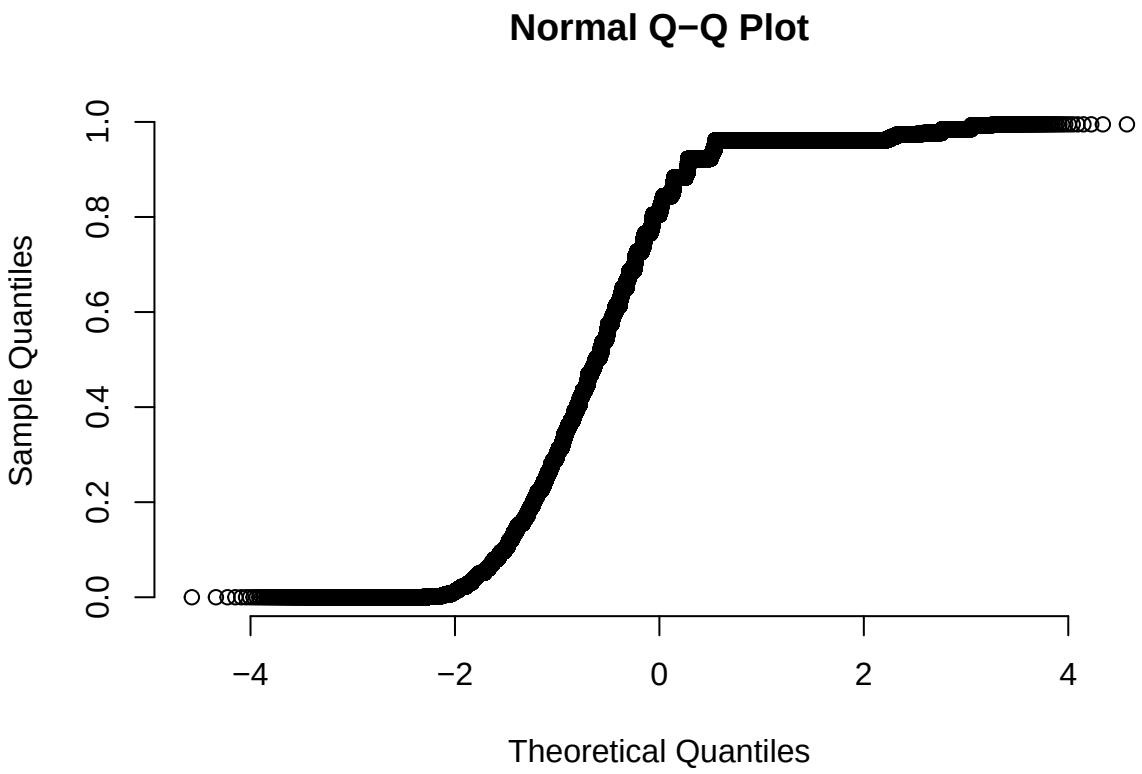
Yes, in this case, a half of the plot is empty because in this dataset the frequency of allele B is very small (always less than 0.5 and mostly it is near to 0).

14. Make a histogram of the p-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? Make a Q-Q plot of the p values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?

```
hist(chisq.stats)
```




```
qqnorm(chisq.stats, pch = 1, frame = FALSE)
```

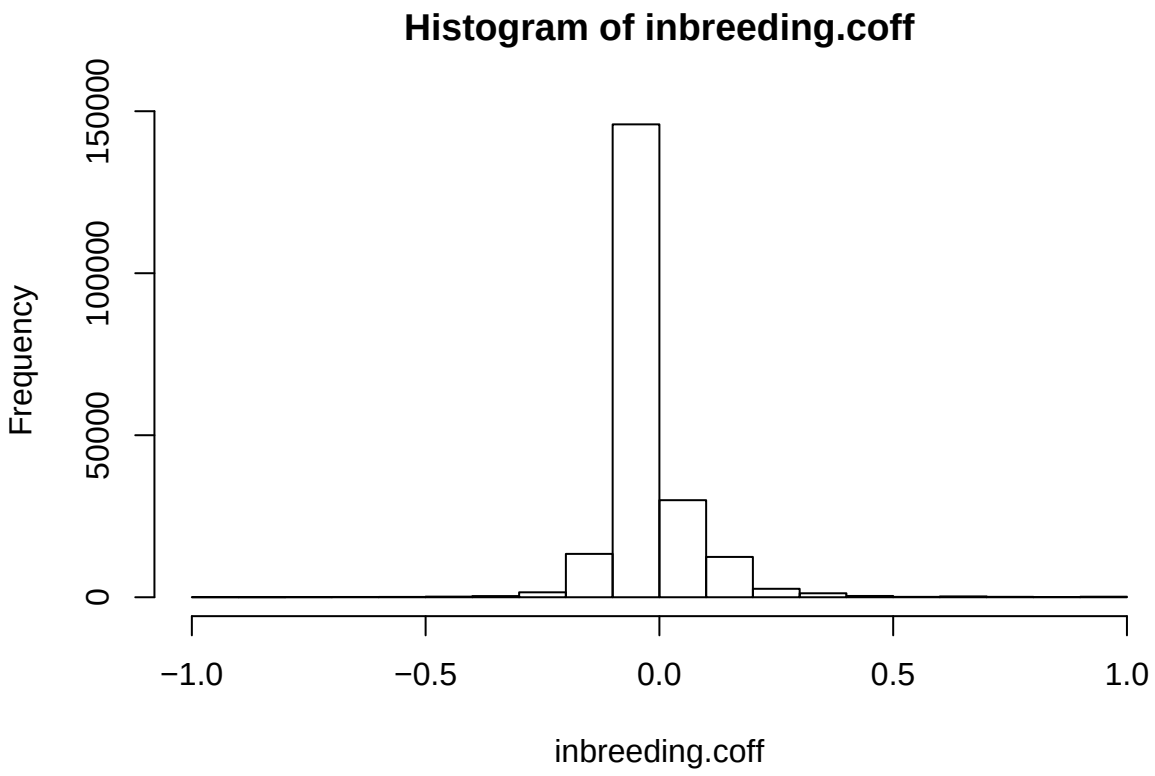


15. Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Try it on the computer if you want. Argue your answer.

This should not affect the statistical tests because we are checking that the HWE law is true: f2 AB = 4fAA fBB. So, interchanging the values of fAA and fBB should not lead to a different result.since swapping the alleles A or B does not affect to the test since p and q are symmetrical

16. Compute the inbreeding coefficient (f) for each SNP, and make a histogram of f. You can use function HWf for this purpose. Give descriptive statistics (mean, standard deviation, etc) of ˆ f calculated over the set of SNPs. What distribution do you expect f to follow theoretically? Use a probability plot to confirm your idea.

```
inbreeding.coff <- apply(matrax,1,function(x) HWf(x))
saveRDS(inbreeding.coff, file="inbreCoff")
inbreeding.coff <- readRDS("inbreCoff")
hist(inbreeding.coff)
```




```
print(mean(inbreeding.coff))
```

```
## [1] -0.004668232
```

The mean is: -0.00466823238593223

```
print(sd(inbreeding.coff))
```

```
## [1] 0.095012
```

The standard deviation is: 0.0950119986299748

```
print(median(inbreeding.coff))
```

```
## [1] -0.004694836
```

The median is: -0.00469483568075117

```
print(min(inbreeding.coff))
```

```
## [1] -0.9814815
```

The minimum is: -0.981481481481482

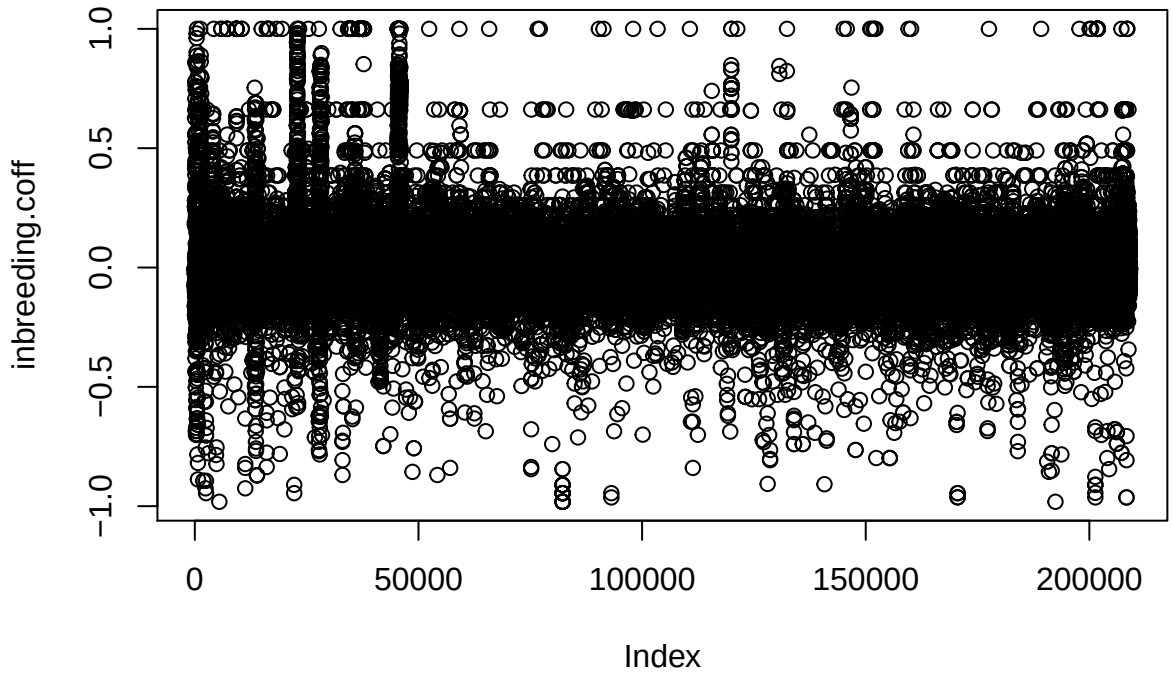
```
print(max(inbreeding.coff))
```

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

The maximum is: 1

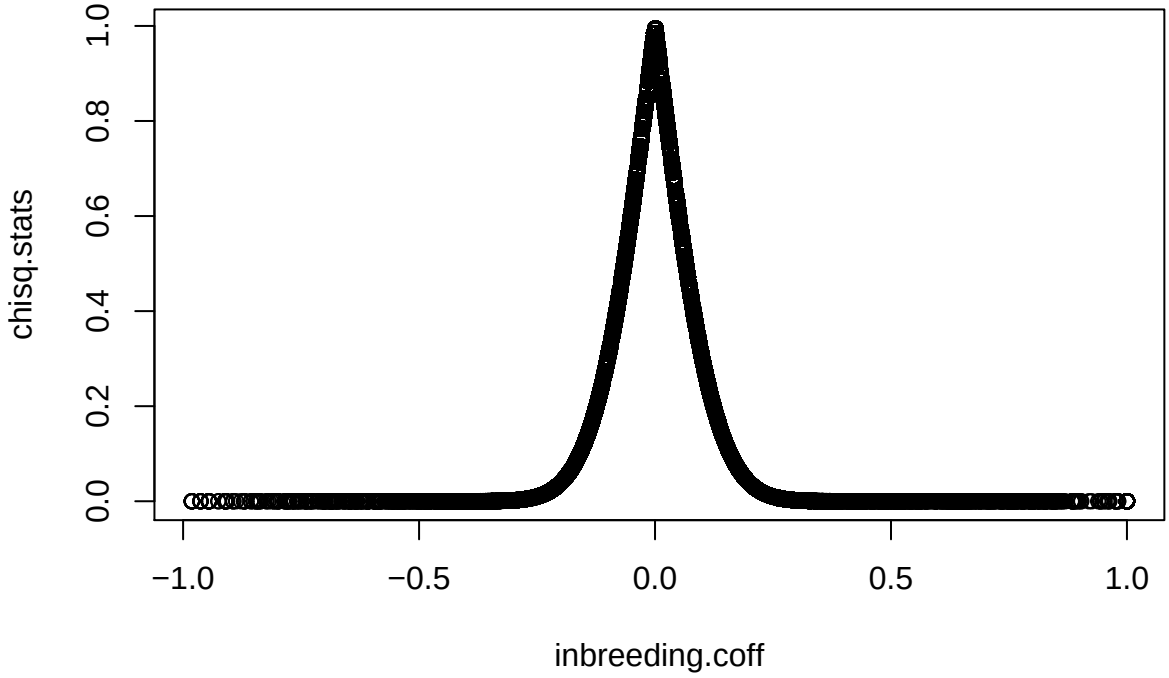
when ˆf is 0 means that HWE holds, we have normally distributed with mean equal 0 and std 0.1.

```
plot(inbreeding.coff)
```



17.Make a plot of the observed chi-square statistics against the inbreeding coefficient (f). What do you observe? Can you give an equation that relates the two statistics?

```
plot(inbreeding.coff,chisq.stats)
```



18. We reconsider the exact test for HWE, using different significant levels. Report the number and percentage of significant variants using an exac test for HWE with alpha = 0:10; 0:05; 0:01 and 0.001. State your conclusions.

as you can see, by decreasing alpha, significant variants are smaller but we can not find any relationship between alpha and the number of significant variants.

```
exact.signi <- function(x){
  lenN <- length(exact.stats)
  lenA <- length(exact.stats[exact.stats <= x])
  return(c(
  print(x),lenA))
}
alphas <- c(0.10, 0.05, 0.01, 0.001)
lapply(alphas, exact.signi)
```

```
## [1] 0.1
## [1] 0.05
## [1] 0.01
## [1] 0.001
```

```
## [[1]]
## [1]     0.1 10049.0
##
## [[2]]
## [1]    0.05 5793.00
##
## [[3]]
```

```
## [1]    0.01 2508.00
##
## [[4]]
## [1]    0.001 1485.000
```

1-We have 10049 with alpha 0.1 The percentage of alpha 0.1 is: 4.80643217234089%

2-We have 5793 with alpha 0.05 The percentage of alpha 0.05 is: 2.77078928991649%

3-We have 2508 with alpha 0.01 The percentage of alpha 0.01 is: 1.1995752700001%

4-We have 1485 with alpha 0.001 The percentage of alpha 0.001 is: 0.710274830921109%