BSG-MDS Practical 2 Statistical Genetics

Biel Caballero and Gerard Gomez

07/11/2023, submission deadline 14/11/2023

```
## Loading required package: combinat
##
## Attaching package: 'combinat'
## The following object is masked from 'package:utils':
##
##     combn
## Loading required package: gdata
##
## Attaching package: 'gdata'
## The following objects are masked from 'package:data.table':
##
##     first, last
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
## Loading required package: gtools
## Loading required package: MASS
## Loading required package: mvtnorm
##
## NOTE: THIS PACKAGE IS NOW OBSOLETE.
##
##   The R-Genetics project has developed an set of enhanced genetics
##   packages to replace 'genetics'. Please visit the project homepage
##   at http://rgenetics.org for informtion.
##
##
## Attaching package: 'genetics'
```

```
## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
```

```
## Loading required package: mice
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
## Loading required package: Rsolnp
```

```
## Loading required package: nnet
```

#Hardy Weinberg Equilibrium

##Create dataset

```
data<-fread("TSIChr22v4.raw", header = TRUE)
geneticData <- as.data.frame(data[,c(-1:-6)])
```

1. How many variants are there in this database? What percentage of the data is missing?

```
print(paste0("There are ",ncol(geneticData), " varaints"))
```

```
## [1] "There are 1102156 varaints"
```

```
sum(is.na(data))
```

```
## [1] 0
```

```
print("No data is missing")
```

```
## [1] "No data is missing"
```

2. Calculate the percentage of monomorphic variants. Exclude all monomorphics from the database for all posterior computations of the practical. How many variants do remain in your database?

```
monomorphicVariants <- c()
for(i in 1:ncol(geneticData)){
  if(length(unique(geneticData[,i])) == 1){
    monomorphicVariants <- append(monomorphicVariants,i)
  }
}
print(paste0("The parcentage of monomorphic variants are ", length(monomorphicVariants)/ncol(geneticData
```

```
## [1] "The parcentage of monomorphic variants are 0.810304530393157"
```

```
noMonomorphic <- geneticData[,-monomorphicVariants]
```

3. Extract polymorphism rs587756191_T from the data, 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 thee functions HWChisq, HWExact and HWPerm for this purpose. Do you think this variant is in equilibrium? Argue your answer.

```
library(HardyWeinberg)
rs587756191_T <- noMonomorphic[,"rs587756191_T"]

counts <- table(rs587756191_T)

HWChisq(c(AA = 106,AB = 1,BB = 0))
```

```
## Warning in HWChisq(c(AA = 106, AB = 1, BB = 0)): Expected counts below 5: chi-
## square approximation may be incorrect
```

```
## 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
```

```
HWExact(c(AA = 106,AB = 1,BB = 0))
```

```
## 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
```

```
HWPerm(c(AA = 106,AB = 1,BB = 0))
```

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

All tests give a p-value of 1, hence they are not in Hardy-Weiberg Equilibrium

4. Determine the genotype counts for all polymorphic variants, and store them in a p $\times$ 3 matrix.

```
polymorphic <- matrix(nrow = ncol(noMonomorphic),ncol = 3)
for(i in 1:ncol(noMonomorphic)){
  tab = table(noMonomorphic[,i])
  na = names(tab)
  for(n in na){
    polymorphic[i,as.numeric(n)+1] <- tab[n]
  }
}
```

5. Apply an exact test for Hardy-Weinberg equilibrium to each SNP. You can use function HWExactStats for fast computation. What is the percentage of significant SNPs (use alpha = 0.05)? Is this the number of markers that you would expect to be out of equilibrium by the effect of chance alone?

```
hwe<-HWExactStats(polymorphic, x.linked = FALSE)
hwe_signSNP<-which(hwe<0.05)
print(paste0("The parcentage of significant SNPs are ", round((length(hwe_signSNP)*100)/length(hwe),1),
```

```
## [1] "The parcentage of significant SNPs are 2.5%"
```

```
print("The number of SNPs we would expect to be out of equilibrium is around 60%")
```

```
## [1] "The number of SNPs we would expect to be out of equilibrium is around 60%"
```

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

```
ms_signSNP<-polymorphic[which.min(hwe),c(1:3)]
print(paste0("The genotype counts are: AA = ",ms_signSNP[1], " AB = ",ms_signSNP[2], " and BB = ",ms_sig
```

```
## [1] "The genotype counts are: AA = 56 AB = 1 and BB = 50"
```

```
print("These genotypic composition is unusual because the predominant genotypes are the monoallelic ones
```

```
## [1] "These genotypic composition is unusual because the predominant genotypes are the monoallelic one
```

7. 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 f calculated over the set of SNPs. What distribution do you expect f to follow theoretically? Use a probability plot to confirm your idea.

```
ic<-HWf(polymorphic)
summary(ic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   -0.96   -0.06    0.01    0.02    0.08    0.98  122389
```

```
print(paste0("The descriptive statistiucs are:"))
```

```
## [1] "The descriptive statistiucs are:"
```

```
print(paste0("    - Mean: ",summary(ic)[4]))
```

```
## [1] "    - Mean: 0.0187182056859644"
```

```
print(paste0("    - Standard deviation: ",sd(ic, na.rm = TRUE)))
```

```
## [1] "    - Standard deviation: 0.130155015109863"
```

```
print(paste0("    - Min: ",summary(ic)[1]))
```

```
## [1] "    - Min: -0.962616822429907"
```

```
print(paste0("    - Max: ",summary(ic)[6]))
```

```
## [1] "    - Max: 0.981249452378866"
```

```
print(paste0("    - Number of NAs: ",summary(ic)[7]))
```

```
## [1] "    - Number of NAs: 122389"
```

```
hist_data <- hist(ic, breaks = 30, col = "skyblue", xlab = "Inbreeding Coefficient (f)",
                  ylab = "Frequency", main = "Histogram of Inbreeding Coefficients", plot = FALSE)
```

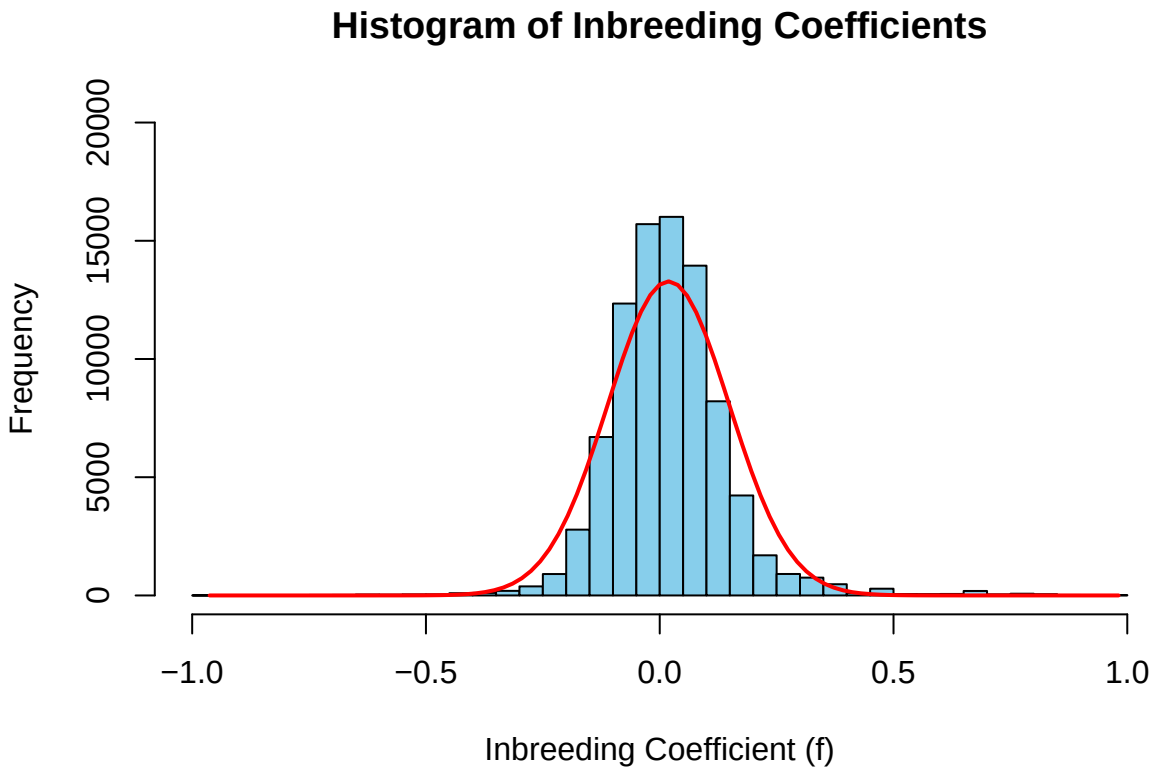
```
## Warning in hist.default(ic, breaks = 30, col = "skyblue", xlab = "Inbreeding
## Coefficient (f)", : arguments 'col', 'main', 'xlab', 'ylab' are not made use of
```

```
breaks <- hist_data$breaks

hist(ic, breaks = breaks, col = "skyblue", xlab = "Inbreeding Coefficient (f)",
```

```
     ylab = "Frequency", main = "Histogram of Inbreeding Coefficients", ylim = c(0,20000))

print("The distribution we expect f to follow theoretically is a normal distribution")
```

```
## [1] "The distribution we expect f to follow theoretically is a normal distribution"
```

```
# Fit a normal distribution curve to the histogram
mu <- mean(ic, na.rm = TRUE)
sigma <- sd(ic, na.rm = TRUE)
x <- seq(min(ic, na.rm = TRUE), max(ic, na.rm = TRUE), length.out = 100)
y <- dnorm(x, mean = mu, sd = sigma) * length(ic[is.finite(ic)]) * diff(breaks[1:2])
lines(x, y, col = "red", lwd = 2)
```




```
print("The distribution we expect f to follow theoretically is a normal distribution")
```

```
## [1] "The distribution we expect f to follow theoretically is a normal distribution"
```

8. Apply the exact test for HWE to each SNP, 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.

```
hwe<-HWExactStats(polymorphic, x.linked = FALSE)
hwe_signSNP<-which(hwe<0.1)
print(paste0("The number of significant SNPs (alpha = 0.1) parcentage of significant SNPs (alpha = 0.1)
```

```
## [1] "The number of significant SNPs (alpha = 0.1) parcentage of significant SNPs (alpha = 0.1) are 4
```

```
hwe_signSNP<-which(hwe<0.05)
print(paste0("The number of significant SNPs (alpha = 0.05) parcentage of significant SNPs (alpha = 0.05
```

```
## [1] "The number of significant SNPs (alpha = 0.05) parcentage of significant SNPs (alpha = 0.05) are
```

```r
hwe_signSNP<-which(hwe<0.01)
print(paste0("The number of significant SNPs (alpha = 0.01) parcentage of significant SNPs (alpha = 0.0
```

```
## [1] "The number of significant SNPs (alpha = 0.01) parcentage of significant SNPs (alpha = 0.01) are
```

```r
hwe_signSNP<-which(hwe<0.001)
print(paste0("The number of significant SNPs (alpha = 0.001) parcentage of significant SNPs (alpha = 0.
```

```
## [1] "The number of significant SNPs (alpha = 0.001) parcentage of significant SNPs (alpha = 0.001) a
```