Lab 2 - Statistical Genetics

Sara Montese

```
library(genetics)
```

```
## Warning: il pacchetto 'genetics' è stato creato con R versione 4.3.2
```

```
## Caricamento del pacchetto richiesto: combinat
```

```
##
## Caricamento pacchetto: 'combinat'
```

```
## Il seguente oggetto è mascherato da 'package:utils':
##
##     combn
```

```
## Caricamento del pacchetto richiesto: gdata
```

```
## Warning: il pacchetto 'gdata' è stato creato con R versione 4.3.2
```

```
##
## Caricamento pacchetto: 'gdata'
```

```
## Il seguente oggetto è mascherato da 'package:stats':
##
##     nobs
```

```
## Il seguente oggetto è mascherato da 'package:utils':
##
##     object.size
```

```
## Il seguente oggetto è mascherato da 'package:base':
##
##     startsWith
```

```
## Caricamento del pacchetto richiesto: gtools
```

```
## Warning: il pacchetto 'gtools' è stato creato con R versione 4.3.2
```

```
## Caricamento del pacchetto richiesto: MASS
```

```
## Caricamento del pacchetto richiesto: mvtnorm
```

```
## Warning: il pacchetto 'mvtnorm' è stato creato con R versione 4.3.2
```

```
##
```

```
## 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.

##

##
## Caricamento pacchetto: 'genetics'

## I seguenti oggetti sono mascherati da 'package:base':
##
##     %in%, as.factor, order

library(data.table)

##
## Caricamento pacchetto: 'data.table'

## I seguenti oggetti sono mascherati da 'package:gdata':
##
##     first, last

library(HardyWeinberg)

## Warning: il pacchetto 'HardyWeinberg' è stato creato con R versione 4.3.2

## Caricamento del pacchetto richiesto: mice

## Warning: il pacchetto 'mice' è stato creato con R versione 4.3.2

##
## Caricamento pacchetto: 'mice'

## Il seguente oggetto è mascherato da 'package:stats':
##
##     filter

## I seguenti oggetti sono mascherati da 'package:base':
##
##     cbind, rbind

## Caricamento del pacchetto richiesto: Rsolnp

## Warning: il pacchetto 'Rsolnp' è stato creato con R versione 4.3.2

## Caricamento del pacchetto richiesto: nnet

chunk_size <- 10000
raw_data <- fread(file="TSIChr22v4.raw", sep = " ", header = TRUE, nThread = chunk
_size)
raw_data_df <- data.frame(raw_data)
```

```r
############ Q1 ##############

SNPdata <- raw_data_df[,7:ncol(raw_data_df)]

# Convert values different from 0, 1, or 2 to NA
SNPdata[!sapply(SNPdata, function(x) x %in% c(0, 1, 2))] <- NA

n <- nrow(SNPdata)# individuals
p <- ncol(SNPdata) #SNPs
cat("1. How many variants are there in this database? \n")

## 1. How many variants are there in this database?

cat(p)

## 1102156

cat("\n")

cat("1. What percentage of the data is missing? \n")

## 1. What percentage of the data is missing?

mis <- 100*sum(is.na(SNPdata))/(n*p)
cat(mis)

## 0

############ Q2 ##############

# 2. Calculate the percentage of monomorphic variants.
mono = which(apply(SNPdata, 2, function(x) length(unique(x[!is.na(x)]))) == 1)

cat("Percentage of monomorphic variants: \n")

## Percentage of monomorphic variants:

cat(100 * length(mono) / ncol(SNPdata))

## 81.03045

# 2. Exclude all monomorphics from the database for all posterior computations of
the practical.
SNPpoly = SNPdata[-mono]

cat("How many variants do remain in your database?\n")

## How many variants do remain in your database?

cat(ncol(SNPpoly))

## 209074
```

```
############ Q3 ##############

rs = SNPpoly[,"rs587756191_T"]

counts <- c(
  AA=sum(rs==0),
  AB=sum(rs==1),
  BB=sum(rs==2)
)

cat("Genotype counts for rs587756191_T: ")#, counts)

## Genotype counts for rs587756191_T:

cat("\n")

cat("AA:", counts[1], "\nAB:", counts[2], "\nBB:", counts[3])

## AA: 106
## AB: 1
## BB: 0

############## # chi-square test ##############
# without continuity correction
results <- HWChisq(counts,cc=0,verbose=TRUE)

## Chi-square test for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  0.002358439 DF =  1 p-value =  0.961267 D =  0.002336449 f =  -0.004694
836

cat("Results chi-square test without continuity correction: ")#

## Results chi-square test without continuity correction:

results

## $chisq
## [1] 0.002358439
##
## $pval
## [1] 0.961267
##
## $D
## [1] 0.002336449
##
## $p
## [1] 0.004672897
##
## $f
## [1] -0.004694836
##
## $expected
##           AA           AB           BB
```

```
## 2.336449e-03 9.953271e-01 1.060023e+02
##
## $chi.contrib
##           AA           AB           BB
## 2.336449e-03 2.193848e-05 5.149879e-08

# with continuity correction
results_cc <- HWChisq(counts,verbose=TRUE)

## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (auto
somal)
## Chi2 =  106.2512 DF =  1 p-value =  6.495738e-25 D =  0.002336449 f =  -0.00469
4836

cat("Results chi-square test with continuity correction: ")

## Results chi-square test with continuity correction:

results_cc

## $chisq
## [1] 106.2512
##
## $pval
## [1] 6.495738e-25
##
## $D
## [1] 0.002336449
##
## $p
## [1] 0.004672897
##
## $f
## [1] -0.004694836
##
## $expected
##           AA           AB           BB
## 2.336449e-03 9.953271e-01 1.060023e+02
##
## $chi.contrib
##           AA           AB           BB
## 1.060023e+02 2.465008e-01 2.336449e-03

############### exact test ###############
results_ex <- HWExact(counts, pvaluetype="selome", verbose=TRUE)

## 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

cat("Results exact test: ")
```

```
## Results exact test:

results_ex

## $pval
## [1] 1
##
## $prob
## 1
## 1
##
## $pofthesample
## 1
## 1

############## permutation test ##############
results_perm <- HWPerm(counts,verbose=TRUE)

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

cat("Results permutation test: ")

## Results permutation test:

results_perm

## $stat
## [1] 0.002358439
##
## $pval
## [1] 1

## Do you think this variant is in equilibrium? Argue your answer

cat("\n")

cat("Since the p-value is 1 or close to 1 for the majority of the tests, we fail t
o reject the null hypothesis.")

## Since the p-value is 1 or close to 1 for the majority of the tests, we fail to
reject the null hypothesis.

cat("The observed distribution is likely under the assumption of HW equilibrium.")

## The observed distribution is likely under the assumption of HW equilibrium.

cat("On the other hand, we notice that for the Chi-Square test with continuity cor
rection, the p-value is very small.")

## On the other hand, we notice that for the Chi-Square test with continuity corre
ction, the p-value is very small.
```

```r
cat("In large sample sizes, the impact of continuity correction is typically less
noticeable, and the test without continuity correction may provide accurate result
s. ")
```

```
## In large sample sizes, the impact of continuity correction is typically less no
ticeable, and the test without continuity correction may provide accurate results.
```

```r
############ Q4 ##############

genotype_counts_matrix <- matrix(0, nrow = ncol(SNPpoly), ncol = 3)

# Loop through each variant
for (variant in 1:ncol(SNPpoly)) {
  # Extract genotype counts for the current variant
  genotype_counts_matrix[variant, 1] <- sum(SNPpoly[, variant] == 0)  # AA
  genotype_counts_matrix[variant, 2] <- sum(SNPpoly[, variant] == 1)  # AB
  genotype_counts_matrix[variant, 3] <- sum(SNPpoly[, variant] == 2)  # BB
}
#print(genotype_counts_matrix)

############ Q5/Q6 ##############

#Apply an exact test for Hardy-Weinberg equilibrium to each SNP.

alpha <-0.05
HW_pvalues <- HWExactStats(genotype_counts_matrix)

# Calculate percentage of variants that are significant
significant_variants <- sum(HW_pvalues < alpha)
perc_sign_variants <- (significant_variants / length(HW_pvalues)) * 100
cat("Percentage of significant SNPs at alpha = 0.05: ", perc_sign_variants, "%\n")
```

```
## Percentage of significant SNPs at alpha = 0.05:  2.770789 %
```

```r
cat("Is this the number of markers that you would expect to be out of equilibrium
by the effect of chance alone?\n")
```

```
## Is this the number of markers that you would expect to be out of equilibrium by
the effect of chance alone?
```

```r
cat("By setting the significance level to 0.05, we would expect 5% of the markers
to show significant deviation from HWE by chance alone.")
```

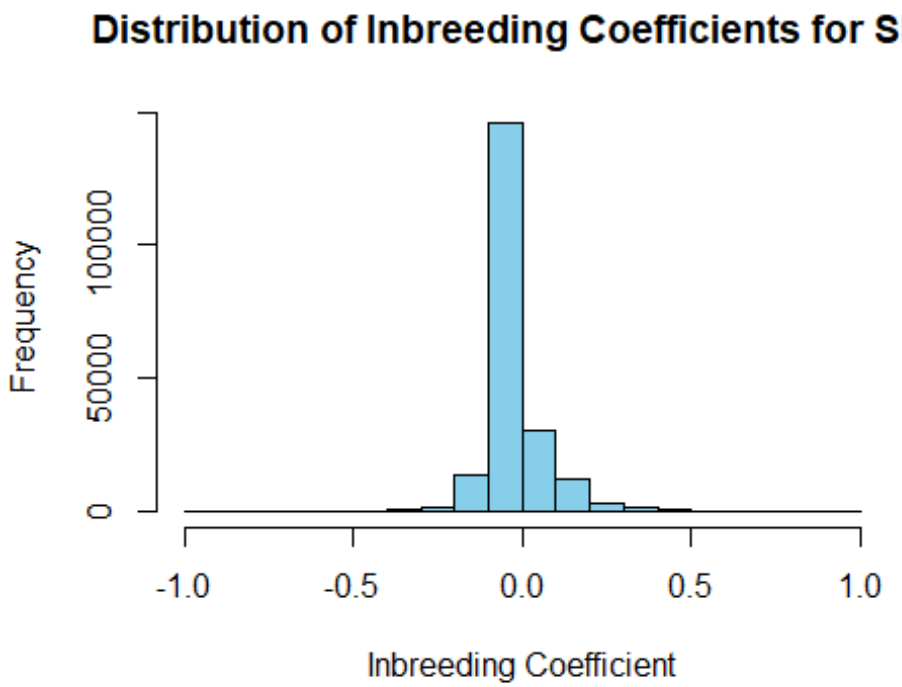
```
## By setting the significance level to 0.05, we would expect 5% of the markers to
show significant deviation from HWE by chance alone.
```

```r
cat("Therefore, since the percentage of significant SNPs is less than 5%, we can c
onsider it in the expected range due to chance.")
```

```
## Therefore, since the percentage of significant SNPs is less than 5%, we can con
sider it in the expected range due to chance.
```

```
############ Q6 ##############

min_pval_index <- which.min(HW_pvalues)
min_pval <-min(HW_pvalues)
most_significant_variant_name <- names(SNPpoly)[min_pval_index]
cat("Most significant variant according to Exact Test:",most_significant_variant_n
ame, "with a p-value of ", min_pval)

## Most significant variant according to Exact Test: rs2629366_C with a p-value of
9.784766e-33

most_significant_variant <- SNPpoly[,most_significant_variant_name]

# genotype counts
genotype_counts <- c(
  sum(most_significant_variant==0),
  sum(most_significant_variant==1),
  sum(most_significant_variant==2)
)
cat("\n")

cat("Genotype counts for most significant SNP:\n")

## Genotype counts for most significant SNP:

cat("AA:", genotype_counts[1], "\nAB:", genotype_counts[2], "\nBB:", genotype_coun
ts[3])

## AA: 56
## AB: 0
## BB: 51

observed_frequencies <- genotype_counts / nrow(SNPpoly)

#compute observed allele frequency for A
p <- ((2*genotype_counts[1])+(genotype_counts[2]))/(2*nrow(SNPpoly))
# observed allele frequency for B
q <- 1 - p



#compute expected genotype frequencies under HWE:
AA_expected_freq <- p^2
AB_expected_freq <- 2*p*q
BB_expected_freq <- q^2



cat("\nHWE expected AA frequncy:", AA_expected_freq, " | observed frequency:", obs
erved_frequencies[1],
    "\nHWE expected AB frequency: ", AB_expected_freq, "| observed frequency:", ob
served_frequencies[2],
    "\nHWE expected BB frequency: ", BB_expected_freq, "| observed frequency:", ob
```

```
served_frequencies[3]
    )

##
## HWE expected AA frequncy: 0.2739104  | observed frequency: 0.5233645
## HWE expected AB frequency:  0.4989082 | observed frequency: 0
## HWE expected BB frequency:  0.2271814 | observed frequency: 0.4766355

cat("\nIn which sense is this genotypic composition unusual?")

##
## In which sense is this genotypic composition unusual?

cat("By comparing the expected genotype frequencies under HWE and the observed gen
otype frequencies, this variant is unusual in the sense that the observed frequenc
y of heterozygous alleles is 0, which we would, in accordance with HWE, expect to
be 0.49, and this suggests how this variant is not consistent with the HWE")

## By comparing the expected genotype frequencies under HWE and the observed genot
ype frequencies, this variant is unusual in the sense that the observed frequency
of heterozygous alleles is 0, which we would, in accordance with HWE, expect to be
0.4989082, and this suggests how this variant is not consistent with the HWE

############ Q7 ##############

inbreeding_factor <- function(genotype_sequence) {
  genotype_counts <- c(
    AA=sum(genotype_sequence==0),
    AB=sum(genotype_sequence==1),
    BB=sum(genotype_sequence==2)
  )
  return(HWf(genotype_counts))
}

inbreeding_coeffs <- apply(SNPpoly, 2, inbreeding_factor)

#descriptive statistics
print(summary(inbreeding_coeffs))

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.981482 -0.033816 -0.004695 -0.004668 -0.004695  1.000000

hist(inbreeding_coeffs, main="Distribution of Inbreeding Coefficients for SNP", xl
ab="Inbreeding Coefficient",col = "skyblue", border = "black")
```



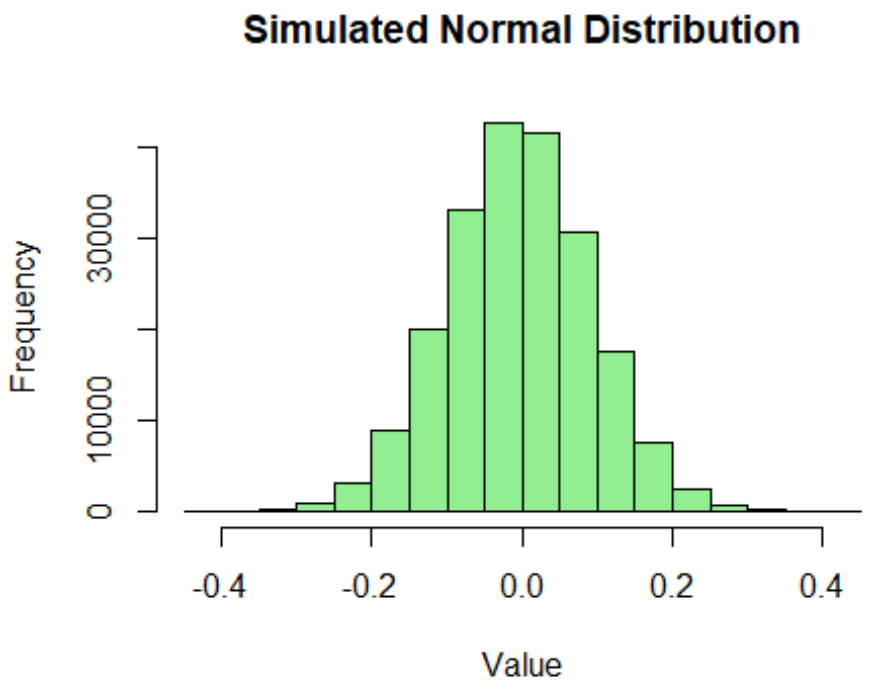

```
cat("What distribution do you expect f to follow theoretically? Use a probability plot to confirm your idea \n")
```

```
## What distribution do you expect f to follow theoretically? Use a probability plot to confirm your idea
```

```
cat("If the population is large and mating is random, as in our example, f can follow an approximately normal distribution. f may be centered in 0 since the observed distribution is likely under the assumption of HW equilibrium.")
```

```
## If the population is large and mating is random, as in our example, f can follow an approximately normal distribution. f may be centered in 0 since the observed distribution is likely under the assumption of HW equilibrium.
```

```
simulated_normal <- rnorm(length(inbreeding_coeffs), mean = mean(inbreeding_coeffs), sd = sd(inbreeding_coeffs))
hist(simulated_normal, main = "Simulated Normal Distribution", xlab = "Value", col = "lightgreen", border = "black")
```




```
alpha_values <- c(0.10, 0.05, 0.01, 0.001)

for (alpha in alpha_values) {
  # Calculate p-values for each SNP
  HW_pvalues <- HWExactStats(genotype_counts_matrix)

  # Calculate the number of significant variants for the current alpha
  significant_variants <- sum(HW_pvalues < alpha)
  perc_sign_variants <- (significant_variants / length(HW_pvalues)) * 100

  cat("Number of significant SNPs at alpha =", alpha, ":", significant_variants, "
\n")
  cat("Percentage of significant SNPs at alpha =", alpha, ":", perc_sign_variants,
"%\n")
  cat("\n")
}
```

```
## Number of significant SNPs at alpha = 0.1 : 10049
## Percentage of significant SNPs at alpha = 0.1 : 4.806432 %
##
## Number of significant SNPs at alpha = 0.05 : 5793
## Percentage of significant SNPs at alpha = 0.05 : 2.770789 %
##
## Number of significant SNPs at alpha = 0.01 : 2508
## Percentage of significant SNPs at alpha = 0.01 : 1.199575 %
##
## Number of significant SNPs at alpha = 0.001 : 1485
## Percentage of significant SNPs at alpha = 0.001 : 0.7102748 %
```

```r
#State your conclusions
cat("The results suggest that when we set a stricter significance level for assess
ing HWE, fewer SNPs are significant. The lower percentages observed at more string
ent alpha levels indicate that the majority of SNPs in the dataset conform to HWE.
We can conclude that the population is in HWE, however, it's important to note tha
t real populations might not meet the assumptions of the Hardy-Weinberg principle
like random mating, no mutation, no migration or large population size.")
```

```
## The results suggest that when we set a stricter significance level for assessin
g HWE, fewer SNPs are significant. The lower percentages observed at more stringen
t alpha levels indicate that the majority of SNPs in the dataset conform to HWE. W
e can conclude that the population is in HWE, however, it's important to note that
real populations might not meet the assumptions of the Hardy-Weinberg principle li
ke random mating, no mutation, no migration or large population size.
```