BSG-MDS practical 2 Statistical Genetics

Eliya Tiram and Ximena Moure

14/11/2023, submission deadline 21/11/2023

```
filename <- "TSIChr22v4.raw"
genetic_data <- fread(filename, drop = c(1:6))
```

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

```
num_variants <- ncol(genetic_data)
missing_data <- sum(is.na(genetic_data))
total_entries <- nrow(genetic_data) * num_variants
missing_data_percentage <- (missing_data / total_entries) * 100

cat("\nNum variants:",num_variants, "\n")
```

```
##
## Num variants: 1102156
```

```
cat("\nPercentage missing:",missing_data_percentage, "\n")
```

```
##
## Percentage missing: 0
```

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?

```
# Identify and remove monomorphic variants
is_monomorphic <- apply(genetic_data, 2, function(x) length(unique(na.omit(x))) == 1)
genetic_data_poly <- genetic_data[, !is_monomorphic, with = FALSE]

# Count remaining variants
remaining_variants <- ncol(genetic_data_poly)
monomorphic_percentage <- (sum(is_monomorphic) / num_variants) * 100

cat("\nPercentage of monomorphic variants:", monomorphic_percentage, "%\n")
```

```
##
## Percentage of monomorphic variants: 81.03045 %
```

```
cat("\nNumber of remaining variants:", remaining_variants, "\n")
```

```
##
## Number of remaining variants: 209074
```

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.

The chi-square test has resulted in a very low p-value (approximately $6.50 \times 10-25$), indicating strong evidence against HWE. However, a crucial note is the warning about expected counts below 5, suggesting that the chi-square approximation might be incorrect.

The exact test reports a p-value of 1, which suggests no evidence against HWE. The exact test is generally more reliable than the chi-square test when dealing with small sample sizes or low genotype frequencies, as it doesn't rely on the approximation that the chi-square test does.

The permutation test also shows a p-value of 1, further supporting the conclusion that there is no evidence against HWE.

Based on the more reliable exact and permutation tests, we can conclude that there is no evidence to suggest that the SNP is not in Hardy-Weinberg equilibrium. The discrepancy in the chi-square test result is likely due to its limitations with small or unevenly distributed samples.

```
specific_snp <- genetic_data_poly[["rs587756191_T"]]

genotype_counts <- table(factor(specific_snp, levels = 0:2))

genotype_counts <- c(AA = sum(genotype_counts[1]), AB = sum(genotype_counts[2]), BB = sum(genotype_coun


library(HardyWeinberg)
chi_square_test <- HWChisq(genotype_counts)
```

```
## Warning in HWChisq(genotype_counts): 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
```

```
chi_square_test_without_correction <- HWChisq(genotype_counts, cc = 0)
```

```
## Warning in HWChisq(genotype_counts, cc = 0): Expected counts below 5: chi-square
## approximation may be incorrect
```

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

```
exact_test <- HWExact(genotype_counts)
```

```
## 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
```

```
permutation_test <- HWPerm(genotype_counts)
```

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

```
print("Chi square test")
```

```
## [1] "Chi square test"
```

```
print(chi_square_test)
```

```
## $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
```

```
print("Chi square test without correction")
```

```
## [1] "Chi square test without correction"
```

```r
print(chi_square_test_without_correction)
```

```
## $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
```

```r
print("Exact test")
```

```
## [1] "Exact test"
```

```r
print(exact_test)
```

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

```r
print("Permutation test")
```

```
## [1] "Permutation test"
```

```r
print(permutation_test)
```

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

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

```
geno.matrix <- matrix(nrow = 3, ncol = ncol(genetic_data_poly))

geno.matrix <- t(sapply(genetic_data_poly, function(x) {
  c(sum(x == 0), sum(x == 1), sum(x == 2))
}))

colnames(geno.matrix) <- c("AA", "AB", "BB")
```

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 = 0.05)? Is this the number of markers that you would expect to be out of equilibrium by the effect of chance alone?

No, the observed rate of significant deviations from HWE is 2.77%, which is lower than the 5% expected by chance at a significance level of 0.05. This suggests that fewer SNPs are deviating from HWE than would be expected if the deviations were occurring randomly, indicating that the population might be more in equilibrium or the test used is conservative for the data.

```
geno.matrix.exact.pval <- HWExactStats(geno.matrix)

# Determine the number of SNPs with significant deviation from HWE at alpha = 0.05
significant.snp.num.exact <- sum(geno.matrix.exact.pval < 0.05)

cat("Number of SNPs significantly deviating from HWE:", significant.snp.num.exact, "\n")
```

```
## Number of SNPs significantly deviating from HWE: 5793
```

```
# Total number of SNPs tested
total_snps_tested <- nrow(geno.matrix)

percentage_significant <- (significant.snp.num.exact / total_snps_tested) * 100


# Expected number of significant SNPs by chance at alpha = 0.05
expected_significant_by_chance <- total_snps_tested * 0.05

cat("Expected number of significant SNPs by chance at alpha = 0.05:", expected_significant_by_chance, "
```

```
## Expected number of significant SNPs by chance at alpha = 0.05: 10453.7
```

```
cat("Percentage of SNPs significantly deviating from HWE at alpha = 0.05:", percentage_significant, "%\n
```

```
## Percentage of SNPs significantly deviating from HWE at alpha = 0.05: 2.770789 %
```

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

The genotypic composition for the most significant SNP (rs2629366_C) shows genotype counts of 56 for AA, 0 for AB, and 51 for BB. This pattern is unusual because it suggests there are no heterozygous individuals (AB) in the sample for this particular SNP, which is very rare.

Under Hardy-Weinberg equilibrium (HWE), you would expect the genotype frequencies to follow the equation

$$p^2 + 2pq + q^2 = 1$$

where: - p is the frequency of one allele (say, A) - q is the frequency of the other allele (say, B) - p^2 is the expected frequency of the AA genotype - 2pq is the expected frequency of the AB genotype - q^2 is the expected frequency of the BB genotype

In a large, randomly mating population, the absence of heterozygotes (AB) is highly unlikely.

```
# Find the index of the smallest p-value, which indicates the most significant SNP
most_significant_index <- which.min(geno.matrix.exact.pval)

# Retrieve the genotype counts for the most significant SNP
most_significant_counts <- geno.matrix[most_significant_index, ]

most_significant_snp_id <- rownames(geno.matrix)[most_significant_index]


cat("Most significant SNP:", most_significant_snp_id, "\n")
```

```
## Most significant SNP: rs2629366_C
```

```
cat("Genotype counts for the most significant SNP [AA, AB, BB]:", most_significant_counts, "\n")
```

```
## Genotype counts for the most significant SNP [AA, AB, BB]: 56 0 51
```

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

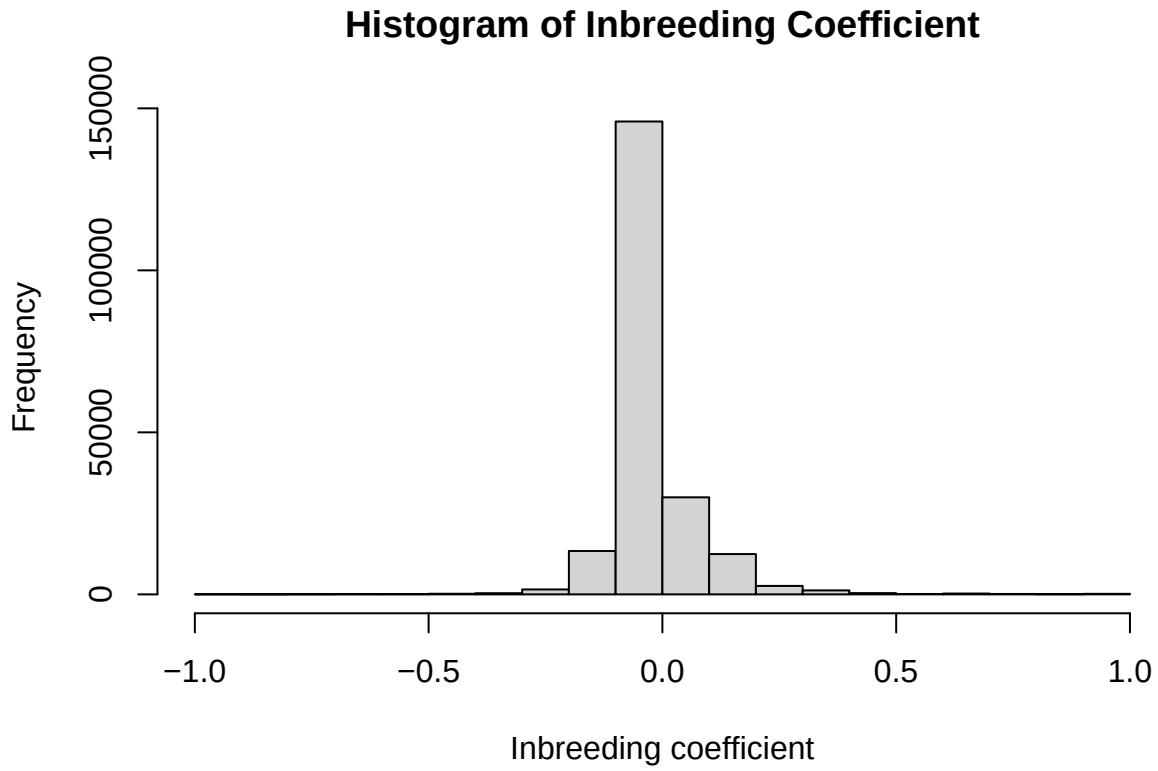
Theoretically, the distribution of the inbreeding coefficient f in a large, randomly mating population where allele frequencies are stable and there is no selection, mutation, migration, or drift, would be centered around zero. This is because most individuals would be expected to be equally likely to inherit alleles from a randomly chosen set of parents, resulting in a low probability of inbreeding and thus a low inbreeding coefficient.

While the concentration of f values around zero does support the theory to some extent, the presence of extreme values and the peak at 1 suggest that other factors might be influencing the genotypic composition of the population.

The Q-Q (quantile-quantile) plot displays the sample quantiles of inbreeding coefficients against the theoretical quantiles of a normal distribution. The plot shows substantial deviation from the line at the lower and upper ends of the distribution. This suggests that the inbreeding coefficients are not normally distributed. The tails are heavier than what would be expected in a normal distribution, indicating a larger number of both low and high extreme values. The Q-Q plot indicates that the assumption of normality for the inbreeding coefficients does not hold.

```
f.coef <- c()
for(i in 1:ncol(genetic_data_poly)) {
  x <- geno.matrix[i,]
  names(x) <- c("AA", "AB", "BB")
  f.coef <- c(f.coef,HWf(x))
}

hist(f.coef, main= "Histogram of Inbreeding Coefficient", xlab = "Inbreeding coefficient")
```



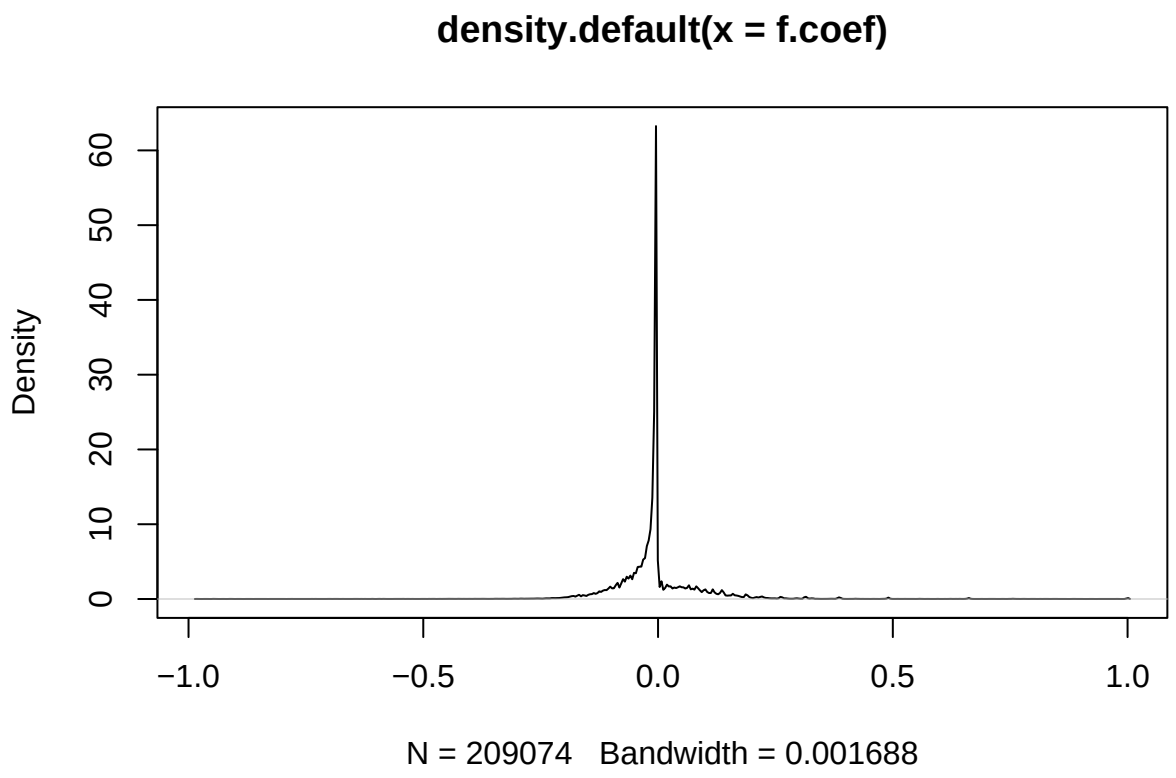

```
f.coef.stats <- sapply(f.coef,mean)
summary(f.coef.stats)
```

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

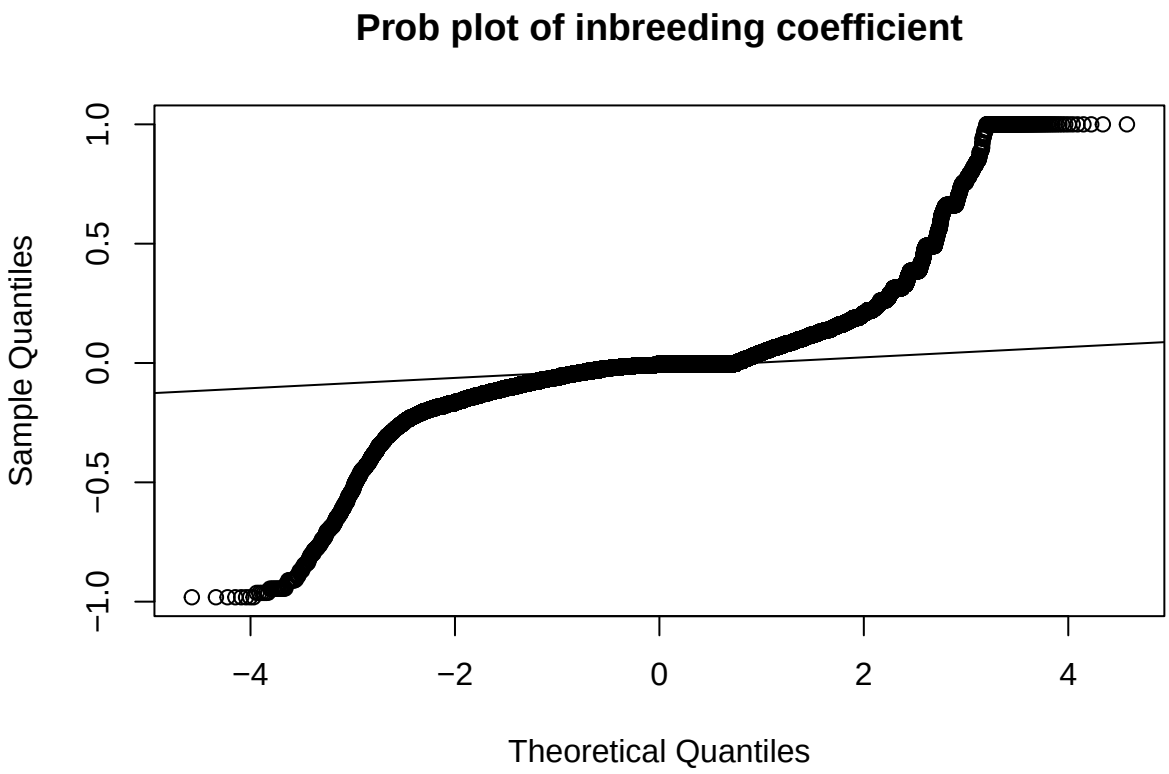
```
sd(f.coef)
```

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

```r
plot(density(f.coef))
```




```r
qqnorm(f.coef, main= "Prob plot of inbreeding coefficient")
qqline(f.coef)
```



#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 = 0.10, 0.05, 0.01 and 0.001. State your conclusions.

Based on the different significance levels tested, we conclude that approximately 2.77% of the SNPs in the dataset have p-values at or below 0.05, suggesting evidence against Hardy-Weinberg equilibrium. About 1.19% of the SNPs have p-values at or below 0.01, providing stronger evidence against equilibrium, and approximately 0.71% of the SNPs have p-values at or below 0.001, which indicates very strong evidence against Hardy-Weinberg equilibrium. These results suggest that a small but notable portion of the SNPs may not be in equilibrium, although the majority do not show significant deviation at the conventional alpha level of 0.05.

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

significant_counts <- numeric(length(alpha_levels))

for (i in seq_along(alpha_levels)) {
  alpha <- alpha_levels[i]
  significant_counts[i] <- sum(geno.matrix.exact.pval < alpha)
}

# Calculate the percentage of significant SNPs for each alpha level
total_snps <- length(geno.matrix.exact.pval)
percentage_significant <- (significant_counts / total_snps) * 100


results_df <- data.frame(
```

```
  Alpha_Level = alpha_levels,
  Significant_SNPs = significant_counts,
  Percentage = percentage_significant
)


print(results_df)
```

```
##   Alpha_Level Significant_SNPs Percentage
## 1       0.100            10049  4.8064322
## 2       0.050             5793  2.7707893
## 3       0.010             2508  1.1995753
## 4       0.001             1485  0.7102748
```