Practical 2

Statistical Genetics: Hardy Weinberg Equilibrium

Anna Putina Marine Mauzeau

2023-11-18

```
library(genetics)
library(ggplot2)
library(data.table)
library(HardyWeinberg)
```

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, with the fread instruction of the package data.table, which is more efficient for reading large datafiles. The first six columns contain non-genetical information. Create a dataframe that only contains the genetic information that is in and beyond the 7th column. Notice that the genetic variants are identifed by an "rs" identifier. The genetic data is coded in the (0, 1, 2) format with 0=AA, 1=AB, 2=BB.

```
data <- fread("TSIChr22v4.raw", data.table = F)
data <- data[, -c(1:6)]
```

Question 1

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

```
n_variants <- ncol(data)

NA.counter <- sum(is.na.data.frame(data))
NA.percentage <- NA.counter*100/prod(dim(data))
```

There are 1102156 variants in this database. The percentage of missing data in this database: 0 %.

Question 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?

```
is.not.monomorphic <- function(col){
  return(!(sum(col) %in% c(0, 2 * nrow(data))))
}

data <- data[, sapply(data, is.not.monomorphic)]
nb.variants <- ncol(data)
```

The percentage of monomorphic variants: 81.03 %.

After removing all monomorphic variants (as defined in previous lab), there are 209077 remaining variants.

Question 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.

```
rs587756191_T <- data$rs587756191_T
genotype_obj <- genotype(as.genotype.allele.count(rs587756191_T, alleles = c("B", "A")))

# Convert the genotype data to a factor with levels "AA," "AB," and "BB"
genotype_factor <- factor(genotype_obj, levels = c("A/A", "A/B", "B/B"))
levels(genotype_factor) <- c("AA", "AB", "BB")

# Create a vector of genotype counts
genotype_counts <- table(genotype_factor)
print(genotype_counts)
```

```
## genotype_factor
##  AA  AB  BB
## 106   1   0
```

```
# Hardy-Weinberg equilibrium test without continuity correction
hwe_test_no_correction <- HWChisq(genotype_counts, cc = 0)
```

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

```
# Hardy-Weinberg equilibrium test with continuity correction
hwe_test_with_correction <- HWChisq(genotype_counts)
```

```
## 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
```

```
# Exact test for Hardy-Weinberg equilibrium
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
```

```
# Perform a permutation test for Hardy-Weinberg equilibrium
perm_test <- HWPerm(genotype_counts)
```

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

According to the Chi-square test without continuity correction: the Chi-square statistic, which is a measure of the difference between the observed and expected genotype frequencies under Hardy-Weinberg equilibrium, is very low, and the p-value is high, which leads us to conclude to a Hardy-Weinberg equilibrium. The 2 other measures of disequilibrium (D and f) are also very low. According to permutation test, there seems to be a Hardy-Weinberg equilibrium : the observed statistic, that is a measure of how much the observed data deviates from the expected Hardy-Weinberg equilibrium, is very low, and p-value is 1. Same conclusions when looking at the result of the exact test : the disequilibrium coefficient D, that quantifies the deviation from Hardy-Weinberg equilibrium, is very close to zero, and p-value is 1.

But the Chi-square test with continuity correction indicates the opposite : the observed genotype frequencies

do not seem to align with the expected frequencies under the assumption of equilibrium, as the chi-squared statistic is here very large, and p-value is almost 0. It seems that it is not an appropriate situation to use correction in the Chi-square test.

We conclude that this variant is **in equilibrium**.

Question 4

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

```
# Function to process a single variant
process_variant <- function(variant_col) {
  genotype_factor <- factor(variant_col, levels = c(0, 1, 2))
  levels(genotype_factor) <- c("AA", "AB", "BB")
  genotype_counts <- table(genotype_factor)
  return(genotype_counts)
}

# Apply the function to each variant column in the data frame
all_genotype_counts <- lapply(data, process_variant)

# Convert the list of counts to a data frame
all_genotype_counts_df <- as.data.frame(do.call(cbind, all_genotype_counts))
all_genotype_counts_df_T <- as.data.frame(t(all_genotype_counts_df))
head(all_genotype_counts_df_T)
```

```
##                AA AB BB
## rs587720402_A 106  1  0
## rs139377059_T 106  1  0
## rs587756191_T 106  1  0
## rs587702478_C 106  1  0
## rs62224609_C   91 16  0
## rs62224611_C   94 13  0
```

Question 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?

```
exact.pvalues <- HWExactStats(all_genotype_counts_df_T, x.linked=FALSE)

# Set the significance level
alpha <- 0.05
# Calculate the percentage of significant p-values
percentage_significant <- mean(exact.pvalues < alpha) * 100
```

Percentage of significant p-values: 2.772 %.

This doesn't necessarily represent the number of markers out of equilibrium by chance alone.

Deviations from HWE can be caused by various factors, such as population structure, selection, or genotyping errors, and not just random chance.

If the markers were in Hardy-Weinberg equilibrium, we wouldn't expect them to be out of equilibrium by the effect of chance, assuming the assumptions of HWE are met. Deviations from HWE can be indicative of

underlying biological or technical factors influencing the genetic variation in the data.

Question 6

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

To identify the SNP that is most significant according to the exact test results, we examine the p-values and select the SNP with the smallest p-value.

```
# Find the SNP with the smallest p-value
most_significant_snp_index <- which.min(exact.pvalues)

# Get the genotype counts for the most significant SNP
most_significant_snp_counts <- all_genotype_counts_df[, most_significant_snp_index,
                                                       drop = FALSE]

# The p-value for the most significant SNP
min_pvalue <- exact.pvalues[most_significant_snp_index]
```

The most significant SNP according to the exact test results: .

For this genotypic composition we have $f_1(AB) = 2pq = 0$.

Thus, in the next generation the allele frequencies will be as follows:

$$f_1(A) = \frac{2f_1(AA) + f_1(AB)}{2} = f_1(AA) + \frac{1}{2}f_1(AB) = p^2 \neq p$$
$$f_1(B) = \frac{2f_1(BB) + f_1(AB)}{2} = f_1(BB) + \frac{1}{2}f_1(AB) = q^2 \neq q$$

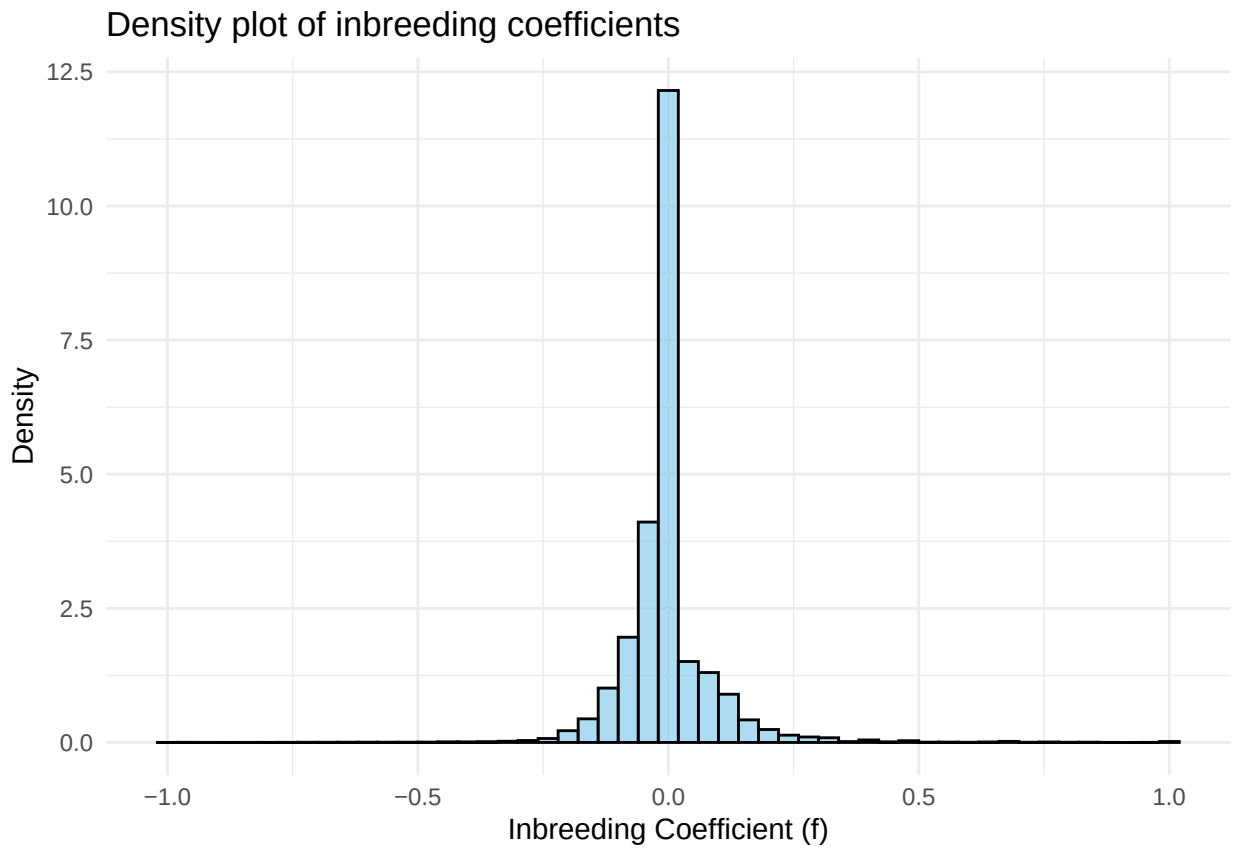
The population is not in Hardy-Weinberg equilibrium for this gene, because the allele frequencies will change across generations.

Question 7

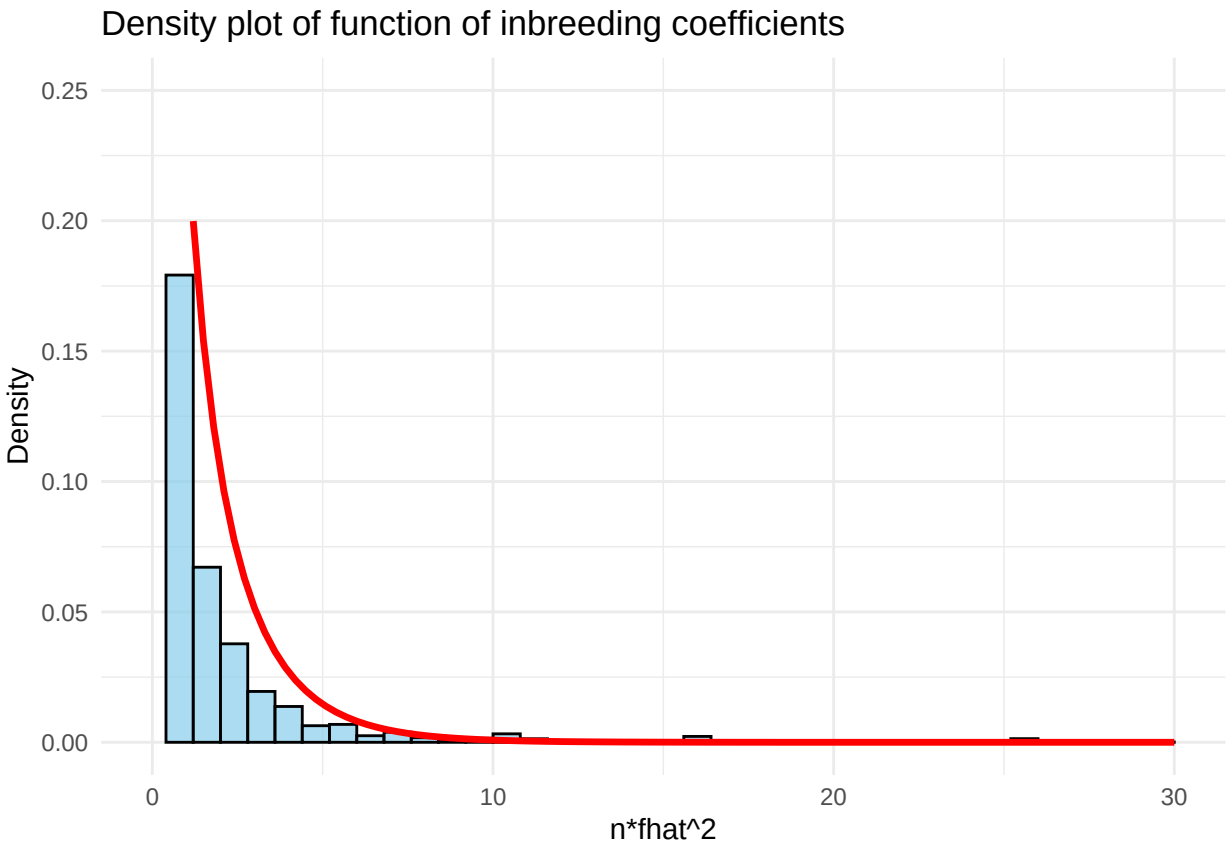
Compute the inbreeding coefficient (f) for each SNP, and make a histogram of f. You can use function `HWf` for this purpose. Give descriptive statistics (mean, standard deviation, etc) of f calculated over the set of SNPs. What distribution do you expect f to follow theoretically? Use a probability plot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

```
fhat <- lapply(all_genotype_counts_df, HWf)
n <- lapply(all_genotype_counts_df, sum)

# Calculate n*f_hat^2
nfhat2 <- unname(unlist(n)) * unname(unlist(fhat))^2

# Dataframes
inbreeding_df <- data.frame(row.names = names(fhat), fhat = unname(unlist(fhat)))
nfhat2_df <- data.frame(row.names = names(fhat), nfhat2 = nfhat2)

# Create a density plot of inbreeding coefficients
inbreeding_plot <- ggplot(inbreeding_df, aes(x = fhat)) +
  geom_histogram(binwidth = 0.04, aes(y = ..density..), fill = "skyblue",
                 color = "black", alpha = 0.7) +
  labs(x = "Inbreeding Coefficient (f)", y = "Density") +
```

```
  theme_minimal()+
  ggtitle("Density plot of inbreeding coefficients")

print(inbreeding_plot)
```




```
# Create a density plot of n*f_hat^2
nfhat2_plot <- ggplot(nfhat2_df, aes(x = nfhat2)) +
  geom_histogram(binwidth = 0.8, aes(y = ..density..), fill = "skyblue",
                 color = "black", alpha = 0.7) +
  labs(x = "n*fhat^2", y = "Density") +
  theme_minimal()+
  ggtitle("Density plot of function of inbreeding coefficients")

# Add a chi-squared distribution on the histogram
with_chi_squared_plot <- nfhat2_plot +
  stat_function(fun = dchisq, args = list(df = 1),
                color = "red", linewidth = 1.2)+
  ylim(0, 0.25) + xlim(0, 30)

print(with_chi_squared_plot)
```



We expect $n\hat{f}^2$ to follow the χ^2 distribution. According to the plot, we could conclude that we received a similar result, even though it would be clearer if we had more data.

```
# Descriptive statistics of inbreeding coefficients
fhat_vector <- unname(unlist(fhat))

# Get specific descriptive statistics
mean_value <- mean(fhat_vector)
median_value <- median(fhat_vector)
sd_value <- sd(fhat_vector)
var_value <- var(fhat_vector)
min_value <- min(fhat_vector)
max_value <- max(fhat_vector)

# Print the results
cat("Mean:", mean_value, "\n")
```

```
## Mean: -0.004682514
```

```
cat("Median:", median_value, "\n")
```

```
## Median: -0.004694836
```

```
cat("Standard Deviation:", sd_value, "\n")
```

```
## Standard Deviation: 0.09508609
```

```
cat("Variance:", var_value, "\n")
```

```
## Variance: 0.009041365
```

```
cat("Minimum:", min_value, "\n")
```

```
## Minimum: -1
```

```
cat("Maximum:", max_value, "\n")
```

```
## Maximum: 1
```

Question 8

Apply the exact test for HWE to each SNP, using different significant levels. Report the number and percentage of significant variants using an exact test for HWE with $\alpha = 0.10$, 0.05, 0.01 and 0.001. State your conclusions.

```
# Set the significance levels
alpha_levels <- c(0.10, 0.05, 0.01, 0.001)

# Iterate through the list of significance levels
for (alpha in alpha_levels) {

  exact.pvalues <- HWExactStats(all_genotype_counts_df_T, x.linked=FALSE)

  # Calculate the number of significant variants
  number_significant <- sum(exact.pvalues < alpha)

  # Calculate the percentage of significant variants
  percentage_significant <- mean(exact.pvalues < alpha) * 100

  # Print the results
  cat("Number of significant variants at alpha =", alpha, ":", number_significant, "\n")
  cat("Percentage of significant variants at alpha =", alpha, ":",
      round(percentage_significant, digits = 3), "%", "\n")
}
```

```
## Number of significant variants at alpha = 0.1 : 10052
## Percentage of significant variants at alpha = 0.1 : 4.808 %
## Number of significant variants at alpha = 0.05 : 5796
## Percentage of significant variants at alpha = 0.05 : 2.772 %
## Number of significant variants at alpha = 0.01 : 2511
## Percentage of significant variants at alpha = 0.01 : 1.201 %
## Number of significant variants at alpha = 0.001 : 1488
## Percentage of significant variants at alpha = 0.001 : 0.712 %
```

The smaller the significant level we take, the smaller the number of significant variants becomes. In other words, we make the condition for equilibrium less strict.