BSG-MDS practical 2 Statistical Genetics

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

Pyry Satama Max de Visser

```
library(HardyWeinberg)
```

```
## 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
```

```
library(data.table)
genotype_data = fread("TSIChr22v4.raw")
df <- data.frame(genotype_data)
df <- df[, (7:ncol(df))] # remove unused columns
```

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

```
n_variants = ncol(df)
cat("Number of variants in the dataset:", n_variants)
```

```
## Number of variants in the dataset: 1102156
```

```
# Convert non-0, 1, or 2 values to NA
df[!sapply(df, function(x) x %in% c(0, 1, 2))] <- NA

n_missing = sum(is.na(df))
total_data_points = ncol(df) * nrow(df)
perc_missing = (n_missing / total_data_points)*100
cat("\nPercentages missing data:", perc_missing)
```

```
##
## Percentages missing data: 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?

```
monomorphic <- sapply(df, function(x) length(unique(na.omit(x))) == 1)
cat("Percentage of monomorphic variants:", (sum(monomorphic)/n_variants)*100, "%\n")
```

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

```
df_poly <- df[, !monomorphic]
cat(ncol(df_poly), "variants remain in the database")
```

```
## 209074 variants remain in the database
```

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?

```
specific_variant <- df_poly[,"rs587756191_T"]
genotype_counts <- c(
  AA=sum(specific_variant==0),
  AB=sum(specific_variant==1),
  BB=sum(specific_variant==2)
)

# Chi-square test without continuity correction (x.linked = FALSE indicates the marker is autosoma)
hw_chisq <- HWChisq(genotype_counts, cc=0, verbose = TRUE, x.linked = FALSE)
```

```
## Warning in HWChisq(genotype_counts, cc = 0, verbose = TRUE, x.linked = FALSE):
## 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
```

```
# Chi-square test with continuity correction
hw_chisq_corrected <- HWChisq(genotype_counts, cc = 0.5, verbose = TRUE, x.linked = FALSE)
```

```
## Warning in HWChisq(genotype_counts, cc = 0.5, verbose = TRUE, x.linked =
## FALSE): 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
```

```
# Exact test for Hardy-Weinberg Equilibrium
hw_exact <- HWExact(genotype_counts, verbose = TRUE, x.linked = FALSE)
```

```
## 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 for Hardy-Weinberg Equilibrium
hw_perm <- HWPerm(genotype_counts, verbose = TRUE, x.linked = FALSE) # uses chi-squared by default
```

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

The chi-square test without continuity correction and the exact and permutation tests suggest that the variant rs587756191_T is in Hardy-Weinberg equilibrium. The chi-square test with continuity correction indicates a deviation and that the variant is not in equilibrium, so this result may not be reliable due to the low genotype counts for AB and BB. Overall, considering the consistency of the other three tests, it seems likely that this variant is in Hardy-Weinberg equilibrium.

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

```
p <- ncol(df_poly)

# Create a matrix to store genotype counts
genotype_counts_all <- matrix(0, nrow = p, ncol = 3)

for (i in 1:p) {
  # Compute genotype couns
  genotype_counts_all[i, 1] <- sum(df_poly[, i] == 0)  # Homozygous for allele 1 AA
  genotype_counts_all[i, 2] <- sum(df_poly[, i] == 1)  # Homozygous for allele 2 AB
  genotype_counts_all[i, 3] <- sum(df_poly[, i] == 2)  # Heterozygous 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 alpha = 0.05)? Is this the number of markers that you would expect to be out of equilibrium by the effect of chance alone?

```
# compute exact test for each variant
alpha <- 0.05
hw_exact_all <- HWExactStats(genotype_counts_all)

# Calculate percentage of variants that are significant with alpha = 0.05
count_below_005 <- sum(hw_exact_all < alpha)
perc_below_005 <- (count_below_005 / length(hw_exact_all)) * 100
cat("Percentage of significant SNPs:", perc_below_005, "%\n")
```

```
## Percentage of significant SNPs: 2.770789 %
```

We find that when using Fisher's exact test to test for HWE, the percentage of significant SNPs is 2.77%, and thus for these variants we will reject the null hypothesis that they are in HWE with a significance level of 0.05. This is the proportion of the variants we would expect to deviate from HWE due to pure chance alone. So if we investigate for HWE with a significance level of 0.05, we would want the percentage of significant variants to be below 0.05, thus indicating that there are other factors influencing the deviations from the

HWE in the variants.

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

```
min_idx <- which.min(hw_exact_all)
most_significant_variant_name <- names(df_poly)[min_idx]
cat("Most significant variant is:",most_significant_variant_name, "with a p-value of ", min(hw_exact_al
```

```
## Most significant variant is: rs2629366_C with a p-value of  9.784766e-33
```

```
most_significant_variant_values <- df_poly[,most_significant_variant_name]

# Get genotype counts
genotype_counts <- c(
  sum(most_significant_variant_values==0),
  sum(most_significant_variant_values==1),
  sum(most_significant_variant_values==2)
)

cat("\nGenotype counts:\nAA:", genotype_counts[1], "\nAB:", genotype_counts[2], "\nBB:", genotype_counts
```

```
##
## Genotype counts:
## AA: 56
## AB: 0
## BB: 51
```

```
observed_frequencies <- genotype_counts / nrow(df_poly)

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

# Expected genotype frequencies under HWE:
exp_AA_freq <- p^2
exp_AB_freq <- 2*p*q
exp_BB_freq <- q^2


cat("\nHWE expected AA freqeuncy:", exp_AA_freq, "vs. observed frequency:", observed_frequencies[1],
    "\nHWE expected AB frequency: ", exp_AB_freq, "vs. observed frequency:", observed_frequencies[2],
    "\nHWE expected BB frequency: ", exp_BB_freq, "vs. observed frequency:", observed_frequencies[3]
    )
```

```
##
## HWE expected AA freqeuncy: 0.2739104 vs. observed frequency: 0.5233645
## HWE expected AB frequency:  0.4989082 vs. observed frequency: 0
## HWE expected BB frequency:  0.2271814 vs. observed frequency: 0.4766355
```

For bi-allelic markers, the expected genotype frequencies under HWE are defined as $f(AA) = p^2$, $f(AB) = 2pq$ and $f(BB) = q^2$ where $f(A) = p$ (frequency of allele A) and $f(B) = q$ (frequency of allele B).

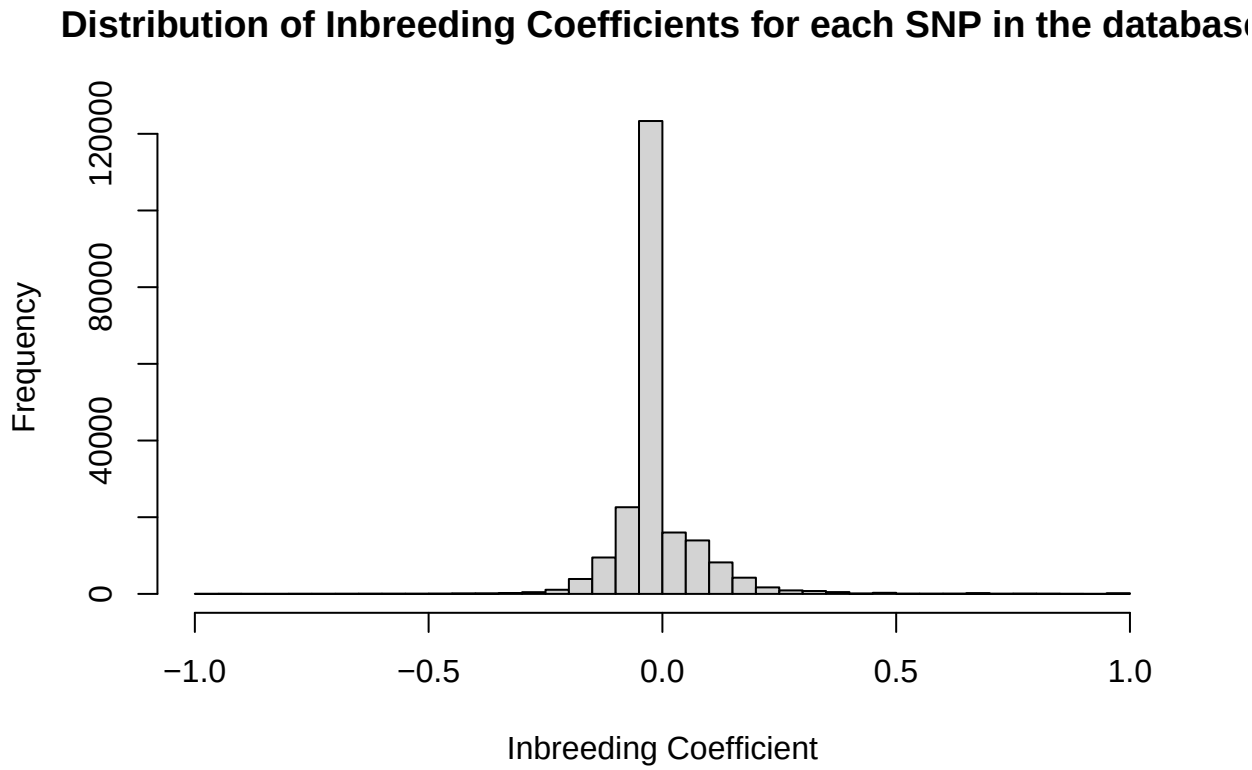
If we calculate the expected genotype frequencies under HWE (as done above) and compare with observed, we see that this variant is unusual (w.r.t. HWE) in that the frequency of heterozygous alleles is 0, which we

would, in accordance with HWE, expect to be 0.49, explaining how this variant violates the HWE significantly.

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.

```
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(df_poly, 2, inbreeding_factor)
hist(inbreeding_coeffs, main="Distribution of Inbreeding Coefficients for each SNP in the database", xla
```




```
# Compute descriptive statistics
print(summary(inbreeding_coeffs))
```

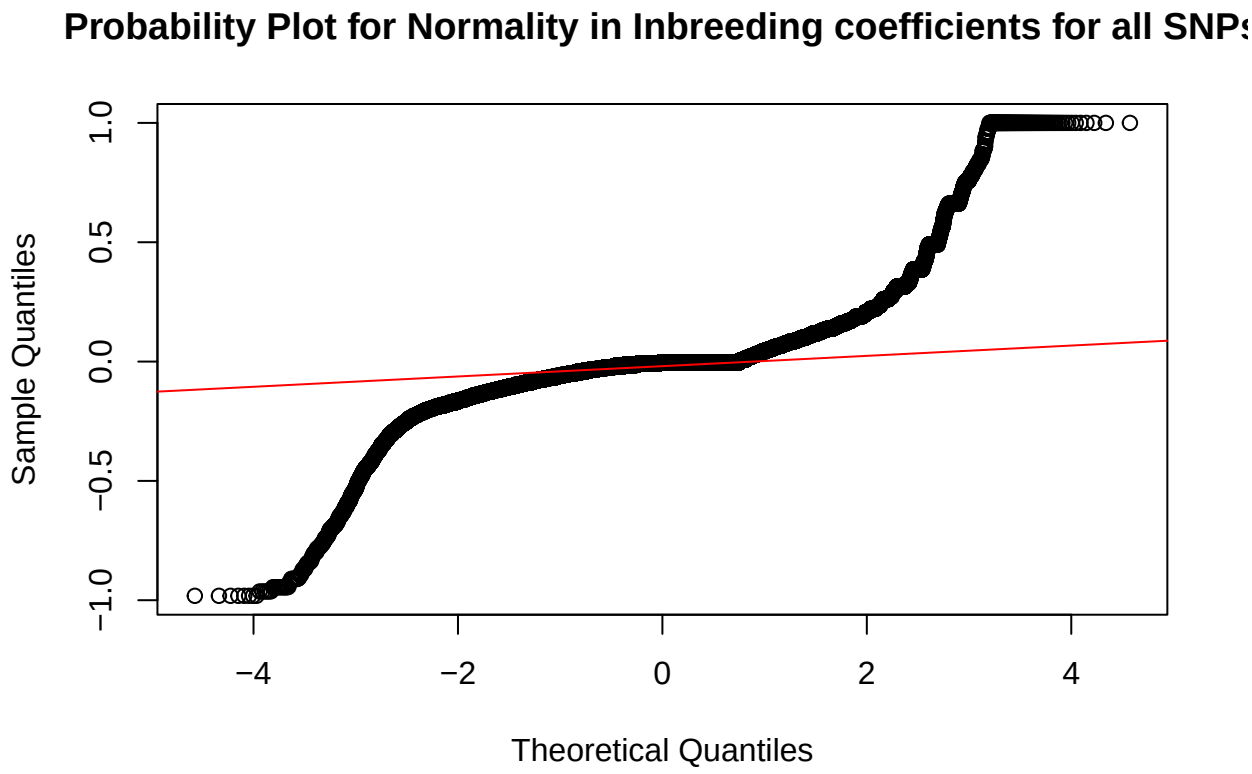
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
```

```
## -0.981482 -0.033816 -0.004695 -0.004668 -0.004695  1.000000
```

```
cat("\nStandard Deviation: ", std <- sd(inbreeding_coeffs))
```

```
##
## Standard Deviation:  0.095012
```

```
# # Create probability plot with assumption of a normal distribution
qqnorm(inbreeding_coeffs, main ="Probability Plot for Normality in Inbreeding coefficients for all SNPs
qqline(inbreeding_coeffs, col = "red")
```



Theoretically, the distribution of inbreeding coefficients is expected to follow a normal distribution. This is because, under HWE, the genotype frequencies stabilise and we expect that the majority of the database is in HWE since only 2.8 (see exercise 5) are statistically significant to deviate from HWE. From inspecting the histogram, we also more or less observe a normal distribution of the data. However, inspecting the probability plot above, it suggests noticeable deviations from normality in the inbreeding coefficients indicating the presence of outliers in the inbreeding coefficients. These are also slightly noticeable in the histogram.

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.

```
hw_exact_all <- HWExactStats(genotype_counts_all)

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

```r
results <- sapply(alpha_levels, function(alpha) {
  significant_count <- sum(hw_exact_all < alpha)
  percent_significant <- (significant_count / length(hw_exact_all)) * 100
  return(c(count = significant_count, percent = percent_significant))
})


results_df <- as.data.frame(t(results))
colnames(results_df) <- c("Count", "Percentage")
rownames(results_df) <- paste("Alpha =", alpha_levels)
results_df
```

```
##               Count Percentage
## Alpha = 0.1   10049  4.8064322
## Alpha = 0.05   5793  2.7707893
## Alpha = 0.01   2508  1.1995753
## Alpha = 0.001  1485  0.7102748
```

The results indicate that as the alpha level becomes more stringent, the number and percentage of SNPs showing significant deviation from HWE decrease. This pattern is expected in large datasets where some SNPs may deviate from HWE due to chance, population structure, selection, or other factors. The relatively low percentages of significant SNPs at stricter alpha levels suggest that most SNPs in the dataset are in Hardy-Weinberg equilibrium.