Practical 02 SG: Hardy-Weinberg equilibrium

Carlos Moyano & Kleber Reyes

2022-11-18

Hardy-Weinberg equilibrium

INFO: https://cran.r-project.org/web/packages/data.table/vignettes/datatable-faq.html

2. How many individuals does the database contain, and how many variants? What percentage of the variants is monomorphic? Remove all monomorphic SNPs from the database. How many variants remain in the database?

```
variants <- ncol(geneticData); variants;
```

```
## [1] 1102156
```

```
individuals <- nrow(geneticData); individuals;
```

```
## [1] 107
```

```
cols <- which(colSums(geneticData == 1, na.rm = TRUE) > 0) # Non monomorphic (contains AB)
variants.poly <-length(cols);
variants.mono <- variants-variants.poly;
perc.mono <- 100*variants.mono/variants; perc.mono
```

```
## [1] 81.04325
```

```
geneticData.poly <- geneticData[, ..cols]

remove(geneticData)
ncol(geneticData.poly) #208933
```

```
## [1] 208933
```

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

We consider this variant is not in equilibrium, the first signs could be seen in the warning of the HWChisq since we have occurences below 5. The p-value obtained is 6.495738e-25 that is lower than 0.05 so we reject the null hypothesis (observed proportions are equal to the expected counts under HWE), this entails that the variant has not reached the HWP.

```
rs587756191_T <- dplyr::recode(geneticData.poly$rs587756191_T, `0`="AA", `1`="AB", `2`="BB")

rs587756191_T.g <- genotype(rs587756191_T,sep="")
rs587756191_T.g.summary <- summary(rs587756191_T.g)
rs587756191_T.g.summary$genotype.freq #genotype counts
```

```
##     Count  Proportion
## A/A   106 0.990654206
## A/B     1 0.009345794
```

```
x <- MakeCounts(geneticData.poly$rs587756191_T)[1,1:3]
results.chi <- HWChisq(x,cc=0.5)
```

```
## Warning in HWChisq(x, cc = 0.5): 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
```

```
results.chi.nocor <- HWChisq(x,cc=0)
```

```
## Warning in HWChisq(x, 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
```

```
results.exact <- HWExact(x)
```

```
## 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
```

```
results.perm <- HWPerm(x)
```

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

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

```
alpha <- 0.05
count_matrix <- MakeCounts(geneticData.poly)[,1:3]
```

5. Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. You can use HWChisqStats for this purpose. How many SNPs are significant (use = 0.05)?

96.16097% of the SNPs have a p-value over 0.05 (that entails that these SNPs have HWE). We can consider the other 3.83903% as imbalanced.

```
Chisq.pvals <- HWChisqStats(count_matrix, pvalues=TRUE)
p <- 100 * length(which(Chisq.pvals > 0.05)) / length(Chisq.pvals);p
```

```
## [1] 96.16097
```

6. How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone?

```
c <- 100 * (1 - (length(which(Chisq.pvals > 0.05)) / length(Chisq.pvals)));c
```

```
## [1] 3.83903
```

7. Which SNP is most significant according to the chi-square test results? Give it genotype counts. In which sense is this genotypic composition unusual?

```
max.ind <- which.max(Chisq.pvals)

most.sign.SNP.chisq <- geneticData.poly[, which.max(Chisq.pvals), with=FALSE]
names(most.sign.SNP.chisq)
```

```
## [1] "rs5748532_T"
```

```
gen.counts <- MakeCounts(most.sign.SNP.chisq)[,1:3];gen.counts
```

```
## AA AB BB
## 32 53 22
```

The composition of the genotype is unusual in the sense that it contains a very high number of AB in comparison to AA or BB (very high number of heterozygotes).

8. Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. You can use function HWExactStats for fast computation. How many SNPs are significant (use = 0.05). Is the result consistent with the chi-square test?

```
exact_test_pvals <- HWExactStats(count_matrix)
sum(exact_test_pvals<=alpha) # num. significant SNPs
```

```
## [1] 5652
```

```
(sum(exact_test_pvals<=alpha) / nrow(count_matrix)) * 100 # %
```

```
## [1] 2.705173
```

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

```
most_significant_SNP <- which.min(exact_test_pvals)
count_matrix[most_significant_SNP,][1:3]
```

```
##  AA  AB  BB
##   0 107   0
```

The composition of the genotype is unusual in the sense that it contains a very high number of AB (heterozygotes), and no AA or BB.

10. Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP, using the HWLratio function. How many SNPs are significant (use = 0.05). Is the result consistent with the chi-square test?

```
m <- ncol(geneticData.poly)
likelihood_ratio_test_pvals <- 0*m
for (i in 1:m) {
  likelihood_ratio_test_pvals[i] <- HWLratio(count_matrix[i,], verbose=FALSE)$pval
}
alpha <- 0.05
sum(likelihood_ratio_test_pvals<=alpha) # num. significant SNPs
```

```
## [1] 7814
```

```
(sum(likelihood_ratio_test_pvals<=alpha) / nrow(count_matrix)) * 100 # %
```

```
## [1] 3.739955
```

The percentage of SNPs in equilibrium is higher than with the chi-square test, so there exists some discrepancy.

11. Apply a permutation test for Hardy-Weinberg equilibrium to the first 10 SNPs, using the classical chi-square test (without continuity correction) as a test statistic. List the 10 p-values, together with the 10 p-values of the exact tests. Are the results consistent?

```
m <- 10
perm_test_pvals <- 0*m
count_matrix_10_first <- MakeCounts(geneticData.poly[,1:m])[,1:3]
for (i in 1:nrow(count_matrix_10_first)) {
  perm_test_pvals[i] <- HWPerm(count_matrix_10_first[i,], verbose=FALSE)$pval
}
perm_test_pvals
```

```
##  [1] 1.000000000 1.000000000 1.000000000 1.000000000 0.639882353 1.000000000
##  [7] 1.000000000 1.000000000 0.122176471 0.008647059
```

```
exact_test_pvals[1:10]
```

```
##  [1] 1.000000000 1.000000000 1.000000000 1.000000000 1.000000000 1.000000000
##  [7] 1.000000000 1.000000000 0.214715301 0.008643867
```

```
sum(perm_test_pvals<=0.05)==sum(exact_test_pvals[1:10]<=0.05)
```

```
## [1] TRUE
```

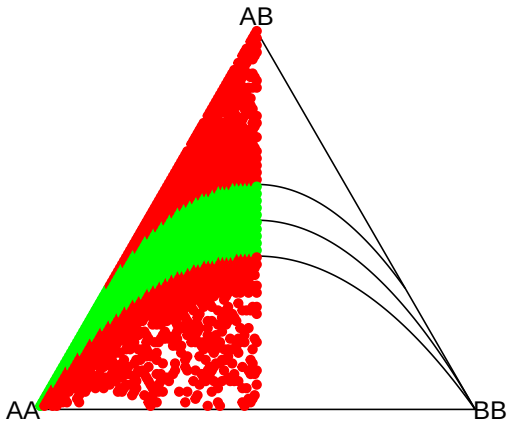
Although the p-values vary a little bit, the results are consistent, since the same SNPs are marked as significant by both tests.

12. Depict all SNPs simultaeneously in a ternary plot with function HWTernaryPlot and comment on your result (because many genotype counts repeat, you may use UniqueGenotypeCounts to speed up the computations)

```
unique_genotypes_count <- UniqueGenotypeCounts(count_matrix)[,1:3]
```

```
## 208933 rows in X
## 1896 unique rows in X
```

```
HWTernaryPlot(unique_genotypes_count)
```

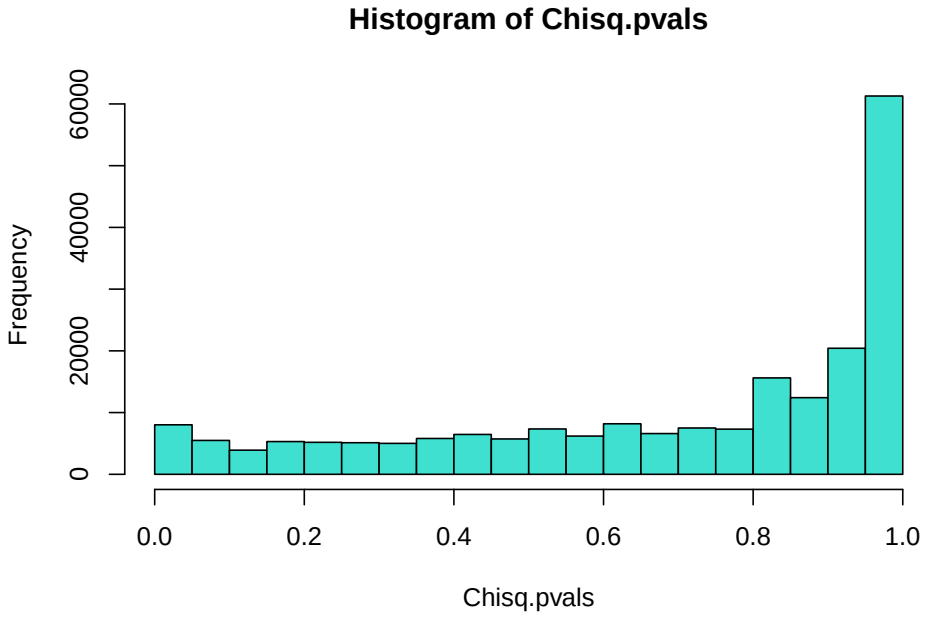


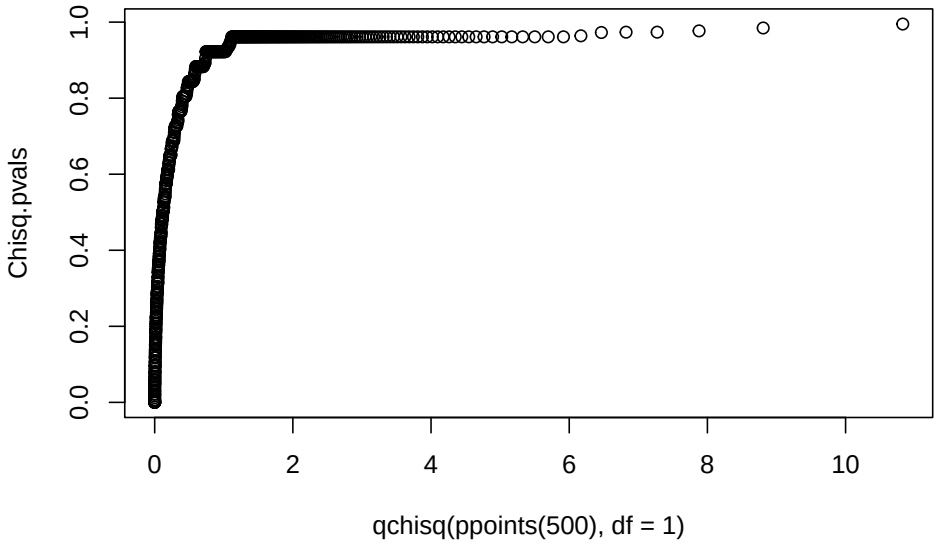
When all the SNPs are shown simultaneously in a ternary plot, the area which represents the acceptance region is much more dense in terms of SNPs projected there that the rejection area. This means that, although there are many SNPs that could be considered as not in HWE equilibrium, we have a great amount of SNPs that are in HWE equilibrium (depicted in green).

13. Can you explain why half of the ternary diagram is empty?

It is half empty due to the fact that the genotypes we're working with have very few counts of 'BB'.

14. Make a histogram of the p-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? Make a Q-Q plot of the p values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?.





We would expect a chi-sq with 1 df, and it seems it's the case (if we consider that the distribution is flipped and the peak is in the region close to 1). The conclusion, observing the histogram and the Q-Q plot is that the HWE holds for the dataset.

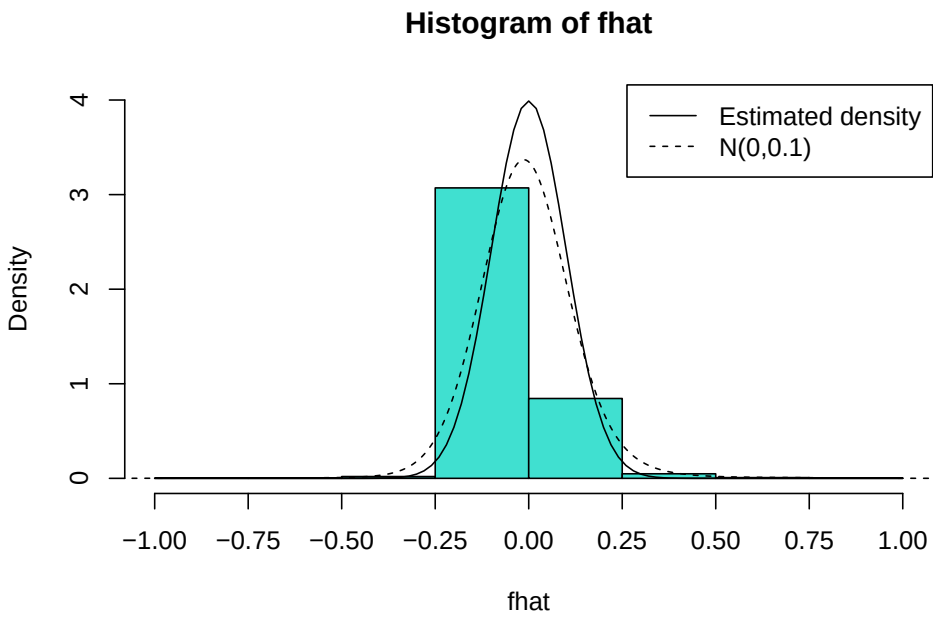
15. Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Try it on the computer if you want. Argue your answer.

Since the proportions under HWE for AA and BB are different, accidentally interchanging the counts would have a direct effect in the result of the statistical tests. This is quite clear in the case of the chi-square statistic proposed in the lectures' notes, due to the fact that it compares observed proportions with expected proportions. If there's an error and the observed proportions are interchanged, the computed value would be very different to the true value.

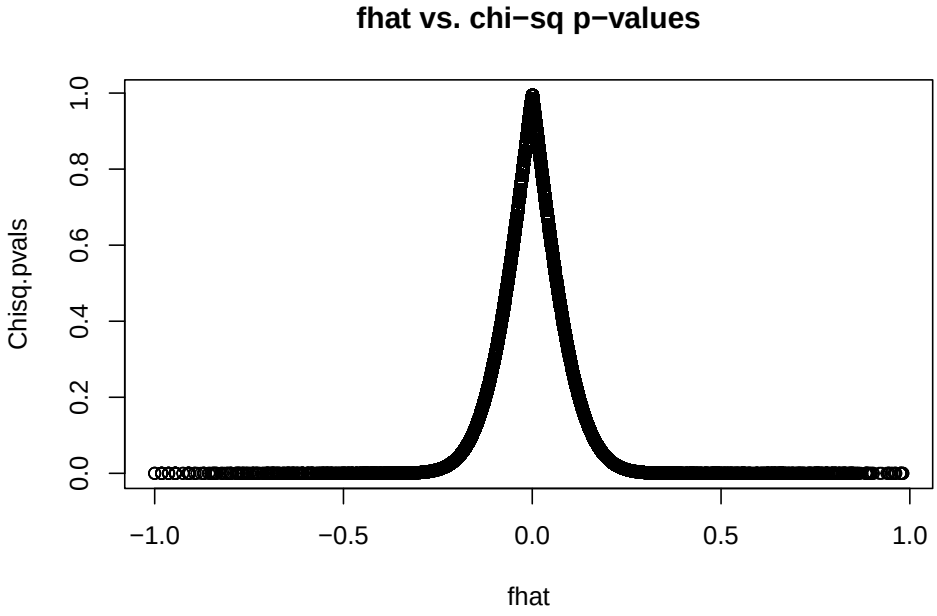
16. Compute the inbreeding coefficient (ˆ f) for each SNP, and make a histogram of ˆ f. You can use function HWf for this purpose. Give descriptive statistics (mean, standard deviation, etc) of ˆ f calculated over the set of SNPs. What distribution do you expect ˆ f to follow theoretically? Use a probability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

```
fhat <- HWf(count_matrix)
psych::describe(fhat)
```

```
##    vars      n  mean   sd median trimmed  mad min  max range skew kurtosis se
## X1    1 208933 -0.01 0.09      0   -0.01 0.03  -1 0.98  1.98 1.19    19.44  0
```



17. Make a plot of the observed chi-square statistics against the inbreeding coefficient (f'). What do you observe? Can you give an equation that relates the two statistics?



In the histogram it is not easy to clearly appreciate the underlying distribution, but it resembles to a normal centered in 0 with a very low variance (this always happens in a dataset where there's HWE, since fhat has value 0 for SNPs when they're under HWE). The inbreeding coefficient is related with a chi-sq by means of the formula n * fhat^2.

18. We reconsider the exact test for HWE, 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.