# Practical 02 SG: Hardy-Weinberg equilibrium

Lovro Katalinić and Ivan Almer

Hand-in: 30/11/2020

Resolve the following exercise in groups of two students. Perform the computations and make the graphics that are asked for in the practical below. Take care to give each graph a title, and clearly label $x$ and $y$ axes, and to answer all questions asked. You can write your solution in a word or Latex document and generate a pdf file with your solution. Alternatively, you may generate a solution pdf file with Markdown. You can use R packages **data.table** and **HardyWeinberg** for the computations. Take care to number your answer exactly as in this exercise. Upload your solution in **pdf format** to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

```
## Loading required package: combinat
##
## Attaching package: 'combinat'
## The following object is masked from 'package:utils':
##
##     combn
## Loading required package: gdata
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
## Loading required package: gtools
## Loading required package: MASS
## Loading required package: mvtnorm
##
## 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.
##
##
## Attaching package: 'genetics'
## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
## 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
##
## Attaching package: 'data.table'
## The following objects are masked from 'package:gdata':
##
##     first, last
```

1. 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. Use the `fread` instruction of the package `data.table`, which is more efficient for reading large datafiles. This data is in (0,1,2) format, where 0 and 2 represent the homozygotes AA and BB, and 1 represents the heterozygote AB. The first six leading columns of the data matrix can be ignored, as they do not contain any genetic information.

```
# Load data
dataset <- fread('TSIChr22v4.raw', data.table=FALSE)

# Remove leading columns and display data
dataset <- dataset[-c(1:6)]
#head(dataset[1:8])
```

2. (1p) 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?

```
# Number of individuals and variants
n <- nrow(dataset)
v <- ncol(dataset)
cat(paste('Number of individuals:', n, '\n'))
```

```
## Number of individuals: 107
```

```r
cat(paste('Number of variants:', v, '\n'))
```

```
## Number of variants: 1102156
```

```r
# Number of polymorphic variants
polymorphic_cols <- apply(dataset, 2, function(x) length(unique(x)) > 1)
dataset_poly <- dataset[polymorphic_cols]
vp <- ncol(dataset_poly)
cat(paste('Number of polymorphic variants:', vp, '\n'))
```

```
## Number of polymorphic variants: 209074
```

3. (3p) 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.

```
## Warning in HWChisq(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
```

```
## Warning in HWChisq(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
```

```
## 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
## Observed statistic: 0.002358439   17000 permutations. p-value: 1
```

If the p-value is small enough, we reject the null hypotesis that the data comes from the specified distribution. We have to keep in mind that small number of counts of one genotype may lead to incorrect test results. In our case, we have **106** AA genotypes, **1** AB genotype and no BB genotypes. Chi-square tests warn us that the results may be incorrect because of the mentioned reason. Exact test and permutation test give better results with p-value equal to 1. So, considering these value, we can say that this variant is in equilibrium.

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

```r
genotype_counts <- data.frame(t(apply(dataset_poly, 2, count_genotypes)))
genotype_counts[1:20,]
```

```
##                    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
## rs192339082_A     106  1  0
## rs587740681_A     106  1  0
## rs4965031_A        77 30  0
## rs375684679_AAAAC  46 38 23
```

```
## rs587646183_C      99  7  1
## rs139918843_C     105  2  0
## rs587743102_T     106  1  0
## rs376238049_T      92 15  0
## rs200777521_A      93 14  0
## rs587710177_A     104  3  0
## rs587598082_T     106  1  0
## rs587701155_A     104  3  0
## rs80167676_T       93 14  0
## rs915675_A         62 34 11
```

5. (1p) 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 $\alpha = 0.05$)?

```
pvalues_chi <- HWChisqStats(genotype_counts, pvalues=TRUE)
pvalues_significant_chi <- pvalues_chi[pvalues_chi <= 0.05]
cat(paste('There are', length(pvalues_significant_chi), 'significant SNPs out of', length(pvalues_chi)))
```

```
## There are 8162 significant SNPs out of 209074
```

6. (1p) How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone?

```
chance = length(pvalues_significant_chi) / length(pvalues_chi )
chance
```

```
## [1] 0.03903881
```

```
cat(paste('We would expect to see', round(chance * 100, digits = 3), '% of markers to be out of equilib
```

```
## We would expect to see 3.904 % of markers to be out of equilibrium.
```

7. (2p) Which SNP is most significant according to the chi-square test results? Give it genotype counts. In which sense is this genotypic composition unusual?

```
most_significant_index <- which.min(pvalues_chi)
genotype_counts[most_significant_index,]
```

```
##                AA AB BB
## rs573187031_T 106  0  1
```

It is unusual because all genotypes of this SNPs are equal to AA besides one which is BB.

8. (1p) 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 $\alpha = 0.05$). Is the result consistent with the chi-square test?

```
pvalues_exact <- HWExactStats(genotype_counts)
pvalues_significant_exact <- pvalues_exact[pvalues_exact <= 0.05]
cat(paste('There are', length(pvalues_significant_exact), 'significant SNPs out of', length(pvalues_exa
```

```
## There are 5793 significant SNPs out of 209074
```

The number of significant SNPs given by an Exact test is smaller than the one given by a chi-square test, but if we compare it to the total number of SNPs, we can say that it is similar.

9. (2p) 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_index <- which.min(pvalues_exact)
genotype_counts[most_significant_index,]
```

```
##              AA AB BB
## rs2629366_C 56  0 51
```

This genotypic composition is unusual because it has almost equal number of AA and BB genotypes, but none of AB genotypes.

10. (1p) Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP, using the `HWLratio` function. How many SNPs are significant (use $\alpha = 0.05$). Is the result consistent with the chi-square test?

```
pvalues_likelihood <- apply(genotype_counts, 1, function(x) HWLratio(x, verbose=FALSE)$pval)
pvalues_significant_likelihood <- pvalues_likelihood[pvalues_likelihood <= 0.05]
cat(paste('There are', length(pvalues_significant_likelihood), 'significant SNPs out of', length(pvalues
```

```
## There are 7955 significant SNPs out of 209074
```

The result of the likelihood ratio test is much closer to chi-square test then exact test was. Likelihood ratio test estimated 7955 significant values, and chi-square test estimated 8162 of them.

11. (1p) 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 result consistent?

```
pvalues_perm_head <- apply(genotype_counts[1:10,], 1, function(x) HWPerm(x, verbose=FALSE)$pval)
pvalues_exact_head <- pvalues_exact[1:10]
rbind(pvalues_perm_head, pvalues_exact_head)
```

```
##                   rs587720402_A rs139377059_T rs587756191_T rs587702478_C
## pvalues_perm_head             1             1             1             1
## pvalues_exact_head            1             1             1             1
##                   rs62224609_C rs62224611_C rs192339082_A rs587740681_A
## pvalues_perm_head    0.6411765            1             1             1
## pvalues_exact_head   1.0000000            1             1             1
##                   rs4965031_A rs375684679_AAAAC
## pvalues_perm_head   0.1243529       0.009235294
## pvalues_exact_head  0.2147153       0.008643867
```

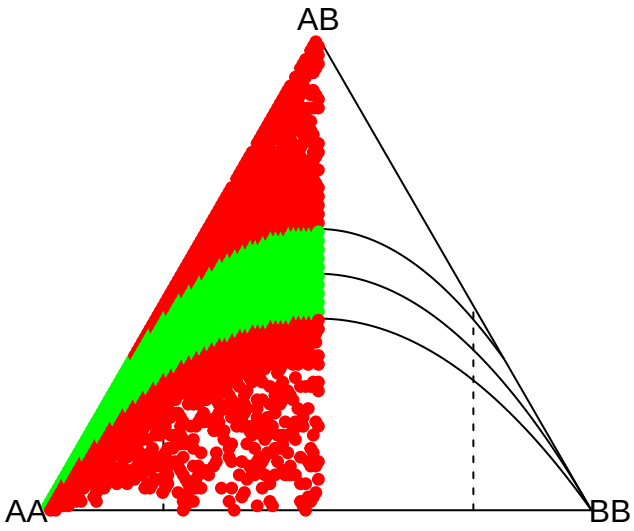
Not all p-values are consisent. We can see that all besides p-values of variant **rs62224609_C** are close.

12. (1p) 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_genotype_counts <- UniqueGenotypeCounts(genotype_counts)[1:3]
```

```
## 209074 rows in X
## 1900 unique rows in X
```

```
HWTernaryPlot(unique_genotype_counts)
```

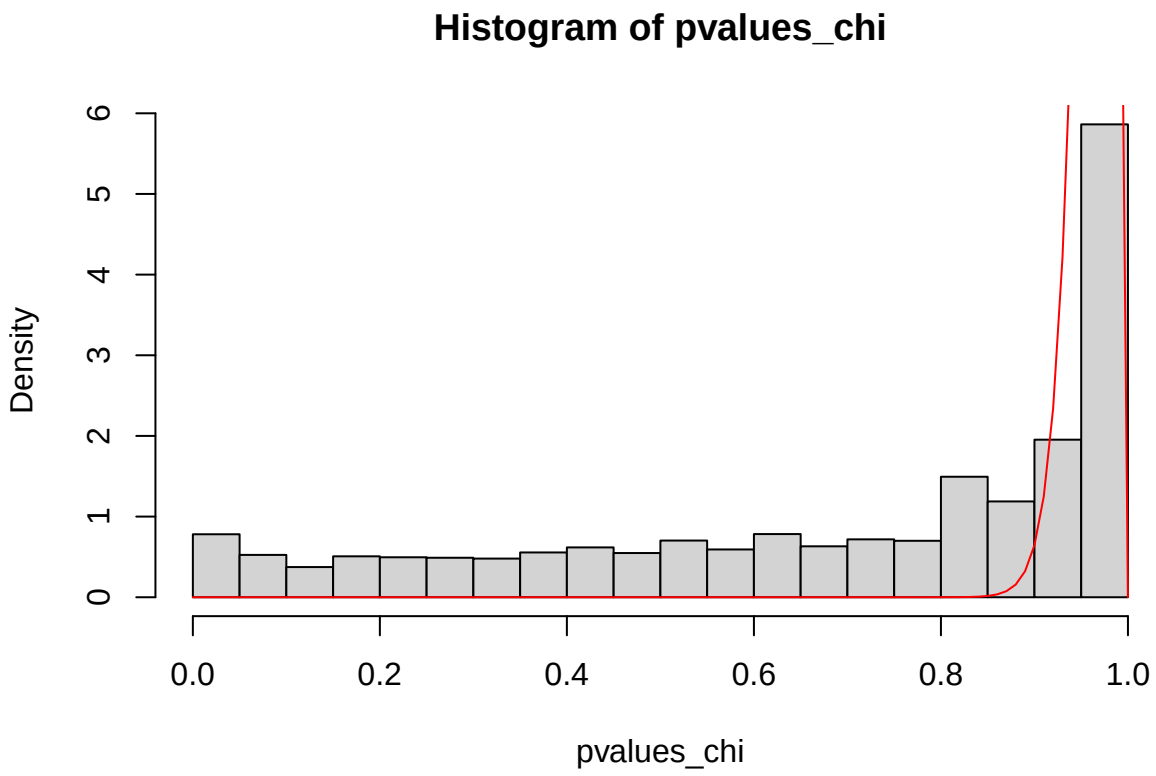



13. (1p) Can you explain why half of the ternary diagram is empty?

Discussing the possibilities why the right part of the ternary diagram is empty we came to the conclusion that A is probably always taken as the more frequent allele and thus there is no way when BB could be a more frequent case than AA.

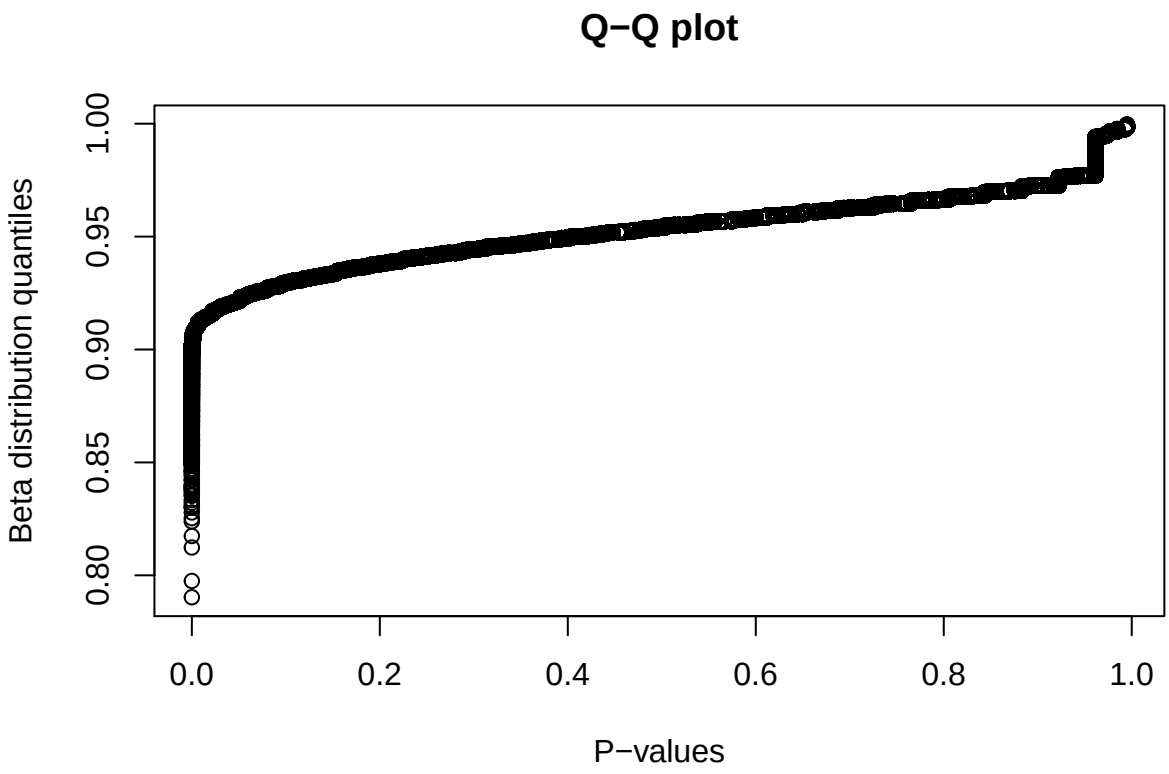
14. (2p) Make a histogram of the $p$-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? Make a Q-Q plot of the $p$ values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?.

```
hist(pvalues_chi, freq=FALSE)
curve (dbeta(x, shape1 = 80, shape2 = 3), add=TRUE, col="red")
```




```
x_qbeta <- seq(0, 1, length.out = length(pvalues_chi))
y_qbeta <- rbeta(x_qbeta, shape1 = 80, shape2 = 3)

qqplot(pvalues_chi, y_qbeta, main="Q-Q plot", xlab = "P-values", ylab = "Beta distribution quantiles")
```




If HWE would hold for the whole data set, we would expect that the left part of the histogram (below 0.05) is 0, i.e. that no row has p-value below 0.05. Ideally we would expect the mass of the histogram to shift as far right as possible.

This histogram is shaped like some very skewed distribution with heavy left tail. We could not find some distribution with a heavy tail like this, do the best fit for this is beta distribution with parameters alpha = 80 and beta = 3. Q-Q plot is presented in the plot, we see that in the middle there is linear behavior, but coming closer to each end of the curve the points start to deviate significantly from the line.

15. (1p) 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.

```
counts <- c(AA=40, AB=10, BB=15)
HWChisq(counts)
```

```
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  24.22897 DF =  1 p-value =  8.553578e-07 D =  -8.846154 f =  0.6388889
```

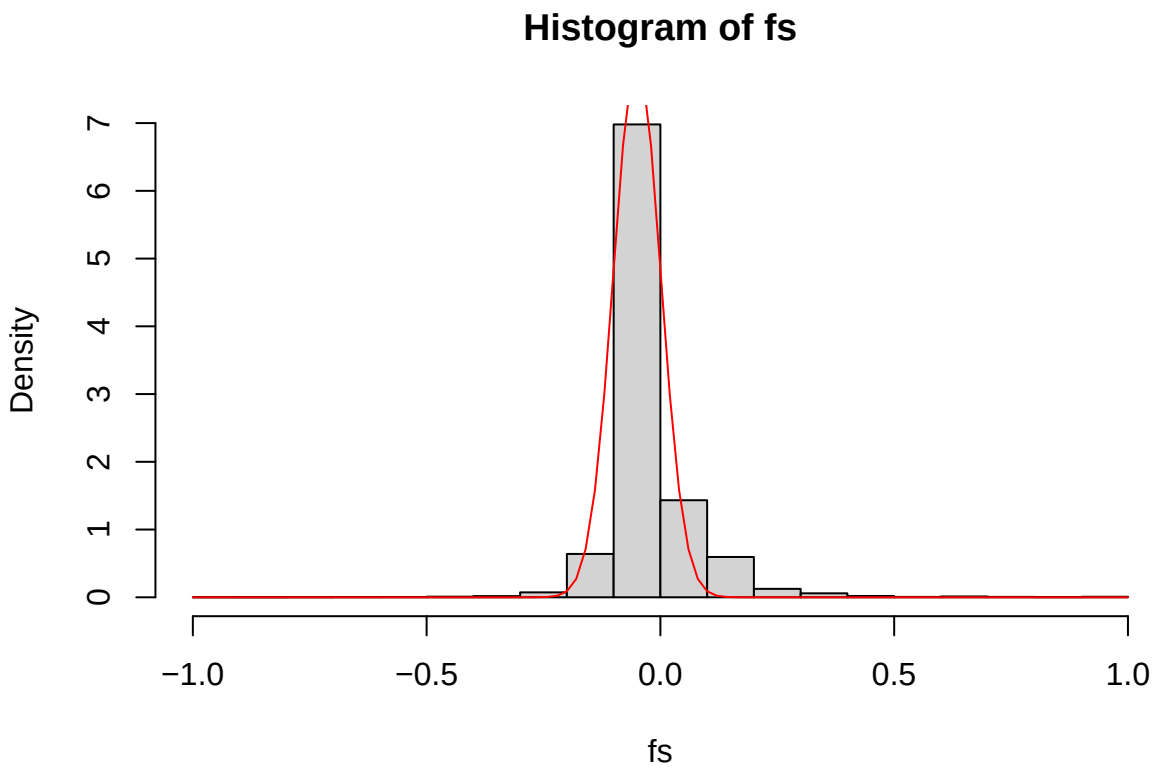
```
interchanged_counts <- c(AA=15, AB=10, BB=40)
HWChisq(interchanged_counts)
```

```
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  24.22897 DF =  1 p-value =  8.553578e-07 D =  -8.846154 f =  0.6388889
```

If counts of two homozygotes were accidentally interchanged, it wouldn't affect the statistical test for HWE. Switching counts of two homozygotes also switches the probabilities p and q, which does not change the test value. It can be seen if we observe formula for Chi-squared test (on slide 19 from the lesson's PDF). Heterozygote count does not change, so the nominator stays equal. Denominator consists of factor *p-squared*, *q-squared* and *n* and stays equal if we switch p and q values.

16. (3p) Compute the inbreeding coefficient ($\hat{f}$) for each SNP, and make a histogram of $\hat{f}$. You can use function `HWf` for this purpose. Give descriptive statistics (mean, standard deviation, etc) of $\hat{f}$ calculated over the set of SNPs. What distribution do you expect $\hat{f}$ to follow theoretically? Use a probability plot to confirm your idea.

```
fs <- HWf(as.matrix(genotype_counts))
hist(fs, freq=FALSE)
curve (dnorm(x, -0.05, 0.05), add=TRUE, col="red")
```



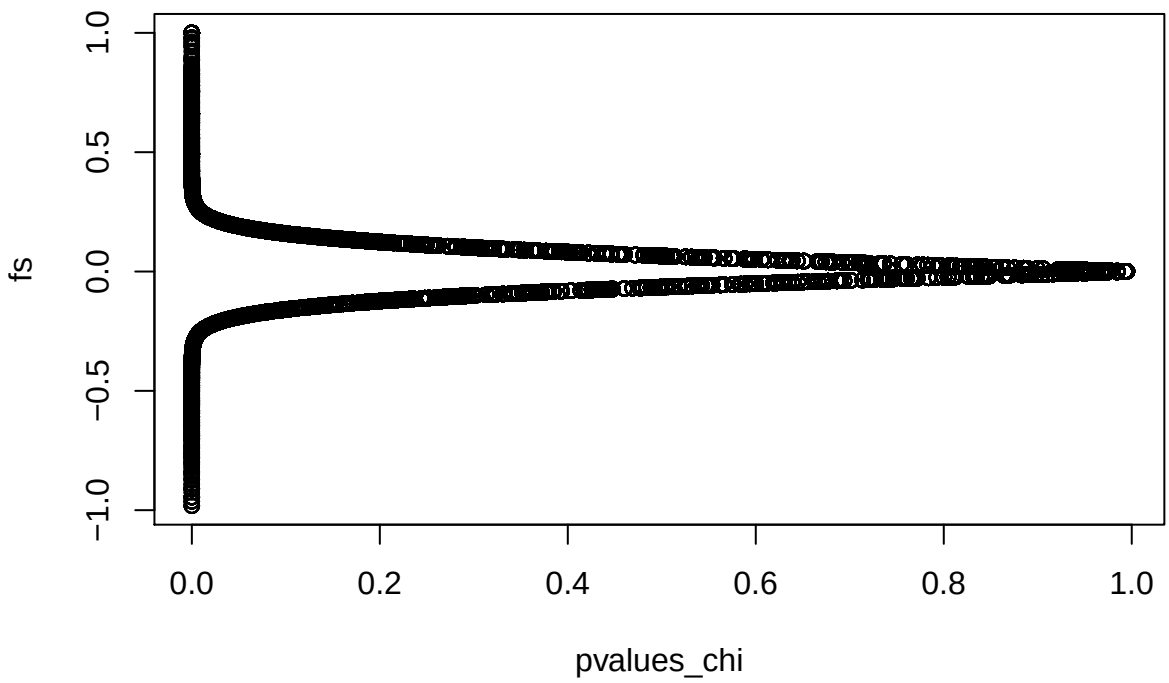

```
cat(paste('Mean:', mean(fs), '\nStandard deviation:', sd(fs), '\nMedian:', median(fs), '\nMin:', min(fs
```

```
## Mean: -0.00466823238593223
## Standard deviation: 0.0950119986299748
## Median: -0.00469483568075117
## Min: -0.981481481481482
## Max: 1
```

Considering that the mean and the median are very similar values, and looking at the histogram, it seems that inbreeding coefficients follow normal distribution. Distribution with parameters N(-0.05, 0.05) seems to make a good explanation of inbreeding coefficients distribution of this data set.

17. (2p) Make a plot of the observed chi-square statistics against the inbreeding coefficient ($\hat{f}$). What do you observe? Can you give an equation that relates the two statistics?

```r
plot(pvalues_chi, fs)
```




We can observe that p-values are closer to zero as inbreeding coefficient is further from zero, and vice versa. That makes sense, because low p-value indicates that variant is not in equilibrium, and it happens when there are no heterozygotes (f = 1) or much more heterozygotes than homozygotes (f = -1).

18. (2p) 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 $\alpha = 0.10, 0.05, 0.01$ and 0.001. State your conclusions.

```r
get_significant_percentage <- function(pvalues, alpha) {
  significant <- pvalues[pvalues <= alpha]
  length(significant) / length(pvalues)
}

significant_percentages <- c(get_significant_percentage(pvalues_exact, 0.10),
                             get_significant_percentage(pvalues_exact, 0.05),
                             get_significant_percentage(pvalues_exact, 0.01),
                             get_significant_percentage(pvalues_exact, 0.001))
names(significant_percentages) <- c('0.1', '0.05', '0.01', '0.001')
significant_percentages
```

```
##         0.1        0.05        0.01       0.001
## 0.048064322 0.027707893 0.011995753 0.007102748
```

We can see that percentage of significant variants decreases as alpha is approaching zero.