Bioinformatics and Statistical Genetic

Hardy-Weinberg equilibrium

Leonardo Ortoleva & Egon Ferri

Contents

```
require(genetics, quietly = T)
require(viridis, quietly = T)
require(knitr, quietly = T)
require(purrr, quietly = T)
require(HardyWeinberg, quietly = T)
require(data.table, quietly = T)
require(fitdistrplus, quietly = T)
require(e1071, quietly = T)
```

1 Load and clean data

FID	IID	PAT	MAT	SEX	PHENOTYPE	rs587697622_G
NA20502	NA20502	0	0	2	-9	0
NA20503	NA20503	0	0	2	-9	0
NA20504	NA20504	0	0	2	-9	0
NA20505	NA20505	0	0	2	-9	0
NA20506	NA20506	0	0	2	-9	0
NA20507	NA20507	0	0	2	-9	0
NA20508	NA20508	0	0	2	-9	0
NA20509	NA20509	0	0	1	-9	0
NA20510	NA20510	0	0	1	-9	0
NA20511	NA20511	0	0	1	-9	0

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?

```
## [1] "The number of individuals is 107"
```

```
## [1] "The number of variants is 1102156"
```

```
## [1] "Monomorphic percentage is 18.97%"
```

```
## [1] "After removing monomorphic SNPs, there are 209074 variants still in the database"
```

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?

```
## [1] "The genotype count is: "
```

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

```
## 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 for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  106.2512 DF =  1 p-value =  6.495738e-25 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
```

As the Chi-square test is not so good with extreme allele frequencies, it's better to analyze the results of Exact and Permutation tests.

In both cases we obtain a p-value=1, so we can conclude that for the polymorphism $rs587756191\_T$ there is no evidence at all for rejecting the null hypothesis of the HW equilibrium.

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

	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

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 ?

```
## [1] "The number of significant SNPs is: 8162"
```

We have 8162 values with a p-value < 0.05. It means that for these SNPs we should reject the null hypothesis of HWE, using a chi squared test (but again, maybe For all the extreme allele frequencies SNPs the result are not fully reliable)

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

```
## [1] "Percentage of out of equilibrium SNPs is: 3.904%"
```

If we consider test result reliable, the value is not so far from the 5% that we would expect from chance alone (assuming that we are using an α of 0.05)

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

```
## [1] "The most significant SNP has a value of : 107 with a p-value: 4.45169388968374e-25"
```

Actually, there are more than one significant SNPs with the same p-value and Chi square value.

```
##   [1] "rs573187031_T"    "rs577591184_T"    "rs148185317_C"
##   [4] "rs147754326_G"    "rs193269614_A"    "rs12157695_A"
##   [7] "rs144259790_T"    "rs547865063_A"    "rs28886615_T"
##  [10] "rs62238771_C"     "rs62238772_C"     "rs112357365_C"
##  [13] "rs62238773_A"     "rs62238774_T"     "rs62238775_A"
##  [16] "rs62238776_A"     "rs62238777_A"     "rs542307783_A"
```

```
##  [19] "rs559158642_T"    "esv3647258_.CN2." "rs34486244_T"
##  [22] "rs117556531_A"    "rs117098863_C"    "rs117082938_A"
##  [25] "rs117160232_C"    "rs139674938_A"    "rs566045660_G"
##  [28] "rs2629366_C"      "rs2930745_C"      "rs371738345_A"
##  [31] "rs181887833_A"    "rs577785977_C"    "rs555611350_T"
##  [34] "rs374429468_A"    "rs552093316_G"    "rs567053049_G"
##  [37] "rs539290184_T"    "rs574187239_T"    "rs200493077_C"
##  [40] "rs361627_C"       "rs361582_C"       "rs150430220_C"
##  [43] "rs79918952_T"     "rs75322635_C"     "rs79749619_A"
##  [46] "rs183242778_G"    "rs529510384_A"    "rs574908975_G"
##  [49] "rs189479110_T"    "rs149439245_A"    "rs139560505_T"
##  [52] "rs144208459_G"    "rs537639516_G"    "rs574372536_A"
##  [55] "rs532126127_C"    "rs181016062_A"    "rs532332815_A"
##  [58] "rs560124699_T"    "rs531322893_A"    "rs142828274_C"
##  [61] "rs6003320_C"      "rs5996398_T"      "rs138179565_G"
##  [64] "rs6003322_T"      "rs6003325_C"      "rs12157900_A"
##  [67] "rs12159107_G"     "rs12157971_A"     "rs7292040_A"
##  [70] "rs6003329_T"      "rs79966650_A"     "rs6003331_C"
##  [73] "rs5996404_T"      "rs6003333_G"      "rs5996405_G"
##  [76] "rs5996406_C"      "rs6003334_T"      "rs141475596_T"
##  [79] "rs78514262_C"     "rs76959583_A"     "rs74652561_C"
##  [82] "rs112340903_A"    "rs113556630_C"    "rs148764145_G"
##  [85] "rs113066312_C"    "rs189957465_A"    "rs193165864_G"
##  [88] "rs190732633_G"    "rs373313150_T"    "rs372896059_A"
##  [91] "rs188286444_G"    "rs570614296_G"    "rs374366570_A"
##  [94] "rs377022957_A"    "rs370158003_T"    "rs587755134_G"
##  [97] "rs146030306_C"    "rs181922026_T"    "rs113610812_C"
## [100] "rs75972762_T"     "rs184581336_A"    "rs143229446_T"
## [103] "rs62240659_C"     "rs73395092_A"     "rs117218058_G"
## [106] "rs144483745_G"    "rs187918976_T"    "rs56664150_T"
## [109] "rs12160426_A"     "rs183128739_A"    "rs79898505_C"
## [112] "rs150032224_G"    "rs187575358_C"    "rs183305135_T"
## [115] "rs117069965_C"    "rs79721418_T"     "rs73889008_T"
## [118] "rs142847785_G"    "rs149716547_C"    "rs150247024_A"
## [121] "rs150903025_T"    "rs556386661_T"    "rs190936998_A"
## [124] "rs148906660_T"    "rs6009287_T"      "rs61384610_A"
## [127] "rs117311464_G"    "rs151217496_G"    "rs148166645_A"
## [130] "rs12159379_G"     "rs74352011_C"     "rs145145738_A"
## [133] "rs143483326_A"    "rs8137450_T"      "rs17247392_A"
## [136] "rs145839574_T"    "rs79366610_C"     "rs143190702_T"
## [139] "rs80266259_G"     "rs117727746_A"    "rs140603885_T"
## [142] "rs562382062_T"    "rs192846303_C"    "rs571742670_A"
```

The genotypic composition of the more significant Chi square test is:

AA 106 AB 0 BB 1

This is an unusual composition because we are in a quasi-monomorphic situation where the only allele which is not 'AA' is a homo-zygote 'BB'. This is a very strange/ extremely rare composition, since if we have only two 'B' alleles, we expect that with probability almost one they will pair with 'A' alleles. We could suppose it is a genotyping error. However since the sample of individual is not so big (only 107 observation) we have to admit than this can also be possible (although, again, extremely rare).

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. Is the result consistent with the chi-square test?

```
## [1] "There are 5793 significant SNPs with a p-value<0.05"
```

The result is quite consistent with Chi-square test.

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

```
## [1] "The most significant SNP is rs2629366_C and has a p-value of 9.7847663185214e-33"
```

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

We have again a very unlikely configuration, and the explanation is similar to the one given in answer 7. It is really really strange to have only Homo-zygotes, because it means that, in the former generation, we had 0 mating between 'AA' and 'BB', and this is in a strong opposition with one of the assumptions of the HW equilibrium that assumes "Random mating (w.r.t the trait under study)".

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

```
## [1] "The number of significant SNPs is: 7955"
```

The result is again consistent with the Chi-Squared test.

```
## [1] "The most significant SNP is rs2629366_C and has a p-value of : 4.51151740563241e-34"
```

```
## AA AB BB
## 62  1 44
```

The situation for the most significant SNP is analogous to previous cases.

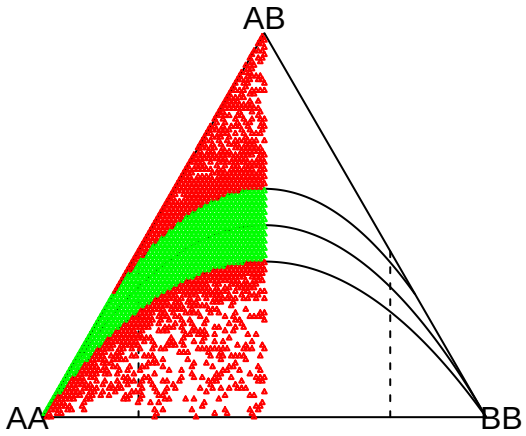
11 Apply a permutation test for Hardy-Weinberg equilibrium to the first 10 SNPs, using the classical chi-square test (without continuity correction) as a test statistic. List the 10 p-values, together with the 10 p-values of the exact tests. Are the result consistent?

	permutation test	exact test
rs587720402_A	1.0000000	1.0000000
rs139377059_T	1.0000000	1.0000000
rs587756191_T	1.0000000	1.0000000
rs587702478_C	1.0000000	1.0000000
rs62224609_C	0.6420000	1.0000000
rs62224611_C	1.0000000	1.0000000
rs192339082_A	1.0000000	1.0000000
rs587740681_A	1.0000000	1.0000000
rs4965031_A	0.1241765	0.2147153
rs375684679_AAAAC	0.0091765	0.0086439

The results are pretty consistent.

11 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)

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

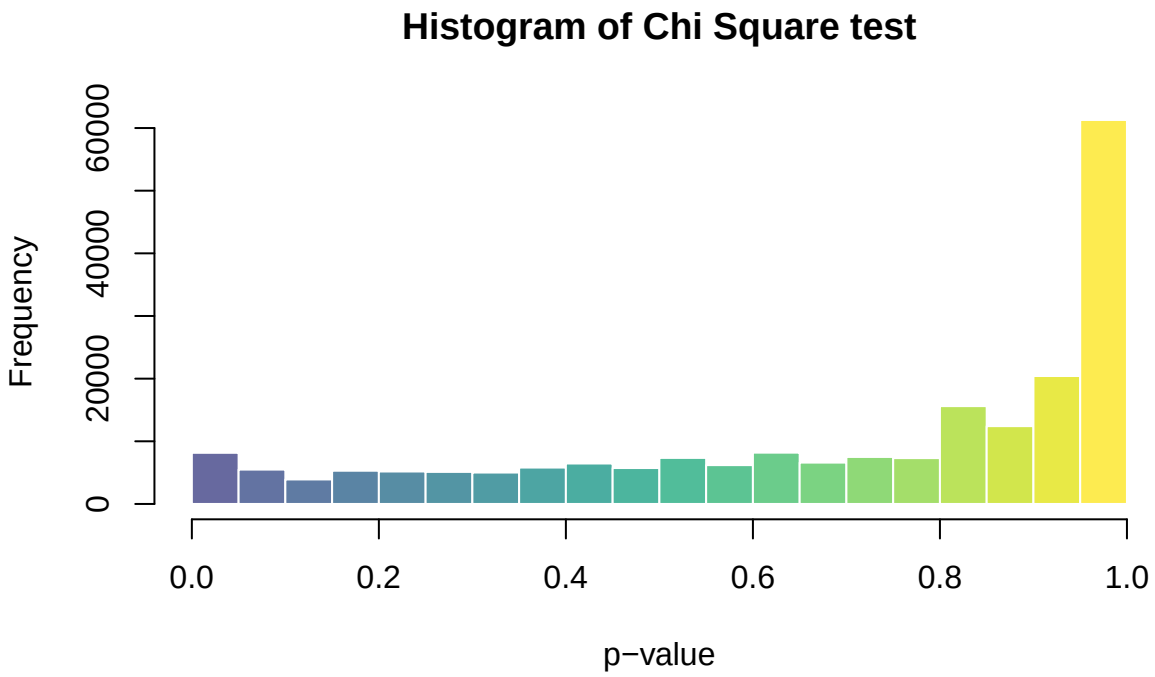


We can see that genotypes concentrate in the zone of equilibrium, and even the genotypes out of equilibrium are more concentrated in the zones around the equilibrium zone.

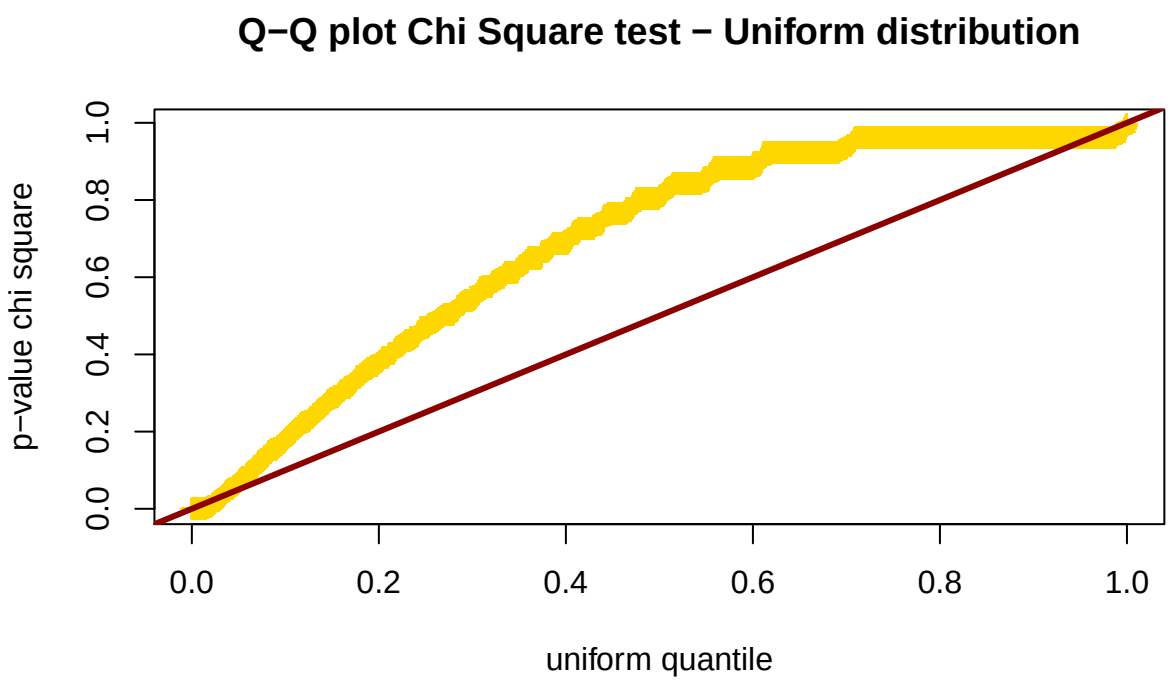
13 Can you explain why half of the ternary diagram is empty?

We guess that this is happening because in this dataset the more frequent allele is always marked as A, so we can't have more 'BB' than 'AA' (and we can't have genotypes without 'AA', except for the practically impossible case of 107 'AB').

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?



For construction, if HWE would hold for the data set we would expect an uniform distribution.



We can conclude clearly that we are not following an uniform distribution.

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.

	normal stat	normal p-value	switched stat	switched p-value
Chi-square test:	0.0023584	0.9612670	0.0023584	0.9612670
Chi-square test with continuity correction:	106.2511737	0.0000000	106.2511737	0.0000000
Likelihood-ratio test:	0.0046949	0.9453725	0.0046949	0.9453725
Exact test with selome p-value:	NA	1.0000000	NA	1.0000000
Exact test with dost p-value:	NA	1.0000000	NA	1.0000000
Exact test with mid p-value:	NA	0.5000000	NA	0.5000000
Permutation test:	NA	NA	NA	NA

Obviously, this will not affect statistical tests for HWE.

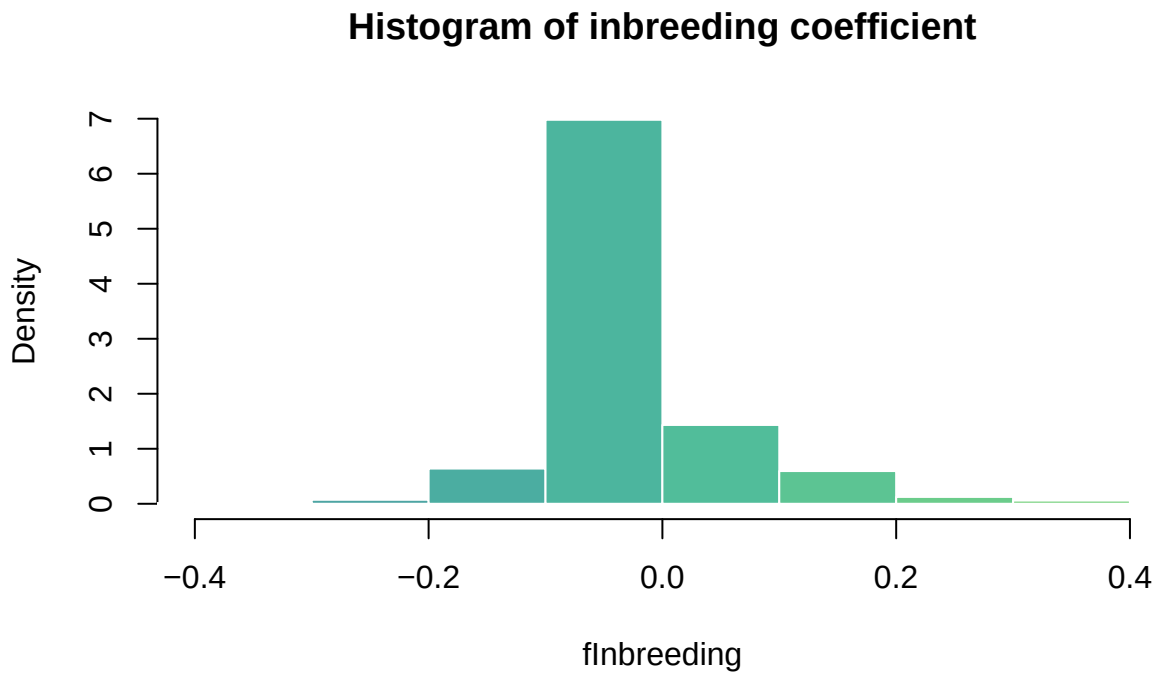
16 Compute the inbreeding coeficient for each SNP, and make a histogram of f. You can use function HWf for this purpose. Give descriptive statistics (mean, standard deviation, etc) of f calculated over the set of SNPs. What distribution do you expect f to follow theoretically? Use a probability plot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

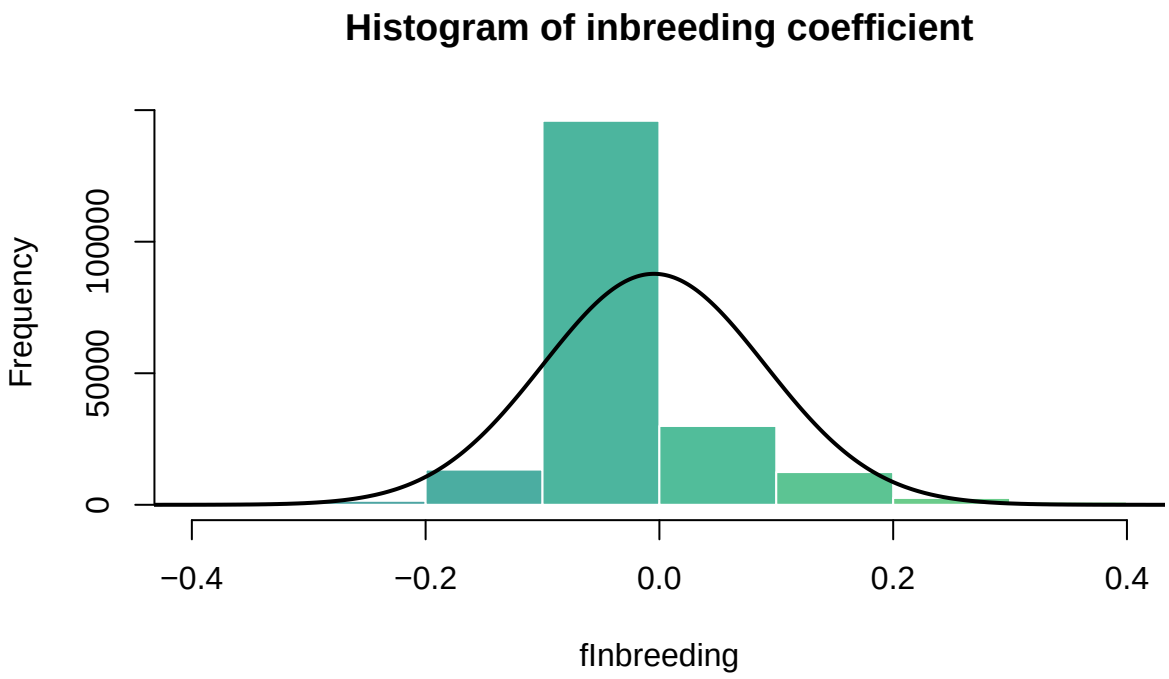
The descriptive statistics are:

fInbreeding

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

```
## [1] "Standard deviation: 0.095"
```

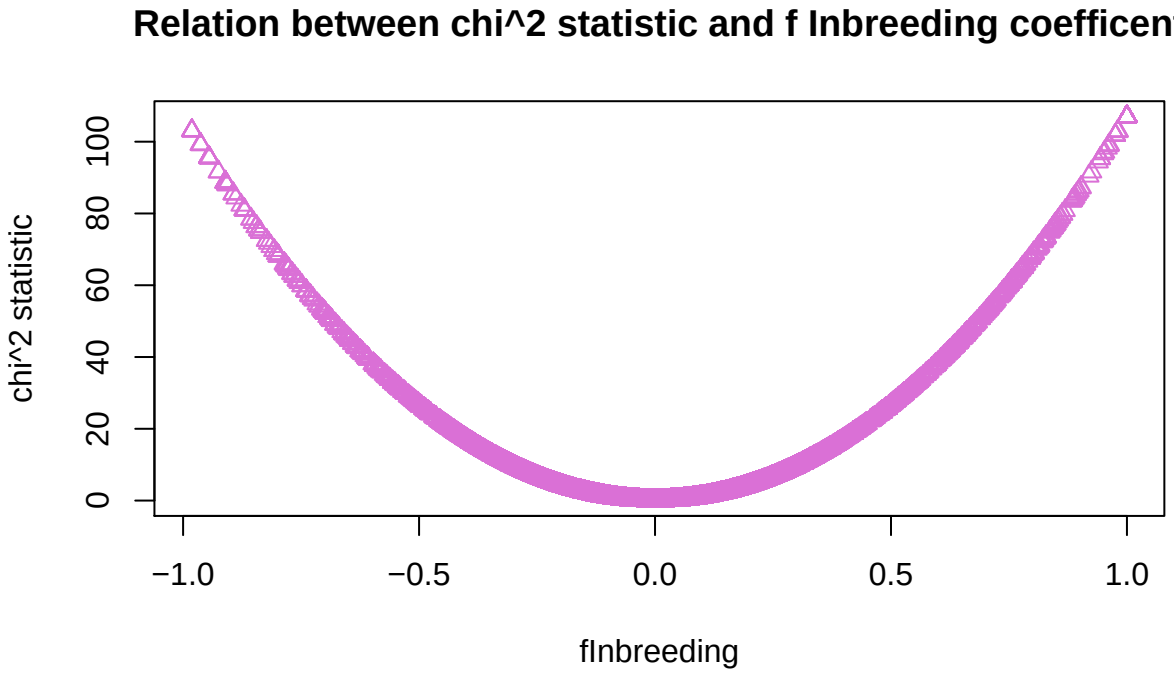




We expect a normal distribution, and apart from the peak near zero, we kinda get that. We see that we have an almost normal distribution, very skewed to the right, and very leptokurtic, since we have a strong peak near the mean.

	f Inbreeding
skewness	1.86543
kurtosis	24.59817

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



We see that the χ^2 statistic is maximized (and so the p-values are minimized) when the f inbreeding coefficent is at his extremes (-1 and 1), and is minimized when the f-statistic is at 0 (as we expected, since if $\hat{f} = 0$, we are in the equilibrium).

the equation is just a parabola with vertex in $[0, 0]$

$$\chi^2 stat = \alpha \hat{f}^2$$

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 alpha = 0.10; 0.05; 0.01 and 0.001. State your conclusions.

```
## [1] "There are 10049 significant SNPs with a p-value<0.10"
## [1] "The percentage of significant variants is 4.80643217234089%"
## [1] "There are 5793 significant SNPs with a p-value<0.05"
## [1] "The percentage of significant variants is 2.77078928991649%"
## [1] "There are 2508 significant SNPs with a p-value<0.01"
## [1] "The percentage of significant variants is 1.1995752700001%"
## [1] "There are 1485 significant SNPs with a p-value<0.001"
## [1] "The percentage of significant variants is 0.710274830921109%"
```

Concerning the Exact test, we can say that for choices of α not so strict, we get less significant variants with respect to the expectation (the ones just by chance), so we could conclude that there is no evidence that we

have more variants out of equilibrium than we would expect. But if we choose smaller α, then things get less defined and we can't be so sure to reject the null-hypotesis of having more out-of-equilibrium variables than we would expect.