BSG - Homework2: Hardy-Weinberg equilibrium

Marti Cardoso, Pau Ferrer

November 20th, 2018

1. The file YRIChr1.rda contains genotype information (10000 SNPs) of individuals from an African population of unrelated individuals. Load this data into the R environment. The file contains a data objects, X, with genotype information. This data is in (0,1,2) format, where 0 and 2 represent the homozygotes AA and BB, and 1 represents the heterozygote AB.

```
load("YRIChr1.rda")
```

2.(2p) How many individuals does the database contain? What percentage of the variants is monomorphic? Remove all monomorphic SNPs from the data bases. How many variants remain in the database? Determine the genotype counts for these variants, and store them in matrix. Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. How many SNPs are significant (use $\alpha = 0.05$)?

```
dim(X)
```

```
## [1]   107 10000
```

```
n <- nrow(X)
p <- ncol(X)
nrow(X)
```

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

```
# All are 0
monomorphic.AA  <- which(apply(X,2,function(col) all(col==0)))
monomorphic.BB  <- which(apply(X,2,function(col) all(col==2)))
monomorphics <- length(monomorphic.AA) + length(monomorphic.BB)
percentage.monomorphics <- (monomorphics/p)*100
X  <- X[,-c(monomorphic.AA,monomorphic.BB)]
dim(X)
```

```
## [1]  107 3035
```

```
#Genotype counts
counts  <- matrix(t(apply(X,2,function(col) c(sum(col==0),sum(col==1),sum(col==2)))), nrow=ncol(X), nco
colnames(counts) <- c("AA", "AB", "BB")
pvalues <- apply(counts, 1, function(row) HWChisq(row,cc = 0,verbose=FALSE)$pval)
passTest  <- pvalues>= 0.05
sum(passTest)
```

```
## [1] 2875
```

There are a total of 107 individuals and 10000 variables (SNPs). 6965 of the variants are monomorphic (all them AA monomorphic), 69.65% of the variants. After removing all the monomorphic variants our dataset has 3035 variants. There are a total of 2875 significant test (equilibrium, we are not rejecting the null hypothesis).

3. (1p) How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone?

We do not expect any deviation from the Hardy Weinberg equilibrium. If there is any kind of deviation from equilibrium it could be due to a genotyping error or some variants related to a disease.

4. (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?

```
passTestExact <- HWExactStats(counts)>=0.05
sum(passTestExact)
```

```
## [1] 2909
```

```
length(passTestExact)
```

```
## [1] 3035
```

```
length(which(passTestExact == passTest))
```

```
## [1] 3001
```

The results for the exact test differ a little bit from those obtained with the chi-squared. There are a total of 2909 significant tests, 34 more than in the chi case. From the 3035 variants, 3001 results of the exact test coincide with the chi-squared test.

5. (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?

```
passTestLRT  <- apply(counts,1,function(col)HWLratio(col,verbose=FALSE)$pval>0.05)
sum(passTestLRT)
```

```
## [1] 2890
```

```
length(which(passTestLRT == passTest))
```

```
## [1] 2994
```

The results are almost the same, in this case, 2890 variants pass the test. From the 3035 variants, 2994 results of the likelihood ratio test coincide with the chi-squared ones.

6. (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?

```
aux <- counts[1:10,]
permutations  <- apply(aux,1,function(row)HWPerm(row,verbose=FALSE,cc=0)$pval)
exact  <- apply(aux,1,function(row)HWExact(row,verbose=FALSE)$pval)
aux <- data.frame(permutations = permutations, exact = exact)
aux
```

```
##    permutations        exact
## 1             0 1.102924e-08
## 2             0 2.841029e-19
## 3             1 1.000000e+00
## 4             1 1.000000e+00
## 5             1 1.000000e+00
## 6             1 1.000000e+00
## 7             1 1.000000e+00
## 8             1 1.000000e+00
## 9             1 1.000000e+00
## 10            1 1.000000e+00
```

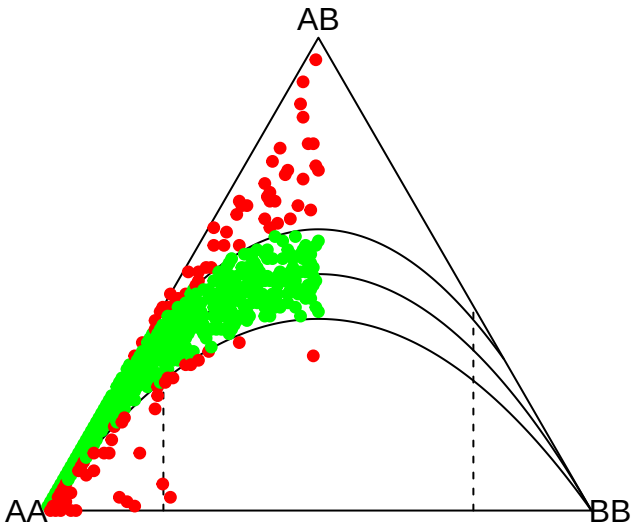
Yes, the results are consistent. In the permutation test, the small p-values are rounded to 0.

7. (1p) Depict all SNPs simultaeneously in a ternary plot, and comment on your result (because many genotype counts repeat, you may use UniqueGenotypeCounts to speed up the computations)

```
Y <- UniqueGenotypeCounts(counts, verbose = TRUE)
```

```
## 3035 rows in X
## 456 unique rows in X
```

```
HWTernaryPlot(Y[,1:3])
```



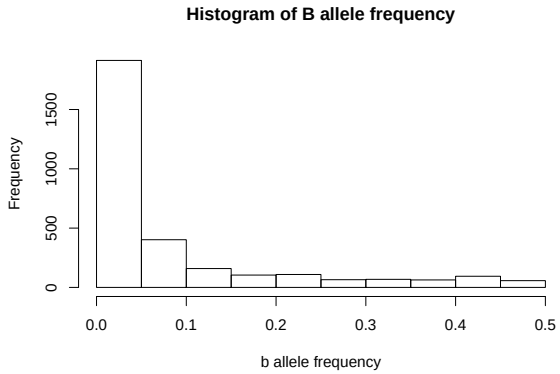
We avoid the repeated frequencies in order to speed up the plot. Then as it can be seen a large proportion of the variants fall inside the banana (meaning they follow the Hardy Weinberg equilibrium), this is the same

result obtained in the tests but graphically. So we have a larger proportion of variants that follow the law. It is also interesting to see that all points fall in one half of the plot.

8. (1p) Can you explain why half of the ternary diagram is empty?

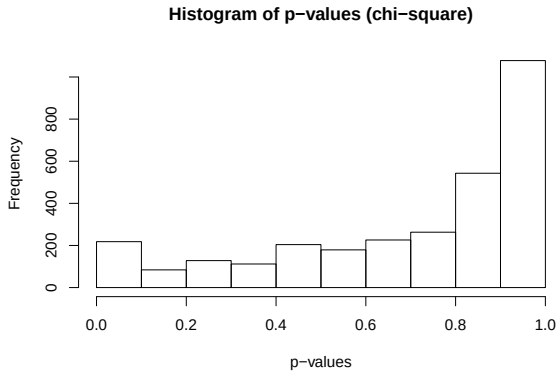
Yes, in this case, a half of the plot is empty because in this dataset the frequency of allele B is very small (always less than 0.5 and mostly it is near to 0). So, we can deduce that in this dataset the B allele is a 'rare' allele and it is always the minor allele. The following plot shows the histogram for the B allele frequencies (or 1 - $freq(A)$).

```
B.freq <- apply(counts, 1 , function(row) (2*row[3]+row[2])/(2*sum(row)))
hist(B.freq, xlab="b allele frequency", main="Histogram of B allele frequency")
```

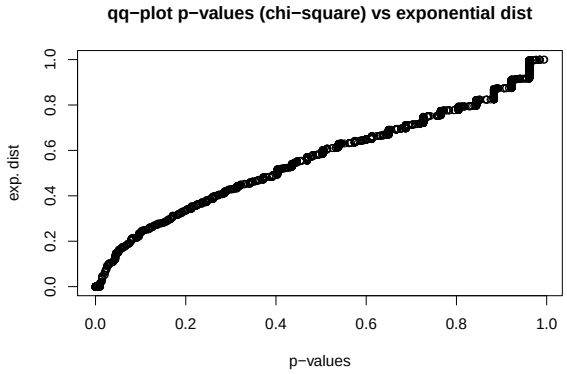


9. (2p) Make a histogram of the p-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? Make a Q-Q plot of the p values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?

```
chi.values <- apply(counts, 1, function(row) HWChisq(row,cc = 0,verbose=FALSE)$pval)
hist(chi.values,main="Histogram of p-values (chi-square)",xlab="p-values")
```




```
#Exponential
exp.dist  <- 1-rexp(length(chi.values),rate=1/sd(chi.values))
exp.dist[exp.dist<0] = 0
qqplot(chi.values, exp.dist, main="qq-plot p-values (chi-square) vs exponential dist", xlab="p-values",
```



We expected the values to be concentrated at 1 (null hypothesis cannot be rejected) but allowing a certain tolerance. By looking at the histogram we can see that more or less the data is grouped at the 1 bin, but not all points have a p-value of 1, it seems that it follows an exponential distribution centered at 1 and the number of SNP decreases exponentially when the p-value is far from 1.

If we perform a qqplot for the exponential distribution we can see that the p-values seem to follow this distribution.

10. (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.

This should not affect the statistical tests because we are checking that the HWE law is true: $f_{AB}^2 = 4 f_{AA} f_{BB}$. So, interchanging the values of f_{AA} and f_{BB} should not lead to a different result.

```
HWChisq(counts[1,], cc =0 , verbose = FALSE)$pval
```

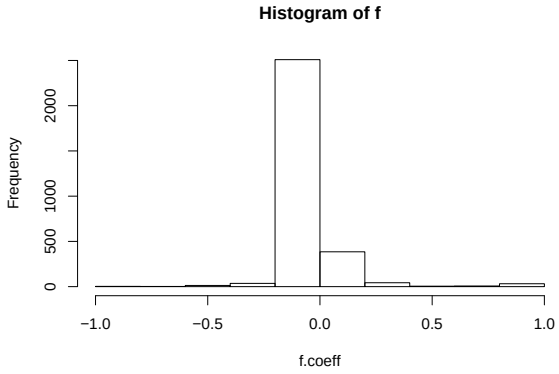
```
## [1] 1.134058e-08
```

```
aux <- c(counts[1,3], counts[1,2], counts[1,1])
HWChisq(aux, cc =0 , verbose = FALSE)$pval
```

```
## [1] 1.134058e-08
```

11. (3p) 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

```
f.coeff <- apply(counts, 1, function(row) HWf(row))
hist(f.coeff, main='Histogram of f')
```



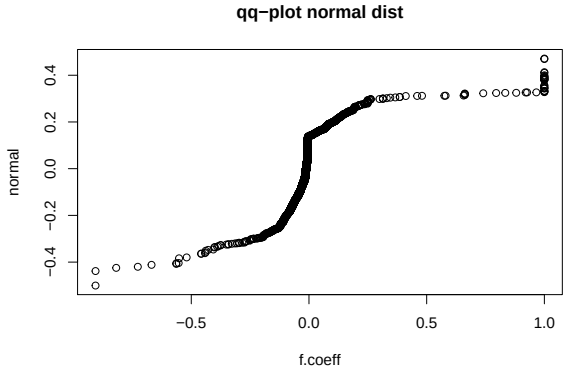

```
summary(f.coeff)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.906709 -0.038835 -0.012979 -0.005368 -0.004695  1.000000
```

```
sd(f.coeff)
```

```
## [1] 0.1372449
```

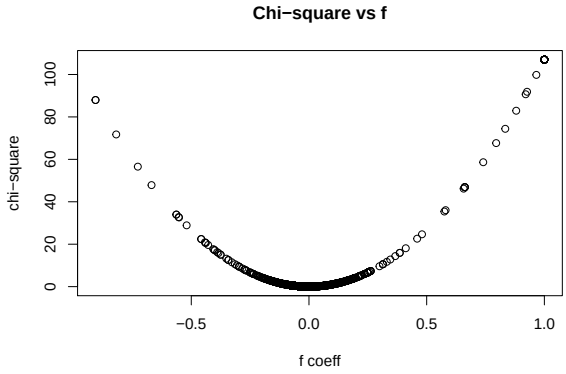
```
#Normal?
normal <- rnorm(length(f.coeff), mean = mean(f.coeff), sd = sd(f.coeff))
qqplot(f.coeff, normal, main='qq-plot normal dist')
```



Above we can observe the histogram and some descriptive statistics for the inbreeding coefficient. The mean is -0.005368 and the standard deviation is 0.137. We expected that the f distribution follows a normal distribution centered at 0. We performed a qq-plot to confirm our idea, but we saw that this plot does not follow a straight diagonal, we thaught that this deviation is caused by some of the SNP that fails the equilibrium tests.

12. (2p) 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?

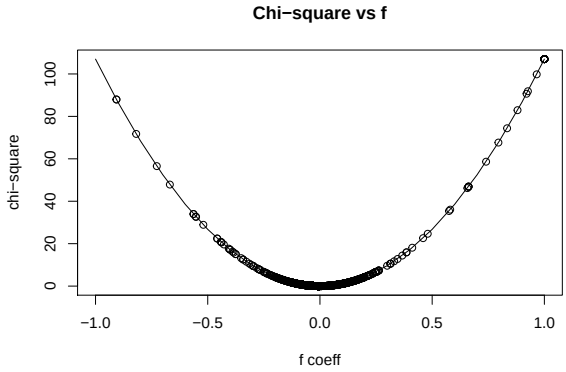
```
chisq <- apply(counts, 1, function(row) HWChisq(row,cc = 0,verbose=FALSE)$chisq)
plot(f.coeff,chisq, main="Chi-square vs f", xlab="f coeff", ylab="chi-square")
```



The plot above shows the chi-square values against the f coefficients. Both values are highly correlated, when the f coefficient is 0, the chi-square is 0 too, and when f coefficient starts to be far from 0, then the chi-square increase very fast.

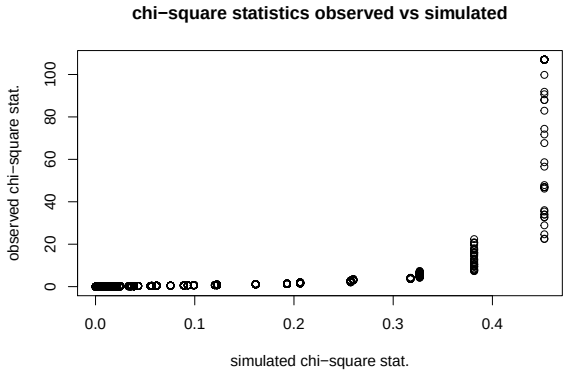
The equation that relates the two coefficients is the following: $chi.square = n * f^2 = 107 * f^2$. The following plot shows the relation between the two statistics.

```
eq = function(f){n*f^2}
x <- seq(-1,1,0.1)
plot(x,eq(x),type='l', main="Chi-square vs f", xlab="f coeff", ylab="chi-square")
points(f.coeff,chisq)
```



13. (1p) Simulate SNPs under the assumption of Hardy-Weinberg equilibrium. Simulate the SNPs of this database, and take care to match each of the SNPs in your database with a simulated SNP that has the same sample size and allele frequency. You can use function HWData of the HardyWeinberg package for this purpose. Compare the distribution of the observed chi-square statistics with the distribution of the chi-square statistics of the simulated SNPs by making a Q-Q plot. What do you observe? State your conclusions.

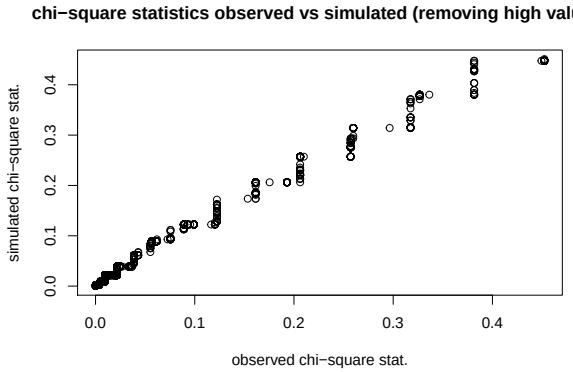
```
nA <- apply(counts, 1 , function(row) (2*row[1]+row[2]))

counts.sim <- HWData(n,nm=ncol(X), nA = nA, exactequilibrium=TRUE)
chisq.sim <- apply(counts.sim, 1, function(row) HWChisq(row,cc = 0,verbose=FALSE)$chisq)
chisq <- apply(counts, 1, function(row) HWChisq(row,cc = 0,verbose=FALSE)$chisq)

qqplot(chisq.sim, chisq, main="chi-square statistics observed vs simulated", xlab="simulated chi-square
```



We simulated the SNPs and we did a Q-Q plot to compare observed and simulated chi-square statistics. The plot above suggests us that the two sets do not follow the same distribution (points are very far from the diagonal). We think that this is caused by the observed SNPs that are not in equilibrium (or p-values near to 0), these SNPs obtains a very high chi-square statistic and deforms the distribution. On the other hand, the simulated ones get lower values because all they are in equilibrium.

We also decided to do a Q-Q plot removing all these observed SNPs that have high chi-square statistic (higher than the minimum of the simulated statistics, 0.5). The plot below shows this new Q-Q plot, where in this case the two sets of statistics seem to come from the same distribution.

```
qqplot(chisq.sim, chisq[chisq<max(chisq.sim)],main="chi-square statistics observed vs simulated (removi
```



14. (2p) We reconsider the exact test for HWE, 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.

```
exact.results <- HWExactStats(counts)
computeSignificantVariants <- function(exact.results, alpha){
  number = sum(exact.results>=alpha)
  perc = number/length(exact.results)
  return(list(number=number,perc=perc))
}
computeSignificantVariants(exact.results,0.10)
```

```
## $number
## [1] 2852
```

```
##
## $perc
## [1] 0.9397035
```

```
computeSignificantVariants(exact.results,0.05)
```

```
## $number
## [1] 2909
##
## $perc
## [1] 0.9584843
```

```
computeSignificantVariants(exact.results,0.01)
```

```
## $number
## [1] 2952
##
## $perc
## [1] 0.9726524
```

```
computeSignificantVariants(exact.results,0.001)
```

```
## $number
## [1] 2986
##
## $perc
## [1] 0.983855
```

With α=0.1, 2852 variants (93.9%) cannot reject the hypothesis (thus it accepts the equilibrium). Using α=0.05 2909 variants (95.8%) accepts equilibrium, with α=0.001 2952 variants (97.26%) and with α=0.001 2986 variants (98.4%) accepts equilibrium. As expected, when we decrease the confidence level more variants will accept the null hypothesis. If we use $\alpha = 0.01$ very few variants can reject the equilibrium, so most of them pass the test, that means that the data seems to be in equilibrium.

15. (1p) Do you think genotyping error is a problem for the database you just studied? Explain your opinion

That's true that some of the SNPs of this dataset does not follow the equilibrium law, but they can be detected with some of the tests that we performed in this homework. Also, the number of SNPs without equilibrium is very low, for example with $\alpha = 0.05$ less than 5% of the variants fail. So, we think that detecting and removing these genotyping errors will not be a problem because a low percentage of SNPs will be dropped.