BSG - Homework 6 - Association analysis

HENRY QIU LO & MEYSAM ZAMANI

December 21, 2019

1. The file rs394221.dat contains genotype information, for cases and controls, of polymorphism rs394221, which is presumably related to Alzheimer's disease.Load the data file into the R environment.

```
initialDataTogether <- read.delim("rs394221.dat", header = FALSE)
```

2. (1p) What is the sample size? What is the number of cases and the number of controls? Construct the contingency table of genotype by case/control status.

```
summaryInTable <- do.call(cbind, lapply(initialDataTogether, summary))
Cases <- c(summaryInTable[3],summaryInTable[2],summaryInTable[1])
Controls <- c(summaryInTable[6],summaryInTable[5],summaryInTable[4])
tableX <- rbind(Cases,Controls)
colnames(tableX) <- c("MM","Mm","mm")
```

```
sum(tableX)
```

```
## [1] 1167
```

The sample size is 1167

```
rowSums(tableX)
```

```
##    Cases Controls
##      509      658
```

The number of cases are 509 and the number of controls are 658

The contingency is the following:

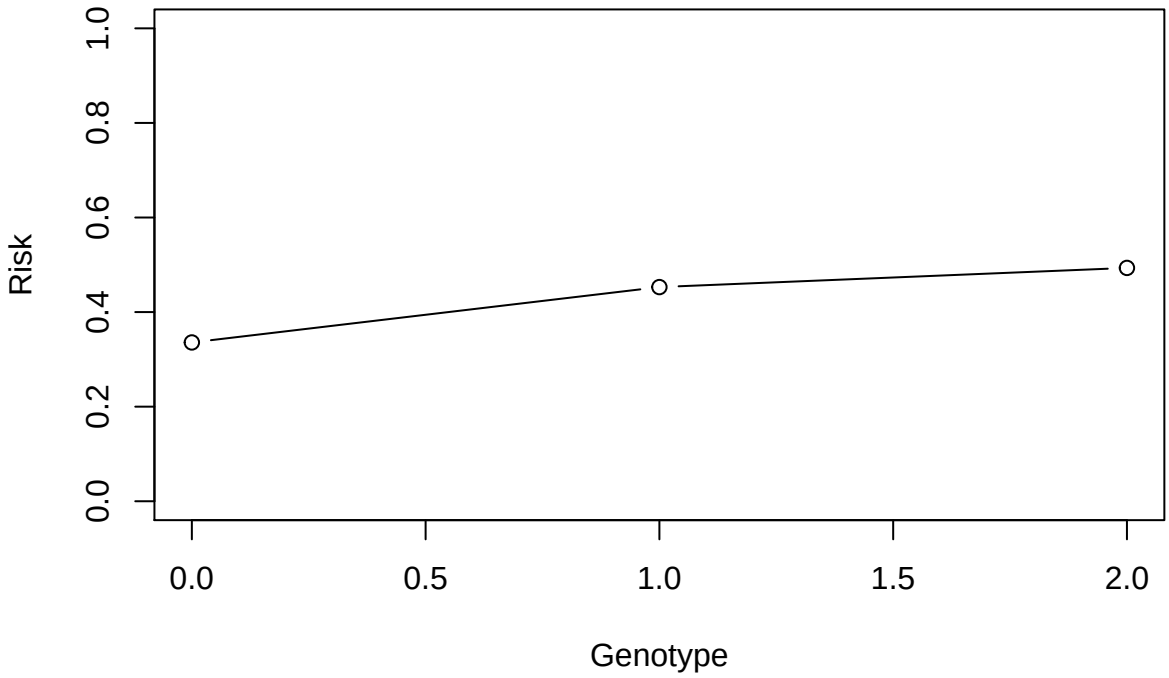
```
tableX
```

```
##           MM  Mm  mm
## Cases    149 269  91
## Controls 153 325 180
```

3. (1p) Explore the data by plotting the percentage of cases as a function of the genotype, ordering the latter according to the number of M alleles. Which allele increases the risk of the disease?

```
totalGenotypeCounts <- colSums(tableX)
riskOfDisease <- Cases/totalGenotypeCounts

plot(c(0,1,2),c(riskOfDisease[3],riskOfDisease[2],riskOfDisease[1]),ylim=c(0,1),type="b", xlab="Genotyp
```



From the plot we can see the allele that much increase the risk of desease seems to be M

4. (2p) Test for equality of allele frequencies in cases and controls by doing an alleles test. Report the test statistic, its reference distribution, and the p-value of the test. Is there evidence for different allele frequencies?

```
tableY <- cbind(2*tableX[,1]+tableX[,2],2*tableX[,3]+tableX[,2])
colnames(tableY) <- c("M","m")

resultsChisqTest <- chisq.test(tableY,correct=FALSE)
```

Test statistic and reference distribution:

```
resultsChisqTest
```

```
##
##  Pearson's Chi-squared test
##
## data:  tableY
## X-squared = 13.797, df = 1, p-value = 0.0002037
```

```
resultsChisqTest$expected
```

```
##                M       m
## Cases    522.521 495.479
## Controls 675.479 640.521
```

```
resultsChisqTest$p.value
```

```
## [1] 0.0002036983
```

The p-value of the test is 0.0002037, as it is very low and less than the significance level, we cannot accept the null hypotesis, concluding that there are evidences for differences of allele frequencies.

5. (2p) Which are the assumptions made by the alleles test? Perform and report any addtional tests you consider adequate to verify the assumptions. Do you think the assumptions of the alleles test are met?

The test for equality of allele frequencies assumes independence.

We will perform the fisher exact test to verify the assumptions.

```
fisher.test(tableY)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tableY
## p-value = 0.0002368
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.154074 1.614038
## sample estimates:
## odds ratio
##   1.364592
```

```
tableY
```

```
##            M   m
## Cases    567 451
## Controls 631 685
```

```
OR <- (tableY[1,1]*tableY[2,2])/(tableY[1,2]*tableY[2,1])
OR
```

```
## [1] 1.364796
```

We can check assumptions of the alleles tests are not met because we know OR $= 1.364592 > 1$ which indicates association.

```
seLnOr <- sqrt(sum(1/tableY))
seLnOr
```

```
## [1] 0.08381887
```

```
llLogOdds <- log(OR) - qnorm(0.975)*seLnOr
ulLogOdds <- log(OR) + qnorm(0.975)*seLnOr
llOdds <- exp(llLogOdds)
ulOdds <- exp(ulLogOdds)
```

```
llOdds
```

```
## [1] 1.158033
```

```
ulOdds
```

```
## [1] 1.608476
```

We also observed that the presence of m raises the odds of M, and the presence of M raises the odds of m.

6. (2p) Perform the Armitage trend test for association between disease and number of M alleles. Report the test statistic, its reference distribution and the p-value of the test. Do you find evidence for association?

```
casReplicas <- rep(c(0,1,2),Cases)
conReplicas <- rep(c(0,1,2),Controls)

x <- c(rep(1, sum(Cases)),
       rep(0, sum(Controls)))

y <- c(casReplicas, conReplicas)
```

```
length(x)
```

```
## [1] 1167
```

```
length(y)
```

```
## [1] 1167
```

```
correlation <- cor(x,y)
n <- sum(tableX)
A <- n*(correlation^2)
```

Test statistic and reference distribution:

```
correlation
```

```
## [1] -0.1097624
```

```
A
```

```
## [1] 14.05977
```

```
pvalue <- pchisq(A,df=1,lower.tail=FALSE)
pvalue
```

```
## [1] 0.0001770917
```

The p-value is 0.0001770917, which is lower than the significance level, this means an evidence of association.

7. (4p) Test for association between genotype and disease status by a logistic regression of disease status on genotype, treating the latter as categorical. Do you find significant evidence for association? Which allele increase the risk for the disease? Give the odds ratios of the genotypes with respect to base line genotype mm. Provide 95% confidence intervals for these odds ratios.

```
newy <- x
newx <- y
x.cat <- rep(NA,length(newx))
x.cat[newx==0] <- "MM"
x.cat[newx==1] <- "Mm"
x.cat[newx==2] <- "mm"
x.cat <- factor(x.cat)
out1.lm <- glm(newy~x.cat, family = binomial(link = "logit"))
summary(out1.lm)
```

```
##
## Call:
## glm(formula = newy ~ x.cat, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1662  -1.0982  -0.9046   1.2587   1.4773
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6821     0.1286  -5.303 1.14e-07 ***
## x.catMm       0.4930     0.1528   3.227 0.001251 **
## x.catMM       0.6556     0.1726   3.798 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1598.7  on 1166  degrees of freedom
## Residual deviance: 1582.7  on 1164  degrees of freedom
## AIC: 1588.7
##
## Number of Fisher Scoring iterations: 4
```

We can observe association due to the standard error of the model is quite small.

Observing the estimated valu the allele M seems to be the one that increase the risk of desease.

```
b <- coefficients(out1.lm)
ORs <- exp(b)
t(data.frame(ORs))
```

```
##     (Intercept)  x.catMm  x.catMM
## ORs   0.5055556 1.637194 1.926309
```

```
varCovar <- vcov(out1.lm)
seDiag <- sqrt(diag(varCovar))
llDiag <- b-qnorm(0.975)*seDiag
ulDiag <- b+qnorm(0.975)*seDiag
llDiag.or <- exp(llDiag)
ulDiag.or <- exp(ulDiag)
```

The odds ratios has values of range with lowerbound:

```
llDiag.or
```

```
## (Intercept)     x.catMm     x.catMM
##   0.3929004   1.2135598   1.3734274
```

And upper bound:

```
ulDiag.or
```

```
## (Intercept)     x.catMm     x.catMM
##   0.6505119   2.2087109   2.7017564
```