Practical 06 BSG: Association analysis

Lovro Katalinic and Ivan Almer

Hand-in: 21/12/2020

In this practical we perform association tests for a binary disease indicator and a genetic polymorphism. 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 MASS, genetics, data.table and others for the computations. Take care to number your answer exactly as in this exercise, preferably by copying each requested item into your solution. Upload your solution to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

```
library(MASS)
library(genetics)
```

```
## 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: 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
```

```
library(HardyWeinberg)
```

```
## 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
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
## The following objects are masked from 'package:gdata':
##
##     first, last
```

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.

1. (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.

```
data <- fread('rs394221.dat', data.table=FALSE, header = FALSE)
n <- nrow(data)
ncases <- sum(data[,2] == 'case')
ncontrol <- n - ncases

cat(paste('Number of rows:', n, '\n'))
```

```
## Number of rows: 1167
```

```
cat(paste('Number of cases:', ncases, '\n'))
```

```
## Number of cases: 509
```

```r
cat(paste('Number of controls:', ncontrol, '\n'))
```

```
## Number of controls: 658
```

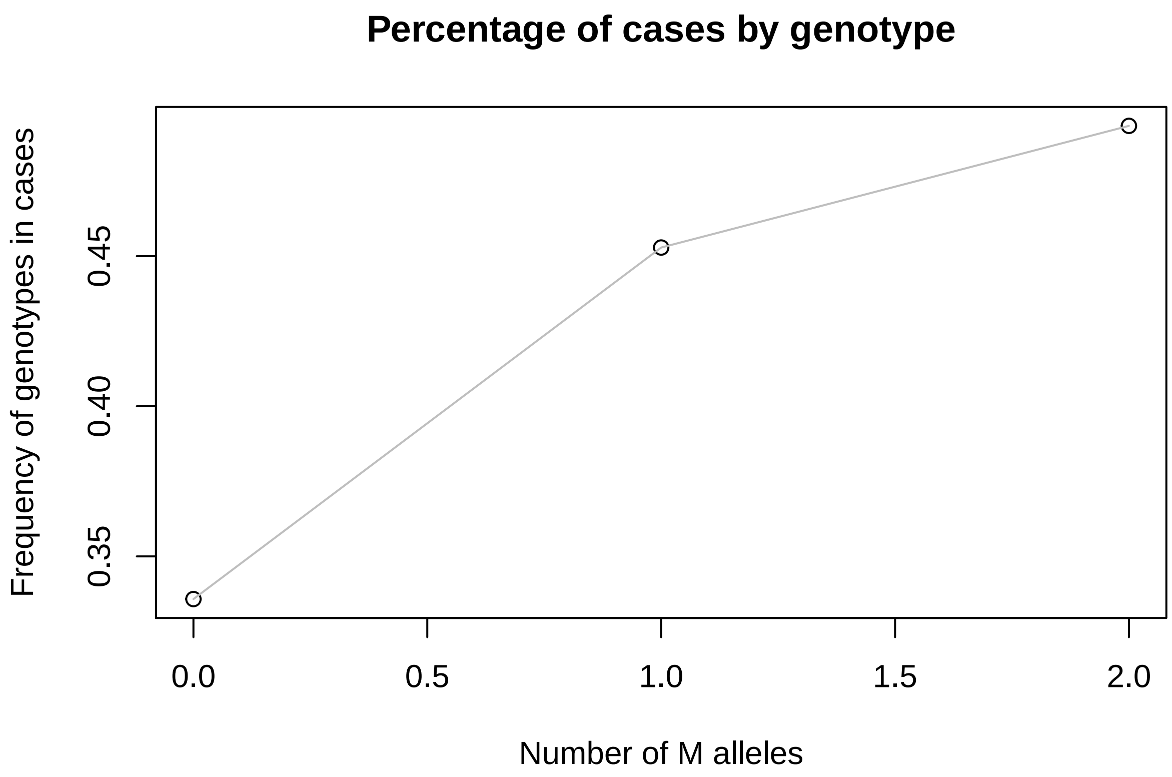
```r
head(data)
```

```
##   V1      V2
## 1 Mm    case
## 2 Mm control
## 3 mm    case
## 4 mm control
## 5 mm control
## 6 MM    case
```

2. (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?

```r
mms <- sum(data[,2] == 'case' & data[,1] == 'mm') / sum(data[,1] == 'mm')
Mms <- sum(data[,2] == 'case' & data[,1] == 'Mm') / sum(data[,1] == 'Mm')
MMs <- sum(data[,2] == 'case' & data[,1] == 'MM') / sum(data[,1] == 'MM')
number_of_alleles <- c(0,1,2)
genotypes_distr <- c(mms, Mms, MMs)

plot(number_of_alleles, genotypes_distr, main="Percentage of cases by genotype", xlab="Number of M allel
lines(number_of_alleles, genotypes_distr, col="gray")
```



Considering this plot, we can observe that the individuals with more alleles M has greater risk of disease than individuals with less such alleles.

3. (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?

```
mcases <- 2*sum(data[,2] == 'case' & data[,1] == 'mm') + sum(data[,2] == 'case' & data[,1] == 'Mm')
Mcases <- 2*sum(data[,2] == 'case' & data[,1] == 'MM') + sum(data[,2] == 'case' & data[,1] == 'Mm')
mcontrol <- 2*sum(data[,2] == 'control' & data[,1] == 'mm') + sum(data[,2] == 'control' & data[,1] == '
Mcontrol <- 2*sum(data[,2] == 'control' & data[,1] == 'MM') + sum(data[,2] == 'control' & data[,1] == '

allele_freq <- rbind(c(mcases, Mcases), c(mcontrol, Mcontrol))
colnames(allele_freq) <- c("m","M")
rownames(allele_freq) <- c("Cases", "Control")
(allele_freq)
```

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

```
chisq.test(allele_freq,correct=FALSE)
```

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

```
fisher.test(allele_freq)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  allele_freq
## p-value = 0.0002368
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6195641 0.8664957
## sample estimates:
## odds ratio
##  0.7328195
```

As the chi-squared statistic of **13.797** which corresponds to p-value **0.0002037** exceeds the critical value, we reject the null hypothesis and conclude that the allele frequencies are biased at 95% significance level. We say that there is enough evidence for different allele frequencies. Fisher's Exact Test gave very similar results.

4. (2p) Which are the assumptions made by the alleles test? Perform and report any additional tests you consider adequate to verify the assumptions. Do you think the assumptions of the alleles test are met?

The assumptions made by chi-squared test are **random sampling**, **sufficiently large size of sample**, **expected cell count greater than 5** and **independence of the observations**. The chi-squared test also relies on HW equilibrium assumption. Sample size seems sufficiently large and expected cell count is great enough. We assume that observations are randomly sampled and independent. What left is to check if the variant is in HW equilibrium.

```
genotypes <- c(sum(data[,1] == 'mm'), sum(data[,1] == 'Mm'), sum(data[,1] == 'MM'))
names(genotypes) <- c("mm","Mm", "MM")
HWChisq(genotypes)
```

```
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  0.3546387 DF =  1 p-value =  0.551499 D =  5.45587 f =  -0.0187137
```

Considering p-value is greater than critical value, we can not reject the null hypotesis which states that the variant is in equilibrium. So, we think assumptions of the alleles test are met

5. (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?

```
x <- as.integer(data[,2] == 'case')
y <- sapply(data[,1], function(x) sum(unlist(strsplit(x, split = "")) == 'M'))

r <- cor(x,y)
A <- n*(r^2)

pvalue <- pchisq(A,df=1,lower.tail=FALSE)
cat(paste('p-value:', pvalue))
```

```
## p-value: 0.000177091650268061
```

Armitage trend test produced p-value of **0.000177091** which is lower than critical p-value of **0.05**. We can reject the null hypotesis and say that there exist a trend and enough evidence for association.

6. (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.

```
y <- x
x <- factor(data[,1])

out.lm <- glm(y~x, family = binomial(link = "logit"))
summ <- summary(out.lm)
summ
```

```
##
## Call:
## glm(formula = y ~ x, 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 ***
## xMm           0.4930     0.1528   3.227 0.001251 **
## xMM           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
```

```r
or <- exp(summ$coefficients[,1])
orl <- exp(summ$coefficients[,1] - 1.96*summ$coefficients[,2])
org <- exp(summ$coefficients[,1] + 1.96*summ$coefficients[,2])

cat(paste('Ratio of odds for diseas with Mm person and mm person: ', or[2], ' (95% confidence interval:
```

```
## Ratio of odds for diseas with Mm person and mm person: 1.63719357565482 (95% confidence interval: [1
```

```r
cat(paste('Ratio of odds for diseas with MM person and mm person: ', or[3], ' (95% confidence interval:
```

```
## Ratio of odds for diseas with MM person and mm person: 1.92630898513217 (95% confidence interval: [1
```

As we can see from the output of **summmary** of **glm** function, each one of the coefficients is significant and therefore which is enough evidence for association between genotype and disease status. We can also see the same results as from the plot in the second task, telling that allel **M** increases the risk for the disease.