Bioinformatics and Statistical Genetic

Association analysis

Leonardo Ortoleva & Egon Ferri

Contents

```
knitr::opts_chunk$set(echo = F)
```

1 Load data into the R environment.

Table 1: Data

	1	2	3	4	5	6	7	8
Genotype	Mm	Mm	mm	mm	mm	MM	MM	mm
Disease	case	control	case	control	control	case	control	control

2 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.

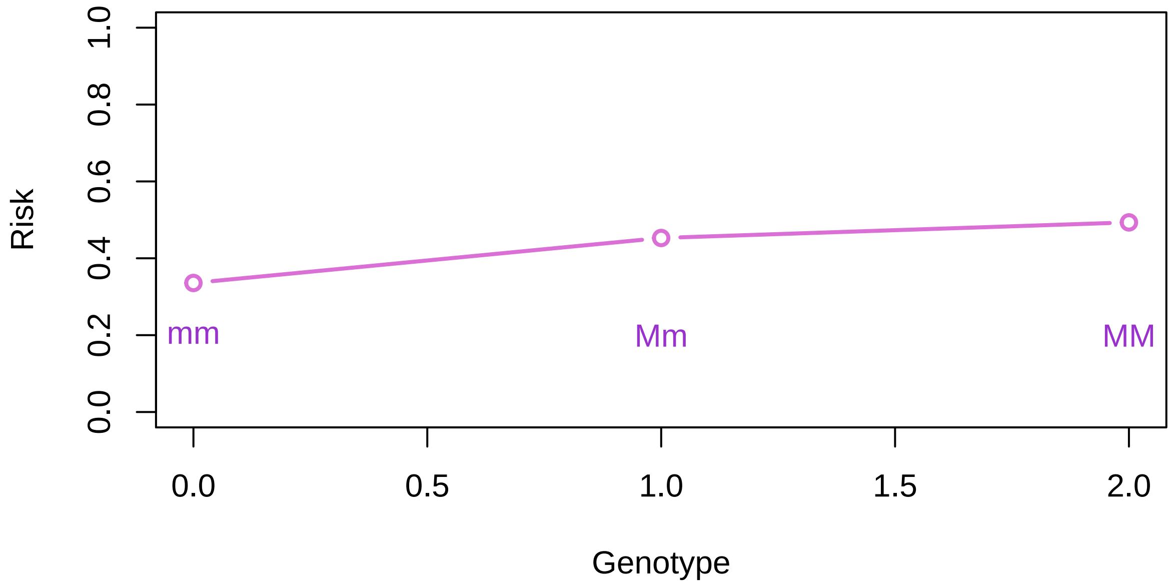
```
## [1] "sample size:  1167"
## [1] "number of cases:  509"
```

```
## [1] "number of controls:  658"
```

Table 2: contingency table

	MM	Mm	mm
Cases	91	269	149
Controls	180	325	153

3 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?



The allele that seems to increase the risk is M.

Table 3: allele contingency table

	m	M
Cases	451	567
Controls	685	631

4 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?

We can conduct more than one allele test to analyze our contingency table.

Fisher's exact test, although in practice it is employed when sample sizes are small, it is valid for all sample sizes.

Since the sample is sufficiently big, and we don't have any cells with really small values, we can also use a χ^2 test with 1 grade of freedom (based on the χ^2 distribution).

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

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  Y
## 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
```

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

Both the test give a very small p-value around 0.0002, so we can strongly reject the null hypotesis that assumes same allele frequency.

5 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 assumes Hardy-Weinberg equilibrium.

```
## Chi-square test for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  0.4086865 DF =  1 p-value =  0.522637 D =  5.45587 f =  -0.0187137
```

```
## 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
```

```
## Haldane Exact test for Hardy-Weinberg equilibrium (autosomal)
## using SELOME p-value
## sample counts: nAA =  302 nAB =  594 nBB =  271 
## H0: HWE (D==0), H1: D <> 0
## D =  5.45587 p-value =  0.5579746
```

```
## Permutation test for Hardy-Weinberg equilibrium
## Observed statistic: 0.4086865   17000 permutations. p-value: 0.5595294
```

Conducting different test over the Hardy-Weimberg equilibrium, we don't have any evidence at all to reject the null hypotesis of it to be respected.

6 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?

The trend test is based on the linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

And the test-statistic is:

$$A = \frac{\hat{\beta}_1^2}{V\left(\hat{\beta}_1\right)} = n \cdot r_{xy}^2$$

```
## [1] "test statistic A:  14.0597678498094"
```

```
## [1] "pvalue:  0.000177091650268061"
```

We find strong evidence for association.

7 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.

```
##
## 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
```

Significant evidence for association is found. As expected from previous results, the ‘M’ allele increase the risk for the disease.

```
## [1] "odds ratios: "
```

```
## (Intercept)     x.catMm     x.catMM
##   0.5055556   1.6371936   1.9263090
```

```
## [1] "confidence intervals:"
```

```
## [1] "lower:"
```

```
## (Intercept)     x.catMm     x.catMM
##    0.374740    1.213560    1.427865
```

```
## [1] "upper"
```

```
## (Intercept)     x.catMm     x.catMM
##   0.6820367   2.2087109   2.5987518
```