BSG-MDS practical 5 Statistical Genetics

Eliya Tiram and Ximena Moure

05/12/2023, submission deadline 12/12/2023

Setting up data

```
cases <- c(112, 278, 150)
controls <- c(206, 348, 150)
```

1. Perform the alleles test for this data set. Provide the p-value and the odds ratio and comment on the results.

We get a p-value of 0.0002100232 which is very small. It is well below the conventional threshold of 0.05 for statistical significance. This indicates that the difference in allele frequencies between the cases and controls in the study is highly unlikely to have occurred by chance. There is a statistically significant association between the SNP and Alzheimer's disease.

Since the odds ratio is less than 1, it suggests that allele A is associated with a decreased risk of Alzheimer's disease compared to allele a. However, it's important to note that the odds ratio is less than 1, which implies a protective effect rather than a risk factor. In other words, individuals with allele A may be less likely to develop Alzheimer's disease compared to those with allele a.

```
allele_A_cases <- 2 * cases[1] + cases[2]
allele_a_cases <- 2 * cases[3] + cases[2]
allele_A_controls <- 2 * controls[1] + controls[2]
allele_a_controls <- 2 * controls[3] + controls[2]


mat <- matrix(c(allele_A_cases, allele_A_controls, allele_a_cases, allele_a_controls), nrow = 2)
colnames(mat) <- c("A","a")
rownames(mat) <- c("Cases","Controls")

# Chi-square test for allele frequencies
chi_test <- chisq.test(mat,correct=FALSE)

fi <- fisher.test(mat)

# Output
cat("P-value:", chi_test$p.value, "\n")
```

```
## P-value: 0.0002100232
```

```r
cat("Odds ratio: ", fi$estimate, "\n")
```

```
## Odds ratio:  0.7406221
```

2. Test for association using a codominant, a dominant and a recessive model. Provide the p-values for all the tests and comment on the results.

All three models show statistical significance, which suggests that the genetic variant is associated with the disease risk. However, the codominant and dominant models have lower p-values, indicating stronger evidence for these models compared to the recessive model.

Given that the p-values for codominant and dominant models are very close, it might suggest that the risk increases with one copy of the 'm' allele and potentially further increases (although not dramatically) with a second copy. This could point towards a codominant pattern where each allele adds to the risk, but the presence of one allele is sufficient for an increased risk, which might also be consistent with a partial dominant effect.

```r
gen_count <- matrix(c(112, 278, 150, 206, 348, 150), nrow = 2, byrow = TRUE)
colnames(gen_count) <- c("AA","Aa","aa")
rownames(gen_count) <- c("Cases","Controls")

# Codominant Test
test_codominant <- chisq.test(gen_count)
p_value_codominant <- test_codominant$p.value
fisher_codominant <- fisher.test(gen_count)
cat("Codominant Test Results:\n")
```

```
## Codominant Test Results:
```

```r
cat("P-value Pearson:", p_value_codominant, "\n")
```

```
## P-value Pearson: 0.0008085403
```

```r
cat("P-value Fisher: ", fisher_codominant$p.value, "\n")
```

```
## P-value Fisher:  0.0007782748
```

```r
# Dominant Test
Y_dominant <- cbind(gen_count[,1], gen_count[,2] + gen_count[,3])
test_dominant <- chisq.test(Y_dominant)
p_value_dominant <- test_dominant$p.value
fisher_dominant <- fisher.test(Y_dominant)
cat("\nDominant Test Results:\n")
```

```
##
## Dominant Test Results:
```

```r
cat("P-value Pearson:", p_value_dominant, "\n")
```

```
## P-value Pearson: 0.0008108124
```

```r
cat("P-value Fisher: ", fisher_dominant$p.value, "\n")
```

```
## P-value Fisher:  0.0006464142
```

```r
# Recessive Test
Y_recessive <- cbind(gen_count[,1] + gen_count[,2], gen_count[,3])
test_recessive <- chisq.test(Y_recessive)
p_value_recessive <- test_recessive$p.value
fisher_recessive <- fisher.test(Y_recessive)
cat("\nRecessive Test Results:\n")
```

```
##
## Recessive Test Results:
```

```r
cat("P-value Pearson:", p_value_recessive, "\n")
```

```
## P-value Pearson: 0.009952705
```

```r
cat("P-value Fisher: ", fisher_recessive$p.value, "\n")
```

```
## P-value Fisher:  0.009094037
```

3. Plot the risk of disease as a function of the number of m alleles. Comment on the results. Which model seems most appropriate?

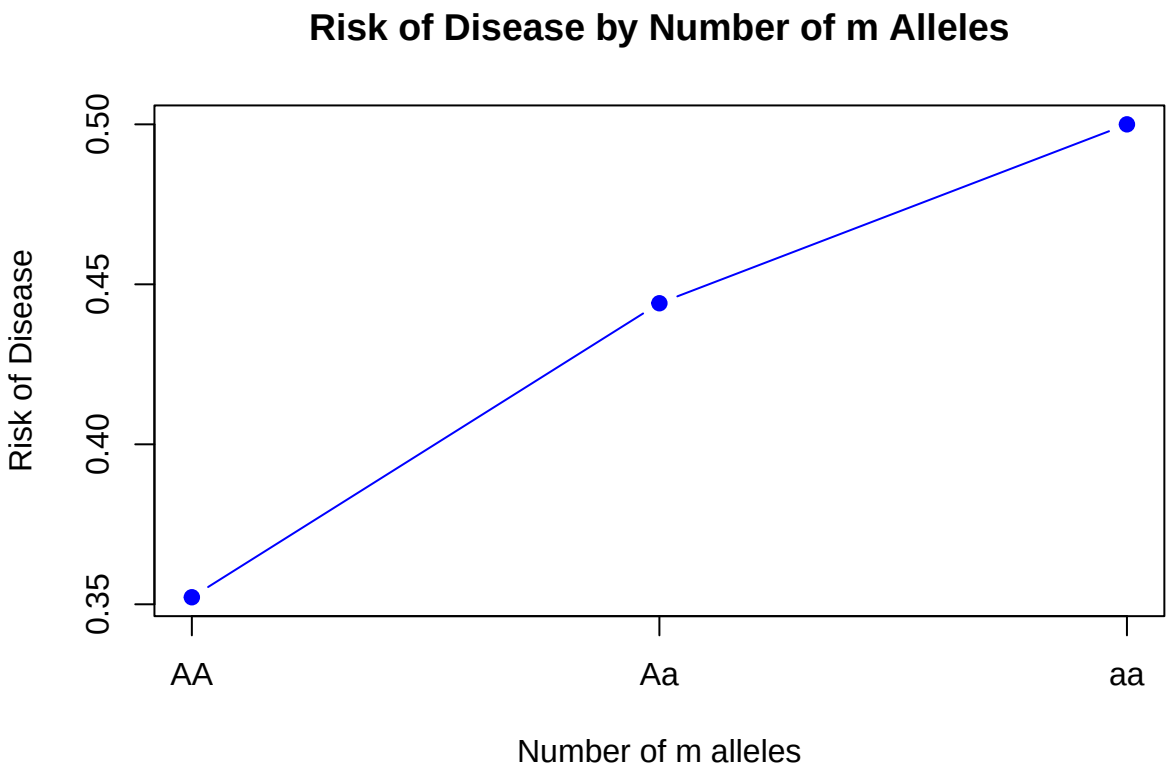
Since the risk of disease increases with the number of m alleles, the additive model seems more appropriate.

The smooth increase in risk from 'AA' to 'aa' suggests that neither a purely dominant nor a purely recessive model may fully explain the genetic influence on disease risk.

```r
risk_AA <- cases[1] / (cases[1] + controls[1])
risk_Aa <- cases[2] / (cases[2] + controls[2])
risk_aa <- cases[3] / (cases[3] + controls[3])
risks <- c(risk_AA, risk_Aa, risk_aa)

labels <- c("AA", "Aa", "aa")
x_pos <- 1:length(labels)

plot(x_pos, risks, type = "b", col = "blue", xlab = "Number of m alleles",
     ylab = "Risk of Disease", pch = 19, xaxt = "n" )
title("Risk of Disease by Number of m Alleles")
axis(1, at = x_pos, labels = labels)
```



4. Perform Armitage trend test for this data set. Does the null hypothesis $1 = 0$ hold? Comment on your response.

Given the p-value is less than the typical threshold for statistical significance (0.05), we reject the null hypothesis. This suggests that there is a statistically significant association between the genetic variant and Alzheimer's disease.

```
X <- rbind(cases,controls)
n <- sum(X)
cas <- rep(c(0,1,2),cases)
con <- rep(c(0,1,2),controls)
y <- c(rep(1,sum(cases)), rep(0,sum(controls)))
x <- c(cas,con)
r <- cor(x,y)
A <- n*(r^2)
pvalue <- pchisq(A,df=1,lower.tail=FALSE)
cat("Armitage trend test:", A,"\n")
```

```
## Armitage trend test: 13.83108
```

```
cat("P-value:", pvalue, "\n")
```

```
## P-value: 0.0002000008
```

5. Is there evidence for association of this marker with the disease? Argument your response.

The plots for the codominant, dominant, and recessive models all show an increase in risk with the presence of the 'm' allele(s), suggesting a relationship between the allele and disease risk. The p-values for all three genetic models are below the standard threshold of 0.05, indicating that the results are statistically significant and the associations observed are unlikely due to chance. The p-value from the Armitage trend test, which is particularly low (0.0002000008), provides strong evidence against the null hypothesis of no association. Together, these elements point to a consistent pattern of association between the genetic marker in question and the disease. The evidence suggests that this marker is likely to be related to the risk of developing the disease.