BIF705 Lab 4: Genetic Association Studies in R

Christopher Eeles

February 6, 2019

1 Introduction

The Hardy-Weinberg Equilibrium (HWE) equation represents the distribution of alleles for a population wherein those alleles are randomly propagated across generations.[1] In the context of genetic association studies, the HWE state of a population is used as a measure of control data quality.[1] To conduct a valid association analysis, the control population must be in HWE as it is the standard by which the genetic association hypothesis is tested; it represents the null hypothesis of no association.[1] Therefore before proceeding to analysis of genetic association, the distribution of the control data must be tested to ensure the model assumptions hold.[1] In this investigation the Pearson goodness-of-fit test (χ^2) will be applied to control data with an α of 0.05. Should HWE be accepted, subsequent analysis will compare the control population to the diseased population at four single nucleotide polymorphisms (SNPs) to test if these mutations are correlated with the diseased phenotype.[2]

Such case-control studies are used to gain a preliminary understanding of which genetic factors may indicate a higher risk of disease in an individual or group of individuals.[1] Testing of both HWE and genetic association will be conducted using the `R` programming language with the `genetics` package. This package extends the functionality of `R` by bundling genotype data into the built-in χ^2 test function.[3] Hypothesis testing for the genetic association of each SNP to the diseased state will also utilize the χ^2 test to determine the p-value of each SNP at an α of 0.05. Results will then be analyzed and conclusions drawn about the predictive relationship of each SNP to the disease phenotype.

2 Results

2.1 Data Summary

Table 1: **Summary of popn.txt Table Columns**

	Gender	**Affected**	**SNP A**
Command	>summary(popn$gender)	>summary(popn$affected)	>summary(popn$A)
Results	Female Male NA's 1037 383 116	Case Control 672 864	1/1 1/2 2/2 NA's 541 704 244 47
	SNP B	**SNP C**	**SNP D**
Command	>summary(popn$B)	>summary(popn$C)	>summary(popn$D)
Result	1/1 1/2 2/2 NA's 531 734 234 37	1/1 1/2 2/2 NA's 507 696 278 55	1/1 2/1 2/2 NA's 296 691 454 95

2.2 Hardy-Weinberg Equilibrium

2.2.1 SNP A

```
>HWE.chisq(genotype(A)[affected == "Control"], simulate.p.value=F)
Pearson's Chi-squared test with Yates' continuity correction
data:  tab
X-squared = 0.0045979, df = 1, p-value = 0.9459
```

2.2.2 SNP B

```
>HWE.chisq(genotype(B)[affected == "Control"], simulate.p.value=F)
Pearson's Chi-squared test with Yates' continuity correction
data:  tab
X-squared = 0.36885, df = 1, p-value = 0.5436
```

2.2.3 SNP C

```
>HWE.chisq(genotype(C)[affected == "Control"], simulate.p.value=F)
Pearson's Chi-squared test with Yates' continuity correction
data:  tab
X-squared = 0.94078, df = 1, p-value = 0.3321
```

2.2.4 SNP D

```
>HWE.chisq(genotype(D)[affected == "Control"], simulate.p.value=F)
Pearson's Chi-squared test with Yates' continuity correction
data:  tab
X-squared = 2.2247, df = 1, p-value = 0.1358
```

2.3 Genetic Associations

2.3.1 SNP A

```
> t1<-table(A,affected) # Show table for A
> t1
A     Case Control
  1/1  280     261
  1/2  298     406
  2/2   83     161
chisq.test(t1,correct=FALSE) #one sided therefore p-val/2
data:  t1
X-squared = 23.738, df = 2, p-value = 7.003e-06
```

2.3.2 SNP B

```
> t2<-table(B,affected) # Show table for B
> t2
B     Case Control
  1/1  202     329
  1/2  332     402
  2/2  123     111
chisq.test(t2,correct=FALSE) #one sided therefore p-val/2
data:  t2
X-squared = 15.063, df = 2, p-value = 0.0005358
```

2.3.3 SNP C

```
> t3<-table(C,affected) # Show table for C
> t3
C     Case Control
  1/1  267     240
  1/2  301     395
  2/2   90     188
chisq.test(t3,correct=FALSE) #one sided therefore p-val/2
data:  t3
X-squared = 30.678, df = 2, p-value = 2.18e-07
```

2.3.4 SNP D

```
> t4<-table(D,affected) # Show table for D
> t4
D     Case Control
  1/1   96     200
  1/2  321     370
  2/2  240     214
chisq.test(t4,correct=FALSE) #one sided therefore p-val/2
data:  t4
X-squared = 30.549, df = 2, p-value = 2.325e-07
```

3 Discussion and Conclusion

3.1 Data

As seen in Table 1, the `popn.txt` file contains a set of allele markers for each of the four SNPs being analyzed. Columns indicate gender, affected, A, B, C, D and subject. The gender column contains 1037 Females, 383 Males and 116 NA's. The affected column designates cases vs controls for the study, with 672 cases and 864 controls respectively. The following four lettered columns—A, B, C, D—specify genotypic frequency for each of the four SNPs being studied; values can be either 1/1, 1/2 or 2/2. The final column, subject, is a unique identifier for each individual in the study. This is included to maintain anonymity of the study participants in accordance with privacy laws and medical ethics, while still allowing identification of participants by authorized professionals.

3.2 Hardy-Weinberg Equilibrium

Given the two-sided default for χ^2 tests in R, calculated p-values must be divided by two to yield the appropriate one-sided value. Therefore the resulting p-values for the A, B, C and D control subjects from section 2.2 were 0.47295, 0.2718, 0.16605 and 0.0679, respectively. Since the selected p-value for this investigation was $\alpha = 0.5$, the null-hypothesis of HWE is not rejected for any of the SNPs—although, SNP D was near the cut-off. Based on this finding it is reasonable to conclude each control population is in HWE, thus the assumptions needed for subsequent genetic association analysis hold.

3.3 Genetic Associations

The R code and output presented in section 2.3 must also be converted from a two-sided to a one-sided test. After halving, the p-values are 3.502e-06 for A, 2.679e-04 for B, 1.09e-06for C and 1.162e-07 for D. Based on the selected α of 0.5, the null hypothesis of no association fo all four of the SNPs is rejected. Since the alternative hypothesis for this χ^2 test is that there is an association between the SNP genotypes and the disease phenotype, it can be concluded that these four SNPs are predictively correlated with the disease from this case-control study. While this suggests a potential role for these SNPs in the development of what ever disease was being studied in this data, further investigation is needed to delineate correlation from causation.

References

[1] School of Biological Sciences and Applied Chemistry. "BIF 705: Topic 4". In: *Seneca College: Toronto, ON* (2019).

[2] School of Biological Sciences and Applied Chemistry. "BIF705 Lab 4: Genetic Association studies in R". In: *Seneca College: Toronto, ON* (2019).

[3] Gregory R. Warnes. "genetics v1.3.8.1". In: *RDocumentation* (2019). URL: https://www.rdocumentation.org/packages/genetics/versions/1.3.8.1.

A R Script

```
1
2       library(genetics)
3
4   #### Load data
5
6   popn <- read.table(file.choose(),header=T) # Reads tab delimited files
7
8   # File.choose dynamically loads the address, allows user to choose file.
9
10  summary(popn) # The summary table is meaningless because subject is not a continuous
        variable.
11
12  # These analyses will ignore the missing values, so the sums may not work out
13
14  attach(popn) # Converts each column into an object of column name with a vector of the
        values in that column
15
16  #### Test HWE in Control Samples
17
18  # Only want control values, so need to pass more than one argument.
19  # By default simulates a p-value from a chisq distribution, we want to specify degrees of
        freedom therefore add simulate.p.value=F
20  # Note that column names are case sensitive
21
22  HWE.chisq(genotype(A)[affected == "Control"], simulate.p.value=F) # This is different from
        normal chisq function because it also
23                                                                    # grabs information about
                                                                        genotypes and
                                                                        incorporates them into
                                                                         a chisq test
24
25  HWE.chisq(genotype(B)[affected == "Control"], simulate.p.value=F)
26
27  HWE.chisq(genotype(C)[affected == "Control"], simulate.p.value=F)
28
29  HWE.chisq(genotype(D)[affected == "Control"], simulate.p.value=F)
30
31
32  # Result indicates that HWE is not rejected
33  # For assignment check HWE for B, C and D as well
34
35  #### Genetic Association
36
37  t1<-table(A,affected) # Show table for A
38  t1
39
40  chisq.test(t1,correct=FALSE) # p-value outputs the two-sided p-val, therefore need to divide
        by 2
41
42  # Output indicates there is an association between the SNP and the disease
43  # Reject null hypothesis of no association
44
45  # Complete this analysis for B, C and D
46  # Learn how to use R markdown for this assignment
47  # This is precursor to genome association studies
48
49  t2<-table(B,affected)
50  t2
51
52  chisq.test(t2,correct=FALSE)
```

```
53
54  t3<-table(C,affected)
55  t3
56
57  chisq.test(t3,correct=FALSE)
58
59  t4<-table(D,affected)
60  t4
61
62  chisq.test(t4,correct=FALSE)
```