MAST30033 Statistical Genomics 2023 S2

Practical 12

Jiadong Mao

Oct, 2023

Contents

What you need for this practical

- install the following libraries:

```r
install.packages('genetics')
install.packages('rrBLUP')
```

- Four data files: `geno_wheat_chro1.txt`, `map_wheat_chro1_cm.txt`, `geno_frog.txt`, `pop_frog.txt`

1 Simulating data under HWE (Week 2 L2)

1. We consider a bi-allelic SNP. We will simulate two groups of data, each having 500 individuals. The first group has allele frequencies $P_A = 0.7$ and $P_B = 0.3$, the second group has allele frequencies $P_A = 0.3$ and $P_B = 0.7$. Both data sets should follow the HWE assumption.

   Hints: For the simulation of each group, you will need to use the function `sample()` with appropriate alleles frequencies as probabilities. Remember that the genotypes are coded here as 1, 0 and -1.

2. Conduct the Chi-squared test on the two simulated genotypes separately. To use our previous `HWEtest()` function presented in Practical 2, you will need to transform your dataset into a matrix (`as.matrix()`), as your simulated genotype will be a single vector.

3. Combine the two data sets as a whole, and conduct a chi-squared test on the pooled data set.

4. Interpret the results.

```r
# same function as in prac 2 but with less comments.
HWEtest <- function(mat.geno){
  #R function to conduct chi-squared test of HWE on a genotype matrix data
  #Input: genotype matrix data where SNPs are in columns
  #Output Chi-squared statistic, and the corresponding p-values for each SNP

  #Genotype counts for each SNP
  nAA = colSums(mat.geno == -1, na.rm = TRUE)
  nAB = colSums(mat.geno == 0, na.rm = TRUE)
  nBB = colSums(mat.geno == 1, na.rm = TRUE)
  # store the observed counts in a data frame
  ObsCount <- data.frame(nAA , nAB , nBB)
  n <- nAA + nAB + nBB

  # for each SNP, we calculate the Freq of allele A
  p <- ncol(mat.geno)
  freqA <- vector(length = p) # initialise
  for (j in 1:p){
    freqA[j] <-  (2*nAA[j] + nAB[j])/(2*n[j])
  }

  #Expected genotype counts of AA, AB and BB under the HWE assumption - based on FreqA
  ExpCount <- data.frame(n*freqA^2, 2*n*freqA*(1-freqA), n*(1-freqA)^2)

  #Chi-squared statistic to compare expected and observed genotype counts.
  ChiSqStat <- apply((ObsCount-ExpCount)^2/ExpCount, 1, sum)

  #p-value of the test statistic ~ Chi-squared distribution wth 1 degree of freedom 1.
  Pval <- 1-pchisq(ChiSqStat, df=1)

  #Output the result
  return(Pval)
}
```

```
# SOLUTIONS

#Specify allele frequencies in the first population
pA1 <- 0.3
pB1 <- 0.7
#Specify allele frequencies in the second population
pA2 <- 0.7
pB2 <- 0.3
#Calculate genothype frequencies of the first population
pAA1 <- pA1*pA1
pBB1 <- pB1*pB1
pAB1 <- 1 - pAA1 - pBB1
#Calculate genothype frequencies of the second population
pAA2 <- pA2*pA2
pBB2 <- pB2*pB2
pAB2 <- 1 - pAA2 - pBB2
n <- 500 #simulate 500 individuals
geno1 <- sample(c(1,0,-1), size=n, replace = TRUE, prob = c(pAA1,pAB1,pBB1))
geno2 <- sample(c(1,0,-1), size=n, replace = TRUE, prob = c(pAA2,pAB2,pBB2))
#Combine the two populations
genoc <- c(geno1,geno2)
#conduct HWE test on the first population
HWEtest(as.matrix(geno1))
```

```
## [1] 0.8382683
```

```
#conduct HWE test on the second population
HWEtest(as.matrix(geno2))
```

```
## [1] 0.7154502
```

```r
#conduct HWE test on the pooled population
HWEtest(as.matrix(genoc))
```

```
## [1] 8.929669e-08
```

```r
#The results show that the population substructure may cause the deviation from the HWE.
```

In assignment 1, you will conduct a PCA on those data.