3858 Tutorial: GLR Hardy-Weinberg Example

Yifan Li, yli2763@uwo.ca

2017-03-06

Ideas and Theory

Suppose someone give me a sequence of 0-1 and he said he tossed a coin to produce this sequence. How do I know this sequence is truly from a tossed coin (which is random and independent), or this a faked data (which is not random or independent)?

One convenient way is to group the data with size k and count the number of heads in each group.

Here are the notations

- N: the total number of times to toss the coin.
- k: group size which need deciding appropriate
- $n\,(=N/k)$: the number of groups in total
- $M\,(=k+1)$: the number of categories in multinomial distn (the number of heads in a group: $0, 1, \cdots, k$.)

One group falls into the jth category means this group has $(j-1)$ heads ($j = 1, \cdots, M$). Suppose we get X_j groups in the jth category ($\sum\limits_{j=1}^{M} X_j = n$). Then the vector $(X_1, X_2, \cdots, X_M) := \underset{\sim}{X}$ will follow the multinomial distribution with n units in total and assigned probability for each category $(p_1, p_2, \cdots, p_M) := \underset{\sim}{p}$ ($\sum\limits_{j=1}^{M} p_j = 1$).

Then the parameter space is $\Theta = S_6 = \{\underset{\sim}{p} | \sum\limits_{j=1}^{M} p_j = 1\}$.

We consider the test:

H_0 : $\underset{\sim}{p}$ follows the setting of $Binomial(k, \beta)$ (or the Hardy-Weinberg model) i.e. $\underset{\sim}{p} \in \Theta_0 = \{\underset{\sim}{p} |\ p_j = \begin{pmatrix} k \\ j-1 \end{pmatrix} \beta^{j-1}(1-\beta)^{k-(j-1)},\ j = 1, \cdots M\}$ versus $H_1 : \underset{\sim}{p} \in \Theta \backslash \Theta_0$.

The MLE for Θ_0 is $\underset{\sim}{p}(\hat{\beta})$, where $\hat{\beta} = \frac{1}{kn}(\sum\limits_{j=1}^{M}(j-1)X_j) = \frac{1}{N}(\sum\limits_{j=1}^{M}(j-1)X_j)$, while the MLE for Θ is $\underset{\sim}{\hat{p}}$, where $\hat{p}_j = \frac{X_j}{n}$. Therefore, we have the GLR,

$$\Lambda(\underset{\sim}{X}) = \prod_{j=1}^{M} \left(\frac{p_j(\hat{\beta})}{\hat{p}_j} \right)^{X_j}$$

and the test statistic (with the approximated form: the chi-squared statistic),

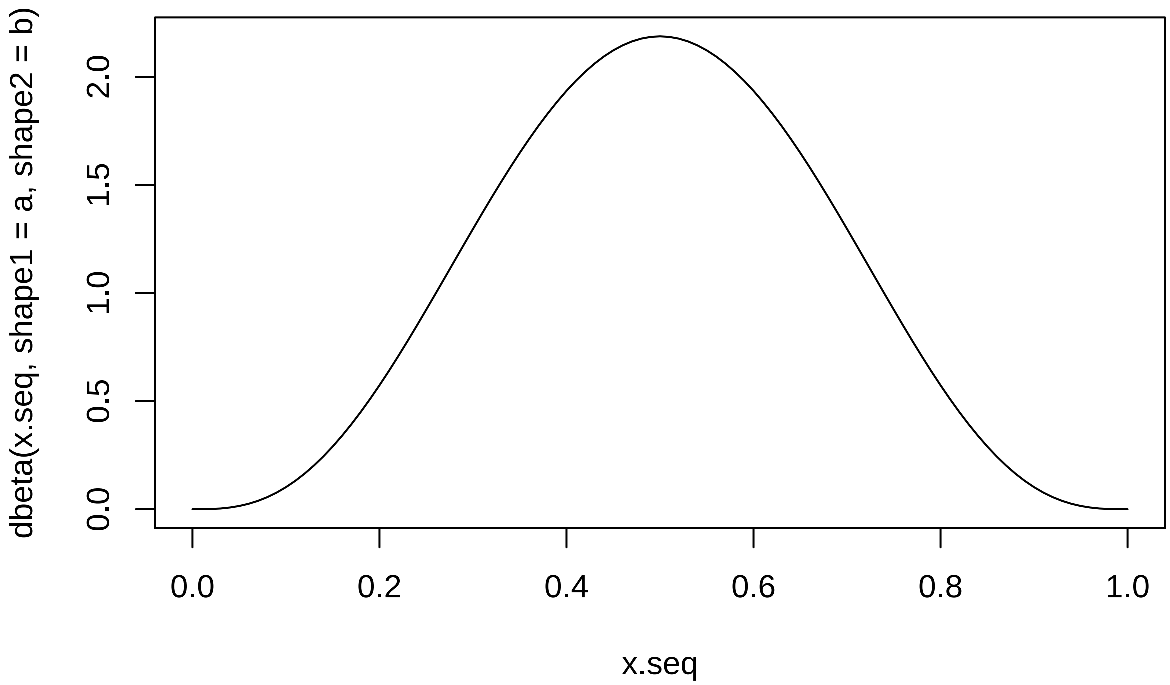
$$-2 \log \Lambda(\underset{\sim}{X}) \approx \sum_{j=1}^{M} \frac{(n\hat{p}_j - np_j(\hat{\beta}))^2}{np_j(\hat{\beta})} \sim \chi^2_{(d)}$$

where $d = dim(\Theta) - dim(\Theta_0) = (M-1) - 1$.

Input the data

```
n = 100
a = 4
b = 4
theta.seq = rbeta( n , shape1 = a , shape2 = b)

x.seq = seq( 0 , 1 , .01)
plot( x.seq , dbeta(x.seq , shape1=a , shape2=b) , type = "l")
```




```
x = rep(0, n)
for(i in 1:n){
  x[i] = rbinom( 1 , size = 5 , prob = theta.seq[i])
}
```

The data is actually generated with a different random theta (= beta) for each set of 5, so the X's are independent but not iid. However for this specific case one needs much more data to detect this.

```
X.multi <- c(6, 12, 31, 24, 22, 5)
x <- X.multi
M <- length(x) #we have 6 categories
k <- M-1 #the group size is 5
n <- sum(x) #the total number of groups is 100
N <- n*k #the total number of times to toss the coin is 500
```

```
#null hypothesis
beta.hat <- sum((0:k)*x)/N #the proportion of heads in total
p.hat0 <- dbinom(0:k, k, beta.hat)
#p.hat0[j] <- choose(k,j)*beta.hat^j*(1-beta.hat)^(k-j)
p.hat <- x/n
LogLambda <- function(x){
  #x is the data of multinomial distribution
  #x=c(x_0,x_1,...,x_k)
```

```r
  -2*sum(x*log(p.hat0/p.hat))
}
Chisquare <- function(x){
  E <- p.hat0*n
  O <- x
  sum((O-E)^2/E)
}
d <- M-1-1
alpha <- .05
cat("The -2*log(Lambda) is ", LogLambda(x), "\n",
    "The Chi-squared statistic is", Chisquare(x), "\n",
    "The lower quantile for", (1-alpha),
    "of chi-squared distn with df", d, "is", qchisq(1-alpha, d))
```

```
## The -2*log(Lambda) is  7.431776
##  The Chi-squared statistic is 8.556909
##  The lower quantile for 0.95 of chi-squared distn with df 4 is 9.487729
```

```r
pchisq(LogLambda(x), d, lower.tail = FALSE)
```

```
## [1] 0.1147556
```

```r
pchisq(Chisquare(x), d, lower.tail = FALSE)
```

```
## [1] 0.07318087
```

Since the value of the statistc is less than the critical value and the p-value here is significantly above 0.1, we will accept the null hypthesis that the parameters $p = (p_1, p_2, \cdots, p_5)$ of the multinomial distribution will follow the Binomial(5,β) setting (or the Hardy-Weinberg model, i.e. $p_i =$). Meanwhile, we may guess the data is coming from a sequence of independent tossing trails.