

### STAT 530 Bioinformatics: Homework 1

#### Problem1:

Risk difference is  $(p_1 - p_2)$ , where  $p_1$  = probability of developing disease if  $X_i = 1$ , that is  $A/(A+B)$ ;  
and  $p_2$  = probability of developing disease if  $X_i = 0$ , that is  $C/(C+D)$ .

(1) Under cross-sectional sampling, we select  $n$  subjects at random from population.

Thus  $a = An/N$ ,  $b = Bn/N$ ,  $c = Cn/N$ ,  $d = Dn/N$ . ( $N=A+B+C+D$ )

Then  $p_1 - p_2 = A/(A+B) - C/(C+D) = a/(a+b) - c/(c+d)$

(2) Under case-control sampling, we choose  $n_1$  cases and  $n_0$  controls at random.

Thus  $a = An_1/(A + C)$ ,  $b = Bn_0/(B + D)$ ,  $c = Cn_1/(A + C)$ ,  $d = Dn_0/(B + D)$ .

Then  $p_1 - p_2 = A/(A+B) - C/(C+D) \neq a/(a+b) - c/(c+d)$

#### Problem2:

Yes, we can use this logistic model to estimate  $\beta_1$  under case-control sampling.

Since  $a = An_1/(A + C)$ ,  $b = Bn_0/(B + D)$ ,  $c = Cn_1/(A + C)$ ,  $d = Dn_0/(B + D)$

Then  $OR = [p_1/(1 - p_1)]/[p_2/(1 - p_2)] = (A/B) / (C/D) = AD/BC = ad/bc$

And the estimation for the logistic model is  $OR = \exp(\beta_1)$

So use the equation  $\exp(\beta_1) = ad/bc$  to estimate  $\beta_1$ .

#### Problem3:

```
gulishana@Gulishana-VirtualBox:~$ head /proc/cpuinfo
```

```
processor      : 0
vendor_id     : GenuineIntel
cpu family    : 6
model         : 61
model name    : Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz
stepping      : 4
cpu MHz       : 2194.920
cache size    : 3072 KB
physical id   : 0
siblings      : 1
```

```
gulishana@Gulishana-VirtualBox:~$ head /proc/meminfo
```

```
MemTotal:      3015124 kB
MemFree:       1489900 kB
MemAvailable:  2147324 kB
Buffers:       69448 kB
Cached:        701712 kB
SwapCached:    0 kB
Active:        1087392 kB
Inactive:      293476 kB
Active(anon):  610632 kB
Inactive(anon): 9060 kB
```

#### Problem4:

There are 45 controls and 44 cases.

The total genotyping rate is 0.99441.

#### Problem5:

I set up a csv file with data of  $X_i$ , which means the number of allele "a". Then I run the R to analyze the data.

## R Notebook

This is an [R Markdown](#) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
setwd("C:/Users/Gulishana/Desktop")
x <- read.csv("hw1.csv")
p <- 187/232
n <- 116
z<-x$x
p.hat <- sum (2*as.numeric(z==2)+as.numeric(z==1))/(2*length(z))
O <- table(z)
E <- length(z)*c((1-p.hat)^2,2*p.hat*(1-p.hat),p.hat^2)
T <- sum((O-E)^2/E)
T

## [1] 11.20194
```

```
pchisq(T,df=1,lower.tail=FALSE)
```

```
## [1] 0.0008171198
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).