Homework 2

Anshuman Narayan

10/9/2018

Q1

Q1a

For this contigency table, we can assume that the random variable follows a Poisson sampling model. We can assume that we observe the number of accidents in Florida over a period of time and record the use and non-use of seatbelts for fatal and nonfatal accidents. Each row of observations(or column) have a total sample size that is random.This can lead to an assumption that the probability of each cell π_{ij} is independent and is the realization of a Poisson random variable with an unkown parameter μ_{ij}. Also,neither of the row totals or column totals are fixed, which further suggests that choosing a Possion sampling model is an appropriate choice.

Q1b

We now need to estimate the probabilty of fatality conditional of seat belt use i.e we need to calculate P(Y=Fatal Accident|X=Seat Belt used) and P(Y=Fatal Accident|X= Seat Belt not used).

```
N21=703 #no of fatal accidents with seat belt used
N22=441239 #no of nonfatal accidents with seat belt used.
N12=55623 #no of non-fatal accidents without seat belt use
N11=1085#no of fatal accidents without seat belt use
Pfatalsb = N21/(N21+N22)
Pfatalnsb=N11/(N11+N12)
cat("\nProbabilty of fatal injury conditional on  wearing seat belt ",Pfatalsb)
```

```
##
## Probabilty of fatal injury conditional on  wearing seat belt  0.001590706
```

```
cat("\nProbability of fatal injury conditional on not  wearing seat belt",Pfatalnsb)
```

```
##
## Probability of fatal injury conditional on not  wearing seat belt 0.0191331
```

Q1c

We now need to estimate the probabilty of wearing a seat belt, conditional on the fatality of the accident i.e we need to calculate P(Y=Seat belt used|X=Fatal) and P(Y=Seat Belt used|X=NonFatal).

```
Psfatal=N21/(N21+N11)
Psnonfatal=N22/(N22+N12)
cat("\nProbabilty of wearing a seat belt conditional on injury being fatal",Psfatal)
```

```
##
## Probabilty of wearing a seat belt conditional on injury being fatal 0.3931767
```

```
cat("\nProbabilty of wearing a seat belt conditional on injury being nonfatal",Psnonfatal)
```

```
##
## Probabilty of wearing a seat belt conditional on injury being nonfatal 0.8880514
```

Q1d

We choose the nature of injury(fatal or nonfatal as the response variable). Let us consider the injury being fatal as a success. With this assumption, we define π_1 as the probabilty of fatality without seat belt use and π_2 as the probabilty of fatality with seat belt use. We now calculate π_1 and π_2 and the difference of sample proportions($ π_2 - π_1), *samplerelativerisk*(π_1/π_2)*andsampleoddsratio*(θ= (π_1/(1-π_1))/(π_2/(1-π_2))$).

```
p1=N11/(N12+N11) #pi 1
p2=N21/(N22+N21)#pi 2
cat("\nThe probability of fatality without seat belt use",p1)
```

```
##
## The probability of fatality without seat belt use 0.0191331
```

```
cat("\nThe probability of fatality with seat belt use",p2)
```

```
##
## The probability of fatality with seat belt use 0.001590706
```

```
cat("\nThe difference of sample proportions",p1-p2)
```

```
##
## The difference of sample proportions 0.0175424
```

```
cat("\nThe sample relative risk ",p1/p2)
```

```
##
## The sample relative risk  12.02805
```

```
theta=(p1/(1-p1))/(p2/(1-p2))
cat("\nThe sample odds ratio",theta)
```

```
##
## The sample odds ratio 12.24317
```

As we can see the sample relative risk and the sample odds are approximately equal.This is because both the probabilties are relatively small(0.02 and 0.0015) which tells us that while the probabilty of having a fatal injury is relatively low, not wearing a seat belt increases this probabilty 12 times which is what the sample relative risk tells us. Also since the pi values are so small, when calculating the odds ratio, both the (1-pi) terms are almost equal to one,because of which the sample odds ratio is approximately pi1/pi2 which is exactly what the sample relative risk is.

Q2

Q2.a

The odds ratio gives us the ratio of the odds of success when X is 1 and the odds of success when X is 2. In this example, that translates to the odds ratio telling us the ratio of the odds of a survivor being being a woman and the odds of a survior being a man. The statement ” The probability of survival for females was 11.4 times that for males” is incorrect because it misinterprets odss for probability. The correct interpretation would be that “The odds for a survior being a woman is 11.4 times more than the odds for a survior being a man”. These two statements would approximately be the same if the π_I's are very small such that $1 - \pi_i$ is almost equal to one, equating the odds ratio to the relative risk.

Q2.b

We have the odds of survival for women given as 2.9. Equating this to the formula we have $\pi_f/(1 - \pi_f)$ we can estimate π_f which is the probability of a woman surviving. The value π_f is given by

```r
pi_f=2.9/3.9
cat("\nThe probability of a woman surviving is ",pi_f)
```

```
##
## The probability of a woman surviving is  0.7435897
```

Since we have the odds of female surviving and the odds ratio of female and male surviving we can plug those values into the formula for odds ratio given in terms of success probabilites. The formula is $ θ= (π_f/(1-π_f))/(π_m/(1-π_m))$ which is equal to 11.4. We can then find the value of π_m.

```r
pi_m=2.9/(11.4+2.9)
cat("The probability of a man surviving is ",pi_m)
```

```
## The probability of a man surviving is  0.2027972
```

Q3

Q3.a

From the question posed, our sample is generated by considering two population sets, one which has people who already have the disease and the second of those who don't. From this we record the tobacco use in both sets. Since the "rows" of the contigency table are from different populations and are fixed, they can be considered independent of one another. This method of sampling is called the independent Multinomial sampling model. The method defined in the question tells us that the resulting model for sampling would be the independent Multinomial sampling.

Q3.b

Given just the realization of this random table, and the method by which it is generated, we cannot estimate the probability of a randomly selected individual would have this disease given that they use tobacco. This is because the table contains counts that would allow us only to calculate the probabilty that given a person has the disease, what is the probability that they smoke. Since we choose from a pool of people who already have the disease, we would be able to infer anything about the probabilty of them having the disease. If we had choosen from pools of Tobacco users and non-users, we would similarly have been unable to state anything about the probability of tobacco usage.

Q3.c

Given the realization of this random table, it would be possible to estimate the ratio of odds a randomly selected individual has this disease given he/she uses Tobacco to the odds o randomly selected individual has this disease given they do not use Tobacco since the odds ratio for our given table i.e the ratio of odds that a randomly selected individual uses tobacco given they have the disease to the odds of a randomly selected individual doesn't use Tobacco given they have this disease and the ratio mentioned above are the same.

$θ= odds(uses tobacco)/odds(doesn't use tobacco) = odds (has disease)/odds(doesnt have disease) $

Q3.d

N_{11} follows a Binomial probabilty distribution with n=200 and π_i=probabilty that they use tobacco given they have the disease. N_{21} also follows a Binomial distrbution with n=200 and the π_j= probability that they use tobacco given they do not have the disease. N_{11} and N_{21} are independent of one another because they are derived from two seperate samples that are independent of one another.

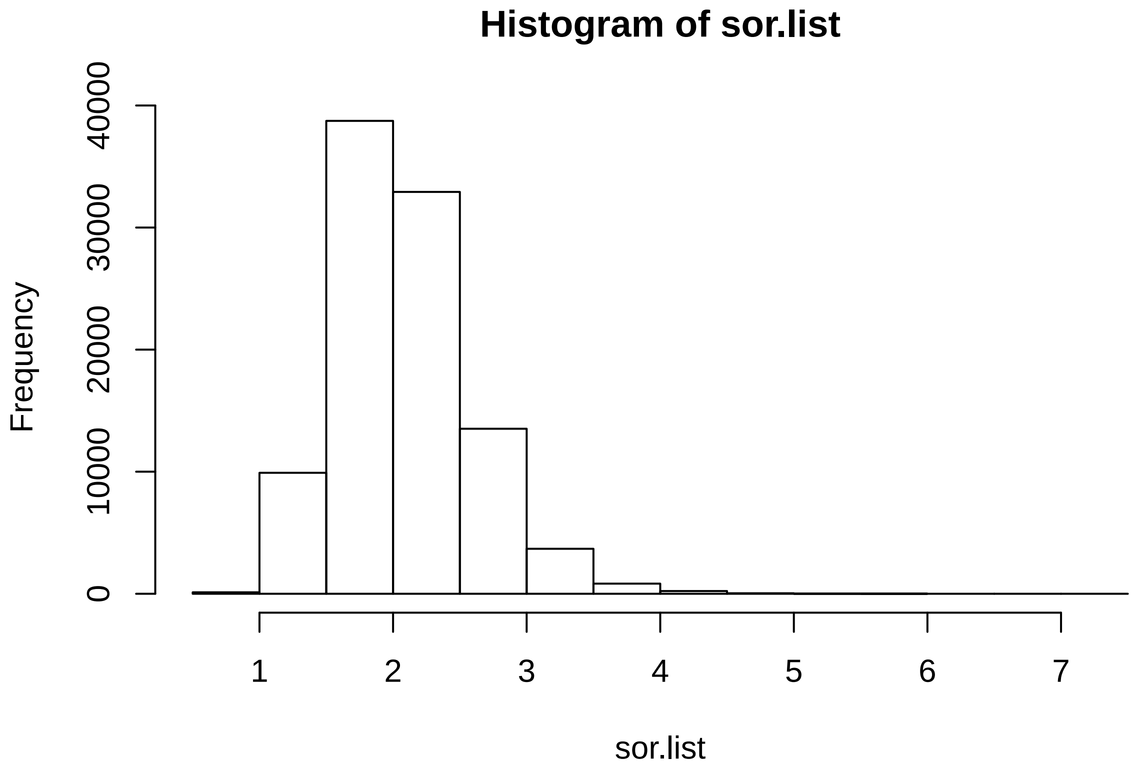
Q3.e

Before we go ahead with simulation study, we need to generate the $\pi_{1|1}$ probability which we can generate from the proportions of a large population provided in the question. We can similarly also generate the $\pi_{2|1}$. Once we do that we can go ahead with generating independent replications of the experiment and generating a series of observed sample odds ratios.

```
## The following code generates a realization
## of reps independent copies of the sample odds
## ratio for a two-by-two contingency table
## with the independent multinomial sampling model

## specify the row totals for the two-by-two table
n1=200
n2=200

## pick the true value for P(Y=1|X=1)
pi1= 0.3333333
## pick the true value for P(Y=1|X=2)
pi2=0.1991952

## the population odds ratio is
theta= ( pi1/(1-pi1) ) / ( pi2/ (1-pi2) )
theta
```

```
## [1] 2.0101
```

```
## specify the number of replications
reps=1e5

## allocate the memory for the vector
## who's rth enry will be the realization
## of the rth independent copy of the sample
## odds ratio
sor.list=numeric(reps)

for(r in 1:reps)
{
  ## generate the counts in the two-by-two contingency table
  ## using the independent multinomial sampling model:
  n11=rbinom(n=1, size=n1, prob=pi1)
  n21=rbinom(n=1, size=n2, prob=pi2)
  tab=rbind( c(n11, n1-n11), c(n21, n2-n21))

  ## compute the realization of the sample odds ratio
  ## for this observed table
  sor=tab[1,1]*tab[2,2] / (tab[2,1]*tab[1,2])

  ## store this sample odds ratio realization
  ## in the rth element of sor.list
  sor.list[r]=sor
}

## get an idea of the shape of the distribution of
## the sample odds ratio
hist(sor.list)
```



```
## compute the arithmetic average of the observed
## sample odds ratios
cat("\nThe arithmetic average of the series of observed sample odds raio",mean(sor.list))
```

```
##
## The arithmetic average of the series of observed sample odds raio 2.079107
```

```
## compute the observed sample standard deviation
## of the observed sample odds ratios
cat("\nThe observed standard deviation of the observed sample odds ratio ",sd(sor.list))
```

```
##
## The observed standard deviation of the observed sample odds ratio  0.4978146
```

The distribution of the 10,000 sample odds ratio is that of a normal distribution centered at the mean of the distribution The value that these sample odds are estimating is the population odds ratio between Disease status(yes,no) and Tobacco use(yes,no) which is given in the question as 2.010101