Homework3

Anshuman Narayan

10/17/2018

Q1

Q1a

Since the data was observed over a period of time where the number of observations were not controlled, assuming that the data represented followed a Poission sampling distribution would be the appropriate sampling model.

Q1b

Before calculate the odds ratio of the first free throw(made/missed) to the odds ratio of the second free throw(made/missed) we need to calculate the success probabilities of the marginal distributions. We state our hypotheses as follows H_null = "successive free throws are independence of one other" H_alternate = "H_null is false" After which we use these probabilities to calculate the odds ratio of the individual successes. With our odds ratio calculated, we look at the 95% confidence interval of the odds ratio.

```
ct = as.table( rbind(c(152,33), c(37,8)) )



rownames(ct) = c("FT1Made", "FT2Missed")
colnames(ct) = c("FT2Made", "FT2Missed")
ct
```

```
##           FT2Made FT2Missed
## FT1Made       152        33
## FT2Missed      37         8
```

```
ct[1,1]*ct[2,2]/(ct[2,1]*ct[1,2])
```

```
## [1] 0.995905
```

```
log(ct[1,1]*ct[2,2]/(ct[2,1]*ct[1,2]))
```

```
## [1] -0.004103412
```

```
sor=function(ct)
{
  val=ct[1,1]*ct[2,2]/(ct[2,1]*ct[1,2])
  return(val)
}
#we use the formula for computing the confidence interval of the log odds raio.
#when we call the exponential function on the confidence interval(CI) of the log odds ratio,
#we get the confidence interval(CI) for odds ratio
sor(ct)
```

```
## [1] 0.995905
```

```
samp.lor=log(sor(ct))
moe=qnorm(0.975)*sqrt(sum(1/ct))
CI = c(samp.lor - moe, samp.lor + moe)
cat("The confidence interval for odds ratio is",exp(CI))
```

```
## The confidence interval for odds ratio is 0.4248685 2.334432
```

As we can see, the value 1 lies in between the odds ratio's 95% confidence interval. Thus , we don't have sufficient evidence against the independence of successive free throws.

Q1c

We now run a likelihood ratio test to see if we get any evidence against our null hypothesis. We calcuate our realization of a chi-squared variable g^2 as $2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}log(n_{ij}/\mu_{ij})$ where $\mu_{ij} = n_{i+}n_{j+}/n$. We then look at the p-value of this variable following a Chi-squared distribution using the pchisq function.

```
expected = apply(ct, 1, sum) %o% apply(ct, 2, sum) / sum(ct)
g.sq = 2*sum( ct * log(ct/expected) )
cat("\nThe value of g-squared ",g.sq)
```

```
##
## The value of g-squared  8.918067e-05
```

```
p.val=1-pchisq(g.sq, 1*1)
cat("\nThe p-value for if g-squared follows the chi-squared distribution",p.val)
```

```
##
## The p-value for if g-squared follows the chi-squared distribution 0.9924652
```

From our p-value we can see that we do not have sufficient evidence against our null hypothesis which states that successive free-throws are independent.

Q2

Q2a

To find out if there is a statistical evidence of an association between Gender and Party ID, we first state our hypothesis. Our hypothesis is defined as H-null : "There is no association between Gender and Party ID" H-alternate : "H-null is false" We can conduct our test for this hypothesis by using the likelihood ratio test or the Pearson's chi-squared test.

```
ct = as.table( rbind(c(422,381,273), c(299,365,232)) )


rownames(ct) = c("Female", "Male")
colnames(ct) = c("Democratic","Independent","Republican")
chisq.test(ct, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 8.2943, df = 2, p-value = 0.01581
```

```
chisq.test(ct)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 8.2943, df = 2, p-value = 0.01581
```

As we can see, both our p-values are lesser than the 5% significance level. Thus we can reject our null hypothesis that there is no association between Gender and Party ID.

Q2b

We generate the Pearson standarized residuals on our expected values of the cells of the contingency table.

```
source("testtable.r")
test.table(ct)
```

```
## -----------------
## Chi-squared tests
## -----------------
## x^2 = 8.294318, df = 2, p-value = 0.01580927
## g^2 = 8.308998, df = 2, p-value = 0.01569365
##
## Pearson standardized residuals
##         Democratic Independent Republican
## Female  2.6852438  -2.4290559 -0.2639868
## Male   -2.6852438   2.4290559  0.2639868
```

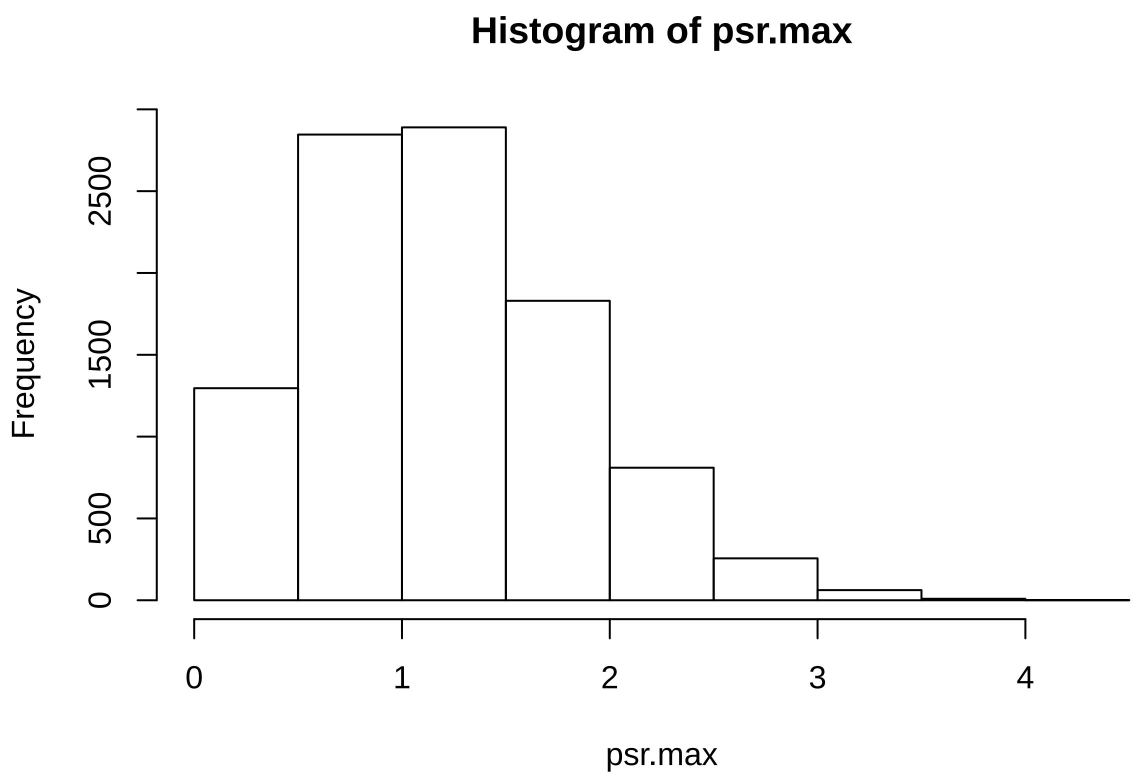
On looking at the Pearson's standardized residuals, the number of females in the Democratic party have a value of 2.68 which means that empirical count is not significantly more than expected count under the null hypothesis model. If the value was greater than 3 then it would have been significantly more than what was expected.

Q2c

```
ct=as.table(rbind(c(422, 381, 273), c(299, 365, 232)))
rownames(ct) = c("female", "male")
colnames(ct) = c("Democrat", "Independent", "Republican")
n=sum(ct)
px=apply(ct, 1, sum)/n
py=apply(ct, 2, sum)/n
px%o%py
```

```
##         Democrat Independent Republican
## female 0.1994958   0.2064131  0.1397301
## male   0.1661229   0.1718830  0.1163551
```

```
n=sum(ct)
probs=as.numeric(px%o%py)
reps=1e4
psr.max=NULL
for(r in 1:reps)
{
  rtab=matrix(rmultinom(1, n, probs), nrow=2,ncol=3)
  expected = apply(rtab, 1, sum) %o% apply(rtab, 2, sum) / n
  col.vec = 1-apply(rtab, 1, sum)/n
  row.vec = 1-apply(rtab, 2, sum)/n
  scale.matrix = col.vec %o% row.vec
  psr=(rtab - expected)/sqrt(expected*scale.matrix)
  psr.max[r]=max(abs(psr))
}
hist(psr.max)
```




```
prop1=length(psr.max[psr.max>2.5])/length(psr.max) #psrmax>2.5
prop2=length(psr.max[psr.max>3])/length(psr.max) #psrmax>3
cat("Probability of psr being greater than 2.5",prop1)
```

```
## Probability of psr being greater than 2.5 0.0328
```

```
cat("\nProbability of psr being greater than 3",prop2)
```

```
##
## Probability of psr being greater than 3 0.0072
```

These two proportions tell us the probability of having empirical count significantly larger than expected at two levels of 2.5 and 3. 2.5 is used for less conservative cases of considering values that do not fit the model, a value more than 3 is a more conservative measure to check if the empirical count doesn't fit the null model.

Q3

Q3.a

We state the null and alternate hypothesis as follows H-null : "There is no dependence between Family income and Education aspirations" H-alternate : "H-null is false"

```
ct = as.table( rbind(c(9,44,13,10), c(11,52,23,22), c(9,41,12,27)) )
rownames(ct) = c("low","middle","high")
colnames(ct) = c("some HS","HS graduate","some college","college graduate")
ct
```

```
##        some HS HS graduate some college college graduate
## low          9          44           13               10
## middle      11          52           23               22
## high         9          41           12               27
```

```r
chisq.test(ct,correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 8.8709, df = 6, p-value = 0.181
```

From the result of our test and looking at the p-value we cannot reject the null hypothesis that there is no dependence between Family income and Education aspirations. ####Q3b Both the variables Family income and Educational aspirations are ordinal. ####Q3c We now conduct the M-squared test to test if there is a linear component in the association trend between Family income and Educational aspirations. We use the row scores(1,2,3) and column scores(1,2,3,4)

```r
ct
```

```
##        some HS HS graduate some college college graduate
## low          9          44           13               10
## middle      11          52           23               22
## high         9          41           12               27
```

```r
source('msqtest.r')
test.out=msq.test(ct, row.scores=1:3 ,col.scores=1:4)
```

```
## -----------------------
## Test for linear trend
## -----------------------
## sample size =  273
## sample correlation value =  0.1321336
## m^2 = 4.748927, df = 1,  p-value = 0.02931658
```

Looking at the p-value of the test performed being less than the significance level of 0.005 we can reject the null hypothesis that states that there is no linear dependence between Family income and Educational aspirations