Stat S-520

Assignment – 11

Parmeet Singh Ahluwalia

Q 1)

Null Hypothesis : H_0: The claim that the M&M's mixing proportion is correct.

Alternate Hypothesis : H_1: The claim that the M&M's mixing proportion is incorrect.

Using R – code to perform the chi-square test:

```
> expected=c(0.13,0.14,0.13,0.24,0.20,0.16)
> observed=c(121,84,118,226,226,123)
> total_amt=sum(observed)
> total_amt
[1] 898
> expected_value=total_amt*(expected)
> expected_value [1] 116.74 125.72 116.74 215.52 179.60 143.68
> alpha=0.05
> df=6-1
> X2=sum((observed-expected_value)^2/expected_value)
> X2 [1] 29.4874
> p=1- pchisq(X2, df)
> p
 [1] 1.860203e-05
```

We are considering alpha as 0.05. P value in the answer above is less than alpha, hence we reject the null hypothesis.

Q 2)

a)

x=c(1,2,3,4,5,6,7,8,9)

The function f(x) is as follows:

$$f(x) = P(X=x) = Log(1+1/x)$$

We get the following f(x) values for different values of x.

f(1)=0.301
f(2)=0.176
f(3)=0.124
f(4)=0.096
f(5)=0.079
f(6)=0.066
f(7)=0.057

f(8)=0.0511
f(9)=0.045

Now $\sum_{i=1}^{9} f(xi) = f(1) + f(2) + f(3) + f(4) + f(5) + f(6) + f(7) + f(8) + f(9) =$
$= 0.9951$

Hence as 0.9951 is rounded to 1. We can say that *f* is a PMF.

b)

Null Hypothesis: H_0: The leading digits follow the Benford's Law.
Alternate Hypothesis: H_1: The leading digits do not follow the Benford's Law.

Using R:

```
> data=c(1,2,3,4,5,6,7,8,9)
>
expected_proportions=c(0.301,0.176,0.124,0.096,0.079,0.066,0.057,0.051
1,0.045)
> observed=c(107,55,39,22,13,18,13,23,15)
> exp_value=305*(expected_proportions)
> exp_value
[1] 91.8050 53.6800 37.8200 29.2800 24.0950 20.1300 17.3850 15.5855
13.7250
```

Since, the significance level is not given, we can assume the value as 0.05.
Degree of freedom = 8

```
> final=sum((observed-exp_value)^2/exp_value)
> final
[1] 14.48037
> p=1 - pchisq(final, df=8)
> p
[1] 0.07007256
```

As the P-value obtained above is greater than 0.05 we do not reject the Null Hypothesis. Hence the leading digits follow Bernford's Law.

Q 3)

H0: A patient's response to treatment is independent of histological type.
H1: A patient's response to treatment is dependent on histological type.

Observed- Histological Type

```
LP=c(74,18,12)
NS=c(68,16,12)
MC=c(154,54,58)
LD=c(18,10,44)
```

Observed-Response

```
positive=c(74,68,154,18)
partial=c(18,16,54,10)
none=c(12,12,58,44)
observed=c(74,18,12,68,16,12,154,54,58,18,10,44)
n=sum(observed)
```

Degrees of Freedom : df =(r -1) * (c-1) = 6

#Expected Values

```
>LP_expected=c((sum(LP)*sum(positive)/n)  ,  (sum(LP)*sum(partial)/n)  ,
(sum(LP)*sum(none)/n))
> LP_expected
[1] 60.69888 18.94424 24.35688

> NS_expected=c((sum(NS)*sum(positive)/n) , (sum(NS)*sum(partial)/n) ,
(sum(NS)*sum(none)/n))
> NS_expected
[1] 56.02974 17.48699 22.48327

> MC_expected=c((sum(MC)*sum(positive)/n) , (sum(MC)*sum(partial)/n) ,
(sum(MC)*sum(none)/n))
> MC_expected
[1] 155.24907  48.45353  62.29740

> LD_expected=c((sum(LD)*sum(positive)/n) , (sum(LD)*sum(partial)/n) ,
(sum(LD)*sum(none)/n))
> LD_expected
[1] 42.02230 13.11524 16.86245

>
expected_values=c(60.69,18.944,24.35,56.02,17.486,22.48,155.24,48.45,6
2.29,42.02,13.11,16.86)

> difference=observed - expected_values

> diff_sq=difference^2

> ind_expectedvalue=diff_sq/expected_values

> ind_expectedvalue

[1]   2.919032790   0.047040541   6.263757700   2.561949304   0.126283655
4.885693950   0.009904664
 [8]   0.635758514   0.295458340 13.730613993   0.737765065 43.687995255

> total=sum(ind_expectedvalue)

> total
[1] 75.90125

> p=1 - pchisq(total, df=6)

> p
[1] 2.498002e-14
```

As the P-value obtained is less than 0.05, we reject the Null Hypothesis and accept the Alternate Hypothesis. Hence a patients response to treatment for Hodgkin's disease is dependent on histological type.

Q 4)

a)

Null Hypothesis : H_0: Anger is not associated with heart disease (i.e, heart disease is independent of anger).

Alternate Hypothesis : H_1: Anger is associated with heart disease.

```
> LA=c(53,3057)
> MA=c(110,4621)
> HA=c(27,606)
> HD=c(53,110,27)
> NHD=c(3057,4621,606)
> observed=c(53,3057,110,4621,27,606)
> N=sum(observed)
> LA_exp=c((sum(LA)*sum(HD)/N) , (sum(LA)*sum(NHD)/N))
> LA_exp
[1]   69.73094 3040.26906
> MA_exp=c((sum(MA)*sum(HD)/N) , (sum(MA)*sum(NHD)/N))
> MA_exp
[1]  106.0762 4624.9238
> HA_exp=c((sum(HA)*sum(HD)/N) , (sum(HA)*sum(NHD)/N))
> HA_exp
[1]  14.19283 618.80717
> expected_values=c(69.73,3040.26,106.07,4624.92,14.19,618.807)
> diff=observed-expected_values
> diff_sq=diff^2
> expect=diff_sq/expected_values
> expect
[1]  4.013952388  0.092172248  0.145610446  0.003322522
11.564207188  0.265057197
> expected_last=sum(expect)
```

```
> expected_last
[1] 16.08432
#df=(no. of rows-1)*(no. of cols -1)
> p=1 - pchisq(expected_last, df=2)
> p
[1] 0.0003216132
```

As the P-value is less than the significance level we reject the null hypothesis and accept the alternate hypothesis.

(b) The P-value obtained above is very less compared to the significance level and hence we reject the null hypothesis. This implies that anger is actually somewhat associated with Heart Disease.

To make the statement that "anger affects the chance of getting heart disease", this analysis is not enough because there might be other causes for heart disease which may or may not include other heart related problems, other causes like diabetes, cholesterol along with anger. Hence this analysis might not be completely true.

Q 5)

Home team Goals

H_0: Poisson model is a good fit for Home Team goals
H_1: Poisson model is not a good fit for Home Team goals.
Since, the significance level is not given ,let us consider alpha to be 0.05

```
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==0)
> x0=92
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==1)
> x1=119
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==2)
>x2=102
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==3)
> x3=46
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==4)
> x4=12
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==5)
> x5=5
```

```
> x=subset(HomeTable, HomeTable$EPL201415.FTHG==6)

> x6=3

> x=subset(HomeTable, HomeTable$EPL201415.FTHG==7)

> x7=0

> x=subset(HomeTable, HomeTable$EPL201415.FTHG==8)

> x8=1

> observed = c(92,119,102,46,12,5,3,0,1)
> frequency=c(0,1,2,3,4,5,6,7,8)
> N=sum(observed)
> N
[1] 380
>
> #goals = sum((1:9)*observed)
> average = sum(frequency*observed)/N
> average
[1] 1.473684
> #expected = rep(NA, 8)
> observed=c(92,119,102,46,12,9)
> expected = rep(NA,6)
> expected[1:5]= N * dpois(0:4, average)
> expected[6]=N*(1-ppois(4,average))
> sum(expected)
[1] 380
> answer = 2 * sum(observed * log(observed/expected))
> 1 - pchisq(answer, df=4)
[1] 0.4015313
```

The p value is greater than 0.05 and hence we accept the null hypothesis.

Away Team Goals:

H_0: Poisson model is a good fit for Away Team goals
H_1: Poisson model is not a good fit for Away Team goals.

Consider alpha as 0.05.

```
> freq=c(0,1,2,3,4,5,6)
> observed = c(132,134,73,32,7,1,1)
> N=sum(observed)
> N
[1] 380
> average = sum(freq*observed)/N
> average
[1] 1.092105
> observed=c(132,134,73,32,9)
> expected = rep(NA,5)
> expected[1:4]= N * dpois(0:3, average)
> expected[5]=N*(1-ppois(3,average))
> sum(expected)
```

```
[1] 380
> answer = 2 * sum(observed * log(observed/expected))
> 1 - pchisq(answer, df=3)
[1] 0.7637317
```

P-value is greater and hence we accept the Null Hypothesis.

Total Goals:

H_0: Poisson model is a good fit for Total Team goals
H_1: Poisson model is not a good fit for Total Team goals.

```
> AwayTable=data.frame(EPL201415$AwayTeam , EPL201415$FTAG)
> HomeTable=data.frame(EPL201415$HomeTeam,EPL201415$FTHG)
> observed=AwayTable$EPL201415.FTAG+HomeTable$EPL201415.FTHG
> table=data.frame(EPL201415$FTAG,EPL201415$FTHG,observed)
> x0=nrow(subset(table,observed==0))
> x1=nrow(subset(table,observed==1))
> x2=nrow(subset(table,observed==2))
> x3=nrow(subset(table,observed==3))
> x4=nrow(subset(table,observed==4))
> x5=nrow(subset(table,observed==5))
> x6=nrow(subset(table,observed==6))
> x7=nrow(subset(table,observed==7))
> x8=nrow(subset(table,observed==8))
> x9=nrow(subset(table,observed==9))

> x0
[1] 31
> x1
[1] 77
> x2
[1] 88
> x3
[1] 85
> x4
[1] 56
> x5
[1] 27
> x6
[1] 9
> x7
[1] 3
> x8
[1] 3
> x9
[1] 1
>
> observed=c(31,77,88,85,56,27,9,3,3,1)
> n=sum(observed)
> n
[1] 380
> freq1=c(0,1,2,3,4,5,6,7,8,9)
```

```
> avg = sum(freq1*observed)/n
> avg
[1] 2.565789

> observed=c(31,77,88,85,56,27,9,7)
> expected = rep(NA,8)
> expected[1:7]= n * dpois(0:6, avg)
> expected[8]=n*(1-ppois(6,avg))
> sum(expected)
[1] 380

> answer = 2 * sum(observed1 * log(observed1/expected))
> p=1 - pchisq(answer, df=6)
>p
[1] 0.9282545
```

Therefore, we can see that since the p value greater than 0.05 we accept the Null Hypothesis.

Q6)