STAT S 520
HOMEWORK 11
RAMPRASAD BOMMAGANTY(rbommaga)

Exercise 13.4 Problem 2:

Solution:

The observed counts are as follows :

Brown	Yellow	Red	Blue	Orange	Green
121	84	118	226	226	123

The expected counts are:

Brown	Yellow	Red	Blue	Orange	Green
116.74	125.72	116.74	215.52	179.60	143.68

The R code for the computation is as follows:

```
> observed= c(121,84,118,226,226,123)
> total <- sum(observed)
> expected<-c(0.13*total,0.14*total,0.13*total,0.24*total,0.20*total,0.16*total)
> G2 <- 2* sum(observed*log(observed/expected))
> 1-pchisq(G2,df=5)
[1] 1.141029e-05
```

Since the p-value is small compared to conventional significance level values, we can reject the null hypothesis that the values are credible and conclude that the claimed proportions are not credible in light of the data.

Exercise 13.4 Problem 5:

Solution:

a. To check that the given function is indeed a pmf we can use the following code:

```
> x1<- log10(1+(1/1))
> x2<- log10(1+(1/2))
> x3<- log10(1+(1/3))
> x4<- log10(1+(1/4))
> x5<- log10(1+(1/5))
> x6<- log10(1+(1/6))
> x7<- log10(1+(1/7))
> x8<- log10(1+(1/8))
> x9<- log10(1+(1/9))
> sum(x1,x2,x3,x4,x5,x6,x7,x8,x9)
[1] 1
```

Since the sum totals to 1, the given function is indeed a pmf.

b. Given observed count values:

Leading digit	1	2	3	4	5	6	7	8	9
Number of towns	107	55	39	22	13	18	13	23	15

Using the null hypothesis that the leading digits of town populations follow Benford's law, we proceed.

The expected counts are as follows:

Leading digit	1	2	3	4	5	6	7	8	9
Number of towns	91.81	53.70	38.10	29.55	24.15	20.41	17.68	15.60	13.95

The R code for the computation of the p-value is as follows:

```
> o <-c(107,55,39,22,13,18,13,23,15)

> e <-c(91.81,53.70,38.10,29.55,24.15,20.41,17.68,15.60,13.95)

> G2 <- 2*sum(o*log(o/e))

> 1-pchisq(G2,df=8)
[1] 0.04767116
```

Since, the p value is less than conventional values of significance level, we can reject our null hypothesis and conclude that it is not plausible that leading digits of town populations follow Benford's law.

Exercise 13.4 Problem 11:

Solution: Given n = 538

The observed counts are as follows:

	Positive	Partial	None
LP	74	18	12
NS	68	16	12
MC	154	54	58
LD	18	10	44

Using the formula $e_{ij} = o_{i+}o_{+j} / n$, we compute the expected counts as follows:

	Positive	Partial	None
LP	60.69	18.94	24.35
NS	56.02	17.48	22.48
MC	155.24	48.45	62.29
LD	42.02	13.11	16.86

The G^2 value is computed by doing $o_{ij} * \log(o_{ij}/ e_{ij})$ for each cell and adding all of them and multiplying by 2.

The computed G^2 value is 68.48.

The R code to compute p-value is :

```
> 1-pchisq(68.48,df=6)
[1] 8.377743e-13
```

The df value is taken as 6, since r = 4 and c = 3, hence (4-1)*(3-1) = 6.

Since the p-value is quite low, we can reject our null hypothesis and conclude that the treatment to Hodgkin's disease does vary by histological type.

Problem 4:

a.
Given n = 8474

The observed values are as follows:

	Heart Disease	No Heart Disease
Low Anger	53	3057
Moderate Anger	110	4621
High Anger	27	606

Using the formula $e_{ij} = o_{i+}o_{+j} / n$, we compute the expected counts as follows:

	Heart Disease	No Heart Disease
Low Anger	69.73	3040.26
Moderate Anger	106.07	4624.92
High Anger	14.19	618.80

The X^2 value is computed by doing $(o_j-e_j)^2/e_j$ for each cell and adding all of them.

The computed X^2 value is 16.06

The degrees of freedom is : (3-1) * (2-1) = 2

The R code for p-value is as follows:

```
> 1-pchisq(16.06,df=2)
[1] 0.0003255482
```

Using conventional significance levels, it can be seen that the p-value is quite low. Hence, we can reject our null hypothesis and conclude that anger is associated with heart disease.

b.
No, this analysis alone does not prove that anger affects the chance of heart disease. Heart disease can occur due to multiple factors apart from anger like obesity, unhealthy lifestyle, stress levels etc. to name a few. While the above test does prove that anger is associated with heart disease, it can be said anger alone is not the cause for heart disease.

Problem 5:

Solution:

a. The R code is as follows:

```
> EPL201415 <- read.csv("http://www.football-data.co.uk/mmz4281/1415/E0.csv")
> home <- EPL201415$FTHG
> away <- EPL201415$FTAG
> N = c(home, away)
> maximum = max(home)
> observed = vector()
> for (i in 0:maximum) {observed = c(observed,length(home[home == i]))}
> observed
[1]  92 119 102  46  12   5   3   0   1
> games = sum(observed)
> goals = sum((0:maximum) * observed)
> average= goals/games
> e = games * dpois(0:20, average)
> o = c(observed[1:5],sum(observed[6:9]))
> e = rep(NA,6)
> e[1:5] = games * dpois(0:4, average)
> e[6] = games * (1-ppois(4,average))
> X2 = sum((o-e)^2/e)
> 1-pchisq(X2, df= 4)
[1] 0.4131955
```

Since the p value is much greater than conventional significance level, we can accept our null hypothesis that the home goals are a fit for the Poisson model.

b.

The R code is as follows:

```
> EPL201415 <- read.csv("http://www.football-data.co.uk/mmz4281/1415/E0.csv")
> away <- EPL201415$FTAG
>  maximum = max(away)
>  observed=vector()
>  for(i in 0:maximum){ observed = c(observed,length(away[away == i]))}
>  games = sum(observed)
>  goals = sum((0:maximum) * observed)
>  average = goals/games
>  e = games * dpois(0:20, average)
>  o = c(observed[1:4],sum(observed[5:7]))
>  e = rep(NA,5)
>  e[1:4] = games * dpois(0:3, average)
>  e[5] = games * (1-ppois(3,average))
>  X2 = sum((o-e)^2/e)
>  1-pchisq(X2, df= 3)
[1] 0.7566041
```

Since the p value is much greater than conventional significance level, we can accept our null hypothesis that the away goals are a fit for the Poisson model.

c.

The R code is as follows:

```
> total = c(home,away)
> maximum = max(total)
>   observed=vector()
>   for(i in 0:maximum){observed = c(observed,length(total[total == i]))}
>   games = sum(observed)
>   goals = sum((0:maximum) * observed)
>   avg = goals/games
>   e = games * dpois(0:20, avg)
>   o = c(observed[1:5],sum(observed[6:8]))
>
>   e = rep(NA,6)
```

```
>  e[1:5] = games * dpois(0:4, avg)
>  e[6] = games * (1-ppois(4,avg))
>  X2 = sum((o-e)^2/e)
>  1-pchisq(X2, df= 4)
[1] 0.4312175
```

Since the p value is much greater than conventional significance level, we can accept our null hypothesis that the total goals are a fit for the Poisson model.

Problem 6: