Solution of assignment 5, ST2304

Problem 1 Download the data and read it in R as done in previous exercises i.e. dataset <- read.csv("survey.csv")

1.1 Testing chi-squares for associations in the data
Sex vs. Political Orientation

```
> political.sex<-table(dataset$political,dataset$sex) #creates cont. table
> political.sex #contingency table for observed data

       female male
  left      20    9
  right      8    4
> chisq.test(political.sex)

        Pearson's Chi-squared test with Yates' continuity correction

data:  political.sex
X-squared = 0.0506, df = 1, p-value = 0.822

Warning message:
In chisq.test(political.sex) : Chi-squared approximation may be incorrect
> chisq.test(political.sex)$exp #contingency table for expected data

          female     male
  left  19.804878 9.195122
  right  8.195122 3.804878
Warning message:
In chisq.test(political.sex) : Chi-squared approximation may be incorrect
```

The warning is related to the relatively low number of observations, and can be ignored here. There is no statistically significant difference between the sexes when it comes to political affiliation ($p=0.822>0.05$). This is also obvious from the similarity of the expected values to the observed values.

Study program vs. Political Orientation

```
> political.program<-table(dataset$political,dataset$studyprogram) #creates cont. table
> political.program #contingency table for observed data

       biology biotech
  left       17      12
  right       4       8
> chisq.test(political.program)

        Pearson's Chi-squared test with Yates' continuity correction

data:  political.program
X-squared = 1.2781, df = 1, p-value = 0.2583

> chisq.test(political.program)$exp #contingency table for expected data
```

```
         biology   biotech
  left  14.853659 14.146341
  right  6.146341  5.853659
```

There is no statistically significant difference between the programs when it comes to political affiliation (p=0.2583>0.05).

Origin vs. Political Orientation
Using the "Region2" column as origin we get

```
> political.region2<-table(dataset$political,dataset$region2) #creates cont. table
> political.region2 #contingency table for observed data

       midtnorge nordnorge ostlandet other sorlandet vestlandet
  left         10         1         7     1         3          7
  right         5         2         3     0         1          1
> chisq.test(political.region2)

        Pearson's Chi-squared test

data:  political.region2
X-squared = 3.6847, df = 5, p-value = 0.5956

Warning message:
In chisq.test(political.region2) :
  Chi-squared approximation may be incorrect
> chisq.test(political.region2)$exp #contingency table for expected data

       midtnorge nordnorge ostlandet     other sorlandet vestlandet
  left  10.609756 2.1219512  7.073171 0.7073171  2.829268   5.658537
  right  4.390244 0.8780488  2.926829 0.2926829  1.170732   2.341463
Warning message:
In chisq.test(political.region2) :
  Chi-squared approximation may be incorrect
```

There is no statistically significant difference between the regions when it comes to political affiliation (p=0.5956>0.05).

1.2 Testing the Study program groups individually
Sex vs. Political Orientation, Biology students
First create the subset from the dataset

```
> biologyset<-dataset[dataset$studyprogram=="biology",]
```

Next is essentially the same as earlier, although the subset is used instead of the full dataset

```
> bio.political.sex<-table(biologyset$political,biologyset$sex) #creates cont. table
> bio.political.sex #contingency table for observed data
       female male
  left      10    7
  right      2    2
> chisq.test(bio.political.sex)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  bio.political.sex
X-squared = 0.0579, df = 1, p-value = 0.8098

Warning message:
In chisq.test(bio.political.sex) :
  Chi-squared approximation may be incorrect
> chisq.test(bio.political.sex)$exp #contingency table for expected data

          female     male
  left  9.714286 7.285714
  right 2.285714 1.714286
Warning message:
In chisq.test(bio.political.sex) :
  Chi-squared approximation may be incorrect
```

There is no statistically significant difference between the sexes among Biology students when it comes to political affiliation (p=0.8098>0.05).

Region vs. Political Orientation, Biology students only

```
> bio.political.region2<-table(biologyset$political,biologyset$region2) #creates cont. t
> bio.political.region2 #contingency table for observed data

        midtnorge nordnorge ostlandet other sorlandet vestlandet
  left          6         1         5     0         1          4
  right         3         0         1     0         0          0
```

The empty group ("other"), will cause an error unless excluded from the chi-square test. This is solved by adding the [,-4] onto the call for the table

```
> chisq.test(bio.political.region2[,-4])

        Pearson's Chi-squared test

data:  bio.political.region2[, -4]
X-squared = 2.625, df = 4, p-value = 0.6224

Warning message:
In chisq.test(bio.political.region2[, -4]) :
  Chi-squared approximation may be incorrect
> chisq.test(bio.political.region2[,-4])$exp #contingency table for expected data

        midtnorge nordnorge ostlandet sorlandet vestlandet
  left   7.285714 0.8095238  4.857143 0.8095238  3.2380952
  right  1.714286 0.1904762  1.142857 0.1904762  0.7619048
Warning message:
In chisq.test(bio.political.region2[, -4]) :
  Chi-squared approximation may be incorrect
```

There is no statistically significant difference between the origins of Biotech students when it comes to political affiliation (p=0.6224>0.05).

Biotech students
The chi-tests are done in the same fashion as for the Biology student subset, although without empty columns in the tables. The resulting data indicates that there is no statistical difference among neither sexes (p=0.9092) nor origin (p=0.4968), when it comes to political affiliation.

1.3 What population?

Might be a bit of a philosophical question, what does the sampled group represent? Definitely not the entire Norwegian populace...
This is to some extent covered in Løvås p. 9-11.

Problem 2 The likelihood function can be simplified to

$$\begin{aligned} L(p) &= \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!} p^{2x_{AA}} 2p^{x_{Aa}} (1-p)^{x_{Aa}} (1-p)^{2x_{aa}} \\ &= \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!} 2p^{2x_{AA}+x_{Aa}} (1-p)^{x_{Aa}+2x_{aa}}. \end{aligned} \tag{1}$$

Taking logs, the log likelihood becomes

$$\begin{aligned} \ln L(p) = &\ln n! - \ln x_{AA}! - \ln x_{Aa}! - \ln x_{aa}! + \ln 2 \\ &+ (2x_{AA} + x_{Aa}) \ln p + (x_{Aa} + 2x_{aa}) \ln(1-p) \end{aligned} \tag{2}$$

The likelihood has its maximum when

$$\begin{aligned} \frac{d}{dp} \ln L(p) &= 0 \\ \frac{2x_{AA} + x_{Aa}}{p} - \frac{x_{Aa} + 2x_{aa}}{1-p} &= 0 \end{aligned} \tag{3}$$

or, letting x_A and x_a denote the total number of A and a-alleles in the sample,

$$\begin{aligned} \frac{x_A}{p} - \frac{x_a}{1-p} &= 0 \\ x_A(1-p) &= x_a p \\ x_A &= (x_a + x_A)p \\ \hat{p} &= \frac{x_A}{x_A + x_a} = \frac{x_A}{2n}. \end{aligned} \tag{4}$$

that is, provided that the population is in Hardy-Weinberg equilibrium, the MLE of p is equal to the sample frequency of A.

Problem 3 The observed genotype data is first made into a vector of data

```
> X<-c(0,8,11,10,26,45)
```

3.1 allele frequencies

Note that the alleles belong to the "allele population", with a population size of n*2, and homozygotes have two copies of a given allele

```
> p1.hat<-(2*X[1]+X[2]+X[3])/(2*sum(X))
> p2.hat<-(2*X[4]+X[2]+X[5])/(2*sum(X))
> p3.hat<-(2*X[6]+X[3]+X[5])/(2*sum(X))
> p1.hat+p2.hat+p3.hat #Just making sure that there has been no obvious mistake, the tot
[1] 1
```

3.2 MLE of genotype frequencies

Now you want to create a vector with the HWE genotype frequencies (the MLE), which has the genotypes in the same order as the vector of observed data

```
> xhat<-c(p1.hat^2,p1.hat*p2.hat*2,p1.hat*p3.hat*2,p2.hat^2,p2.hat*p3.hat*2,p3.hat^2)
```

3.3 Expected number of observations

This is calculated by multiplying the expected HWE frequencies with the population size

```
> Xhat<-xhat*sum(X)
> Xhat
[1]  0.9025  5.1300 12.0650  7.2900 34.2900 40.3225
> sum(Xhat) # Should equal sum(X), in this case 100
[1] 100
```

3.4 Chi-square statistic

The chi-square statistic is calculated as

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i} \tag{5}$$

Which here translates to

```
> chi.HWE<-sum((X-Xhat)^2/Xhat)
> chi.HWE
[1] 6.156367
```

3.5 p-value

The df for the test is calculated from (number of cells-1)-(parameters estimated) = (possible genotypes-1)-(alleles-1), here (6-1)-(3-1)=3 /newline Note that you only estimate two parameters, the third allele frequence can always be written as a function of the first two.

```
> pchisq(chi.HWE,df=3,lower.tail=F)
[1] 0.1042456
```

P>0.05, hence the null hypothesis cannot be rejected, deviation from HWE is not statistically significant.