Solution of assignment 5, ST2304

Problem 1

Download the data and read it in R as done in previous exercises e.g.

```
dataset <- read.csv("student survey.csv")
```

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

```
> political.sex<-table(student.survey$political,student.survey$sex) #create
s cont.table
> political.sex   #contingency table for observed data

       female male
  left      65   17
  right     23   12
> chisq.test(political.sex)

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

data:  political.sex
X-squared = 1.7449, df = 1, p-value = 0.1865
> chisq.test(political.sex)$exp

          female      male
  left  61.67521 20.324786
  right 26.32479  8.675214
```

There is no statistically significant difference between the sexes when it comes to political affiliation (p=0.1865>0.05).

Study program vs. Political Orientation

```
> political.program<-table(student.survey$political,student.survey$studypro
gram) #creates cont.table
> political.program   #contingency table for observed data

       biology biotech
  left       56      26
  right      19      16
> chisq.test(political.program)

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

data:  political.program
X-squared = 1.527, df = 1, p-value = 0.2166

> chisq.test(political.program)$exp

        biology biotech
  left  52.5641 29.4359
  right 22.4359 12.5641
```

There is no statistically significant difference between the programs when it comes to political affiliation (p=0.2166>0.05).

Origin vs. Political Orientation

Using the ”Region2” column as origin we get

```
> political.region2<-table(student.survey$political,student.survey$region2)
> political.region2   #contingency table for observed data

       midtnorge nordnorge ostlandet other sornorge vestlandet
  left        21         2        31     3        6         19
  right       11         5        10     1        3          5
> chisq.test(political.region2)

        Pearson's Chi-squared test

data:  political.region2
X-squared = 7.6949, df = 5, p-value = 0.1739

Warning message:
In chisq.test(political.region2) :
  Chi-squared approximation may be incorrect
> chisq.test(political.region2)$exp

       midtnorge nordnorge ostlandet    other sornorge vestlandet
  left   22.42735  4.905983  28.73504 2.803419 6.307692  16.820513
  right   9.57265  2.094017  12.26496 1.196581 2.692308   7.179487
Warning message:
In chisq.test(political.region2) :
  Chi-squared approximation may be incorrect
```

The warning is related to the relatively low number of observations in some categories. There is no statistically significant difference between the regions when it comes to political affiliation (p=0.1739>0.05).

1.2 Testing the Study program groups individually

Sex vs. Political Orientation, Biology students

First, create the subset from the data. The rest is essentially the same as earlier, although the subset is used instead of the full dataset.

```
> biologyset<-student.survey[student.survey$studyprogram=="biology",]
> bio.political.sex<-table(biologyset$political,biologyset$sex)
> bio.political.sex

       female male
  left       42   14
  right      12    7
> chisq.test(bio.political.sex)

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

data:  bio.political.sex
X-squared = 0.4868, df = 1, p-value = 0.4853

> chisq.test(bio.political.sex)$exp

       female  male
  left   40.32 15.68
  right  13.68  5.32
```

There is no statistically significant difference between the sexes among Biology students when it comes to political affiliation (p=0.4853>0.05).

Region vs. Political Orientation, Biology students only

```
> bio.political.region2<-table(biologyset$political,biologyset$region2)
> bio.political.region2

       midtnorge nordnorge ostlandet other sornorge vestlandet
  left         15         1        21     1        4         14
  right         7         1         7     0        1          3
> chisq.test(bio.political.region2)

        Pearson's Chi-squared test

data:  bio.political.region2
X-squared = 2.0795, df = 5, p-value = 0.838

Warning message:
In chisq.test(bio.political.region2) :
  Chi-squared approximation may be incorrect
> chisq.test(bio.political.region2)$exp

       midtnorge nordnorge ostlandet     other sornorge vestlandet
  left  16.426667 1.4933333 20.906667 0.7466667 3.733333  12.693333
  right  5.573333 0.5066667  7.093333 0.2533333 1.266667   4.306667
Warning message:
In chisq.test(bio.political.region2) :
  Chi-squared approximation may be incorrect
```

There is no statistically significant difference between the origins of Biology students when it comes to political affiliation (p=0.838>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=0.2399) nor origin (p=0.3557), when it comes to political affiliation.

1.3 What population?

This 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

$$L(p) = \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!} p^{2x_{AA}} 2p^{x_{Aa}} (1-p)^{x_{Aa}} (1-p)^{2x_{aa}}$$

$$= \frac{n!}{x_{AA}!x_{Aa}!x_{aa}!} 2p^{2x_{AA}+x_{Aa}} (1-p)^{x_{Aa}+2x_{aa}}$$

The log likelihood becomes

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

The likelihood has its maximum when

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

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

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

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 $n \times 2$, 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
[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)
> xhat
[1] 0.009025 0.051300 0.120650 0.072900 0.342900 0.403225
```

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) = 100
[1] 100
```

3.4 Chi-square statistic

The chi-square statistic is calculated as

$$\chi^2 = \sum_{i=1}^{k} \frac{(obs_i - \exp_i)^2}{\exp_i}$$

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

Note that you only estimate two parameters, the third allele frequency 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.