Assignment 5, ST2304

Problem 1

Open this spreadsheet (http://www.math.ntnu.no/~diserud/ST2304/student survey.csv) and load it into R using the same method as in earlier assignments. This survey among previous ST2304 students describe which study program they belong to (biology, biotech or other), their sex (male, female), if the party they voted for in the last election was on the political right (FrP, H, V or KrF) or left (Sp, Ap, Sv, Rv etc..) and which part of Norway they grew up in (region2 = sørlandet, østlandet, vestlandet, midtnorge and other).

1. Test for association between the sex and political orientation, between study program and political orientation, and between region and political orientation. Include contingency tables of observed and expected values for each of the three tests. See `?chisq.test`. Discuss your findings briefly.

2. Redo the analysis separately for subsets of data containing all biology and all biotechnology students, respectively. You may want to do this by first selecting the desired portions of the original data frame as follows

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

   You also need to use `detach( )` and `attach( )` as appropriate on the various data frames before doing the different chi-square tests.

   Discuss your findings.

3. All statistical hypotheses are statements about a "population" or a "random process". Statistical inferences about such statements are based on a sample of observations from the population or of the random process. Is it clear which population or random process we are referring to in the above hypothesis tests?

Problem 2

Suppose that we observe x_{AA}, x_{Aa}, x_{aa} individuals of genotype *AA, Aa* and *aa* in a sample of n individuals from a population with genotype frequencies

$$P_{AA} = p^2$$
$$P_{Aa} = 2p(1-p)$$
$$P_{aa} = (1-p)^2$$

Since the observed counts are multinomially distributed the likelihood function is

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

Based on the above likelihood function, show that the MLE of population allele frequency p is equal to the frequency of the allele in the sample, that is, $\hat{p} = x_A / (2n)$ where $x_A = 2x_{AA} + x_{Aa}$.

Problem 3

In a sample of $n = 100$ individuals we observe the following number of different genotypes

Genotype	No. of individuals
A_1A_1	0
A_1A_2	8
A_1A_3	11
A_2A_2	10
A_2A_3	26
A_3A_3	45

Write a short R script which computes the following:

1. The MLE of the population allele frequencies p_1, p_2, p_3 of allele A_1, A_2 and A_3.

2. The MLE of the population genotype frequncies assuming that the population is in Hardy-Weinberg equilibrium.

3. Estimates of the expected number of observations of each genotype in the sample.

4. The observed value of the chi-square statistic for the test of the null-hypothesis that the population is in Hardy-Weinberg equilibrium.

5. The P value of the test.

Use vectors of length 6 to represent the observed counts and the estimated genotype frequencies.