Assignment 11, ST2304

Problem 1 Crossing-over events during meiosis in diploid organisms can be modelled as a Poisson process along the chromosomes. Between two given loci, the number of crossing over events X is then Poisson distributed with some expectation d depending on the physical distance between the loci. Recombination between a given pair of loci occurs if an odd number of crossing-over events X occurs.

1. Using simulations, compute an estimate of the probability of recombination for $d = 0.01$, $d = 0.1$, $d = 0.5$ and $d = 5$.
2. Does the recombination probability seem to go towards a limiting value as r goes to infinity? You could optionally make a graph showing how the recombination probability depends on d.

Hint: To check if a number X is odd, compute X modulo 2 and test if the result is equal to 0 or 1. The modulo operator is available as `%%` in R (see `?"%%"`).

Problem 2 If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then

$$\left(\frac{S^2(n-1)}{\chi^2_{\alpha/2}}, \frac{S^2(n-1)}{\chi^2_{1-\alpha/2}}\right) \tag{1}$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2}$$

is a $(1-\alpha)$-confidence interval for σ^2.

1. What does it mean that (1) is a confidence interval?
2. Verify that the coverage is equal to nominal level of 95% using simulations (see handout 5, sect. 4.2).

Problem 3 (Difficult conceptually) Consider the alternative model for the moose ovulation data from assignment 10, that is, the model in which the probability that an individual has ovulated when observed at a given time $time_i$ is

$$p_i = q\phi(\beta_0 + \beta_1 \text{time}_i), \tag{3}$$

and that the number of individuals having ovulated out of the total number of individuals n_i at time time_i, $x_i \sim \text{bin}(n_i, p_i)$.

In assignment 10 fitted this model to the observed data by maximising the likelihood with respect to q, β_0, β_1. The goodness-of-fit of this model can be tested using the deviance D of the model as the test statistic. Under the null hypothesis that the model is correct D is chi-square with $n - p$ degrees of freedom. The deviance is defined as

$$D = 2(\ln L_{\text{full}} - \ln L) \tag{4}$$

where $\ln L$ is the maximum log likelihood under the full and under the fitted model. Under the full model, all the parameters $p_1, p_2, \ldots, p_n$ are unknown parameters.

1. Compute the MLEs of each p_i, that is, $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$ under the full model and store the result in a vector `phat`.

2. Compute the corresponding log likelihood under the full model. Hint: If you form an R expression based on equation (23) in handout 5, this will fail because some of the $\hat{p}_i$'s are zero. Taking logs will then produce `NaN` so that the whole expressions will evaluate to `NaN`. Instead use the expression

   ```
   sum(dbinom(x,size=n,prob=phat,log=T))
   ```

3. Examine the output from `optim` in assignment 10 to find the maximum likelihood of model (3).

4. Use this to compute the observed deviance of this model.

5. What is the critical value and the P value for the goodness-of-fit test of model (3)?

6. What is the expected value of D given that the model is correct? Is there any sign of over- or under-dispersion in the data?