STAT5703 HW2 Ex4

Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

Exercise 4

Q1

Since it's a multinomial model, the joint distribution is,

$$f_\theta(N_A, N_C, N_G, N_T) = \frac{n!}{N_A!N_C!N_G!N_T!} p_A^{N_A} \cdot p_C^{N_C} \cdot p_G^{N_G} \cdot P_T^{N_T}$$

Q2

By leaving out those terms which are not influenced by the value of θ, we can derive the log-likelihood as,

$$L_\theta = \log f_\theta \equiv \sum_{x \in \{A,C,G,T\}} N_x \log P_x$$

By letting the first derivative to be 0, we have,

$$\begin{aligned}
\frac{dL_\theta}{d\theta} &= \sum_x N_x \cdot \frac{1}{p_x} \cdot \frac{dp_x}{d\theta} \\
&= N_A \cdot \frac{-1}{1-\theta} + N_C \cdot \frac{1-2\theta}{\theta-\theta^2} + N_G \cdot \frac{2\theta-3\theta^2}{\theta^2-\theta^3} + N_T \cdot \frac{\cdot 3\theta^2}{\theta^3} \\
&= 0
\end{aligned}$$

Simplying the equation, we get,

$$-N_A \cdot \theta + N_c(1-2\theta) + N_G(2-3\theta) + 3N_T(1-\theta) = 0\theta \, (N_A + 2N_C + N_G + 3N_T) = N_C + 2N_G + 3N_T$$

which is exactly what we want to prove.

Q3

Since this distribution follows those regularity conditions, its asymptotic distribution is a normal distribution with the mean as θ and variance as the inverse of Fisher information. We have,

$$I(\theta) = -E[\frac{d^2 L_\theta}{d\theta^2}] = -\sum_{x \in A,C,G,T} E[N_x] \left(-\frac{1}{p_x^2} \cdot \left(\frac{dp_x}{d\theta} \right)^2 + \frac{1}{p_x} \cdot \frac{d^2 p_x}{d\theta^2} \right)$$

Since for a specific base x, its marginal distribution is a binomial distribution with $p = p_x$ (since we can view all the other bases as a large group and regard one occurance of base x as a success). So using $E[N_x] = n \cdot p_x$, we can simplify the above formula as,

$$I(\theta) = n \sum_{x \in A,C,G,T} \left(\frac{1}{p_x} \cdot \left(\frac{dp_x}{d\theta} \right)^2 - \frac{d^2 p_x}{d\theta^2} \right)$$
$$= n \cdot \frac{1 + \theta + \theta^2}{\theta(1-\theta)}$$

Therefore, the asymptotic distribution is $N(\theta, \frac{\theta(1-\theta)}{n(1+\theta+\theta^2)})$.

Q4

According to the definition, we need to solve the following equation,

$$E[T] = \sum_{x \in A,C,G,T} a_x E[N_x] = n \sum_{x \in A,C,G,T} a_x p_x$$

And after solving it, we get,

$$a_A = 0, \quad a_C = a_G = a_T = 1/n$$

Q5

$$Var[T] = Var[\frac{N_C + N_T + N_G}{n}] = Var[1 - \frac{N_A}{n}] = \frac{\theta(1-\theta)}{n}$$

Since both estimators are unbiased, the relative efficacy would be the ratio of variance,

$$e(T, \hat{\theta}) = \frac{Var[T]}{Var[\theta]} = 1 + \theta + \theta^2$$

Q6

Similar to Q2, we can first write out the log-likelihood and let the gradient to be 0 to find the MLE. Here we relevel the bases as (A:1, C:2, G:3, T:4).

$$L = \sum_{i=1}^{3} N_i \log R_i + N_4 \log \left(1 - \sum_{i=1}^{3} p_i \right)$$

And the gradient would be,

$$\frac{\partial L}{\partial p_i} = \frac{N_i}{p_i} - \frac{N_4}{1 - \sum p_i} = 0$$

By solving it, we get,

$$p_x = \frac{N_x}{n}, \forall x \in \{A, C, G, T\}$$

Since there are three unknown (and free) parameters, the Fisher information is a matrix. Using the result for $E[N_i]$, we have,

$$I[i,j] = \begin{cases} n[p_i + (1 - \sum_i p_i)^{-1}], & \text{if } i = j \\ n\,(1 - \sum_i p_i)^{-1}, & \text{otherwise} \end{cases}$$

And the covariance matrix would be the inverse matrix of the fisher information matrix.

Compare with those two unbiased estimator, this estimator is also unbiased but with two more unknown parameters. Also, the covariance is now a matrix instead of a scalar.

Q7

We can use the likelihood ratio test to test the hypothesis,

$$W = 2\sum_{i=1}^{4} N_i \log \frac{p_i}{p_i(\theta)} = 2\sum_{i=1}^{4} N_i \log \frac{N_i}{n \cdot p_i(\theta)} \sim \chi^2_{3-1} = \chi^2_2$$

The degree of freedom of the chi-square distribution is 2 since the difference of the number of free parameters in these two models is 2.