STAT5703 HW2 Exercise 4

Wen Fan(wf2255), Banruo Xie(bx2168), Hanjun Li(hl3339)

Exercise 4.

Question 1.

The independent random variables N is in multinomial distribution, the joint distribution is

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

Question 2.

log likelihood:

$$L_\theta = \log P_\theta = \log(n!) - \sum_{x\in\{A,C,G,T\}} N_x! + \sum_{x\in\{A,C,G,T\}} N_x \log p_x$$

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

$$\begin{gathered}-N_A + N_C \frac{1-2\theta}{\theta} + N_G \frac{2-3\theta}{\theta} + N_T \frac{3(1-\theta)}{\theta} = 0 \\ -N_A\theta + N_c(1-2\theta) + N_G(2-3\theta) + 3N_T(1-\theta) = 0 \\ \theta\left(N_A + 2N_C + 3N_G + 3N_T\right) = N_C + 2N_G + 3N_T \\ \hat{\theta} = \frac{N_C + 2N_G + 3N_T}{N_A + 2N_C + N_G + 3N_T}\end{gathered}$$

Question 3.

In this case we have$\hat{\theta} \to N(\theta, \frac{1}{nI(\theta)})$, where $I(\theta)$ is the Fisher Information.

$$\begin{aligned}I(\theta) &= -E[\frac{d^2 L_\theta}{d\theta^2}] \\ &= -E[\frac{2\theta a - a - b\theta}{\theta^2(1-\theta)^2}] \\ &= n \cdot \frac{1+\theta+\theta^2}{\theta(1-\theta)}\end{aligned}$$

where $a = N_C + 2N_G + 3N_T,\ b = N_A + 2N_C + N_G + 3N_T\ \ E[a] = n(\theta+\theta^2+\theta^3),\ E[b] = n(1+\theta+\theta^2)$ Therefore, the asymptotic distribution is $N(\theta, \frac{\theta(1-\theta)}{n(1+\theta+\theta^2)})$

Question 4.

We want $E[T] = \theta = n(a_A(1-\theta) + a_C(\theta-\theta^2) + a_G(\theta^2-\theta^3) + a_T(\theta^3))$ Therefore$a_A = 0$, $a_C = 1/n$, $a_G = 1/n$, $a_T = 1/n$

Question 5.

$$Var[T] = Var[\frac{N_C + N_T + N_G}{n}] = Var[1 - \frac{N_A}{n}] = \frac{Var[N_A]}{n^2} = \frac{\theta(1-\theta)}{n}$$
$$efficiency(T, \hat{\theta}) = \frac{Var[T]}{Var[\hat{\theta}]} = 1 + \theta + \theta^2$$

Question 6.

log-likelihood if p_i doesn't depend on θ, and using Lagrange:

$$Lagrange_{p_x;\lambda} = \sum_{x\in\{A,C,G,T\}} N_x \log p_x - \lambda(\sum_{x\in\{A,C,G,T\}} p_x - 1)$$

$$\frac{Lagrange_{p_x;\lambda}}{\partial p_x} = \frac{N_x}{p_x} - \lambda = 0$$

By solving it, we get,

$$p_x = N_x/\lambda \sum_{x\in\{A,C,G,T\}} p_x = \sum_{x\in\{A,C,G,T\}} \frac{N_x}{\lambda} = 1$$

Therefore

$$\lambda = n$$
$$\hat{p}_x = \frac{N_x}{n}, \forall x \in \{A, C, G, T\}$$

Compare with p_x depends on θ :

$$\hat{p}_A = 1 - \hat{\theta}, \hat{p}_C = \hat{\theta} - \hat{\theta}^2, \hat{p}_G = \hat{\theta}^2 - \hat{\theta}^3, \hat{p}_T = \hat{\theta}^3$$

Both are unbiased estimator, but p_A depends on θ needs observed occurrences for all bases, p_A not depends on θ noly need one of them but has 2 more free parameters.

Question 7.

The likelihood ratio test for testing the hypothesis:$P = P(\theta)$,

$$\Lambda = 2 \sum_{x\in\{A,C,G,T\}} N_x \log \frac{\hat{p}_x}{\hat{p}_x(\theta)} = 2 \sum_{x\in\{A,C,G,T\}} N_x \log \frac{N_x}{n\hat{p}_x(\theta)} \sim \chi_2$$