Maximum Likelihood Estimation and Model Fitting, CMEE MSc. Tin-Yu Hui

Practical 2 (Tuesday)

Question 1

i. If r.v. X and Y are independent, then the joint pdf is the ______ of their marginal pdfs.

ii. Given the _____ and the statistical model MLE provides estimates to the _______ of interest.

iii. [Circle the correct answer] The likelihood function is a function of parameters/data.

Question 2 [Marginal and conditional distributions]

Given a pair of r.v. X and Y with their joint pdf $f_{XY}(x, y) = y\left(\frac{1}{2} - x\right) + x$, where $0 < x < 1$ and $0 < y < 2$.

i. Show that $f_{XY}(x, y)$ is a valid joint pdf.

ii. Find the two marginal distributions, $f_X(x)$ and $f_Y(y)$. What can you say about the mean and variance of X?

iii. Find the conditional distribution of X given Y.

iv. Now assume we know $Y = 1/2$, what is the conditional distribution of X?

v. Calculate the conditional mean of X given $Y = 1/2$.

Question 3 [Law of total variance of the Wright-Fisher model]

Genetic drift does not change the mean but increases the variance of allele frequency. Statistically, the process can be modelled by binomial sampling. Assume p_0 is known, and that N is the constant diploid population size. The allele frequency at the next generation p_1 is of course random, with $E(p_1) = p_0$ and $Var(p_1) = p_0(1 - p_0)/2N$. The latter holds because the allele *count* follows a binomial distribution with probability p_0 and size $2N$ (with variance $2Np_0(1 - p_0)$), and the allele *frequency* is simply the scaled version of it.

Similarly, we also know p_2 but only on the assumption that p_1 is known. It is because WF sampling is an iterative process, that offspring are always sampled from the parental generation. Statistically, we say $Var(p_2|p_1) = p_1(1 - p_1)/2N$. Try to find $Var(p_2)$ without referencing p_1, using the Eve's formula.

In fact the general formula is $Var(p_t) = p_0(1 - p_0)[1 - \left(1 - \frac{1}{2N}\right)^t]$ which can be derived via Mathematical Induction with the same argument (exercise).

Question 4

i. $X_1, X_2, \ldots, X_n$ follow i.i.d. $Exponential(\lambda)$. What is the likelihood function $L(\lambda)$?

ii. Please also find the log-likelihood function $l(\lambda)$.

iii. Find $\lambda = \hat{\lambda}$ such that the log-likelihood function is maximised.

Question 5

i. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $Poisson(\lambda)$. Find the MLE for λ.

ii. If I observed 5, 3, 2, and 6 events (independently), all within a fixed period of time, what would be the best guess for λ?

Question 6

The exercise left to you in class. Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, both μ and σ^2 are not known. Find the MLE for the two parameters.

Question 7 [Linear regression exercise]

It is a spin-off exercise from the original marked-release-recapture experiment for census population size estimation. We measured the difference in their body lengths and how long (in days) they had been hanging around before falling back into our hands. We would like to investigate the relationship between the two variables. There is a short note on the use of `optim()` in today's presentation. You will need the dataset `recapture.csv`