CM224 HW 1 Solution

Ashish Kumar Singh (UID:105479019)

October 14, 2021

Problem 1. GWAS non causal SNP. What is the most likely reason?

Answer 1. iii) The SNP is a tag and simply correlated with the causal SNP.

Solution 1. Since both the cases and controls come from the same population, any ancestry confounders might affect both cases and controls. Also cell-type composition wont create a difference in cases and controls from same population. Genotyping error is possible but again it would affect analysis of both cases and controls, so no difference. Hence, the only option left is that it is a tag region correlated with causal SNP.

Problem 2. 16S RNA data. Number of time bacteria i appears.

Sub-Problem 2a. Most appropriate generative distribution?

Answer 2a. ii) Poisson

Solution 2a. Multinomial is not possible as it is a single bacteria count. Binomial is also not possible as count is simply number of bacteria in one's microbiome and its not similar to coin toss and we also don't know maximum count N. Uniform distribution assumes that counts are equally probable and after a certain count, probability is 0. It seems logical to me that different count would have different probabilities. Hence the only distribution left is Poisson. Moreover since the samples are from different individual they are independent. So, I think Poisson distribution is the most appropriate generative distribution.

Sub-Problem 2b. Maximum Likelihood Estimate of θ?

Answer 2b. i) 7

Solution 2b. Likelihood function for the given n data samples c_i ,

$$L(D, \lambda) = \prod_{i=1}^{n} \lambda^{c_i} \frac{e^{-\lambda}}{c_i!}$$

Log Likelihood function would be,

$$LL(D,\lambda) = \sum_{i=1}^{n} (c_i \log\lambda - \lambda - \log c_i!)$$

Taking derivative wrt λ and setting to zero,

$$\sum_{i=1}^{n} \Big(\frac{c_i}{\lambda} - 1\Big) = 0$$

$$\lambda_{MLE} = \frac{\sum_{i=1}^{n} c_i}{n}$$

So for given data {12,2,0,14,7} the MLE estimate is 7.

Problem 3. $f(x,\lambda,k) = (\frac{k}{\lambda})(\frac{x}{\lambda})^{k-1}e^{-(\frac{x}{\lambda})^k}$. What is the MLE of λ?

Answer 3. i) $\Big(\frac{\sum_i {x_i}^k}{n}\Big)^{\frac{1}{k}}$

Solution 3. Likelihood function for the given data,

$$L(D,\lambda) = \prod_{i=1}^{n} f(x_i,\lambda,k)$$

$$L(D,\lambda) = \prod_{i=1}^{n} (\tfrac{k}{\lambda})(\tfrac{x_i}{\lambda})^{k-1}e^{-(\frac{x_i}{\lambda})^k}$$

Taking log of likelihood,

$$LL(D,\lambda) = \sum_{i=1}^{n} \big(\log k + (k-1)\log x_i - k\log\lambda - (\tfrac{x_i}{\lambda})^k\big)$$

Taking derivative wrt λ and setting to zero,

$$\sum_{i=1}^{n} \Big(-\tfrac{k}{\lambda} + k(\tfrac{{x_i}^k}{\lambda^{k+1}})\Big) = 0$$

$$-\tfrac{nk}{\lambda} + k(\tfrac{\sum_i {x_i}^k}{\lambda^{k+1}}) = 0$$

$$(\tfrac{\sum_i {x_i}^k}{\lambda^k}) = n$$

$$\lambda_{MLE} = \Big(\tfrac{\sum_i {x_i}^k}{n}\Big)^{\frac{1}{k}}$$

Problem 4. Birthday Problem

Sub-Problem 4a. Minimum value of N, such that probability is greater than 0.5?

Answer 4a. iii) 23

Solution 4a. Let $P(A)$ be the probability that two person share the same birthday, $P(A')$ be the probability the that no one shares birthday, since these are mutually exclusive and exhaustive,

$$P(A) + P(A') = 1$$

$P(A')$ can be written as product of prob p1; p2 not matching with p1; p3 not matching with (p1,p2) and so on as birthdays are independent

$$P(A') = \frac{365}{365} * \frac{364}{365} * \frac{363}{365} * ... \frac{365 - N + 1}{365} = \frac{365!}{365^n * (365 - N)!}$$

$$P(A) = 1 - P(A') = 1 - \frac{365!}{365^n * (365 - N)!}$$

Checking value of $P(A)$ with a computer program,
for N=22, $P(A) = 0.475695$
for N=23, $P(A) = 0.507297$

Sub-Problem 4b. $N = 50$, probability of two people same birthday?

Answer 4b. i) 0.97

Solution 4b. Checking value of $P(A)$ with a computer program as in $4a$,
for N=50, $P(A) = 0.970374$