1. A study conducted in the Maternal and Child Health Department asks each of n randomly selected mothers of infants less than 6 months of age a set of k "yes" or "no" questions. Assuming the probability of answering "yes" in each question is denoted by π, the researcher in interested in estimating the probability of "perfect care", which is defined by answering "yes" in all of the k questions, i.e., $\theta = \pi^k$.

   (a) Let X_i denote the number of questions mother i answering "yes" among those k questions. One can assume that X_i follows a binomial distribution with pdf

   $$f(x|\pi) = \left(\begin{array}{c} k \\ x \end{array} \right) \pi^x (1-\pi)^{k-x}.$$

   Assuming $X_1, \ldots, X_n$ is a random sample of size n, derive the maximum likelihood estimator (MLE) of θ.

   Solution: The likelihood function of π can be written as

   $$\begin{aligned} L(\pi|x_1, \ldots, x_n) &= \prod_{i=1}^{n} \left(\begin{array}{c} k \\ x_i \end{array} \right) \pi^{x_i} (1-\pi)^{k-x_i} \\ &= \prod_{i=1}^{n} \left(\begin{array}{c} k \\ x_i \end{array} \right) \pi^{\sum_{i=1}^n x_i} (1-\pi)^{nk - \sum_{i=1}^n x_i}. \end{aligned}$$

   The log-likelihood function can then be written as

   $$\ell(\pi|x_1, \ldots, x_n) \propto \log(\pi) \sum_{i=1}^{n} x_i + \log(1-\pi)(nk - \sum_{i=1}^{n} x_i).$$

   Taking the first derivative, we have

   $$\frac{\partial}{\partial \pi} \ell(\pi|x_1, \ldots, x_n) = \pi^{-1} \sum_{i=1}^{n} x_i - (1-\pi)^{-1}(nk - \sum_{i=1}^{n} x_i).$$

   Setting the first derivative as 0, we can show that the MLE of π is $\hat{\pi} = k^{-1}\bar{X}$, where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Since the second derivative of the log-likelihood function

   $$\frac{\partial^2}{\partial \pi^2} \ell(\pi|x_1, \ldots, x_n) = -\pi^{-2} \sum_{i=1}^{n} x_i - (1-\pi)^{-2}(nk - \sum_{i=1}^{n} x_i) \leq 0,$$

   we can conclude the MLE of π is indeed the maximizer. By invariance property of MLE, the MLE of θ is $\hat{\pi}^k$.

(b) Derive Cramér-Rao Lower Bound (CRLB) of the variance of any unbiased estimator of θ.

Solution: We know that the numerator of CRLB is

$$(\frac{d}{d\pi}\pi^k)^2 = k^2\pi^{2(k-1)},$$

and the denominator can be written as

$$\begin{aligned} -E\left\{\frac{\partial^2}{\partial\pi^2}\ell(\pi|x_1,\ldots,x_n)\right\} &= \pi^{-2}nk\pi + (1-\pi)^{-2}(nk - nk\pi) \\ &= nk\pi^{-1} + nk(1-\pi)^{-1}. \end{aligned}$$

Combing these two results, we have

$$\text{CRLB} = k^2\pi^{2(k-1)}\frac{\pi(1-\pi)}{nk}.$$

(c) Show that $T = \sum_{i=1}^n X_i$ is a complete and sufficient statistic for π (as well as for θ), and that

$$W = \begin{cases} 1 & \text{if } X_1 = k \\ 0 & \text{otherwise,} \end{cases}$$

is an unbiased estimator of θ.

Solution: The statistic $T = \sum_{i=1}^n X_i$ can be shown as a complete and sufficient statistic via exponential family. Since $E(W) = P(X_1 = k) = \pi^k = \theta$. W is an unbiased estimator of θ.

(d) By Lehmann-Sheffe Theorem, the estimator $\phi(T) = E(W|T)$ is the uniformly minimum variance unbiased estimator (UMVUE). Show that

$$\phi\left(\sum_{i=1}^n X_i\right) = \frac{\begin{pmatrix}(n-1)k \\ \sum_{i=1}^n X_i - k\end{pmatrix}}{\begin{pmatrix}nk \\ \sum_{i=1}^n X_i\end{pmatrix}}$$

if $\sum_{i=1}^n X_i = k, k+1, \ldots, nk$, and $\phi(\sum_{i=1}^n X_i) = 0$ otherwise. Here you may use the fact that $\sum_{i=1}^n X_i$ follows a binomial distribution with parameters nk and π, denoted by $\text{Bin}(nk, \pi)$.

Solution: We have

$$\begin{aligned}
\phi(t) = E(W|T=t) &= P\left(X_1 = k | \sum_{i=1}^{n} X_i = t\right) \\
&= \frac{P\left(X_1 = k, \sum_{i=1}^{n} X_i = t\right)}{P\left(\sum_{i=1}^{n} X_i = t\right)} = \frac{P(X_1 = k) P\left(\sum_{i=2}^{n} X_i = t-k\right)}{P\left(\sum_{i=1}^{n} X_i = t\right)} \\
&= \frac{\pi^k \left(\begin{array}{c} (n-1)k \\ t-k \end{array} \right) \pi^{t-k}(1-\pi)^{nk-t}}{\left(\begin{array}{c} nk \\ t \end{array} \right) \pi^t (1-\pi)^{nk-t}} \\
&= \frac{\left(\begin{array}{c} (n-1)k \\ t-k \end{array} \right)}{\left(\begin{array}{c} nk \\ t \end{array} \right)}
\end{aligned}$$

(e) Without using $E(E(W|T)) = E(W)$, show that $\phi(\sum_{i=1}^{n} X_i)$ is indeed an unbiased estimator of θ.

Solution:

$$\begin{aligned}
E\left\{\phi\left(\sum_{i=1}^{n} X_i\right)\right\} &= \sum_{t=k}^{nk} \phi(t) \left(\begin{array}{c} nk \\ t \end{array} \right) \pi^t (1-\pi)^{nk-t} \\
&= \sum_{t=k}^{nk} \left(\begin{array}{c} (n-1)k \\ t-k \end{array} \right) \pi^t (1-\pi)^{nk-t} \\
&= \sum_{t'=0}^{(n-1)k} \left(\begin{array}{c} (n-1)k \\ t' \end{array} \right) \pi^{t'+k} (1-\pi)^{(n-1)k-t'} \\
&= \pi^k \sum_{t'=0}^{(n-1)k} \left(\begin{array}{c} (n-1)k \\ t' \end{array} \right) \pi^{t'} (1-\pi)^{(n-1)k-t'} \\
&= \pi^k,
\end{aligned}$$

which shows $\phi(\sum_{i=1}^{n} X_i)$ is an unbiased estimator of θ.

(f) One may see that the variance of $\phi(\sum_{i=1}^{n} X_i)$ is quite complicated and difficult to derive, which makes the estimator less useful because one cannot make any inference on θ (e.g., does a mother with private insurance more likely have a "prefect care"?) A biostatistician suggests using the MLE in (a) to make inference on θ since the large sample property of the estimator can be easily derived.

Comment on the approach of the biostatistician and derive the large sample distribution of the MLE of θ in (a). [Hint: Use Central Limit Theorem (CLT) and Delta method].

Solution: By Central Limit Theorem, we can have

$$\sqrt{n}(\bar{X} - k\pi) \to_d N(0, k\pi(1-\pi)),$$

which means

$$\sqrt{n}(\hat{\pi} - \pi) \to_d N(0, k^{-1}\pi(1-\pi)).$$

By Delta method, we can conclude the large sample property of the MLE $\hat{\pi}^k$ is

$$\sqrt{n}(\hat{\pi}^k - \pi^k) \to_d N(0, \{g'(\pi)\}^2 k^{-1}\pi(1-\pi)),$$

where $g'(\pi) = k\pi^{k-1}$. Notice that, if we naively move $\sqrt{n}$ to the right hand side, the variance is the same as CRLB. That means, when n is large, the variance of the MLE is very close to the lower bound of the variance of any unbiased estimator. This is the reason why people like maximum likelihood estimator.

2. In malaria study, human-to-mosquito transmission is mediated by sexual stage parasites called gametocytes. *P. falciparum* gametocytes often represent only about 1% of the total parasite load in blood, which makes the detection of the gametocytes difficult. Suppose the amount of the gametocytes detected by a current technology, e.g., Pfs25 RT-PCR, distribute like a location-shifted exponential distribution with pdf

$$f(x|\theta) = \frac{1}{\beta} \exp\left(-\frac{x-\theta}{\beta}\right),$$

where $x > \theta$ and $\theta > 0$ is so-called detection limit. Assuming β is *known*, a researcher is eager to know how low is the amount of gametocytes the current technology can detect.

(a) If the research collects a random sample $X_1, \ldots, X_n$ of size n from an essay, derive the maximum likelihood estimator $\hat{\theta}$ of θ.

Solution: The likelihood function is

$$L(\theta|\boldsymbol{x}) = \left(\frac{1}{\beta}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i - n\theta}{\beta}\right) I(\theta < x_{(1)}) I(x_{(n)} < \infty),$$

where $x_{(1)}$ and $x_{(n)}$ are minimum and maximum order statistics respectively. Since $\exp\{-\beta^{-1}(\sum_{i=1}^n x_i - n\theta)\}$ is a monotone increasing function of θ and maximized at $x_{(1)}$. We can claim that the MLE $\hat{\theta} = X_{(1)}$.

(b) To test a null hypothesis $H_0 : \theta \leq \theta_0$ versus the alternative hypothesis $H_0 : \theta > \theta_0$, one can derive a likelihood ratio test (LRT) statistic $\lambda(\boldsymbol{x})$ and claim that the null hypothesis is rejected if $\lambda(\boldsymbol{x}) \leq c$ for some cutoff c. Derive the LRT statistic $\lambda(\boldsymbol{x})$.

Solution: The LRT is defined by

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= \frac{\sup_{\theta\in\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\theta\in\Theta} L(\theta|\boldsymbol{x})} \\
&= \frac{\sup_{\theta\in\Theta_0} \left(\frac{1}{\beta}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i - n\theta}{\beta}\right) I(\theta < x_{(1)}) I(x_{(n)} < \infty)}{\left(\frac{1}{\beta}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i - n x_{(1)}}{\beta}\right) I(x_{(n)} < \infty)} \\
&= \begin{cases} 1 & \text{if } \; x_{(1)} \leq \theta_0 \\ \exp\{-n(x_{(1)} - \theta_0)/\beta\} & \text{if } \; x_{(1)} > \theta_0. \end{cases}
\end{aligned}$$

(c) One may find the cutoff c is difficult to find since one may not know the distribution of $\lambda(\boldsymbol{X})$. Instead, we can find an equivalent rejection region using the MLE in (a). Show that the equivalent region is $\{\boldsymbol{x}; \hat{\theta} \geq c^*\}$.

Solution: Since $\lambda(\boldsymbol{x})$ is a monotone non-increasing function of $x_{(1)}$, we know the rejection region $\{\boldsymbol{x}; \lambda(\boldsymbol{x}) \leq c\}$ is equivalent to $\{\boldsymbol{x}; x_{(1)} \geq c^*\}$.

(d) Find the cutoff c^* such that the test has a size α, i.e., $\alpha = \sup_{\theta\in\Theta_0} P(\hat{\theta} \geq c^*)$.

Solution: Now, we have

$$\alpha = \sup_{\theta\in\Theta_0} P(X_{(1)} \geq c^*) = \sup_{\theta\in\Theta_0} \exp\{-n(c^* - \theta)/\beta\},$$

since the cdf of $X_{(1)}$ is

$$F_{X_{(1)}}(x) = 1 - P(X_{(1)} > x) = 1 - \{P(X_1 > x)\}^n = 1 - \exp\{-n(x - \theta)/\beta\}.$$

One can see that $\exp\{-n(x - \theta)/\beta\}$ is an increasing function of θ, so the supremum happens in the boundary, $\theta = \theta_0$, which makes

$$\alpha = \exp\{-n(c^* - \theta_0)/\beta\},$$

and

$$c^* = \theta_0 - \beta \log(\alpha)/n.$$