Problem Set 2.

In the population of people who visit nursing homes, the outcome Y_i has density denoted by $\text{pr}(Y_i = y \mid \theta) = f(y, \theta)$, where θ is unknown. To learn about θ we conduct a study, for which assume the following:

1. Originally, a Simple Random Sample of n people, is "eligible" for sampling, so that $Y_1, ..., Y_n$ are iid from the population density $f(y, \theta)$.

2. For each person i in the "eligible" sample, we may observe the data Y_i, in which case we let an indicator $I_i = 1$, or we may not observe the data Y_i, in which case we let $I_i = 0$. Let $\text{pr}(I_i = 1 \mid Y_i = y, \alpha)$ be denoted by $\pi(y, \alpha)$.

Let Y^{obs} denote the vector $(Y_1, ..., Y_n)$ except that Y_i is replaced by NA (for "not available") if $I_i = 0$; let Y^{mis} be the missing outcomes; and let $I = (I_1, ..., I_n)$. Then, the likelihood of the data (Y^{obs}, I) is:

$$\begin{aligned}
\text{pr}(Y^{obs}, I \mid \theta, \alpha) &= \int \text{pr}(Y^{obs}, Y^{mis}, I \mid \theta, \alpha) dY^{mis} \\
&= \int \prod_i \text{pr}(Y_i \mid \theta) \text{pr}(I \mid Y_i, \alpha) dY^{mis} \\
&= \int \prod_i f(Y_i, \theta) \text{pr}(I_i \mid Y_i, \alpha) dY^{mis} \\
&= \prod_{i:I_i=1} f(Y_i, \theta)\pi(Y_i, \alpha) \prod_{i:I_i=0} \int f(Y_i, \theta)(1 - \pi(Y_i, \alpha)) dY_i \qquad (1)
\end{aligned}$$

The probability mechanism $\pi()$ by which some data become missing is called the missing data mechanism (Rubin, 1976).

Questions. Assume that n "eligible" persons are starting their stay to nursing homes in a time window around the present time; assume that our study is actually conducted by visiting a simple random sample of people who right now are at nursing homes; assume that Y_i is the total length that person i has stayed and will stay at the home; and assume that all those we visited now are followed-up and we find out Y_i for these people. The latter sample of Y_i is only a subset of the "eligible persons" and is more likely to include an "eligible" person with a longer than a shorter stay Y_i. To address this phenomenon, known in Biometry as length bias, assume here that the probability, $\pi(Y_i, \alpha)$, of getting an "eligible" Y_i in our study sample is Y_i/α, where α is the maximum length of stay that can occur (i.e., $f(y; \theta)$ =0 for $y > \alpha$).

(i) Using this model, and (1) above, write down the likelihood of the data $D_0 = (Y^{obs}, I_1, ..., I_n)$ in terms of $f()$ and α, simplifying where possible.

(ii) In practice, we do not know the number of "eligible" persons, but we know the number of people, n_1, with $I_i = 1$ in step 2. Suppose we observe Y_i from $n_1 = 500$ people at step 2. Write down the likelihood of the data $\{Y_i : i = 1, ..., n_1\}$ given $\{I_i = 1 : i = 1, ..., n_1\}$ and given $n_1 = 500$.

(iii) Assume that, in the target population of people who go to nursing homes, the length of stay Y is a Gamma random variable with mean θ_1 and variance θ_2. What is the expectation of Y_i given $I_i = 1$?

(iv) Find a minimal sufficient statistic (possible a vector) from the likelihood in (ii) and the assumption in (iii) for the mean θ_1 and variance θ_2.

(v) What would be the likelihood in (ii) using the assumption in (iii), and what would be the minimal sufficient statistic, if we had mistakenly assumed that $\pi(Y_i, \alpha)$ is not a function of Y_i ? Would we end up with the same inference for θ_1 and θ_2 in that case as in the case where we assumed the length-biased $\pi(Y_i, \alpha)$, and why ?