James K. Pringle
Statistical Theory
Dr. Constantine Frangakis
Problem Set 2
March 14, 2013

Problem Set 2

Problems (i) through (v)

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:

$$\text{pr}(Y^{obs}, I \mid \theta, \alpha) = \prod_{i:I_i=1} f(Y_i, \theta)\pi(Y_i, \alpha) \prod_{i:I_i=0} \int f(Y_i, \theta)(1 - \pi(Y_i, \alpha))dY_i \tag{1}$$

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 $Y_i = Y_i/\alpha$, where 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 $D_0 = (Y^{obs}, I_1, \cdots, I_n)$ in terms of $f()$ and α, simplifying where possible.

Proof. From (1), we start calculating

$$\begin{aligned}
\mathrm{pr}(D_0 \mid \theta, \alpha) &= \mathrm{pr}(Y^{obs}, I \mid \theta, \alpha) && (2)\\
&= \prod_{i:I_i=1} f(Y_i, \theta)\pi(Y_i, \alpha) \prod_{i:I_i=0} \int f(Y_i, \theta)(1 - \pi(Y_i, \alpha))dY_i && (3)\\
&= \prod_{i:I_i=1} f(Y_i, \theta)\frac{Y_i}{\alpha} \prod_{i:I_i=0} \int f(Y_i, \theta) - \frac{Y_i}{\alpha} f(Y_i, \theta)dY_i && (4)\\
&= \prod_{i:I_i=1} f(Y_i, \theta)\frac{Y_i}{\alpha} \prod_{i:I_i=0} \left(\int f(Y_i, \theta)dY_i - \int \frac{Y_i}{\alpha} f(Y_i, \theta)dY_i\right) && (5)\\
&= \prod_{i:I_i=1} f(Y_i, \theta)\frac{Y_i}{\alpha} \prod_{i:I_i=0} \left(1 - \frac{1}{\alpha} E_\theta[Y_i]\right) && (6)\\
&= \left(\prod_{i:I_i=1} f(Y_i, \theta)\frac{Y_i}{\alpha}\right) \left(1 - \frac{1}{\alpha} E_\theta[Y_i]\right)^{n_2} && (7)
\end{aligned}$$

Equation (4) follows from the preceding one because $\pi(Y_i, \alpha) = Y_i/\alpha$. In (6), E_θ is the conditional expectation, given θ. The n_2 in (7) is the total number of unobserved people—in other words, $|\{i : I_i = 0\}|$. □

(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 $\{Y_i : i = 1, \cdots, n_1\}$ given $\{I_i = 1 : i = 1, \cdots, n_1\}$ and given $n_1 = 500$.

Proof. The likelihood of the data $\{Y_i : i = 1, \cdots, n_1\}$ given $\{I_i = 1 : i = 1, \cdots, n_1\}$ is

$$\begin{aligned}
\mathrm{pr}(Y_i \mid I_i = 1, \theta, \alpha) &= \frac{\mathrm{pr}(Y_i, I_i = 1 \mid \theta, \alpha)}{\mathrm{pr}(I_i = 1 \mid \theta, \alpha)} && (8)\\
&= \frac{\prod_{i=1}^{500} f(Y_i, \theta)\frac{Y_i}{\alpha}}{\mathrm{pr}(I_i = 1 \mid \alpha)} && (9)
\end{aligned}$$

Equation (8) follows from the definition of conditional probability. The numerator in (9) comes from (7) and the hypothesis that $n_2 = 0$. The denominator in (9) is equal to the denominator in (8) since $\mathrm{pr}(I_i = 1 \mid \alpha)$ does not depend on θ. Now we find by the law of total probability

$$\begin{aligned}
\mathrm{pr}(I_i = 1 \mid \alpha) &= \mathrm{pr}(I_i = 1 \mid \alpha, \theta) && (10)\\
&= \int \mathrm{pr}(I_i = 1 \mid Y_i, \alpha, \theta) f(Y_i, \theta) dY_i && (11)\\
&= \int \pi(y, \alpha) f(Y_i, \theta) dY_i && (12)\\
&= \int \frac{Y_i}{\alpha} f(Y_i, \theta) dY_i && (13)\\
&= E_\theta[Y_i]/\alpha && (14)
\end{aligned}$$

Hence, from (9) and (14) we have by independence of Y_i's and indendence with I_i that

$$\begin{aligned}
\mathrm{pr}(Y_i \mid I_i = 1, \theta, \alpha) &= \frac{\prod_{i=1}^{500} f(Y_i, \theta) \frac{Y_i}{\alpha}}{\prod_{i=1}^{500} E_\theta[Y_i]/\alpha} && (15)\\
&= \prod_{i=1}^{500} \frac{f(Y_i, \theta) Y_i}{E_\theta[Y_i]} && (16)
\end{aligned}$$

This is the likelihood equation we seek. □

(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 $I_i = 1$?

Proof. By (16), we have

$$\mathrm{pr}(Y_i \mid I_i = 1, \theta, \alpha) = f(Y_i, \theta) Y_i / E_\theta[Y_i] \quad (17)$$

So the expectation of (17) is

$$\begin{aligned}
E[Y_i | I_i = 1, \alpha] &= \int \frac{Y_i^2}{E[Y_i]} f(Y_i) dY_i && (18)\\
&= \frac{1}{E[Y_i]} E[Y_i^2] && (19)\\
&= \frac{1}{E[Y_i]} (\mathrm{var}(Y_i) + E[Y_i]^2) && (20)\\
&= \frac{1}{\theta_1} (\theta_2 + \theta_1^2) && (21)\\
&= \frac{\theta_2}{\theta_1} + \theta_1 && (22)
\end{aligned}$$

□

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

Proof. According to Wikipedia the density of the gamma distribution is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{23}$$

with mean α/β and variance α/β^2. Since we assume Y_i has a gamma distribution with mean θ_1 and variance θ_2, we can use the substitutions that $\alpha = \theta_1^2/\theta_2$ and $\beta = \theta_1/\theta_2$. Hence,

$$f(Y_i) = \frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2}}{\Gamma(\theta_1^2/\theta_2)} Y_i^{\theta_1^2/\theta_2 - 1} e^{-Y_i\theta_1/\theta_2} \tag{24}$$

Plugging this in to (16), and noting that under our assumptions, $f(Y_i, \theta) = f(Y_i)$ we have

$$\begin{aligned}
\mathrm{pr}(Y_i \mid I_i = 1, \theta_1, \theta_2) &= \prod_{i=1}^{500} \frac{\frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2}}{\Gamma(\theta_1^2/\theta_2)} Y_i^{\theta_1^2/\theta_2 - 1} e^{-Y_i\theta_1/\theta_2} Y_i}{E[Y_i]} &(25)\\
&= \prod_{i=1}^{500} \frac{\frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2}}{\Gamma(\theta_1^2/\theta_2)} Y_i^{\theta_1^2/\theta_2} e^{-Y_i\theta_1/\theta_2}}{\theta_1} &(26)\\
&= \prod_{i=1}^{500} \frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2} Y_i^{\theta_1^2/\theta_2} e^{-Y_i\theta_1/\theta_2}}{\theta_1\Gamma(\theta_1^2/\theta_2)} &(27)
\end{aligned}$$

From our class definitions, a statistic $T()$ is a minimally sufficient statistic for θ_1, θ_2 if and only if

$$T(x) = T(y) \quad \text{if and only if} \quad \frac{\mathrm{pr}(x \mid \theta_1, \theta_2)}{\mathrm{pr}(y \mid \theta_1, \theta_2)} \quad \text{is free of } \theta \tag{28}$$

So assume we have another random variable family X_i with the same distribution as Y_i, then

$$\begin{aligned}
\frac{\mathrm{pr}(Y_i \mid \theta_1, \theta_2)}{\mathrm{pr}(X \mid \theta_1, \theta_2)} &= \frac{\prod_{i=1}^{500} \frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2} Y_i^{\theta_1^2/\theta_2} e^{-Y_i\theta_1/\theta_2}}{\theta_1\Gamma(\theta_1^2/\theta_2)}}{\prod_{i=1}^{500} \frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2} X_i^{\theta_1^2/\theta_2} e^{-X_i\theta_1/\theta_2}}{\theta_1\Gamma(\theta_1^2/\theta_2)}} &(29)\\
&= \prod_{i=1}^{500} \frac{Y_i^{\theta_1^2/\theta_2} e^{-Y_i\theta_1/\theta_2}}{X_i^{\theta_1^2/\theta_2} e^{-X_i\theta_1/\theta_2}} &(30)\\
&= \prod_{i=1}^{500} \left(\frac{Y_i}{X_i}\right)^{\theta_1^2/\theta_2} e^{(-Y_i+X_i)\theta_1/\theta_2} &(31)\\
&= \left(\prod_{i=1}^{500} \frac{Y_i}{X_i}\right)^{\theta_1^2/\theta_2} e^{(\theta_1/\theta_2)\sum_{i=1}^{500}(-Y_i+X_i)} &(32)
\end{aligned}$$

Thus, if $\prod_{i=1}^{500} Y_i = \prod_{i=1}^{500} X_i$ and $\sum_{i=1}^{500} Y_i = \sum_{i=1}^{500} X_i$, the conditions will be met. Hence $T(\prod_{i=1}^{500} Y_i, \sum_{i=1}^{500} Y_i)$ is a minimally sufficient statistic. □

(v) What would the likelihood in (iii) be 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 where we assumed the length-biased $\pi(Y_i, \alpha)$, and why?

Proof. Now we assume that $pi(Y_i, \alpha) = g(\alpha)$ some function of α. Then (14) becomes

$$\begin{aligned} \mathrm{pr}(I_i = 1 \mid \alpha) &= \int \pi(y, \alpha) f(Y_i, \theta) dY_i && (33) \\ &= \int g(\alpha) f(Y_i, \theta) dY_i && (34) \\ &= g(\alpha) \int f(Y_i, \theta) dY_i && (35) \\ &= g(\alpha) && (36) \end{aligned}$$

Then (16) becomes

$$\begin{aligned} \mathrm{pr}(Y_i \mid I_i = 1, \theta, \alpha) &= \prod_{i=1}^{500} \frac{f(Y_i, \theta) g(\alpha)}{g(\alpha)} && (37) \\ &= \prod_{i=1}^{500} f(Y_i, \theta) && (38) \end{aligned}$$

And that is the likelihood from (ii). Then (18) with (38) become the new (22)

$$\begin{aligned} E[Y_i \mid I_i = 1, \alpha] &= \int Y_i f(Y_i) dY_i && (39) \\ &= E[Y_i] && (40) \\ &= \theta_1 && (41) \end{aligned}$$

□

Since we used (16) to find the answer to (iv) we use (38). That gives

$$\begin{aligned} \frac{\mathrm{pr}(Y_i \mid \theta_1, \theta_2)}{\mathrm{pr}(X \mid \theta_1, \theta_2)} &= \frac{\prod_{i=1}^{500} \frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2} Y_i^{\theta_1^2/\theta_2 - 1} e^{-Y_i \theta_1/\theta_2}}{\Gamma(\theta_1^2/\theta_2)}}{\prod_{i=1}^{500} \frac{(\theta_1/\theta_2)^{\theta_1^2/\theta_2} X_i^{\theta_1^2/\theta_2 - 1} e^{-X_i \theta_1/\theta_2}}{\Gamma(\theta_1^2/\theta_2)}} && (42) \\ &= \prod_{i=1}^{500} \frac{Y_i^{\theta_1^2/\theta_2 - 1} e^{-Y_i \theta_1/\theta_2}}{X_i^{\theta_1^2/\theta_2 - 1} e^{-X_i \theta_1/\theta_2}} && (43) \\ &= \prod_{i=1}^{500} \left(\frac{Y_i}{X_i} \right)^{\theta_1^2/\theta_2 - 1} e^{(-Y_i + X_i)\theta_1/\theta_2} && (44) \\ &= \left(\prod_{i=1}^{500} \frac{Y_i}{X_i} \right)^{\theta_1^2/\theta_2 - 1} e^{(\theta_1/\theta_2) \sum_{i=1}^{500} (-Y_i + X_i)} && (45) \end{aligned}$$

And thus we have the same minimally sufficient statistic, $T(\prod_{i=1}^{500} Y_i, \sum_{i=1}^{500} Y_i)$.

It is not clear what this says about the inference on θ.