Prosper Aime Tchoumo 91.5/100

STAT 520 Midterm 1

This midterm has 2 parts and a total of 10 questions. Please show your work. I will assign partial credit.

1 Part 1

Designing effective mouse traps is more difficult than one may think. If the trap is too slow to close, then the mouse can escape. If the trap does not close hard enough, then the mouse can run away, while still alive and attached to the trap.

A study is conducted to evaluate the effectiveness of three types of mouse traps. Two mice are randomly assigned to each type of trap. The mice are released in a cage containing the trap. The traps are baited with the same amount and type of peanut-butter. The investigator records a binary indicator of whether or not the trap successfully captures and kills the mouse.

10/10 1. Define random variables appropriate for this experiment, along with the associated sample spaces. Let i index the mice
Let Y_i be a random variable associated with a binary indicator of wether or not the trap successfully captures and kills the mouse
Sample space $S_Y = \{0, 1\}$ since the data is binary

10/10 2. What is one reason not to use the normal distribution in this context?
Because the data is binary

10/10 3. Name one way in which the investigator exercised control over the study conditions.
The traps are baited with the same amount and type of peanut-butter

5/10 4. Is the operation of defining random variables part of *scientific or statistical abstraction*?
yes
traps? Which?

2 Part 2

An agronomist is interested in the weights of plants in a region. A random sample of 10 plants is selected. A plant is either diseased or not diseased. The agronomist measures the weights of all plants in the sample. Table 1 displays the data.

Let $i = 1, \ldots, 10$ index the plants in the sample. Let Y_i be a random variable connected with the weight of plant i. Let D_i be a random variable connected with the disease status of the plant. The variable D_i assumes the value of 1 if the plant is diseased, and D_i is zero if the plant is not diseased.

Plant (i)	D_i	Y_i
1	1	2.7
2	1	0.2
3	1	0.6
4	0	1.0
5	0	2.0
6	0	0.8
7	1	4.3
8	0	2.7
9	1	3.4
10	0	0.4

Table 1: Disease status (D_i) and weight (Y_i) of randomly sampled plants.

The agronomist postulates a statistical model for (Y_i, D_i). Let (y_i, d_i) denote a realized value for the pair (Y_i, D_i). The density of the joint distribution of (Y_i, D_i) is given by

$$f(y_i, d_i \mid \theta_1, \theta_2, p) = \left(\frac{1}{\theta_1}\exp(-y_i/\theta_1)\right)^{d_i} p^{d_i} \left(\frac{1}{\theta_2}\exp(-y_i/\theta_2)\right)^{1-d_i} (1-p)^{1-d_i}. \quad (1)$$

Use the density in (1) and the data in Table 1 to answer the following questions.

1. Give a form for the density of the joint distribution of $\{(Y_i, D_i) : i = 1, \ldots, n\}$.

$$f(y_i, d_i \mid \theta_1, \theta_2, p) = \prod_{i=1}^{10} \left[\frac{1}{\theta_1} \exp(-y_i/\theta_1)^{d_i} p^{d_i} \left(\frac{1}{\theta_2} \exp(-y_i/\theta_2) \right)^{1-d_i} (1-p)^{1-d_i} \right].$$

(since they are i.i.d)

2. Give a formula for the maximum likelihood estimator of θ_1.

$$f(y_i, d_i \mid \theta_1, \theta_2, p) = \left(\frac{1}{\theta_1} \exp(-y_i/\theta_1) \right)^{d_i} p^{d_i} \left(\frac{1}{\theta_2} \exp(-y_i/\theta_2) \right)^{1-d_i} (1-p)^{1-d_i}$$

$$L(\theta_1, \theta_2, p) = \prod_{i=1}^{10} \left(\frac{1}{\theta_1} \exp(-y_i/\theta_1) \right)^{d_i} p^{d_i} \left(\frac{1}{\theta_2} \exp(-y_i/\theta_2) \right)^{1-d_i} (1-p)^{1-d_i}$$

$$\Rightarrow \ell(\theta_1, \theta_2, p) = \sum_{i=1}^{10} \left[d_i \log\left(\frac{1}{\theta_1} \exp(-y_i/\theta_1) \right) + d_i \log p + (1-d_i) \log\left(\frac{1}{\theta_2} \exp(-y_i/\theta_2) \right) + (1-d_i) \log(1-p) \right]$$

$$= \sum_{i=1}^{10} \left(d_i(-\log\theta_1 - y_i/\theta_1) + d_i \log p + (1-d_i)(-\log\theta_2 - y_i/\theta_2) + (1-d_i)\log(1-p) \right)$$

$$\frac{\partial \ell(\theta_1, \theta_2, p)}{\partial \theta_1} = \sum_{i=1}^{10} d_i \left(-\frac{1}{\theta_1} + \frac{y_i}{\theta_1^2} \right)$$

$$= -\frac{1}{\theta_1} \sum_{i=1}^{10} d_i + \frac{1}{\theta_1^2} \sum_{i=1}^{10} d_i y_i$$

$$\frac{\partial \ell(\theta_1, \theta_2, p)}{\partial \theta_1}\Big|_{\theta_1 = \hat{\theta}_1} = 0 \iff \frac{1}{\theta_1}\left(-\sum_{i=1}^{10} d_i + \frac{1}{\theta_1} \sum_{i=1}^{10} d_i y_i \right) = 0 \Rightarrow \hat{\theta}_1 = \frac{\sum_{i=1}^{10} d_i y_i}{\sum_{i=1}^{10} d_i}$$

3. Give a formula for the maximum likelihood estimator of $\theta_1 - \theta_2$.

The MLE of $\theta_1 - \theta_2$ is $\hat{\theta}_1 - \hat{\theta}_2$

Now $\frac{\partial \ell(\theta_1, \theta_2, p)}{\partial \theta_2} = \sum_{i=1}^{10} (1-d_i)\left(-\frac{1}{\theta_2} + \frac{y_i}{\theta_2^2} \right)$

$$\frac{\partial \ell(\theta_1, \theta_2, p)}{\partial \theta_2}\Big|_{\theta_2 = \hat{\theta}_2} = 0 \iff \frac{1}{\theta_2} \sum_{i=1}^{10} (1-d_i)\left(-1 + \frac{y_i}{\theta_2} \right) = 0$$

$$\Rightarrow \sum_{i=1}^{10} \left(-1 + \frac{y_i}{\theta_2} + d_i - \frac{d_i y_i}{\theta_2} \right) = 0$$

$$\Rightarrow -10 + \frac{1}{\theta_2}\left(\sum_{i=1}^{10} y_i - \sum_{i=1}^{10} d_i y_i \right) + \sum_{i=1}^{10} d_i = 0$$

$$\Rightarrow \theta_2 = \frac{\sum_{i=1}^{10} (y_i - d_i y_i)}{10 - \sum_{i=1}^{10} d_i}$$

So $\hat{\theta}_1 - \hat{\theta}_2 = \left(\frac{\sum_{i=1}^{10} y_i d_i}{\sum_{i=1}^{10} d_i} \right) - \left(\frac{\sum_{i=1}^{10} y_i - d_i y_i}{10 - \sum_{i=1}^{10} d_i} \right)$

10/10

4. Using the data in Table 1, calculate the maximum likelihood estimate of $\theta_1 - \theta_2$.

$\hat{\theta}_1 = \frac{\sum_{i=1}^{10} y_i d_i}{\sum_{i=1}^{10} d_i}$

$= \frac{(2.7 + 0.2 + 0.6 + 4.3 + 3.4)}{(1+1+1+1+1)} = \frac{11.2}{5} = 2.24$

Now let's compute $\hat{\theta}_2$

$\sum_{i=1}^{10} d_i y_i = 2.7 + 0.2 + 0.6 + 4.3 + 3.4 = 11.2$

So $\hat{\theta}_2 = \frac{\sum_{i=1}^{10} y_i - \sum_{i=1}^{10} d_i y_i}{10 - \sum_{i=1}^{10} d_i}$

$= \frac{18.1 - 11.2}{10 - 5} = 1.38$

So $\hat{\theta}_1 - \hat{\theta}_2 = 2.24 - 1.38 = 0.86$

8.5/10

5. One can define a reduced model by imposing a restriction that $\theta_1 = \theta_2$. Construct a likelihood ratio test of the null hypothesis, $H_0 : \theta_1 = \theta_2$ against the alternative $H_1 : \theta_1 \neq \theta_2$.

$H_0: \theta_1 = \theta_2 = \theta$ (θ, p)

$H_0: \theta_1 = \theta_2 \Leftrightarrow H_0: \theta_1 - \theta_2 = 0$

The loglikelihood of the reduced model is

$l(\theta) = \sum_{i=1}^{10} [d_i(-\log\theta - y_i/\theta) + d_i \log p + (1-d_i)(-\log\theta - y_i/\theta) + (1-d_i)\log(1-p)]$

$= \sum_{i=1}^{10} d_i(-\log\theta - y_i/\theta) + (1-d_i)(-\log\theta - y_i/\theta)$

The test statistic is

$T_n = -2(l(\theta_0) - l(\hat{\theta}))$

The loglikelihood of the full model is $l(\hat{\theta}_1 - \hat{\theta}_2) = \sum_{i=1}^{10} [d_i(-\log(2.24) - \frac{y_i}{2.24}) + d_i \log p + (1-d_i)(-\log(1.38 - \frac{y_i}{1.38})) + (1-d_i)\log(1-p)]$

$= \sum_{i=1}^{10} [d_i(-\log(2.24) - \frac{y_i}{2.24}) + (-d_i)(-\log(1.38 - \frac{y_i}{1.38}))]$

$\chi^2_{1,1-\alpha} = \chi^2_{1,1-0.5}$

$= \chi^2_{1,0.95}$

$= 3.843$

We will reject the null hypothesis.

6. Can you express the density (1) in canonical exponential family form? If so, give an expression for the density (1) in canonical exponential family form. If not, explain why not.

8/10

$$h(y_i, d_i | \theta_1, \theta_2) = \left(\frac{1}{\theta_1}\exp(-y_i/\theta_1)\right)^{d_i} p^{d_i}\left(\frac{1}{\theta_2}\exp(-y_i/\theta_2)\right)^{1-d_i} (1-p)^{1-d_i}$$

$$= \exp\left\{\log\left[\left(\frac{1}{\theta_1}\exp(-\frac{y_i}{\theta_1})\right)^{d_i} p^{d_i}\left(\frac{1}{\theta_2}\exp(-y_i/\theta_2)\right)^{1-d_i} (1-p)^{1-d_i}\right]\right\}$$

$$= \exp\left\{d_i\left(-\log\theta_1 - \frac{y_i}{\theta_1}\right) + d_i\log p + (1-d_i)\left(-\log\theta_2 - \frac{y_i}{\theta_2}\right) + (1-d_i)\log(1-p)\right\}$$

$$= \exp\left\{-d_i\log\theta_1 - \frac{d_i}{\theta_1}y_i + d_i\log p - \log\theta_2 - \frac{y_i}{\theta_2} + d_i\log\theta_2 + \frac{d_i y_i}{\theta_2} + (1-d_i)\log(1-p)\right\}$$

$$= \exp\left\{-\frac{d_i}{\theta_1}y_i + \frac{1}{\theta_2}(-y_i + d_i y_i) - \left(d_i\log\theta_1 + \log\theta_2 - d_i\log\theta_2 - d_i\log p - (1-d_i)\log(1-p)\right)\right\}$$

We only have 2 sufficient statistics instead of 3.