1

Statistics 520 Midterm Exam I

Sept. 22, 2023

Name DEBARSHI CHAKRABORTY ISU ID 247664478

INSTRUCTIONS: Read the questions carefully and completely. Answer each question and show all your work in the space provided. Credit cannot be given if work is not shown. Good luck!

Analysis of Survival Data for Cancer Treatment (25 pointes)

A local hospital is analyzing the effectiveness of two different cancer treatments: A and B. The hospital wants to assess whether the survival rate six months after treatment is significantly different between the two treatments. Treatment is considered effective if a patient survives for at least six months after receiving it.

Data has been collected from 12 patients who received treatment, with survival after 6 months recorded as 1 (Survived) or 0 (Did not survive). 6 patients are randomly selected to receive treatment A, and the other six patients receive treatment B. The data is tabulated below:

	Patient A1	Patient A2	Patient A3	Patient A4	Patient A5	Patient A6
Treatment A	0	1	0	1	1	1

	Patient B1	Patient B2	Patient B3	Patient B4	Patient B5	Patient B6
Treatment B	1	0	1	0	1	0

1. (1pts) What are the measures the hospital took to exercise control over the study conditions?

→ The experiment of all patients conducted in same hospital (same environment).
→ Treatments are given randomly to patients.
→ All 12 participants are affected by cancer.

2. (1pts) Define appropriate random variables and corresponding sample spaces.

$X_i \equiv$ survival status of ith patient receiving treatment A ($i = 1(1)6$)

$Y_j \equiv$ survival status of jth patient receiving trt B ($j = 1(1)6$) | sample space of $X = \Omega_X = \{0, 1\}$ (0: not survived, 1: survived)

$\Omega_Y = \{0,1\}$ (same as X)

3. (1pts) Can we use a normal distribution as a model for the data collected in this study? Explain briefly.

We can not use normal distn as a model for the data collected since the data consists of binary random variables.

26

For the rest of the problems, we assume that the random variables associated with the survival of each treatment are *iid* realizations from a Bernoulli distribution. The probability mass function (PMF) of a Bernoulli distribution is given by

$$P(Y = y; p) = p^y(1-p)^{1-y}, \quad y \in \{0, 1\}; \quad 0 < p < 1. \qquad (1)$$

4. (1pts) Let p_A be the probability of survival under treatment A, and $\ell(p_A)$ be the log likelihood function of p_A based on the data from treatment A. Give a mathematical expression for $\ell(p_A)$.

$$\text{Likelihood} = L(p_A) = p_A^{\sum y_i}(1-p_A)^{n-\sum y_i}$$
$$= p_A^4(1-p_A)^2$$
$$\text{log likelihood} = l(p_A) = 4\log(p_A) + 2\log(1-p_A)$$

$n = 6$
$y_1 = 0$
$y_2 = 1$
$y_3 = 0$
$y_4 = 1$
$y_5 = 1$
$y_6 = 1$

5. (2pts) Let $U(p_A)$ be the score function for p_A. Give a mathematical expression for $U(p_A)$.

$$U(p_A) = \frac{\partial}{\partial p_A} l(p_A)$$
$$= \frac{4}{p_A} + \frac{2}{(1-p_A)}(-1) = \frac{4}{p_A} - \frac{2}{1-p_A}$$

6. (2pts) Derive the maximum likelihood estimate of p_A.

$$U(\hat{p}_A) = 0 \Rightarrow \frac{\cancel{4}\,2}{\hat{p}_A} = \frac{\cancel{2}\,1}{1-\hat{p}_A}$$
$$\Rightarrow 2 - 2\hat{p}_A = \hat{p}_A \Rightarrow 3\hat{p}_A = 2$$
$$\Rightarrow (\hat{p}_A)_{MLE} = 2/3$$

7. (2pts) Give a mathematical expression for the expected information in a random sample of size 6 from the probability density function (1).

$$\cancel{I(p_A) = -\frac{\partial^2}{\partial p_A^2}(l(p_A)) = -\frac{\partial}{\partial p_A}(U(p_A)) = -\frac{\partial}{\partial p_A}\left(\frac{\quad}{p_A} - \right)}$$

$$L(p) = p^y(1-p)^{1-y}$$
$$\Rightarrow l(p) = y\log(p) + (1-y)\log(1-p)$$
$$\Rightarrow \frac{\partial}{\partial p} l(p) = \frac{y}{p} - \frac{1-y}{1-p}$$
$$\Rightarrow -\frac{\partial^2}{\partial p^2} l(p) = +\frac{y}{p^2} + \frac{1-y}{(1-p)^2}$$
$$I(p) = \frac{y}{p^2} + \frac{1-y}{(1-p)^2}$$

we know $E(Y) = p$

$$\Rightarrow E(I(p)) = \frac{1}{p^2}\cdot p + \frac{1}{(1-p)^2}(1-p)$$
$$= \frac{1}{p} + \frac{1}{1-p} = \frac{1-p+p}{p(1-p)}$$

$\therefore E(I(p)) = \frac{1}{p(1-p)}$ | $\therefore E(I_6(p)) = 6 \cdot I(p) = \frac{6}{p(1-p)}$ (ans)

2

$\frac{p}{1-p} = e^{\theta}$

$\frac{p}{1-p} + 1 = e^{\theta} + 1$

$\frac{1}{1-p} = e^{\theta} + 1$

$V(\hat{p}_A) = \frac{1}{n^2} V(\Sigma y_i)$

$= \frac{1}{n^2} \cdot n \cdot V(y_i)$

$= \frac{1}{n} \cdot p(1-p)$

$g(x) = \log\left(\frac{x}{1-x}\right)$

$= \log x - \log(1-x)$

$g'(x) = \frac{1}{x} + \frac{1}{1-x}$

$= \frac{1-x+x}{x(1-x)}$

$= \frac{1}{x(1-x)}$

8. (2pts) Construct a 95% Wald interval for p_A.

$(\hat{p}_A)_{MLE} = 2/3$

$I(p_A) = -\frac{\partial}{\partial p_A}\left(\frac{4}{p_A} - \frac{2}{(1-p_A)}\right)$

$= -\left(-\frac{4}{p_A^2} - \frac{2}{(1-p_A)^2}\right)$

$= \frac{4}{p_A^2} + \frac{2}{(1-p_A)^2}$

$2(1-p)^2 + 2p^2$

$\therefore$ 95% Wald Interval for p_A (asymptotic) where $\alpha = 0.5$

$= \frac{2}{3} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(\frac{4}{(\hat{p}_A)^2} + \frac{2}{(1-\hat{p}_A)^2}\right)^{-1}}$

$= \frac{2}{3} \pm 1.96\sqrt{\left(\frac{4}{4/9} + \frac{2}{1/9}\right)^{-1}}$

$= \frac{2}{3} \pm 1.96\sqrt{(9+18)^{-1}}$

$= \frac{2}{3} \pm 1.96 \times \sqrt{\frac{1}{27}}$

$= \frac{2}{3} \pm 0.38$

$= (0.28, 1.04)$

$\because 0 < p_A < 1$

$\therefore$ The Wald interval is $(0.28, 1)$ (A)

9. (2pts) The log odds of survival is defined as

$$\theta = \log\left(\frac{p_A}{1-p_A}\right).$$

Estimate θ and provide a corresponding standard error for your estimator.

$\theta = g(p_A)$ where $g(x) = \log\left(\frac{x}{1-x}\right)$

$\hat{\theta}_{MLE} = g(\hat{p}_{A MLE})$ (by invariance)

$= \log\left(\frac{\hat{p}_A}{1-\hat{p}_A}\right) = \log\left(\frac{2/3}{1/3}\right) = \log_e(2)$ (A)

$\hat{V}(\hat{\theta}_{MLE}) = (g'(p_A))^2\, \hat{V}(\hat{p}_A)$

$= \left(\frac{1}{\hat{p}_A(1-\hat{p}_A)}\right)^2 \cdot \frac{1}{6}\, \hat{p}_A(1-\hat{p}_A) = \frac{1}{6 \times \frac{2}{3} \times \frac{1}{3}}$

$= \frac{9}{12} = \frac{3}{4}$

$\therefore SE(\hat{\theta}_{MLE}) = \sqrt{\hat{V}(\hat{\theta}_{MLE})} = \frac{\sqrt{3}}{2}$ (A)

10. (2pts) Express the pmf (1) in canonical form. What is the canonical parameter? What statistic is sufficient for the canonical parameter?

$P(Y=y|p) = p^y(1-p)^{1-y}$

$= \exp\{y\log(p) + (1-y)\log(1-p)\}$

$= \exp\{y\log(p) - y\log(1-p) + \log(1-p)\}$

$= \exp\left\{y\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\}$ Let $\theta = \log\frac{p}{1-p}$

$= \exp\{T_1(y)\cdot\theta + B(\theta)\}$

$T_1(y) = y$

$B(\theta) = \log(1-p)$

$= \log(e^{\theta}+1)$

$c(y) = 0$

Canonical parameter $(\theta) = \log\left(\frac{p}{1-p}\right)$

Sufficient statistic for $\theta = y$.

11. (2pts) Use your canonical parametrization to derive an expression for $E[Y]$ as a function of the canonical parameter, where Y is a Bernoulli random variable with pmf (1). Show the steps of your derivation.

We know by the properties of exponential family

$$E(T_1(y)) = \frac{\partial}{\partial\theta} B(\theta)$$

$$\Rightarrow E(Y) = \frac{\partial}{\partial\theta} \log(e^{\theta}+1) = \frac{e^{\theta}}{e^{\theta}+1}$$

$$= \frac{p/1-p}{p/1-p+1} = \frac{p/1-p}{1/1-p}$$

$$= p. \text{ (Ans)}.$$

12. (1pts) Is it true that the probability density function (1) is a member of the natural exponential family? You only need to write "Yes" or "No."

YES (since $T_1(Y) = Y$).

13. (1pts) Is it true that the probability density function (1) is a member of the exponential dispersion family? You only need to write "Yes" or "No."

YES. $\left(\exp\{\phi(y\theta - b(\theta)) + c(y,\phi)\} \right)$
PA $\phi = 1$, $c(y,\phi) = 0$)

14. (1pts) Is it true that the probability density function (1) forms a location/scale family? You only need to write "Yes" or "No."

NO.

-0.5

15. (2pts) Do you think that there is a significant difference in effectiveness between Treatment A and Treatment B? Justify your answer through a likelihood ratio test for the relevant parameter.

$H_0: p_A = p_B$, $H_1: p_A \neq p_B$.

$$L(p_A, p_B) = p_A^{4}(1-p_A)^2 \, p_B^3 (1-p_B)^3$$

Under H_1: $\hat{p}_A = \frac{2}{3}$, $\hat{p}_B = \frac{1}{2}$

Under H_0: $\hat{p}_A = \hat{p}_B = \frac{7}{12}$ $\longrightarrow$ $l(p) = 7\log(p) + 5\log(1-p)$

$$T = -2(\ln(\hat{p}) - \ln(\hat{p}_A, \hat{p}_B)) \xrightarrow{d} \chi^2_{(1)}$$

$$= -2(\{7\log(\tfrac{7}{12}) + 5\log(\tfrac{5}{12})\} - \{4\log(\tfrac{2}{3}) + 2\log(\tfrac{1}{3}) + 3\log(\tfrac{1}{2}) + 3\log(\tfrac{1}{2})\})$$

Reject H_0 if $T_{obs} > \chi^2_{(1)}(1-\alpha)$.

$\frac{4}{6} - \frac{3}{6}$
$= \frac{1}{6}$

16. (2pts) Construct a 95% Wald interval for the difference in proportions of survival between Treatment A and Treatment B.

we want Wald interval for $p_A - p_B$.
p_A and p_B are ind.

$$(\hat{p}_A, \hat{p}_B) \sim N_2\left(\begin{pmatrix} p_A \\ p_B \end{pmatrix}, \Sigma \right)$$

where $\Sigma = \begin{pmatrix} \left(\frac{4}{\hat{p}_A^2} + \frac{2}{(1-\hat{p}_A)^2}\right)^{-1} & 0 \\ 0 & \left(\frac{3}{(\hat{p}_B)^2} + \frac{3}{(1-\hat{p}_B)^2}\right)^{-1} \end{pmatrix}$

$g(x,y) = x - y$

by delta method

$$\therefore \hat{p}_A - \hat{p}_B \sim N\left(p_A - p_B, (1 \;\; 0)\, \Sigma \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)$$

$$\Rightarrow \hat{p}_A - \hat{p}_B \sim N\left(p_A - p_B, \left(\frac{4}{\hat{p}_A^2} + \frac{2}{(1-\hat{p}_A)^2}\right)^{-1} + \left(\frac{3}{(\hat{p}_B)^2} + \frac{3}{(1-\hat{p}_B)^2}\right)^{-1}\right)$$

$\therefore$ Wald C.I. for $p_A - p_B$ is given by

$$(\hat{p}_A - \hat{p}_B) \pm 1.96 \sqrt{\left(\frac{4}{\hat{p}_A^2} + \frac{2}{(1-\hat{p}_A)^2}\right)^{-1} + \left(\frac{3}{(\hat{p}_B)^2} + \frac{3}{(1-\hat{p}_B)^2}\right)^{-1}}$$

$$= \left(\frac{2}{3} - \frac{1}{2}\right) \pm 1.96 \sqrt{\frac{1}{27} + \frac{1}{24}} = \frac{1}{6} \pm 1.96 \times 0.28$$
$$= \frac{1}{6} \pm 0.55 = (-0.38, 0.71)$$

(2)

$$T = -2\left[7\log 7 - 7\log 12 + 5\log 5 - 5\log 12 - 4\log 2 + 4\log 3 + 2\log 3 + 6\log 2\right]$$

$$= -2\left[7\log 7 + 5\log 5 - 12\log 12 + 6\log 3 + 4\log 2\right]$$

$$= 2.42$$

$\chi^2_{(1)}(1-\alpha) = 5.99$

$\because T_{obs} < \chi^2_{(1)}(1-\alpha)$

$\therefore H_0$ is accepted i.e. no significant diff.

$P(\chi^2_{(1)} > x) = 0.05$

$\int_x^\infty \frac{1}{2} e^{-\frac{u}{2}}\, du = 0.05$

$-e^{-\frac{u}{2}}\Big|_x^\infty = 0.05$

$e^{-\frac{x}{2}} = 0.05$

$-\frac{x}{2} = \log(0.05)$

$\Rightarrow x = -2\log(0.05)$

$x = 5.99$

5

+1 1/5

17. (2 Extra Credit) The hospital also recorded the size of the tumor $X_{i,j}$ for treatment i and patient j when the treatment began, and would like to investigate the relationship between the size of the tumor and the probability of survival of the patients. Write a statistical model for this investigation, and discuss what are the relevant parameter(s) to estimate, how to estimate the parameter(s), and how to make statistical inference.

$X_{ij} \equiv$ size of tumor belonging to i^{th} trt and j^{th} patient

$i = 1, 2$

$j = 1, 2, 3, 4, 5, 6 \ldots$

We can use the model

$$E(X_i) = \mu_1(X_{ij}) = \frac{e^{\beta_{01} + \beta_{11} x_{ij}}}{1 + e^{\beta_{01} + \beta_{11} x_{ij}}}$$

$$E(Y_i) = \mu_2(X_{ij}) = \frac{e^{\beta_{02} + \beta_{12} x_{ij}}}{1 + e^{\beta_{02} + \beta_{12} x_{ij}}}$$

β_{11}, β_{12} are the parameters of interest.

We can use logistic regression to estimate the β's.