5. Censored data

In a given population, n individuals are sampled randomly, with replacement, and each sampled individual is asked whether his/her salary is greater than some fixed threshold z. Assume that the salary of a randomly chosen individual has the exponential distribution with unknown parameter λ. Asking whether the salary overcomes a given threshold rathen than directly asking for the salary increases the number people that are willing to answer and decreases the number of mistakes in the collected answers.

Denote by $X_1, \ldots, X_n$ the binary responses of the n sampled individuals, so that $X_i \in \{0, 1\}$. We call the X_i **censored data** .

(a)

2/2 points (graded)
What kind of distribution do the X_i s follow?

- ( ) Exponential distribution with parameter $\mu(\lambda)$
- (•) Bernoulli with parameter $\mu(\lambda)$
- ( ) Poisson with parameter $\mu(\lambda)$

Give the parameter of this distribution in terms of λ and z:

Parameter $\mu(\lambda) =$ exp(-lambda*z) **Answer:** exp(-lambda * z)

$\exp(-\lambda \cdot z)$

Solution:

If $Y_1, \ldots, Y_n$ denote the salaries of the sampled individuals, then

$$Y_i \sim \text{Exp}(\lambda), \quad 1 \le i \le n,$$

and

$$X_i = \mathbf{1}\{Y_i \ge z\}, \quad 1 \le i \le n.$$

Hence, X_i follows a Bernoulli distribution with parameter

= 1 - Fx(z)

$$\mu(\lambda) = p(\lambda) = \mathbb{E}[X_1] = \mathbf{P}_\lambda(Y_i \ge z) = e^{-\lambda z}.$$

提交 你已经尝试了2次（总共可以尝试2次）

Answers are displayed within the problem

(b)

1/1 point (graded)

Let $\overline{X}_n$ be the proportion of sampled individuals whose response was 1 (corresponding to *Yes*). Convince yourself that $\overline{X}_n$ is asymptotically normal.

What is its asymptotic variance?

$V(\overline{X}_n) =$ exp(-lambda*z)*(1 - exp(**Answer:** exp(-lambda * z)*(1-exp(-lambda * z))

$\exp(-\lambda \cdot z) \cdot (1 - \exp(-\lambda \cdot z))$

Solution:

$\overline{X}_n$ is just the sample average and hence asymptotically normal by the Central Limit Theorem. As a Bernoulli variable, the variance of X_i is

$$\mathrm{Var}(X_i) = p(\lambda)(1 - p(\lambda)) = e^{-\lambda z}(1 - e^{-\lambda z}).$$

Hence, we have

$$\sqrt{n}(\overline{X}_n - e^{-\lambda z}) \xrightarrow[n\to\infty]{(D)} \mathcal{N}(0, e^{-\lambda z}(1 - e^{-\lambda z})).$$

提交 你已经尝试了1次（总共可以尝试3次）

Answers are displayed within the problem

(c)

1/1 point (graded)

Find a function f such that $f(\overline{X}_n)$ is a consistent estimator of λ.

Write **barX_n** for the sample average $\overline{X}_n$.

$f(\overline{X}_n) =$ ln(barX_n)/(-z) **Answer:** -ln(barX_n)/z

$\frac{\ln(barX_n)}{-z}$

Solution:

By the Law of Large Numbers, we have

$$\overline{X}_n \xrightarrow[n\to\infty]{\mathbf{P}} \mathbb{E}[X_1] = e^{-\lambda z}.$$

Hence, we can solve for λ with a continuous function,

$$\lambda = -\frac{1}{z}\ln(\mathbb{E}[X_1]),$$

and obtain a consistent estimator by setting

$$f(\overline{X}_n) = -\frac{1}{z}\ln(\overline{X}_n).$$

提交 你已经尝试了1次（总共可以尝试3次）

Answers are displayed within the problem

(d)

1/1 point (graded)

Convince yourself that $f\left(\overline{X}_n\right)$ is asymptotically normal and compute its asymptotic variance.

$V\left(f\left(\overline{X}_n\right)\right) =$ (1 - exp(-lambda*z))/(z^2 **Answer:** (exp(lambda * z) - 1)/z^2

$\frac{1-\exp(-\lambda \cdot z)}{z^2 \cdot \exp(-\lambda \cdot z)}$

Solution:

Use part (b) together with the Delta Method to obtain

$$\sqrt{n}\left(f\left(\overline{X}_n\right) - f\left(e^{-\lambda z}\right)\right) \xrightarrow[n\to\infty]{(D)} \mathcal{N}\left(0, \left(f'\left(e^{-\lambda z}\right)\right)^2 e^{-\lambda z}\left(1 - e^{-\lambda z}\right)\right),$$

with

$$f(u) = -\frac{1}{z}\ln(u).$$

Computing the first derivative yields

$$f'(u) = -\frac{1}{zu}, \quad \text{so } f'\left(e^{-\lambda z}\right) = -\frac{1}{ze^{-\lambda z}}.$$

Plugging this into the above Delta Method formula gives

$$\sqrt{n}\left(f\left(\overline{X}_n\right) - \lambda\right) \xrightarrow[n\to\infty]{(D)} \mathcal{N}\left(0, e^{2\lambda z}e^{-\lambda z}\left(1 - e^{-\lambda z}\right)\frac{1}{z^2}\right),$$

so the asymptotic variance is

$$V\left(f\left(\overline{X}_n\right)\right) = \frac{e^{\lambda z} - 1}{z^2}.$$

提交 你已经尝试了1次（总共可以尝试3次）

Answers are displayed within the problem

(e)

0/1 point (graded)

What equation must z satisfy in order to minimize the asymptotic variance computed in part (d)? Write this equation in the form $g_\lambda(z) = z$, where g_λ is a function that depends on the unknown parameter λ.

$g_\lambda(z) =$ 1.59362/lambda **Answer:** 2*(1 - exp(-lambda * z))/lambda

$\frac{1.59362}{\lambda}$

Solution:

Writing $V(z)$ for the asymptotic variance if the parameter is z, from part (d), we have

$$V(z) = \frac{e^{\lambda z} - 1}{z^2}.$$

Differentiating yields

$$V'(z) = \frac{2 + e^{\lambda z}(-2 + \lambda z)}{z^3}.$$

We solve for stationarity by setting $V'(z) = 0$, which is equivalent to

$$\begin{aligned} 0 = &\ 2 + e^{\lambda z}(-2 + \lambda z) \\ z = &\ \frac{2}{\lambda}(1 - e^{-\lambda z}) = g_\lambda(z). \end{aligned}$$

Since

$$\begin{aligned} \lim_{z \downarrow 0} \frac{e^{\lambda z} - 1}{z^2} &= \infty, \\ \lim_{z \uparrow \infty} \frac{e^{\lambda z} - 1}{z^2} &= \infty, \end{aligned}$$

one of the solutions to

$$z = g_\lambda(z)$$

will have to be the global minimizer of V. In the following, we show that there is only one solution apart from $z = 0$.

Let

$$h(z) = z - \frac{2}{\lambda}(1 - e^{-\lambda z}).$$

Then $h(0) = 0$ and

$$h'(z) = 1 - 2e^{-\lambda z}.$$

The function h' has a unique zero at

$$z^* = -\frac{1}{\lambda}\ln\left(\frac{1}{2}\right).$$

Hence, h is first monotonically decreasing from 0 and then strictly monotonically increasing. That means there can only be a unique crossing point with 0 apart from $z = 0$.

提交 你已经尝试了2次（总共可以尝试3次）

Answers are displayed within the problem

(f)

1/1 point (graded)

Let $Y_1, \ldots, Y_n$ be the salaries of the n sampled people. If one could actually observe $Y_1, \ldots, Y_n$, what would be the Fisher information of Y, $I_Y(\lambda)$, depending on λ?

$I_Y(\lambda) =$ 1/lambda^2

$\frac{1}{\lambda^2}$

Answer: 1/(lambda^2)

Solution:

The likelihood for one sample can be written as

$$L_1(Y_1, \lambda) = \lambda e^{-\lambda Y_1}.$$

That means that the log likelihood for one sample is

$$\ell_1(Y_1, \lambda) = \ln(\lambda) - \lambda Y_1.$$

The second derivative is then given by

$$\frac{\partial^2}{\partial \lambda^2} \ell_1(\lambda) = -\frac{1}{\lambda^2}.$$

and hence

$$I(\lambda) = -\mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} \ell_1(\lambda)\right] = \frac{1}{\lambda^2}.$$

提交 你已经尝试了1次（总共可以尝试3次）

Answers are displayed within the problem

(g)

1/1 point (graded)

In the model where only the X_i's are observed (with fixed threshold z), what is the Fisher information? Denote it by $I_X(\lambda)$.

$I_X(\lambda)$ z^2/(exp(lambda*z) - 1)

$\frac{z^2}{\exp(\lambda \cdot z) - 1}$

Answer: z^2/(exp(lambda * z) - 1)

Solution:

The likelihood for one sample can be written as

$$L_1(X_1, \lambda) = e^{-\lambda z X_1}\left(1 - e^{-\lambda z}\right)^{1 - X_1}$$

That means that the log likelihood for one sample is

$$\ell_1(X_1, \lambda) = -\lambda z X_1 + (1 - X_1)\ln\left(1 - e^{-\lambda z}\right)$$

Its first derivative is

$$\frac{\partial}{\partial \lambda} \ell_1 (X_1, \lambda) = -zX_1 + \frac{ze^{-\lambda z}(1 - X_1)}{1 - e^{-\lambda z}}.$$

The second derivative is then given by

$$\frac{\partial^2}{\partial \lambda^2} \ell_1 (X_1, \lambda) = -\frac{z^2 (1 - X_1) e^{\lambda z}}{(e^{\lambda z} - 1)^2},$$

and hence

$$\begin{aligned} I(\lambda) = & -\mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} \ell_1 (\lambda)\right] \\ = & \frac{(1 - e^{-\lambda z}) z^2}{(e^{\lambda z} - 1)(1 - e^{-\lambda z})} \\ = & \frac{z^2}{(e^{\lambda z} - 1)}. \end{aligned}$$

提交 你已经尝试了3次（总共可以尝试3次）

Answers are displayed within the problem

(h)

2/2 points (graded)

Compare $I_Y(\lambda)$ and $I_X(\lambda)$:

- [x] $I_Y(\lambda) \geq I_X(\lambda)$ for all λ
- [ ] $I_Y(\lambda) \leq I_X(\lambda)$ for all λ
- [ ] $I_Y(\lambda) \geq I_X(\lambda)$ for some λ, $I_Y(\lambda) < I_X(\lambda)$ for others.

How do you interpret this in this model?

- [ ] It depends on the parameter λ whether the censored data or the actual data provides a better estimate.
- [x] The actual data always provides a better estimate
- [ ] The censored data always provides a better estimate.

Solution:

We claim that

$$I_Y(\lambda) \geq I_X(\lambda), \quad \text{for all } \lambda > 0.$$

In order to show this, note that it is enough to show

$$e^u - 1 - u^2 \geq 0, \quad \text{for all } u > 0,$$

by setting $u = \lambda z$.

To see that this is true, repeat the argument from Problem Set 3:

$$\exp(u) - 1 \geq u, \quad \text{for all } u > 0,$$

and since $u^2 < u$ for $u \in (0,1)$, we have

$$\exp(u) - 1 \geq u^2, \quad \text{for } u \in (0,1).$$

Moreover,

$$\exp(1) - 1 = e > 1 = 1^2,$$

and

$$\frac{d}{du}(\exp(u) - 1) = \exp(u), \quad \frac{d}{du}u^2 = 2u,$$

so that

$$\frac{d}{du}(\exp(u) - 1) = \exp(u) \geq 1 + u + \frac{u^2}{2} > 2u = \frac{d}{du}u^2, \quad \text{for all } u > 0,$$

which can be checked by the quadratic formula. This means that for $u \geq 1$,

$$\exp(u) - 1 = e + \int_1^u \exp(t)\ dt > 1 + \int_1^u 2t\,dt = u^2.$$

This means that in terms of asymptotic statistical performance, the actualy observations beat the censored data, which is what we expected. On the other hand, if the actual data is not available (or at a much lower sample size), it might still be better to use the X_i.

提交 你已经尝试了2次（总共可以尝试2次）

Answers are displayed within the problem

讨论

显示讨论

主题： Unit 3 Methods of Estimation:Homework 6 Maximum Likelihood Estimation and Method of Moments / 5. Censored data