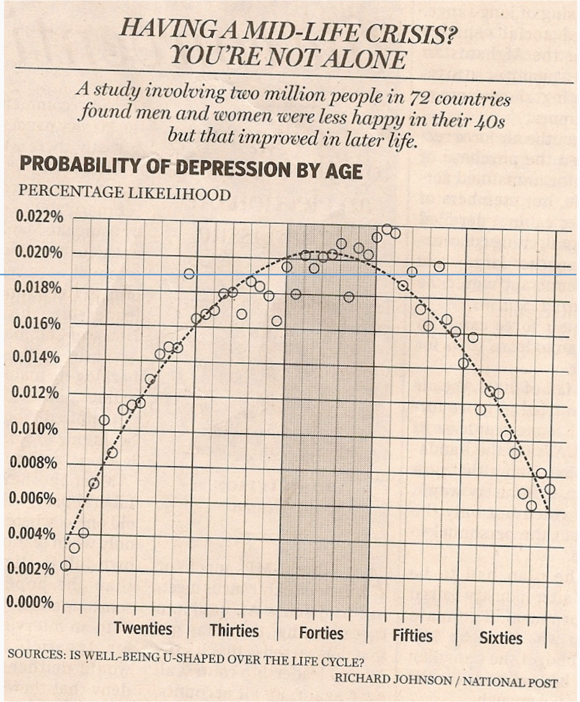
Mathematical Statistics I

STA2212H S LEC9101

Week 1

January 13 2021



Today

1. Course Overview
2. Review of Likelihood AoS Ch 9
3. HW3 from MS I

- January 13 3.15 – 4.15 Ruoqi Yu Zoom Link
  “Matching Methods for Observational Studies Derived from Large Administrative Databases”
- January 18 5.15 – 6.15 “Statistical Learning with Electronic Health Records Data” Jesse Gronsbell

Link

STA 2212S: Mathematical Statistics II
Wednesday, 10-12 am; Friday 10-11 am Eastern January 13 – April 12 2021

From the calendar:

This course is a continuation of STA2112H. It is designed for graduate students in statistics and biostatistics. Topics include: Likelihood inference, Bayesian methods, Significance testing, Linear and generalized linear models, Goodness-of-fit, Computational methods

Prerequisite: STA2112H

I will definitely cover the first 3 topics, and the 5th, and we'll see how time goes for the others. "Computational methods" was probably meant to be shorthand for "bootstrap" and "MCMC", and will be touched on in the other topics.

Course Delivery:

… Course Overview

S STA 2212S: Mathematical Statistics II Link

Syllabus Spring 2021

Week	**Date**	**Methods**	**References**
1	Jan 13/15	Review of parametric inference	AoS Ch 9
2	Jan 20/22	Significance testing	AoS Ch 10.1,2,6,7; SM Ch 7.3
3	Jan 27/29	Significance testing	
4	Feb 3/5	Goodness of fit testing	AoS Ch 10.3,4,5,8
5	Feb 10/12	Multiple testing and FDR	AoS Ch 10.7, EH Ch 15.1,2
6	Feb 17/19	Break	
7	Feb 24/26	Bayesian Inference	AoS Ch 11.1-4; SM Ch 11.1,2; EH

Review of likelihood

AoS §9.3–7

Link

STA2212: Inference and Likelihood

A. Notation

One random variable: Given a model for X which assumes X has a density $f(x;\theta), \quad \theta \in \Theta \subset \mathbb{R}^k$, we have the following definitions:

likelihood function	$L(\theta; y) = c(y) f(y;\theta)$	
log-likelihood function	$\ell(\theta; y) = \log L(\theta; y) = \log f(y;\theta) + a(y)$	
score function	$u(\theta) = \partial \ell(\theta; x)/\partial\theta$	$s(x;\theta)$
observed information function	$j(\theta) = -\partial^2 \ell(\theta; x)/\partial\theta\partial\theta^T$	
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}^1$	$I(\theta)$ (9.11)

Independent observations: When we have X_i independent, identically distributed from $f(x_i;\theta)$, then, denoting the observed sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ we have:

likelihood function	$L(\theta; \boldsymbol{x}) = \prod_{i=1}^n f(x_i;\theta)$	$\mathcal{L}_n(\theta)$

Notes A Notation

X r.v. $f(x;\underline{\theta})$ $\underline{\theta} \in \mathbb{R}^k$ $(\theta_1 \cdots \theta_k)$ $\theta \in \Theta \subseteq \mathbb{R}^k$

par. sp

$L(\theta;x) = f(x;\theta)\,c(x)$

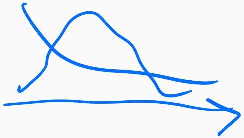
$L(\theta_1;x)/L(\theta_2;x)$ relative

$\ell(\theta;x) = \log f(x;\theta) + a(x)$

$\ell(\theta_1;x) - \ell(\theta_2;x)$

$\ell'(\theta) = \frac{\partial}{\partial\theta}\ell(\theta;\underset{\sim}{x}) = u(\theta) = u(\theta;x)$

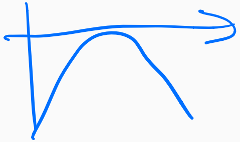
score f=

$f(x)$ 

1 obs^r

$i(\theta) = E_\theta\{u(\theta)u(\theta)^T\}$

$U(\theta) =$ r.v. $\ell'(\theta;X)$

$\ell(\theta)$ 

$I(\theta) = \mathrm{var}_\theta\{U(\theta)\}$

$\theta \in \mathbb{R}^k$ vector U $k \times 1$ vector observed (Fisher) info

$$-\frac{\partial^2 \ell(\theta;x)}{\partial\theta\,\partial\theta^T}_{k\times k} = j(\theta)$$

expected (Fisher) info $i(\theta)$

$k\times k$ matrix

$X_1,\dots,X_n$ iid $f(x;\theta)$

obs $(x_1 \dots x_n) = \underset{\sim}{x}$

$$L(\theta;\underset{\sim}{x}) \overset{iid}{=} \prod_{i=1}^n f(x_i;\theta)$$

$$\ell(\theta;\underset{\sim}{x}) = \sum_{i=1}^n \log f(x_i;\theta)$$

$$\ell'(\theta;\underset{\sim}{X}) = \underset{\sim}{U}(\theta)$$

$k\times 1$

$$= \sum_{i=1}^n U_i(\theta)$$

$$\ell'(\theta;\underset{\sim}{x}) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i;\theta)$$

$$= \sum u_i(\theta) = \sum_{i=1}^n u(\theta;x_i)$$

$\theta \in \mathbb{R}$ $X_1, \dots, X_n$ iid $f(x_i; \theta)$

$$\boxed{\frac{1}{\sqrt{n}} \frac{\sum Z_i}{\sigma_z} \xrightarrow{d} N(0,1)}$$

var $Z_i = \sigma_z^2$

$$\underline{U(\theta)} = \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial \theta} \log f(X_i; \theta)}_{} = \sum U_i(\theta) \quad \text{iid}$$

$$E U(\theta) = 0 \qquad \text{var}(U(\theta)) = \underset{\text{AoS}}{I_n(\theta)} = \underset{\text{AS}}{n I(\theta)} = n i_1(\theta) = i_n(\theta)$$

scalar

$$\frac{1}{\sqrt{n}} U(\theta) I^{-1/2}(\theta) \xrightarrow{d} N(0,1) \qquad \rightsquigarrow$$

under regularity cond's

$$\boxed{(\hat{\theta}_n - \theta) I_n^{1/2}(\theta)} = \frac{U(\theta)}{\{I_n(\theta)\}^{\frac{1}{2}}} + o_p(1)$$

$o_p(1) \xrightarrow{p} 0$ as $n \to \infty$

Approx (Scalar) $\to \underbrace{\sqrt{n}(\hat\theta-\theta)\{I(\hat\theta)\}^{1/2}}_{\text{left has } n} \xrightarrow[n\to\infty]{d} N(0,1)$

no n

" $\hat\theta - \theta \mathrel{\dot\sim} N(0, \underline{I_n^{-1}(\hat\theta)})$ "

$\approx$ ← n fixed

e.g. $\hat\theta \pm 1.96 \cdot I_n^{-1/2}(\hat\theta)$ is approx. 95% CI

$\bar x \pm z_{\alpha/2}\{\hat\sigma/\sqrt{n}\}$

↖ $J_n^{1/2}(\hat\theta)$ bec. $J_n(\theta) \equiv I_n^{-1}(\theta)$

$I_n(\theta) = E(U_1 U_1^T) \cdot n$

Notes B Asymptotics, vector

$$\boxed{\hat{\theta} \pm 1.96\{\hat{j}_n(\hat{\theta})\}^{-1/2}}$$

$*$ $\hat{j}_n(\theta) = -\ell''(\theta)$

obs'd

Notes B Asymptotics, vector

$$(\hat{\theta}-\theta)^T I_n(\theta)(\hat{\theta}-\theta) \xrightarrow[n\to\infty]{d} \chi^2_k$$

(regularity cond's)

$j=1,\dots,k \quad \hat{\theta}_j-\theta_j \;\dot\sim\; N(0,\{j^{-1}(\hat{\theta})\}_{j,j})$

$$\underline{\text{or}}\ \approx N(0, J_n(\hat{\theta})_{j,j})$$

maximum likelihood estimators are asymptotically normally distributed ✓

maximum likelihood estimators are "equivariant"

$(\hat\beta_0, \hat\beta_1, \hat\beta_2)$

If $\tau = g(\theta)$ then $\hat\tau = g(\hat\theta)$

(diff'ble)

$(\hat\beta_1 / \hat\beta_2) \leftarrow$

$(\hat\beta_1 - \hat\beta_2) \leftarrow$

$$\mathrm{var}\,\hat\theta_n = I_n^{-1}(\theta)$$

$$\widehat{\mathrm{var}}_\theta\{g(\hat\theta_n)\} \doteq \boxed{\underbrace{\left(\frac{\partial g}{\partial \theta}\right)^T}_{m\times k}\Big|_{\theta=\hat\theta} \cdot \underbrace{I_n^{-1}(\hat\theta)}_{k\times k} \cdot \underbrace{\left(\frac{\partial g}{\partial \theta}\right)}_{k\times m}\Big|_{\theta=\hat\theta}}$$

Th.9.28

$$\frac{\partial g}{\partial \theta} = \begin{bmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_1} \\ \vdots & & \\ \frac{\partial g_1}{\partial \theta_k} & & \frac{\partial g_m}{\partial \theta_k} \end{bmatrix}_{k\times m}$$

Ex. 9.29

$m \times m$

$g: \mathbb{R}^k \to \mathbb{R}^m$, $m \le k$

~~More about likelihood~~

- ~~maximum likelihood estimators are consistent~~ Th. 9.13

delta method

$\mathbb{R}^1$

Taylor series

$$\mathrm{var}\{g(\hat\theta)\} = \mathrm{var}\left[\underline{g(\theta)} + (\hat\theta-\theta)g'(\theta) + \text{remainder}\right]$$

$$= \mathrm{var}\{(\hat\theta-\theta)g'(\theta) + \text{rem.}\}$$

$$\doteq I_n^{-1}(\hat\theta)\{g'(\hat\theta)\}^2$$

$$\doteq \{j_n(\hat\theta)\}^{-1}\{g'(\hat\theta)\}^2$$

- ~~maximum likelihood estimators are consistent~~ Th. 9.13

3 examples: 1. $X_1, \dots, X_n$ iid $N(\mu, \sigma^2)$ $\theta = (\mu, \sigma^2)$

$g(\theta) = \sigma/\mu$ Ex. 9.20

2. logistic regression X_i's indep.t

$X_i \sim Ber(p_i)$, $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_i$ $\theta = (\beta_0, \beta_1)$

z for which LHS $= 0$ $g(\theta) = -\beta_0/\beta_1 \Leftarrow$ LD50

$g(\hat\theta) = -\hat\beta_0/\hat\beta_1$

- maximum likelihood estimators are consistent

Th. 9.13

*' when I get $(\hat{\theta}-\theta)I_n^{1/2}(\theta) \xrightarrow{d} N(0,1)$

[pf. need $\hat{\theta} \xrightarrow{P} \theta$]

consistency

$L(\theta; x_1,\dots,x_n)$? $\to E$

$\ell(\theta)$

$L(\theta; X_1,\dots X_n)$ ← r.v.

$E\{\ \uparrow\ \}$ is max'd at "true" θ

More about likelihood

- maximum likelihood estimators are consistent Th. 9.13
- maximum likelihood estimators are asymptotically normally distributed Th. 9.18

usually $\hat{\theta}$ def'd

$$l'(\hat{\theta}) = 0$$

(thm that the root of $l'(\theta) = 0$ is consistent)

More about likelihood

- maximum likelihood estimators are consistent — Th. 9.13
- maximum likelihood estimators are asymptotically normally distributed — Th. 9.18
- among all consistent estimators, maximum likelihood estimators have the smallest asymptotic variance — Th. 9.23

More about likelihood

But $k \to \infty$ with n, ✗

asy. normal

- maximum likelihood estimators are consistent ✓ Th. 9.13
- maximum likelihood estimators are asymptotically normally distributed ✓ Th. 9.18
- among all consistent estimators of θ, maximum likelihood estimators have the smallest asymptotic variance Th. 9.23
- i.e., maximum likelihood estimators are asymptotically efficient p. 131

"mle is BAN"

$$\text{a.var}\{\tilde{\theta}_n(\underset{\sim}{X})\} \geq \underline{\underline{I_n^{-1}(\theta)}} \qquad \text{a.var } \hat{\theta}_n(\underset{\sim}{X})$$

$k=1$ ~~$k \to n$~~   mle

where $\tilde{\theta}_n(\underset{\sim}{X})$ is a consistent est. of θ $\text{var}\,\hat{\mu} = \sigma^2/n$

Example: $X_1, \ldots X_n$ $N(\mu, \sigma^2)$ $\hat{\mu} = \bar{X} \to \mu$ (const)

$\boxed{\text{var}\,\tilde{\mu} = \frac{\sigma^2}{n} \times \frac{\pi}{2}}$ $\tilde{\mu} = \text{med}(X_1, \ldots, X_n) \to \mu$

- likelihood functions depend on the data only through the sufficient statistic[1]

forget it for $\underline{\theta}$

- sufficient statistics have all the information about the parameters in the model

for the pars in model

- algebraically

$L(\theta; \underset{\sim}{x}) \propto L(\theta; t) \times f(a \mid t)$

$$f(\boldsymbol{x};\theta) \propto f(t;\theta) f(a \mid t),$$

where $(t, a) \leftrightarrow \boldsymbol{x}$ is a one-to-one transformation of $\boldsymbol{x} = (x_1, \ldots, x_n)$

- examples Poisson, normal, gamma, logistic regression

$|J|$

$\underset{\sim}{x} \overset{1-1}{\longleftrightarrow} (\underset{\sim}{t}, \underset{\sim}{a})$: $T = t(\underset{\sim}{X})$ is suff't if

$f(\underset{\sim}{x}; \theta) = f(t, a; \theta)$

Poisson : $f(x_i; \theta) = \theta^{x_i} e^{-\theta} / x_i!$

$f(\underset{\sim}{x}; \theta) = \theta^{\Sigma x_i} e^{-n\theta} / \Pi(x_i!)$

$L(\theta; \underset{\sim}{x}) = \theta^{\Sigma x_i} e^{-n\theta}$ Σx_i suff't for θ

[1] In fact AoS defines sufficiency this way

 m.s.s. (minimal) not unique, but 1-1

$$\prod_i f(x_i;\alpha,\beta) = \prod_i \left[\frac{1}{\Gamma(\alpha)\beta^{\alpha}} \cdot x_i^{\alpha-1} e^{-x_i/\beta} \right]$$

$\theta = (\alpha, \beta)$

gamma density

$$\log \prod f(x_i;\alpha,\beta)$$

$$\ell(\alpha,\beta) = \underline{\underline{-n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum\log(x_i) - \sum x_i \cdot \frac{1}{\beta}}}$$

$$t = t(x) = \left(\sum\log x_i, \sum x_i\right)$$

$x_i > 0$

1-1 equiv. to $\left(\prod_{i=1}^{n} x_i, \sum x_i\right)$

Exponential families have sufficient statistics

Exponential families are ‘smooth’

Not all families are smooth

Family

$U(0, \theta)$

Calculating maximum likelihood estimators

Newton-Raphson

EM-algorithm

STAT 2112: Homework 3

The homework is worth 60 pts. Problem 2c and Problem 4 are optional and worth up to 15 points of extra credit.

Problem 1

Suppose that the number of accidents per week at an intersection follows a Poisson distribution with parameter μ. Over the past 52 weeks, there were 0 accidents in 30 weeks and one or more accidents in 22 weeks. Assume that the 52 weeks are independent.

(a) 10 pts

Given this data, find the method of moments estimator of μ. Keep in mind that we only know if there were 0 or > 0 accidents for any week.

Profile likelihood

$$f(x_i;\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x_i^{\alpha-1}e^{-x/\beta}$$