AMS 132

7 Feb 17

this time: more about maximum likelihood & Bayesian inference

next time: yet more about Bayes

Fisher's recipe for maximum likelihood (large-sample) inference in "nice" problems (regular)

①

Step 0 with a sampling model for your data set

$\underset{\sim}{y} = (y_1, \ldots, y_n)$ observed data

$\underset{\sim}{Y} = (Y_1, \ldots, Y_n)$ random variables representing the sampling process before it's occurred

Ex. Kaiser Ich Case Study

marginal sampling dist. of obs.

$(Y_i \mid \theta\, \mathcal{B}) \overset{IID}{\sim} \text{Bernoulli}(\theta)$

$(i = 1, \ldots, n)$

$\mathcal{B}$ = background information, assumptions & judgments

i.e., for $y_i = \{0, 1\}$

$P(Y_i = y_i \mid \theta\, \mathcal{B}) = \theta^{y_i} (1-\theta)^{1-y_i}$

Step 1 write out the joint sampling distribution of $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)$

Ex. (Kaiser) By independence (IID), joint dist. of $(Y_1, \dots, Y_n)$ is product of marginals:

$$P(Y_1 = y_1, \overset{\text{and}}{\dots}, Y_n = y_n \mid \theta\, B) \overset{\text{IID}}{=} \prod_{i=1}^{n} P(Y_i = y_i \mid \theta\, B)$$

$$= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{y_1 + \dots + y_n} (1-\theta)^{n-(y_1+\dots+y_n)}$$

$$= \theta^{s} (1-\theta)^{n-s}, \text{ where } s = \sum_{i=1}^{n} y_i.$$

Before data arrives, this is a function of $\underline{y} = (y_1, \dots, y_n)$ for fixed θ.

Step 2 After data arrives, think of it (the joint sampling dist.) as a function of θ for fixed y, and multiply it by an arbitrary positive constant c:

$$\underbrace{\ell(\theta \mid y\, B)}_{\text{likelihood function}} = c\, \underbrace{P(\underline{Y} = y \mid \theta\, B)}_{\text{joint sampling dist.}}$$

$\ell(\theta; y)$

Ex. (Kaiser) $\ell(\theta \mid \underset{\sim}{y}, B) = c\, \theta^{s} (1-\theta)^{n-s}$. (3)

Important point, which I forgot to make earlier

Notice that the likelihood function depends on the data vector $\underset{\sim}{y} = (y_1, \dots, y_n)$ only through the sum $s = \sum_{i=1}^{n} y_i$.

Definition (Fisher)

If $\ell(\theta \mid \underset{\sim}{y})$ depends on $\underset{\sim}{y}$ only through some function $s(\underset{\sim}{y})$, s is said to be a <u>sufficient statistic</u> for θ in the sampling model that gave rise to $\ell(\theta \mid y)$;

<u>shorthand</u>: s is <u>sufficient</u> for θ

Ex. (Kaiser) Some sufficient statistics in the Bernoulli sampling model:

(1) $s_1(\underset{\sim}{y}) = \underset{\sim}{y}$; i.e., the entire data vector is sufficient (Fisher was trying to achieve <u>dimensionality reduction</u>, from n observations down to a smaller number of "pieces of information"; for this purpose $s_1(\underset{\sim}{y}) = \underset{\sim}{y}$ is useless)

---

(2) Suppose that n is even, and define $s_2(\underset{\sim}{y}) = \left(\sum_{i=1}^{n/2} y_i , \sum_{i=\frac{n}{2}+1}^{n} y_i \right)$; this is also sufficient & is a big dimensionality-reduction improvement on $s(\underset{\sim}{y}) = \underset{\sim}{y}$, from n to 2.

---

(3) $s_3(y) = \sum_{i=1}^{n} y_i$ is even better (n down to 1)!

If (in format Definition) no sufficient statistic

exists of dimension lower than the one (s^*, say) you're currently considering, s^* is said to be <u>minimal</u> sufficient.

Ex. (Kaiser) $s = \sum_{i=1}^{n} y_i$ is a minimal suff. stat. in the Bernoulli sampling model

---

Technical note: strictly speaking a minimal sufficient stat. in this model is based on $\left(\sum_{i=1}^{n} y_i, n\right)$, but people usually omit mention of n

---

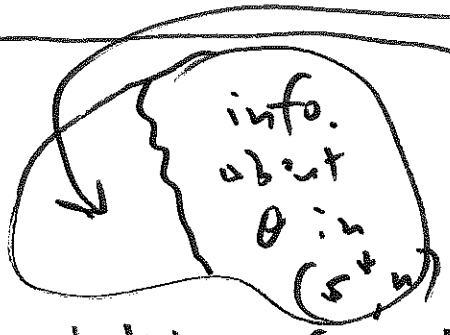
Note: Minimal sufficiency is not unique; $s = \sum_{i=1}^{n} y_i$ is minimal sufficient in Kaiser case study, but so is $s+1$

Generally speaking (but not always),

if the unknown quantity θ in the sampling model is of dimension k & a (minimal) suff. stat. exists, it will also be of dimension k (weird exceptions exist)

Important limitation of the idea of sufficiency

Fisher claimed that, if a (minimal) suff. stat. s^* exists in your sampling model, you can throw y away & just go forward with (s^*, n). This is false:



total information about θ in y

info useful to criticize your sampling dist. (& ∴ your likelihood f'n).

ex. we pretend $(Y_i \mid \theta, B) \overset{IID}{\sim} \text{Bernoulli}(\theta)$ ⑦

$(i = 1, \ldots, n)$

$n = 10$

$s = 5$	1 0 1 1 0 0 1 0 1 0	IID plausible
$s = 5$	0 0 0 0 0 1 1 1 1 1	← IID implausible
$s = 5$	0 1 0 1 0 1 0 1 0 1	

return to Fisher's recipe

recall: step 2: get lik. fn. from joint dist.

(kaiser)

$$\ell(\theta \mid y) = \ell(\theta \mid s, n)$$

$$= \theta^s (1-\theta)^{n-s}$$

fig 2.5 Also compute

log likelihood function:

$$\ell\ell(\theta \mid y) = \ell\ell(\theta \mid s, n)$$

$$= \log \ell(\theta \mid y)$$

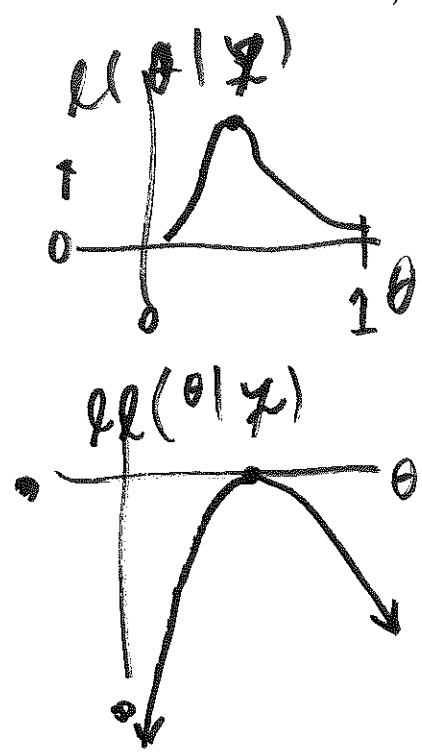
here $\ell\ell(\theta \mid \underline{x}) = \log \theta^{s} (1-\theta)^{n-s}$

$$= (\log \theta^{s}) + (\log (1-\theta)^{n-s})$$

$$= s \log \theta + (n-s) \log (1-\theta)$$

Step 3 | Find θ to Maximize $\ell(\theta \mid s, n)$ or equivalently to maximize $\ell\ell(\theta \mid s\ n)$.

This is usually done by differentiating once with respect to θ, setting first partial derivative (w.r.t. θ) to 0 & solving



Note: Always easier to diff. $\ell\ell(\theta|y)$ because log converts products into sums

lik. fn is always a product

ex. (Kaiser)

$$\ell\ell(\theta|s,n) = s\log\theta + (n-s)\log(1-\theta)$$

$$\frac{d}{d\theta}\ell\ell(\theta|s,n) = \frac{s}{\theta} + (n-s)\left(\frac{1}{1-\theta}\right)(-1)$$

$$= \frac{s}{\theta} - \frac{n-s}{1-\theta} = \frac{s(1-\theta) - (n-s)\theta}{\theta(1-\theta)}$$

$$= \frac{s - \cancel{s\theta} - n\theta + \cancel{s\theta}}{\theta(1-\theta)} = 0 \quad \text{iff} \quad s = n\theta$$

$$\therefore \boxed{\hat{\theta}_{MLE} = \frac{s}{n} = \bar{y}}$$

population sample

$$\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \xrightarrow{IID} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} n = 112$$

mean $\theta = ?$

mean $\bar{y} = \frac{s}{n} = \frac{4}{112} \doteq 0.036$

intuitively sample mean should be good est. of pop. mean

here $\hat{\theta}_{MLE} = \frac{4}{112} \doteq 0.036$ | step 4 (10)

Compute $\hat{SE}(\hat{\theta}_{MLE}) = ?$

$$\hat{SE}_{RS}(\hat{\theta}_{MLE}) = \sqrt{\hat{V}_{RS}(\hat{\theta}_{MLE})}$$

repeated sampling

Note: we already know in kaiser problem that

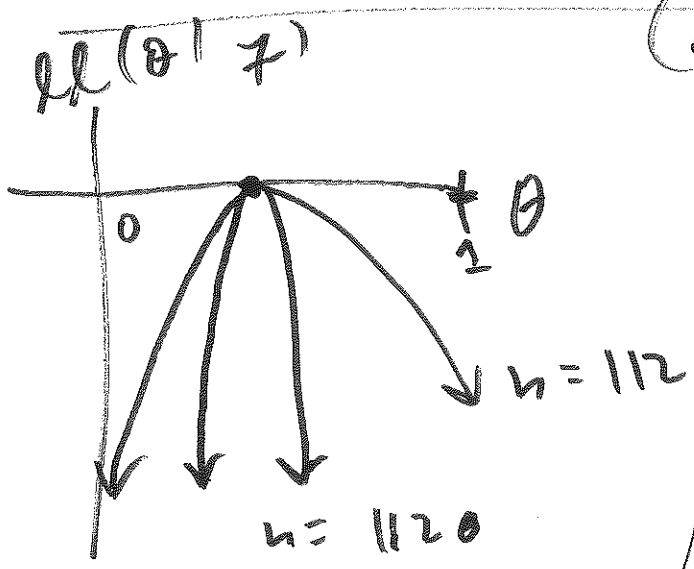
$$\hat{V}_{RS}(\hat{\theta}_{MLE}) = \hat{V}_{RS}(\bar{Y})$$

$$= \frac{\hat{\theta}(1-\hat{\theta})}{n}$$

so

$$\hat{SE}_{RS}(\hat{\theta}_{MLE}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \doteq 0.017$$

Fisher had a general-purpose method for computing $\hat{V}_{RS}(\hat{\theta}_{MLE})$:



(observed) Fisher information

$$\hat{I}(\hat{\theta}_{MLE}) = \left[-\frac{d^2}{d\theta^2} \ell\ell(\theta \mid y) \right]_{\theta = \hat{\theta}_{MLE}}$$

Fisher showed that

$$\hat{V}_{RS}(\hat{\theta}_{MLE}) = \hat{I}^{-1}(\hat{\theta}_{MLE})$$

& $$\hat{SE}_{RS}(\hat{\theta}_{MLE}) = \sqrt{\hat{I}^{-1}(\hat{\theta}_{MLE})}$$

$$\ell\ell(\theta \mid s\ n) = s \log \theta + (n-s) \log(1-\theta)$$

$$\frac{d}{d\theta} \ell\ell(\theta \mid s\ n) = \frac{s}{\theta} - \frac{n-s}{1-\theta}$$

$$\frac{d^2}{d\theta^2} \ell\ell(\theta \mid s\ n) = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}$$

simplify

$$= -\frac{n}{\theta(1-\theta)}$$

So

$$\hat{I}(\hat{\theta}_{MLE}) = \frac{n}{\hat{\theta}_{MLE}(1-\hat{\theta}_{MLE})}$$

$$= \underline{\underline{O}}(n)$$

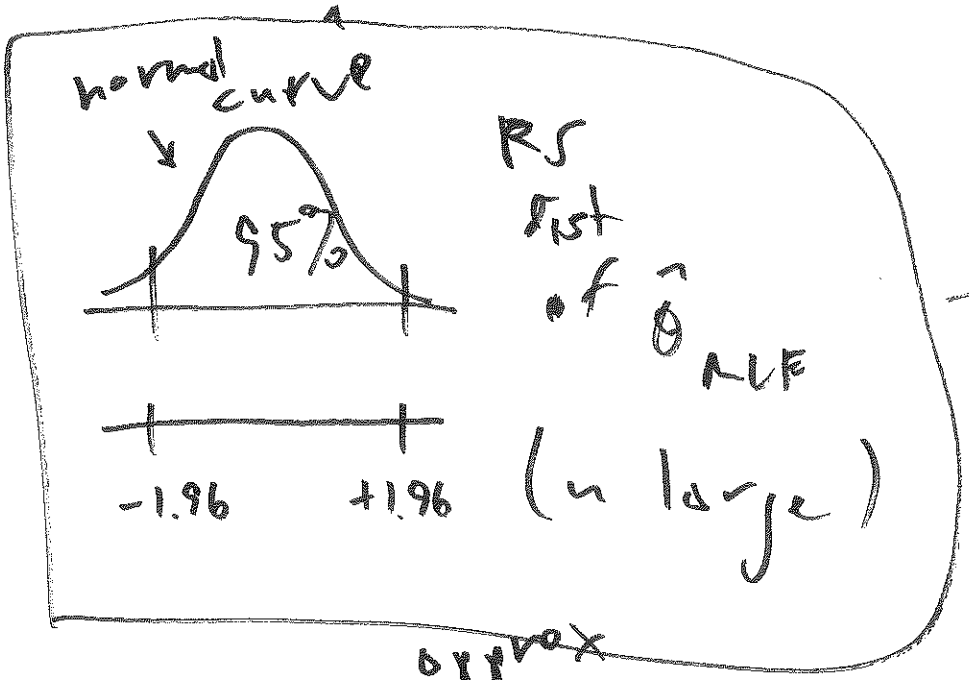
as $n \uparrow$, $\hat{I} \uparrow$ at a rate proportional to n

& $\hat{SE}_{RS}(\hat{\theta}_{MLE})$

$$= \sqrt{\frac{\hat{\theta}_{MLE}(1-\hat{\theta}_{MLE})}{n}}$$

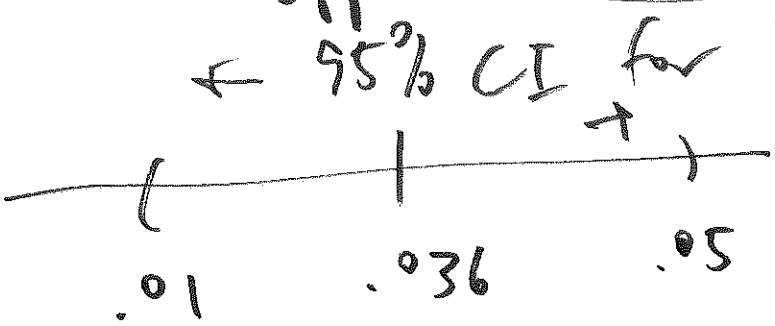
$= 0.017$ for s ive or take for $\hat{\theta}_{MLE}$

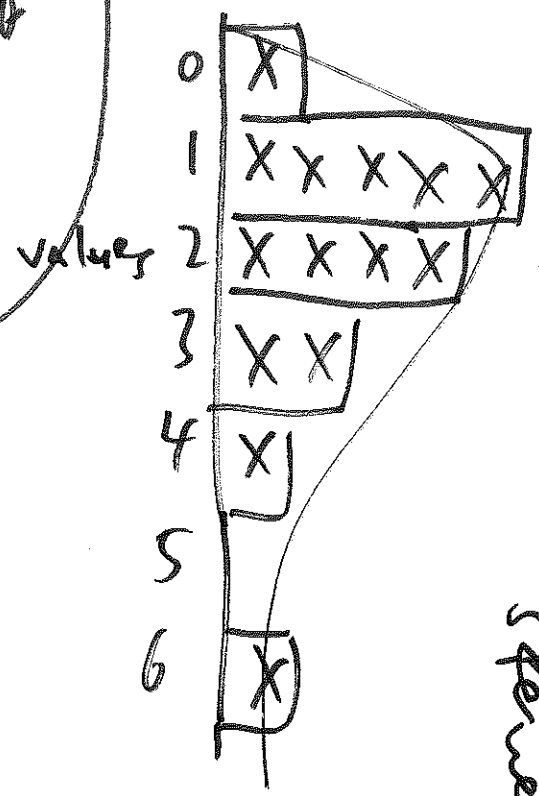
Step 5 | approx. $100(1-\alpha)\%$ confidence (CI) interval for θ: $\alpha = .05 \rightarrow 95\%$ CI



$$\hat{\theta}_{MLE} \pm 1.96\, \widehat{SE}(\hat{\theta}_{MLE})$$

$$.036 \pm 2(.017)$$





1 2 1 1 4 1 2 2 0

3 6 2 1 3

stemed