(The Fisher Information in X about θ at θ_0 is not only the variance of the slope of the log-likelihood at θ_0, but it is also the negative expected curvature of the log-likelihood at θ — the more curved the log-likelihood tends to be at θ_0, the more discriminating power one has to distinguish between θ_0 and θ's near θ_0, the better one's information about θ_0

"Pf" (Outline for the cont's case)

$$l''(\theta) = \frac{d}{d\theta} l'(\theta) = \frac{d}{d\theta}\left(\frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)}\right)$$

$$= \frac{f(x|\theta)\frac{d^2}{d\theta^2} f(x|\theta) - \left(\frac{d}{d\theta} f(x|\theta)\right)^2}{(f(x|\theta))^2}$$

So $$E_{\theta_0} l''(\theta_0) = \int \frac{d^2}{d\theta^2} f(x|\theta)\Big|_{\theta=\theta_0} dx$$

$$- \int \frac{\left(\frac{d}{d\theta} f(x|\theta)\Big|_{\theta=\theta_0}\right)^2}{(f(x|\theta_0))^2} f(x|\theta_0)\, dx$$

$$\overset{\text{AIP}}{=} \frac{d^2}{d\theta^2}\int f(x|\theta)\, dx\Big|_{\theta=\theta_0} - E_{\theta_0}(l'(\theta_0))^2$$

$$= 0 - I(\theta_0)$$

Example $X \sim Bi(n, p)$

$$l''(p) = -\frac{X}{p^2} - \frac{n-X}{(1-p)^2}$$

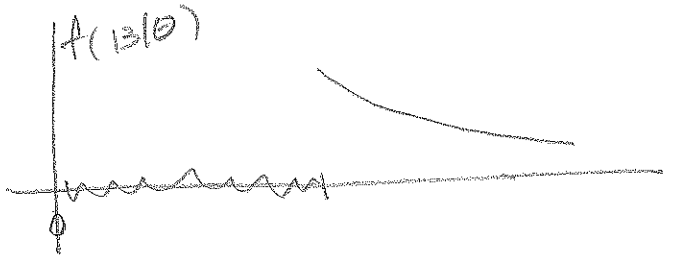
$$E_{p_0} l''(p_0) = -\frac{np_0}{p_0^2} - \frac{n(1-p_0)}{(1-p_0)^2}$$

$$= -n\left(\frac{1}{p_0} + \frac{1}{1-p_0}\right)$$

$$= -\frac{n}{p_0(1-p_0)} = -I(p_0)$$

BTW, regularity conditions in a careful development of this stuff are meant to ensure that for all x the density $f(x|\theta)$ changes smoothly in θ at θ_0

e.g., they outlaw models like $U(0,\theta)$ dsn - for $x = 13$ one has



$f(x|\theta)$ is not positive $\forall \theta$

There is a discontinuity in $f(13|\theta)$ at $\theta = 13$

(so, obviously, $f(13|\theta)$ isn't differentiable at $\theta = 13$)

A second useful property of the FI is that for models of independence, it is additive, i.e. if

$$X = (X_1, X_2, \ldots, X_n)$$

has dsn depending upon θ and $(\forall \theta)$ $X_1, X_2, \ldots, X_n$ are independent

$$I_X(\theta) = \sum_{i=1}^{n} I_{X_i}(\theta)$$

why? Take, for example, the iid case with marginal $f(x|\theta)$

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\ell'(\theta) = \sum_{i=1}^{n} \frac{\frac{d}{d\theta} f(X_i|\theta)}{f(X_i|\theta)}$$

for any θ a sum of iid r.v.'s under any θ_0

$$I_X(\theta_0) = \mathrm{Var}_{\theta_0}\, \ell'(\theta_0) = \sum_{i=1}^{n} \mathrm{Var}_{\theta_0} \frac{\frac{d}{d\theta} f(X_i|\theta)\Big|_{\theta=\theta_0}}{f(X_i|\theta_0)}$$

$$= \sum_{i=1}^{n} I_{X_i}(\theta_0)$$

$$= n\, I_{X_i}(\theta_0)$$

$$= n I_{X_1}(\theta_0)$$

Example $X_1, X_2, \ldots, X_n$ iid $\mathrm{Ber}(p)$

$X = (X_1, \ldots, X_n)$

Clearly $I_X(p) = \frac{n}{p(1-p)}$ and $I_{X_1}(p) = \frac{1}{p(1-p)}$

$$\| \quad n\, I_{X_1}(p)$$

day 12

Another (sensible) fact is that $I_{T(X)}(\theta) \le I_X(\theta)$

i.e. you can't increase FI by taking a function of X — see the "handout" on the web page

Further $I_{T(X)}(\theta) = I_X(\theta) \quad \forall \theta$

$\iff T(X)$ is sufficient for θ

(provided all is well-defined, etc.)

] will be a HW problem

Example X_1, X_2 independent, $X_1 \sim N(\mu, 1)$ variance
$X_2 \sim N(\mu, 4)$

FI in (X_1, X_2) about μ?

$$I_{X_1}(\mu) \overset{+}{\oplus} I_{X_2}(\mu)$$

↑ independence

$$\ln f(x|\mu,\sigma^2) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\frac{\partial}{\partial\mu}(\quad) = \frac{1}{\sigma^2}(x-\mu)$$

$$\frac{\partial^2}{\partial\mu^2}(\quad) = -\frac{1}{\sigma^2}$$

$$\therefore \quad I(\mu) = \frac{1}{\sigma^2}$$

$$I_{X_1}(\mu) + I_{X_2}(\mu) = 1 + \frac{1}{4} = \frac{5}{4}$$

Related to FI is another information measure called Kullback-Leibler information/divergence (for 2 dens P, Q specified by densities/pmfs f, g)

Def If f, g are either pdf's on $\mathbb{R}^k$ or pmf's on some discrete space, The K-L divergence/information (regarding f relative to g) is

$$I(f,g) = E_f\left(\ln\frac{f}{g}(X)\right) = \int \ln\left(\frac{f}{g}(x)\right) f(x)\,dx$$

$$= \sum \ln\left(\frac{f}{g}(x)\right) f(x)$$

Non-negativity of K-L Information

$$I(f,g) = E_f\left(\ln \frac{f(x)}{g(x)}\right) \qquad -\ln(\cdot) \text{ convex}$$

$$= E_f\left(-\ln \frac{g(x)}{f(x)}\right)$$

$$\geq -\ln\left(E_f \frac{g(x)}{f(x)}\right)$$

$$E_f \frac{g(x)}{f(x)} = \int \frac{g(x)}{f(x)} f(x)\,dx = 1$$

$$= -\ln(1) = 0$$

The fact is that $I(f,g) \geq 0$ — It is the (f) average log likelihood ratio (compared to g) — it's plausible that it gives a measure of how effectively X will let us discriminate f from g

Example f normal $(\mu_1, 1)$ g normal $(\mu_2, 1)$

$$\ln\left(\frac{f}{g}\right)(x) = -\frac{1}{2}(x-\mu_1)^2 + \frac{1}{2}(x-\mu_2)^2$$

$$I(f,g) = E_f\left(-\frac{1}{2}(X-\mu_1)^2 + \frac{1}{2}(x-\mu_2)^2\right)$$

$$= -\frac{1}{2}\operatorname{Var}_f X + \frac{1}{2}\left(\operatorname{Var}_f(X) + (\mu_2-\mu_1)^2\right)$$

$$= \frac{1}{2}(\mu_2-\mu_1)^2$$

In this particular example, $I(f,g) = I(g,f)$ — in general $I(f,g)$ is not symmetric, i.e. typically

$$I(g,f) \neq I(f,g)$$

Kullback-Leibler information has an additive property — i.e. if

$$g(x,y) = g_1(x)\, g_2(y)$$
$$f(x,y) = f_1(x)\, f_2(y)$$

$$I(f,g) = I(f_1,g_1) + I(f_2,g_2)$$

and it has a connection to sufficiency just like that of FI

"Result" Suppose that under $\wedge f$, $T(x)$ has a dsn specified by joint density or pmf f^* and under g, $T(x)$ has a dsn specified by joint density or pmf g^*. Then

$$I_X(f,g) \geq I_{T(X)}(f^*, g^*)$$

and there is equality above iff $T(X)$ is sufficient for the 2-dsn model $\mathcal{P} = \{P_f, P_g\}$ for X

In the case of a parametric model where $\theta \in \mathbb{R}^1$, $f = f_\theta$ and $g = f_{\theta_0}$. There is a connection between K-L information and FI

"Result" Under appropriate regularity conditions

$$\frac{d^2}{d\theta^2} I_X(f_{\theta_0}, f_\theta)\Big|_{\theta=\theta_0} = I_X(\theta_0)$$

Why?

$$\frac{d^2}{d\theta^2} I_X(f_{\theta_0}, f_\theta) = \frac{d^2}{d\theta^2} \int \ln \frac{f_{\theta_0}}{f_\theta}(x)\, f_{\theta_0}(x)\, dx$$

$$\overset{\text{AIP}}{=} \int \frac{d^2}{d\theta^2}\left[f_{\theta_0}(x) \ln f_{\theta_0}(x) - f_{\theta_0}(x) \ln f_\theta(x)\right] dx$$

$$= \int \left[0 - f_{\theta_0}(x) \frac{d^2}{d\theta^2} \ln f_\theta(x)\right] dx$$

when you evaluate this at $\theta = \theta_0$ you get $I_X(\theta_0)$

Example f_{μ_0} normal $(\mu_0, 1)$ g normal $(\mu, 1)$

From before

$$I(f,g) = \frac{1}{2}(\mu_0 - \mu)^2$$

$\frac{d^2}{d\mu^2} = 1$ which when evaluated at μ_0 is 1

day 13 and $FI(\mu) = 1$

Begin "Point Estimation"

The basic notion here is to use data X to guess at the value of some function of θ, say $\gamma(\theta) \in \mathbb{R}^k$

Def a statistic $\delta(X)$ taking values in $\mathbb{R}^k$ is a point estimator of $\gamma(\theta)$

Example $X_1, X_2, \ldots, X_n$ iid $N(\mu, \sigma^2)$

$$\delta(X) = (\bar{X}, S^2)$$

is an "obvious" point estimator of $(\mu, \sigma^2) \in \mathbb{R}^2$

$$\delta(X) = \Phi\left(\frac{c - \bar{X}}{S}\right)$$

is an "obvious" point estimator of $\gamma(\mu, \sigma^2) = \Phi\left(\frac{c-\mu}{\sigma}\right)$

Example $X_1, X_2, \ldots, X_n$ iid $Bi(k, p)$ with $P[X_1 < c]$

both k, p to be estimated --- $\delta(X) = ?$

Standard methods of producing estimators that we'll consider are

Method of Moments
Maximum Likelihood ("ML") > generate estimators that subsequently need to be shown to be "good"
Bayes > a principled criterion that with MCMC had computational issues

Method of Moments (For iid models)

This is based on LLN motivation

LLN If $W_1, W_2, \ldots$ are iid P with $E|W_1| < \infty$
Then $\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{P} EW_1$

So, for $X_1, X_2, \ldots$ iid $f(x|\theta)$ if $E_\theta|X|^r < \infty$ and $j \le r$, with

$$M_j(\theta) = EX_1^j$$

can expect

$$\hat{M}_{jn} = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

to approximate $M_j(\theta)$ (typically small) — Then, if for some r $\exists$ $h: \mathbb{R}^r \to \mathbb{R}^k$ s.t.

$$\gamma(\theta) = h(M_1(\theta), M_2(\theta), \ldots, M_r(\theta))$$

then a plausible estimator is

$$\delta_n(X) = h(\hat{M}_{1n}, \hat{M}_{2n}, \ldots, \hat{M}_{rn})$$

(a method-of-moments estimator)

Example $X_1, X_2, \ldots, X_n$ iid $Bi(k,p)$

$$\mu_1(k,p) = kp$$

$$\mu_2(k,p) = kp(1-p) + k^2p^2$$
$$= kp - kp^2 + k^2p^2$$

So $k = \frac{\mu_1}{p}$ and $\mu_2 = \mu_1 - \mu_1 p + \mu_1^2$

$$p = \mu_1 + 1 - \frac{\mu_2}{\mu_1}$$

and then $$k = \frac{\mu_1}{\mu_1 + 1 - \frac{\mu_2}{\mu_1}}$$

This suggests the MOM estimator

$$\delta_n(X) = \left(\frac{\hat{\mu}_{1n}}{\hat{\mu}_{1n} + 1 - \frac{\hat{\mu}_{2n}}{\hat{\mu}_{1n}}}, \hat{\mu}_{1n} + 1 - \frac{\hat{\mu}_{2n}}{\hat{\mu}_{1n}} \right)$$

Thm If $E_\theta |X_1|^r < \infty$ and

$$\gamma(\theta) = h(\mu_1(\theta), \ldots, \mu_r(\theta))$$

for $h: \mathbb{R}^r \to \mathbb{R}^k$ a cont^s function then

$$\delta_n(X) \xrightarrow{P_\theta} \gamma(\theta)$$

This is an important theoretical property called "consistency"

Def If $\{\delta_n(X)\}$ is a sequence of estimators of $\gamma(\theta)$ and

$$\delta_n(X) \xrightarrow{P_{\theta_0}} \gamma(\theta_0)$$

we'll say that the sequence is consistent for $\gamma(\theta)$ at θ_0

Example $X_1, X_2, \ldots, X_n$ iid $Bi(k,p)$

$$h_2(m_1, m_2) = m_1 + 1 - \frac{m_2}{m_1}$$

and $$h_1(m_1, m_2) = \frac{m_1}{h_2(m_1, m_2)}$$

are both cont's on $(0,\infty)^2$ — Since the Binomial dsns have 2nd moments, as long as $p > 0$ and $k \geq 1$

$$\hat{m}_{1n} \xrightarrow{P_{k,p}} m_1(k,p) \text{ and } \hat{m}_{2n} \xrightarrow{P_{k,p}} m_2(k,p)$$

and $\delta_n(X)$ is consistent for (k,p)

day 14

Example "0-inflated Poisson" model

$X_1, X_2, \ldots, X_n$ iid with pmf

$$f(x|\lambda,p) = p\,\frac{e^{-\lambda}\lambda^x}{x!} + (1-p)\,I[x=0]$$

for $x = 0, 1, 2, \ldots$

Another way to think about this is that for

independent $W \sim$ Bernoulli (p)

$Y \sim$ Poisson λ

$X = WY$ has this dsn

For this model

$$m_1 = EX = EWY = p\lambda$$

$$m_2 = EX^2 = EW^2 EY^2$$

$$= p(\lambda + \lambda^2)$$

$$= p\lambda + p\lambda^2$$

$$\frac{\mu_2}{\mu_1} = 1+\lambda \quad \text{i.e.} \quad \lambda = \frac{\mu_2}{\mu_1} - 1 = \frac{\mu_2 - \mu_1}{\mu_1}$$

$$\text{and} \quad p = \frac{\mu_1}{\lambda} = \frac{\mu_1^2}{\mu_2 - \mu_1}$$

and λ and p are cts functions of μ_1, μ_2 so a consistent MOM estimator of (p,λ) is

$$S_n(X) = \left(\frac{\hat{\mu}_{1n}^2}{\hat{\mu}_{2n} - \hat{\mu}_{1n}}, \frac{\hat{\mu}_{2n} - \hat{\mu}_{1n}}{\hat{\mu}_{1n}} \right)$$

$$\| \quad (\hat{p}_n, \hat{\lambda}_n)$$

Example "Truncated Poisson" model

$X_1, X_2, \ldots, X_n$ iid with pmf

$$f(x|\lambda) = \frac{e^{-\lambda}}{1-e^{-\lambda}} \frac{\lambda^x}{x!} \qquad x = 1, 2, \ldots$$

$\mu_1(\lambda) = EX = \frac{\lambda}{1-e^{-\lambda}}$ and a MOM estimator of λ is found by setting

$$\bar{X}_n = \hat{\mu}_{1n} = \frac{\lambda}{1-e^{-\lambda}}$$

There is no obvious explicit formula for $\hat{\lambda}_n$ — we want $h(\bar{X}_n)$ where h is the inverse function for

$$h^*(\lambda) = \frac{\lambda}{1-e^{-\lambda}}$$

Jargon: $\delta_n(X) = \hat{\lambda}_n$ is a solution of the "estimating equation"

$$\bar{X}_n = \frac{\lambda}{1 - e^{-\lambda}}$$

Related alternative description of the method of moments: If

$$\mu_1(\theta) = h_1^*(\theta)$$
$$\mu_2(\theta) = h_2^*(\theta)$$
$$\vdots$$
$$\mu_r(\theta) = h_r^*(\theta)$$

and in an iid model $\hat{\theta}_n$ solves

$$(h_1^*(\theta), \ldots, h_r^*(\theta)) = (\hat{\mu}_{1n}, \ldots, \hat{\mu}_{rn})$$

Then $\hat{\theta}_n$ is a MOM estimate of θ

MOM often produces sensible estimators iid models, but these often lack optimality properties and besides, the method is really not all that general (depending upon making averages)

"Maximum Likelihood" is a more general methodology that often produces estimators that have good properties

Def If $\hat{\theta}$ maximizes a Likelihood $L(\theta)$, it is called a maximum likelihood estimate of θ

Example $X_1, X_2, \ldots, X_n$ iid $U(\theta_1, \theta_2)$

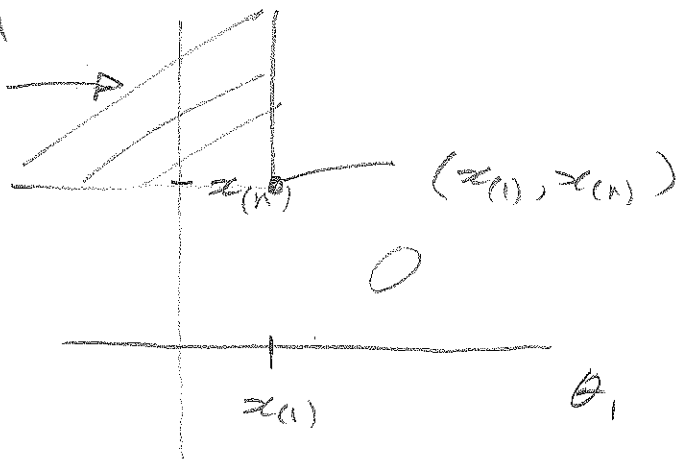
$X = (X_1, X_2, \ldots, X_n)$

joint pdf

$$f(x|\theta) = \begin{cases} \left(\frac{1}{\theta_2-\theta_1}\right)^n & \text{if } \underline{\text{each } x_i \in [\theta_1,\theta_2]} \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_1 \le x_{(1)} \le x_{(n)} \le \theta_2$$

$$L(\theta) = \left(\frac{1}{\theta_2-\theta_1}\right)^n$$



$\left(\frac{1}{(\theta_2-\theta_1)}\right)^n$ increases in θ_1 up to $x_{(1)}$ and decreases in θ_2 down to $x_{(n)}$ ... so

$$\delta_n(X) = (X_{(1)}, X_{(n)})$$

is the MLE of $\theta = (\theta_1, \theta_2)$

Example $X_1, X_2, \ldots, X_n$ iid Poisson(λ)

$X = (X_1, \ldots, X_n)$ — joint pmf is

$$f(x|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \quad \text{on } \{0,1,\ldots,\}^n$$

So

$$\log L(\lambda) = \log f(x|\lambda)$$
$$= -n\lambda + \left(\sum x_i\right)\log\lambda - \sum \log(x_i!)$$

If $\sum x_i = 0$ This is clearly maximized when $\lambda = 0$

If $\sum x_i > 0$

$$\frac{d}{d\lambda} \log L(\lambda) = -n + \frac{\sum x_i}{\lambda}$$

which is 0 if $\lambda = \bar{x}$

< 0 if $\lambda > \bar{x}$

> 0 if $\lambda < \bar{x}$

$\therefore \log L(\lambda)$ and Thus $L(\lambda)$ is maximized at $\lambda = \bar{x}$ i.e. $\delta(X) = \bar{X}$ is the MLE of λ

"The MLE" is a problematic phrase because for a given data set and $f(x|\theta)$

1) There need not be any maximizer of $L(\theta)$
   - in the case that $f(x|\theta)$ is a density, it's not hard at all to generate situations where $L(\theta)$ is unbounded
   - even when $L(\theta)$ is bounded The sup of $L(\theta)$ need not be attainable as a value of $L(\theta)$
2) There can be multiple maximizers of $L(\theta)$

am15