3/5/09 Math 152

Sufficiency

$X_1, \ldots, X_n$ a sample

from $f(x|\theta)$

Is there a statistic

$T(X_1, \ldots, X_n)$

that contains all information in the sample about θ?

Def: $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ

iff

$$f_{X_1 \ldots X_n | T}(x_1, \ldots, x_n | t)$$

does not depend on θ for any value of $T = t$.

i.e. given a value $T = t$
the conditional dist. of $X_1, \ldots, X_n$
contains no further information on θ.

e.g. Bernoulli Trials. (n)

unknown prob success $= \theta$.

$T =$ total # of successes

turns out to be sufficient.

Can we throw away all other info on X_i and just record T.?.

formally: yes.

realistically: no.

e.g. $X_1, \dots X_n$ ind. Bernoulli

$$T = \sum_{i=1}^{n} X_i \qquad P(X_i = 1) = \theta.$$

$$P(X_1 = x_1, \dots \quad X_n = x_n \mid T = t)$$

$$= \frac{P(X_1 = x_1, \dots X_n = x_n, T = t)}{P(T = t)}.$$

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

Given that the numbers of successes is t the probability that they occur on any particular set of t outcomes is the same for any value of θ.

So T is <u>sufficient</u> <u>for</u> <u>θ</u>.

<u>Thm:</u> $T(X_1, \ldots, X_n)$ is sufficient for θ

iff

$$(*)\ f(x_1, \ldots, x_n \mid \theta) = g(T(x_1, \ldots, x_n), \theta) \cdot h(x_1, \ldots x_n).$$

<u>Pf:</u> (Discrete Case).

let $\vec{X} = (X_1, \ldots, X_n)$ $\quad \vec{x} = (x_1, \ldots, x_n)$

and suppose $f(x_1, \ldots x_n \mid \theta)$ factors as in $(*)$

then

$$P(T=t) = \sum_{T(\vec{x})=t} P(\vec{X}=\vec{x}).$$

$$= g(t,\theta) \sum_{T(\vec{x})=t} h(\vec{x})$$

and $\therefore$

$$P(\vec{X}=\vec{x} \mid T=t) = \frac{P(\vec{X}=\vec{x}, T=t)}{P(T=t)}$$

$$= \frac{\cancel{g(t,\theta)}\, h(\vec{x})}{\cancel{g(t,\theta)} \sum_{T(\vec{x})=t} h(\vec{x})}$$

$$= \frac{h(\vec{x})}{\sum_{T(\vec{x})=t} h(\vec{x})}.$$

which does not depend on θ.

<u>Conversely</u>, Suppose $f_{X|T}(x|t)$ has no dependence on θ.

let $g(t,\theta) = P(T=t\,|\,\theta)$

$h(\vec{x}) = P(\vec{X}=\vec{x}\,|\,T=t)$

Then

$$P(\vec{X}=\vec{x}\,|\,\theta) = P(\vec{X}=\vec{x}\,|\,T=t)\,P(T=t\,|\,\theta)$$

$$= g(t,\theta)\,\underbrace{h(\vec{x})}_{\text{ind. of } \theta.}$$

$\blacksquare$

#18 $X_1, \ldots X_n$ i.i.d. on $[0,1]$.

$$f(x|\alpha) = \frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} x^{\alpha-1}(1-x)^{2\alpha-1}$$

$\alpha > 0.$

We know $E(X) = 1/3$

$Var(X) = \dfrac{2}{9(3\alpha+1)}.$

method of moments for α

a) $$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Set $$\hat{\mu}_2 - \left(\frac{1}{3}\right)^2 = \frac{2}{9(3\alpha+1)}$$

and solve for α.

$$\frac{1}{\frac{9}{2}\left(\hat{\mu}_2 - 1/9\right)} = 3\hat{\alpha} + 1$$

$$\hat{\alpha} = \frac{1}{3}\left(\frac{2}{9\hat{\mu}_2 - 1} - 1\right)$$

$$= \frac{2}{27\hat{\mu}_2 - 3} - \frac{1}{3}.$$

b). m.l.e for α.

$$l(\alpha) = \sum_{i=1}^{n}\Big(\log\Gamma(3\alpha) - \log\Gamma(2\alpha) - \log\Gamma(\alpha) + (\alpha-1)\log X_i + (2\alpha-1)\log(1-X_i)\Big)$$

$$\ell'(\alpha) = 3n\,\frac{\Gamma'(3\alpha)}{\Gamma(3\alpha)} - \frac{2n\,\Gamma'(2\alpha)}{\Gamma(2\alpha)} - n\,\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$+ \sum_{i=1}^{n} \log X_i + 2\sum_{i=1}^{n} \log(1-X_i).$$

$$\ell'(\alpha) = 0$$

$$\Longleftrightarrow \quad 3\,\frac{\Gamma'(3\alpha)}{\Gamma(3\alpha)} - 2\,\frac{\Gamma'(2\alpha)}{\Gamma(2\alpha)} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$\underbrace{+ \frac{1}{n}\sum_{i=1}^{n}\log X_i + 2\left(\frac{1}{n}\sum_{i=1}^{n}\log(1-X_i)\right)}_{+\frac{1}{n}\left(\sum_{i=1}^{n}\log\left(X_i(1-X_i)\right)\right)} = 0.$$

c). $\ell''(\alpha) = \ldots$

let $g(x) = \dfrac{\Gamma'(x)}{\Gamma(x)}$.

$$\ell''(\alpha) = 9 \cdot g'(3\alpha) - 4g'(2\alpha) - g'(\alpha)$$

$$-\frac{1}{E(\ell''(\alpha))} = \frac{1}{9g'(3\alpha) - 4g'(2\alpha) - g'(\alpha)}$$

and take $\alpha = \hat{\alpha}$.

d) $f(x_1, \dots, x_n \mid \alpha)$

$$= \prod_{i=1}^{n} \frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} x_i^{\alpha-1} (1-x_i)^{2\alpha-1}$$

$$= \left(\frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} \right)^n \frac{\prod_{i=1}^{n} x_i^{\alpha} (1-x_i)^{2\alpha}}{\prod_{i=1}^{n} x_i (1-x_i)}.$$

$$= \left(\frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} \right)^n g(T(\vec{x}), \alpha) \cdot h(\vec{x}).$$

where $T(\vec{x}) = \prod_{i=1}^{n} x_i (1-x_i)^2.$

$$g(\vec{T}, \alpha) = \vec{T}^{\alpha}$$

$$h(\vec{x}) = \frac{1}{\prod_{i=1}^{n} x_i (1-x_i)}.$$

Some Exercises from Chapter 8.

\#7. $P(X=k) = p(1-p)^{k-1}$ $\quad 0 < p < 1$.

i.i.d sample size n.

a). $E(X) = \frac{1}{p}$

So the method of moments estimate for p is

$$\hat{p} = \frac{1}{\overline{X}}$$

b) $$\text{lik}(p) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n \prod_{i=1}^{n} (1-p)^{x_i-1}$$

$$l(p) = n \log p + \sum_{i=1}^{n} (x_i - 1) \log(1-p).$$

$$l'(p) = \frac{n}{p} + \left(\sum_{i=1}^{n} (x_i - 1)\right)\left(\frac{-1}{1-p}\right)$$

$$= \frac{n}{p} + n(\overline{X} - 1)\left(-\frac{1}{1-p}\right) = 0.$$

$$\Rightarrow \overline{X} - 1 = \frac{1-p}{p} = \frac{1}{p} - 1.$$

$$\Rightarrow \overline{X} = \frac{1}{p} \qquad \text{so} \quad \hat{p} = \frac{1}{\overline{X}}$$

as before.

c) The asymptotic variance of the MLE is

$$-\frac{1}{E(\ell''(p))}.$$

Now,

$$\ell''(p) = -\frac{n}{p^2} - \frac{n(\bar{x}-1)}{(1-p)^2}$$

and ∴

$$-E(\ell''(p)) = \frac{n}{p^2} + \frac{n}{(1-p)^2}\left(\frac{1}{p}-1\right).$$

$$= \frac{n}{p^2} + \frac{n}{(1-p)^2}\left(\frac{1-p}{p}\right).$$

$$= \frac{n}{p^2} + \frac{n}{p(1-p)}. = \frac{(1-p)n + pn}{p^2(1-p)}$$

$$= \frac{n}{p^2(1-p)}.$$

So the asymptotic variance is

$$\frac{p^2(1-p)}{n}.$$

d).

$$f_{p|x}(p|x) = \frac{f_{x|p}(x|p)\, f_p(p)}{\int_0^1 f_{x|p}(x|p)\, f_p(p)\, dp}.$$

Take $f_p(p) = \begin{cases} 1 & 0 \le p \le 1 \\ 0 & \text{else.} \end{cases}$

$$f_{X|p}(x|p) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{n(\bar{X}-1)}$$

So

$$f_{P|X}(p|x) = \frac{p^n (1-p)^{n(\bar{X}-1)}}{\int_0^1 p^n (1-p)^{n(\bar{X}-1)} \, dp}.$$

The Beta density is (pg 58 in Rice).

$$f(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1} \qquad 0 \le u \le 1.$$

$$[\beta(a,b)].$$

and our posterior has this form.

If X is $\sim \beta(a,b)$ (let $C_{a,b} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$)

$$E(X) = \int_0^1 u f(u) \, du$$

$$= C_{a,b} \int_0^1 u^a (1-u)^{b-1} \, du$$

$$= \frac{C_{a,b}}{C_{a+1,b}} = \frac{\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}{\dfrac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)}}.$$

$$= \frac{\Gamma(a+b)}{\Gamma(a+b+1)} \frac{\Gamma(a+1)}{\Gamma(a)}.$$

Since $\Gamma(x+1) = x\Gamma(x)$, this is $\frac{a}{a+b}$.

Our ~~e~~ posterior density is

$$\beta(n+1,\ n(\bar{x}-1)+1)$$

So the posterior mean is.

$$\frac{n+1}{n+1+n(\bar{x}-1)+1} = \frac{n+1}{n\bar{x}+2}$$

$$= \frac{1+\frac{1}{n}}{\bar{x}+\frac{2}{n}}.$$

See problems 2.49, 2.50 and also (47, 48) for information on Γ.

#73 Find a sufficient statistic for the Raleigh distribution.

Suppose X, Y are independent $N(0,\sigma^2)$.

What is the distribution of $R = \sqrt{X^2+Y^2}$?

$$\text{Prob}(R \le r) = \frac{1}{2\pi\sigma^2} \iint_{x^2+y^2 \le r^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}\, dx\, dy.$$

$$= \frac{1}{2\pi\sigma^2} \int_0^{2\pi} \int_0^{r} e^{-\rho^2/2\sigma^2} \rho\, d\rho\, d\theta.$$

$$= \frac{1}{\sigma^2} \int_0^{r} \rho e^{-\rho^2/2\sigma^2}\, d\rho.$$

$$\boxed{u = \rho^2/2\sigma^2 \qquad du = \frac{\rho}{\sigma^2}\, d\rho.}$$

$$= \int_0^{r^2/2\sigma^2} e^{-u}\, du.$$

$$= -e^{-u} \Big|_0^{r^2/2\sigma^2}$$

$$= 1 - e^{-r^2/2\sigma^2}.$$

$\Longrightarrow$

$$f_R(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2} \qquad r \ge 0.$$

For this problem we write

$$f(x/\theta) = \frac{x}{\theta^2} e^{-x^2/2\theta^2} \qquad x \ge 0.$$

Assuming an i.i.d sample of size n

$$lik(\theta) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} e^{-x_i^2/2\theta^2}$$

$$= \frac{1}{\theta^{2n}} \left(\prod_{i=1}^{n} x_i\right) e^{-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2}$$

$$= \left(\prod_{i=1}^{n} x_i\right) \frac{1}{\theta^{2n}} e^{-\frac{1}{2\theta^2}\left(\sum_{i=1}^{n} x_i^2\right)}.$$

$$= h(\vec{x})\, g(\theta, T).$$

$$h(\vec{x}) = \prod_{i=1}^{n} x_i \qquad T = \sum_{i=1}^{n} x_i^2$$

$$g(\theta, T) = \frac{1}{\theta^{2n}} e^{-\frac{1}{2\theta^2}(T)}$$

$T = \sum_{i=1}^{n} x_i^2$ is sufficient for θ.

#45

a) m.l.e. for θ (as above).

$$l(\theta) = -2n \log\theta + \sum_{i=1}^{n} \log x_i - \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2$$

$$l'(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n} x_i^2$$

$\hat{\theta}$ satisfies $\frac{2n}{\hat{\theta}} = \frac{1}{\hat{\theta}^3}\sum_{i=1}^{n} x_i^2$, $\hat{\theta}^2 = \frac{1}{2n}\sum_{i=1}^{n} x_i^2$

$$\hat{\theta} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}.$$

b) method of moments estimate for θ?

$$E(X) = \int_0^\infty \frac{r^2}{\theta^2} e^{-r^2/2\theta^2}\, dr$$

$u = r \qquad dv = \frac{r}{\theta^2} e^{-r^2/2\theta^2} dr$

$du = dr \qquad v = -e^{-r^2/2\theta^2}$

$$-re^{-r^2/2\theta^2}\Big|_0^\infty + \int_0^\infty e^{-r^2/2\theta^2}\, dr$$

$$= \int_0^\infty e^{-r^2/2\theta^2}\, dr = \frac{\sqrt{2\pi}\,\theta}{2} = \left(\sqrt{\frac{\pi}{2}}\right)\theta$$

$$\hat{\theta} = \bar{X}\sqrt{\frac{2}{\pi}}$$

c) approximate variances of m.l.e and method of moments estimates?

$$l''(\theta) = \frac{2n}{\theta^2} - \frac{3}{\theta^4}\sum_{i=1}^{n} X_i^2$$

$$\therefore E[l''(\theta)] = -\frac{2n}{\theta^2} + \frac{3}{\theta^4}\left(\sum_{i=1}^{n} E(X_i^2)\right)$$

$$= -\frac{2n}{\theta^2} + \frac{3n}{\theta^4} E(X^2)$$

where $X \sim f(x|\theta)$. (Raleigh).

$$E(X^2) = \int_0^\infty \frac{r^3}{\theta^2} e^{-r^2/2\theta^2} dr.$$

$u = r^2$ $\quad dv = \frac{r}{\theta^2} e^{-r^2/2\theta^2}$

$du = 2r\,dr$

$v = -e^{-r^2/2\theta^2}$

$$= -r^2 e^{-r^2/2\theta^2}\Big|_0^\infty + 2\int_0^\infty r e^{-r^2/2\theta^2} dr$$

$$= 2\int_0^\infty r e^{-r^2/2\theta^2} dr = 2\theta^2.$$

So $-E[\ell''(\theta)] = -\frac{2n}{\theta^2} + \frac{6n}{\theta^2} = \frac{4n}{\theta^2}$

And the asymptotic variance is $\frac{\theta^2}{4n}$.

Since $\hat{\theta}_{moment} = \bar{X}\sqrt{\frac{2}{\pi}}$

$$Var(\hat{\theta}_m) = \frac{2}{\pi n} Var(X)$$

where X is Raleigh $f(x|\theta)$.

Now $Var(X) = E(X^2) - E(X)^2$

$$= 2\theta^2 - \left(\sqrt{\frac{\pi}{2}}\,\theta\right)^2$$

$$= \left(2 - \frac{\pi}{2}\right)\theta^2.$$

So $Var(\hat{\theta}_m) = \frac{2}{\pi n}\left(2 - \pi/2\right)\theta^2.$

$$= \frac{1}{n}\left(\frac{4}{\pi} - 1\right)\theta^2.$$

See the computer file Chromatin.R for the rest of the problem