Jiapeng Wang

6.13 Suppose X_1 and X_2 are iid observations from the pdf $f(x|\alpha) = \alpha x^{\alpha-1} e^{-x^\alpha}, x > 0,\ \alpha > 0$. Show that $(\log X_1)/(\log X_2)$ is an ancillary statistic.

6.13

Let $Y_1 = \log X_1$ and $Y_2 = \log X_2$. Then we have the pdf of Y_1 and Y_2, $f(y|\alpha) = \alpha \exp(\alpha y - e^{\alpha y})$ (From Thm 2.1.5)

We can also write that $f(y|\alpha) = \frac{1}{1/\alpha} \cdot \{\frac{y}{1/\alpha} - e^{y/1/\alpha}\}$

which means the family of distributions of Y_i is scale family with scale parameter $\frac{1}{\alpha}$. Hence if we let $Z_1 = \alpha Y_1$ $Z_2 = \alpha Y_2$

Then the pdf of Z_1 and Z_2 is $f(z) = \exp(z - e^z)$. Z_1, Z_2 are i.i.d

$\frac{\log X_1}{\log X_2} = \frac{Y_1}{Y_2} = \frac{Z_1}{Z_2}$ which does not depend on α. Thus,

$(\log X_1)/(\log X_2)$ is an ancillary statistic

$\square$

6.16 A famous example in genetic modeling (Tanner, 1996 or Dempster, Laird, and Rubin 1977) is a genetic linkage multinomial model, where we observe the multinomial vector (x_1, x_2, x_3, x_4) with cell probabilities given by $(\frac{1}{2}+\frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4})$.

(a) Show that this is a curved exponential family.

(b) Find a sufficient statistic for θ.

(c) Find a minimal sufficient statistic for θ.

6.16

(a). the pmf of multinomial vector (X_1, X_2, X_3, X_4) is

$$f(\vec{x}|\theta) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!} \cdot \left(\frac{1}{2}+\frac{\theta}{4}\right)^{x_1} \cdot \left(\frac{1}{4}(1-\theta)\right)^{x_2} \cdot \left(\frac{1}{4}(1-\theta)\right)^{x_3} \cdot \left(\frac{\theta}{4}\right)^{x_4}$$

$$= \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!} \left[\exp\left(x_1 \log\left(\frac{1}{2}+\frac{\theta}{4}\right) + (x_2+x_3)\log\left(\frac{1}{4}-\frac{\theta}{4}\right) + x_4 \log\left(\frac{\theta}{4}\right)\right)\right]$$

We suppose $h(x) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!}$, $c(\theta)=1$, $t_1(x) = x_1$, $t_2(x) = x_2+x_3$, $t_3(x) = x_4$, $w_1(\theta) = \log\left(\frac{1}{2}+\frac{\theta}{4}\right)$, $w_2(\theta) = \log\left(\frac{1}{4}-\frac{\theta}{4}\right)$, $w_3(\theta) = \log\left(\frac{\theta}{4}\right)$

which means $k=3$, and the dimension of the parameter is 1

We have $f(\vec{x}|\theta) = h(x)\cdot c(\theta)\cdot \exp\left[\sum_{i=1}^{k} w_i(\theta)\, t_i(x)\right]$, which is a curved expoential family.

(b). We now have $f(\vec{x}|\theta) = h(x)\cdot\exp[\eta(\theta)\cdot T(x)]$, $T(x) = (x_1, (x_2+x_3), x_4)$

$\eta(\theta) = \left(\log\left(\frac{1}{2}+\frac{\theta}{4}\right), \log\left(\frac{1}{4}-\frac{\theta}{4}\right), \log\left(\frac{\theta}{4}\right)\right)$. Thus we have $T(x) = (x_1, (x_2+x_3), x_4)$ is a sufficient statistic for θ. (natural sufficient statistic)

(c). $$\frac{f(x,\theta)}{f(y,\theta)} = \frac{y_1!\,y_2!\,y_3!\,y_4!}{x_1!\,x_2!\,x_3!\,x_4!} \exp\left((x_1-y_1)\log\left(\frac{1}{2}+\frac{\theta}{4}\right) + (x_2+x_3-y_2-y_3)\log\left(\frac{1}{4}-\frac{\theta}{4}\right) + (x_4-y_4)\log\left(\frac{\theta}{4}\right)\right)$$

, Now, we prove $T(x) = (x_1, (x_2+x_3), x_4)$ is a minimal sufficient statistic for θ.

$T(X)=T(Y) \Longleftrightarrow X_1=Y_1,\ X_2+X_3=Y_2+Y_3,\ X_4=Y_4$

$\Longleftrightarrow \dfrac{f(x,\theta)}{f(y,\theta)}=\dfrac{y_1!\,y_2!\,y_3!\,y_4!}{x_1!\,x_2!\,x_3!\,x_4!} \Longleftrightarrow \dfrac{f(x,\theta)}{f(y,\theta)}$ is constant as a function of θ. Hence. $T(X)=(X_1,(X_2+X_3),X_4)$ is a minimal sufficient statistic for θ.

$\square$

6.17 Let $X_1, \ldots, X_n$ be iid with geometric distribution

$$P_\theta(X = x) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots, \quad 0 < \theta < 1.$$

Show that ΣX_i is sufficient for θ, and find the family of distributions of ΣX_i. Is the family complete?

6.17. the pmf of X_i is $f(x|\theta) = \theta(1-\theta)^{x-1} = \frac{\theta}{1-\theta} e^{x \log(1-\theta)}$

Then we suppose $c(\theta) = \frac{\theta}{1-\theta}$, $h(x) = 1$, $t(x) = x$, $w(\theta) = \log(1-\theta)$.

Hence from Thm 6.2.10 we get that ΣX_i is sufficient for θ.

From the definition of geometric distribution, we know that

$\Sigma X_i \sim$ negative binomial (n, θ) And the parameter space Θ contains an open set in $\mathbb{R}^1$ so from Thm 6.2.25 we know the family is complete. $\square$

6.20 For each of the following pdfs let $X_1, \ldots, X_n$ be iid observations. Find a complete sufficient statistic, or show that one does not exist.

(a) $f(x|\theta) = \frac{2x}{\theta^2}, \quad 0 < x < \theta, \quad \theta > 0$

(b) $f(x|\theta) = \frac{\theta}{(1+x)^{1+\theta}}, \quad 0 < x < \infty, \quad \theta > 0$

(c) $f(x|\theta) = \frac{(\log\theta)\theta^x}{\theta - 1}, \quad 0 < x < 1, \quad \theta > 1$

(d) $f(x|\theta) = e^{-(x-\theta)}\exp(-e^{-(x-\theta)}), \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$

(e) $f(x|\theta) = \binom{2}{x}\theta^x(1-\theta)^{2-x}, \quad x = 0, 1, 2, \quad 0 \le \theta \le 1$

(a). $f(\vec{x}|\theta) = \frac{2^n}{\theta^{2n}} x_1 \cdots x_n$

Let $T(X) = \max\{X_i\}$, Then the pdf of $T(X)$ is $g(t|\theta) = \frac{2n}{\theta^{2n}} t^{2n-1}$

Then $f(\vec{x}|\theta) = g(T(\vec{x}), \theta) \cdot h(x)$, $h(x) = \frac{2^{n-1}}{n} \cdot \frac{x_1 \cdots x_n}{\max\{x_i\}^{2n-1}}$

Thus from factorization Thm, $T(X)$ sufficient

Then for a function $p(t)$. if $E_\theta(p(T)) = 0$. we have

$\int_0^\theta p(t) \cdot \frac{2n}{\theta^{2n}} t^{2n-1} dt = 0 \Rightarrow \int_0^\theta p(t) \cdot t^{2n-1} dt = 0.$

We notice that if $p(t) \neq 0$ then $\int_0^\theta t^{2n-1} dt \neq 0$ so $p(t) = 0$

i.e $P_\theta(g(T) = 0) = 1$. Thus, $T(X) = \max\{X_i\}$ is a complete sufficient statistic.

(b). $f(x|\theta) = \theta \cdot \exp(-(1+\theta)\log(1+x))$. which are exponential families

So from Thm 6.2.10 and Thm 6.2.25. We know that $\eta(\theta) = -(1+\theta)$ is full ranked

$\Rightarrow T(X) = \sum_{i=1}^n \log(1+X_i)$ is a complete sufficient statistic

(c). $f(x|\theta) = \frac{(\log\theta)\theta^x}{\theta - 1} = \frac{\log\theta}{\theta - 1} \cdot \exp(x \cdot \log\theta)$, exponential families,

So again from Thm 6.2.10 and Thm 6.2.25. $\eta(\theta) = \log\theta$ is full ranked

which means $T(X) = \sum_{i=1}^n X_i$ is a complete sufficient statistic

(d). $f(x|\theta) = e^{-(x-\theta)} \exp(-e^{-(x-\theta)}) = e^{-x} \cdot e^{\theta} \cdot \exp(-e^{-x} \cdot e^{\theta})$,

which is in the form of full-ranked exponential family, so from Thm 6.2.10 and 6.2.25, we get that

$T(X) = \sum_{i=1}^{n} e^{-X_i}$ is a sufficient statistic for θ.

(e). $f(x|\theta) = \binom{2}{x} \theta^{x} \cdot (1-\theta)^{2-x}$, $x = 0, 1, 2$. $0 \leq \theta \leq 1$

$$= \binom{2}{x} \exp(x \cdot \log\theta + (2-x) \cdot \log(1-\theta))$$

$$= \binom{2}{x} \exp(x(\log\theta - \log(1-\theta)) + 2\log(1-\theta))$$

$$= \binom{2}{x} \cdot (1-\theta)^2 \cdot \exp(x(\log\theta - \log(1-\theta)))$$

So it is exponential family. And we also the parameter space Θ contains an open set in R^k. Thus, from Thm 6.2.10 and Thm 6.2.25, $\sum X_i$ is a complete sufficient statistic $\square$

(b) Find a complete sufficient statistic for θ.

6.23 Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $(\theta, 2\theta)$, $\theta > 0$. Find a minimal sufficient statistic for θ. Is the statistic complete?

From definition 6.2.13. We can consider $\frac{f(x|\theta)}{f(y|\theta)}$

$$= \frac{\frac{1}{\theta^n} \mathbb{1}\{x_{(1)} \le \theta \le \frac{x_{(n)}}{2}\}}{\frac{1}{\theta^n} \mathbb{1}\{y_{(1)} \le \theta \le \frac{y_{(n)}}{2}\}}$$

, which is constant if and only if $x_{(1)} = y_{(1)}$, $x_{(n)} = y_{(n)}$ thus, $(X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for θ. Now we assume the statistic is complete. Then by Basu's thm, $(X_{(1)}, X_{(n)})$ is independent of every ancillary statistic. We already know that the uniform $(\theta, 2\theta)$ family is a scale family, which means $X_{(1)}/X_{(n)}$ is an ancillary statistic. However, $(X_{(1)}, X_{(n)})$ is clearly not independent of $X_{(1)}/X_{(n)}$, so the statistic $(X_{(1)}, X_{(n)})$ is not complete.

$\square$

6.26 Use Theorem 6.6.5 to establish that, given a sample $X_1, \ldots, X_n$, the following statistics are minimal sufficient.

	Statistic	Distribution
(a)	$\bar{X}$	$\mathrm{n}(\theta, 1)$
(b)	$\sum X_i$	$\mathrm{gamma}(\alpha, \beta)$, α known
(c)	$\max X_i$	$\mathrm{uniform}(0, \theta)$
(d)	$X_{(1)}, \ldots, X_{(n)}$	$\mathrm{Cauchy}(\theta, 1)$
(e)	$X_{(1)}, \ldots, X_{(n)}$	$\mathrm{logistic}(\mu, \beta)$

Theorem 6.6.5 (Minimal sufficient statistics) *Suppose that the family of densities $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$ all have common support. Then*

a. *The statistic*

$$T(\mathbf{X}) = \left(\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})}, \frac{f_2(\mathbf{X})}{f_0(\mathbf{X})}, \ldots, \frac{f_k(\mathbf{X})}{f_0(\mathbf{X})} \right)$$

is minimal sufficient for the family $\{f_0(\mathbf{x}), \ldots, f_k(\mathbf{x})\}$.

b. *If $\mathcal{F}$ is a family of densities with common support, and*

(i) *$f_i(\mathbf{x}) \in \mathcal{F}$, $i = 0, 1, \ldots, k$,*
(ii) *$T(\mathbf{x})$ is sufficient for $\mathcal{F}$,*

then $T(\mathbf{x})$ is minimal sufficient for $\mathcal{F}$.

6.26. (a). First, we can write the pdf of $n(\theta, 1)$.

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} = h(x) \cdot c(\theta) \cdot \exp x\theta$$

$h(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}$, $c(\theta) = e^{-\frac{\theta^2}{2}}$. Hence, from Thm 6.2.10,

$T(X) = \sum_{i=1}^{n} X_i$ is sufficient, i.e. $\bar{X}$ is sufficient.

for any θ $f(x|\theta)$ has common support. then we can let $\mathcal{F} = \{ f(x|\theta) \}$ and $\bar{X}$ is sufficient for $\mathcal{F}$

thus, $T(x)$ is minimal sufficient.

(b). the pdf of $\mathrm{gamma}(\alpha, \beta)$ is $f(x|\beta) = \dfrac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} = \Gamma(\alpha) \cdot x^{\alpha-1} \cdot \beta^\alpha \cdot e^{-\beta x}$

The from Thm 6.2.10 $T(X) = \sum X_i$ is sufficient.

$\dfrac{f(\vec{x}|\beta)}{f(\vec{y}|\beta)} \propto \exp\left(-\beta\left(\sum_{i=1}^{n} X_i - Y_i\right)\right)$ is constant as a function of β if and only if $\sum X_i = \sum Y_i$

Thus, $\sum_{i=1}^{n} X_i$ is minimal sufficient (c). the pdf of uniform $(0,\theta)$ is $p(x|\theta) = \frac{1}{\theta^n} 1\{0 \le \theta \le X_{(n)}\}$

Same as 6.23 we get $T(X) = \max X_i$ is sufficient.

$\frac{f(\vec{x}|\theta)}{f(\vec{y}|\theta)} = 1 \Leftrightarrow X_{(n)} = Y_{(n)}$, which means $T(X)$ is minimal sufficient.

(d). by Ex 6.25 we know that $X_{(1)}, \cdots, X_{(n)}$ is sufficient, and we also know the pdf of Cauchy $(\theta,1)$ has common support, which means $T(X) = (X_{(1)}, \cdots, X_{(n)})$ is minimal sufficient.

(e). Also, $X_{(1)}, \cdots, X_{(n)}$ is clearly sufficient. the pdf of logistic(μ, β) is $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}$, so it has common support for any θ.

Thus, $T(X) = (X_{(1)}, \cdots, X_{(n)})$ is minimal sufficient.

$\square$

6.36 One advantage of using a minimal sufficient statistic is that unbiased estimators will have smaller variance, as the following exercise will show. Suppose that T_1 is sufficient and T_2 is minimal sufficient, U is an unbiased estimator of θ, and define $U_1 = E(U|T_1)$ and $U_2 = E(U|T_2)$.

(a) Show that $U_2 = E(U_1|T_2)$.

(b) Now use the conditional variance formula (Theorem 4.4.7) to show that $\operatorname{Var} U_2 \le \operatorname{Var} U_1$.

6.36 (a). $E(U_1|T_2) = E(E(U|T_1)|T_2)$, since T_1 is sufficient and T_2 is minimal sufficient, so we can find $g(\cdot)$ s.t. $T_2 = g(T_1)$. thus by the smoothness property of conditional expectation, we have $E(E(U|T_1)|T_2) = E(U|T_2) = U_2$

i.e $U_2 = E(U_1|T_2)$.

Lemma: the smothness: Let $\mathscr{C}_1, \mathscr{C}_2 \subset \mathscr{F}$ be σ-algebra, and $\mathscr{C}_1 \subset \mathscr{C}_2$

Then we have : $E(\xi|\mathscr{C}_1) = E[E(\xi|\mathscr{C}_2)|\mathscr{C}_1]$, a.s

(b).

Theorem 4.4.7 (Conditional variance identity) *For any two random variables X and Y,*

$$\operatorname{Var} X = \operatorname{E}(\operatorname{Var}(X|Y)) + \operatorname{Var}(\operatorname{E}(X|Y)), \tag{4.4.4}$$

$\operatorname{Var}(U_1) = \operatorname{Var}(E(U_1|T_2)) + E(\operatorname{Var}(U_1|T_2)) = \operatorname{Var}(U_2) + E(\operatorname{Var}(U_1|T_2)) \geq \operatorname{Var}(U_2)$

$\square$

6.40 Let $X_1, \ldots, X_n$ be iid observations from a location–scale family. Let $T_1(X_1, \ldots, X_n)$ and $T_2(X_1, \ldots, X_n)$ be two statistics that both satisfy

$$T_i(ax_1 + b, \ldots, ax_n + b) = aT_i(x_1, \ldots, x_n)$$

for all values of $x_1, \ldots, x_n$ and b and for any $a > 0$.

(a) Show that T_1/T_2 is an ancillary statistic.

(b) Let R be the sample range and S be the sample standard deviation. Verify that R and S satisfy the above condition so that R/S is an ancillary statistic.

(a). Since $X_1, \cdots, X_n$ from a location-scale family, from Thm 3.5.6, we can write $X_i = \sigma Z_i + \mu$, where $Z_1, \cdots, Z_n$ is a random sample from the standard pdf $f(z)$. Then $\frac{T_1(X_1, \cdots, X_n)}{T_2(X_1, \cdots, X_n)} = \frac{T_1(\sigma Z_1 + \mu, \cdots, \sigma Z_n + \mu)}{T_2(\sigma Z_1 + \mu, \cdots, \sigma Z_n + \mu)}$

$= \frac{\sigma T_1(Z_1, \cdots, Z_n)}{\sigma T_2(Z_1, \cdots, Z_n)} = \frac{T_1(Z_1, \cdots, Z_n)}{T_2(Z_1, \cdots, Z_n)}$. So T_1/T_2 is of only $Z_1, \cdots, Z_n$, which does not depend on μ or σ.

Thus T_1/T_2 is an ancillary statistic

(b). $R(aX_1+b, \cdots, aX_n+b) = aX_{(n)} + b - aX_{(1)} - b = a(X_{(n)} - X_{(1)}) = aR(X_1, \cdots, X_n)$. $(a>0)$

We also have $S^2(aX_1+b, \cdots, aX_n+b) = \frac{1}{n-1}\sum\left((aX_i+b) - (a\bar{X}+b)\right)^2$
$= \frac{a^2}{n-1}\sum(X_i - \bar{X})^2 = a^2 S^2(X_1, \cdots, X_n)$

So $S(aX_1+b, \cdots, aX_n+b) = S(X_1, \cdots, X_n)$. Hence R and S satisfy the above condition so that R/S is an ancillary statistic

$\square$