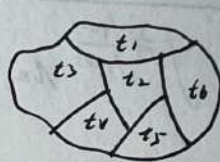
Principles of Data Reduction

- Experimenter uses the information in $X_1, X_2 \cdots X_n$ to infer an unknown parameter θ. If n is large. Then the observed sample is too large to interpret.

$\Rightarrow$ We will use $T(\vec{X})$ rather than the entire sample $\vec{X}$.

For $T(\cdot)$, it can partition sample space into different sets $A_t = [\vec{x} \in \mathcal{X} \mid T(\vec{x}) = t]$:



$\Rightarrow$ For θ, we're interested in the information relevant θ. discard information irrelevant θ.

(1) The Sufficient Principle:

- $T(\vec{X})$ is sufficient statistic for θ. Then it captures all information of θ, θ only depends on X through $T(\vec{X})$. If $T(\vec{x}) = T(\vec{y})$, then inference of θ are same.

Pf: For $X, Y \sim f(x|\theta)$, $T(\vec{x}) = T(\vec{y})$

$$\text{Then: } P_\theta(X = \vec{x}) = P_\theta(X = \vec{x} \mid T(X) = T(\vec{x}))\, P(T(X) = T(\vec{x}))$$

$$= P_\theta(Y = \vec{x} \mid T(Y) = T(\vec{y}))\, P(T(Y) = T(\vec{y}))$$

$$= P_\theta(Y = \vec{x})$$

① Criteria:

By definition: $P_\theta(X=x \mid T(X)=T(x)) = \dfrac{P_\theta(X=x, T(X)=T(x))}{P_\theta(T(X)=T(x))}$

$$= \frac{p(x|\theta)}{q(T(x)|\theta)} = h(x). \quad q(\cdot) \text{ is pdf of } T(X).$$

$\Rightarrow$ Thm. $T(X)$ is sufficient statistic $\Rightarrow \dfrac{p(x|\theta)}{q(T(x)|\theta)} = h(x)$

☆ Factorization Thm:

$X \sim f(x|\theta)$. $T(X)$ is sufficient statistic for θ.

$\Leftrightarrow \quad f(\vec{x}|\theta) = g(T(\vec{x})|\theta)\, h(\vec{x})$

Pf: ($\Rightarrow$). Proved above

($\Leftarrow$) $\dfrac{f(x|\theta)}{q(T(x)|\theta)} = \dfrac{g(T(\vec{x})|\theta)\, h(\vec{x})}{\int_{A_{T(\vec{x})}} f(x|\theta)\, dx} = \dfrac{g(T(\vec{x})|\theta)\, h(\vec{x})}{\int_{A_{T(\vec{x})}} g(T(\vec{y})|\theta)\, h(\vec{y})\, dy}$

$$= \frac{g(T(\vec{x})|\theta)\, h(\vec{x})}{g(T(\vec{x})|\theta) \int_{A_{T(\vec{x})}} h(\vec{y})\, dy} = r(\vec{x})$$

② Sufficient statistic vector:

· It's usually for multiple parameters.

Thm. (For exponential family)

$X_k \sim f(x|\vec{\theta})$. $1 \le k \le n$. i.i.d. $f(x|\vec{\theta}) = h(x)\, c(\vec{\theta})\, e^{\sum_1^k w_i(\vec{\theta})\, t_i(x)}$

$\vec{\theta} = (\theta_1, \cdots \theta_d)$, $d \le k$. Then $T(\vec{x}) = \left(\sum_1^n t_1(x_j), \cdots \sum_1^n t_k(x_j)\right)$

is sufficient statistic vector for $\vec{\theta}$.

③ Minimal sufficient Statistic :

· Note that: $\vec{X}$ is the Largest sufficient statistic.

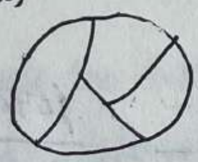
$\forall g$. $g(x)$ is one-to-one func. Then for $T(\vec{x})$ S-S.

$g(T(\vec{x}))$ is S-S. too.

$\Rightarrow$ We want to find a statistic carrying little information. but retaining all information of θ.

<u>Def</u>: $T(x)$ is minimal sufficient statistic. if any other S.S. $T'(x)$, $T'(x) = T'(y) \Rightarrow T(x) = T(y)$. ($T = f(T')$)

$A(x)$



$T(x)$ submits a coarsest partition!
$\Rightarrow$ Achieve greatest reduction.

<u>Remark</u>: i) If MLE is a sufficient statistic. Then it's the minimal.

<u>Pf</u>: For $\forall$ sufficient statistic $T(x)$

$$f(x|\theta) = g(T(x)|\theta)\, h(x)$$

$$\lambda(x) = h(x) \sup_{\theta} g(T(x)|\theta) = h(x)\, g(T(x)|\hat{\theta})$$

where $\hat{\theta}$ satisfies $g'(T(x)|\hat{\theta}) = 0$. $\therefore \hat{\theta} = r(T(x))$

ii) Most location family $f(x-\theta)$. The most reduction is $(X_{(1)}, X_{(2)} \cdots X_{(n)})$. So it's the minimal S.S.

iii) Minimal S.S isn't unique. Since $\forall g$ one-to-one if T is minimal. Then $g(T)$ is minimal, too.

Thm. (Criterion)

$f(x|\theta)$ is pmf or pdf of X. For $T(\vec{x})$. $\forall \vec{x}, \vec{\eta} \in \Omega$.

$\frac{f(\vec{x}|\theta)}{f(\vec{\eta}|\theta)}$ is irrelevant with $\theta \Leftrightarrow T(x) = T(\eta)$

Then $T(\vec{x})$ is minimal s.s. for θ.

Pf: 1') $T(\vec{x})$ is s.s:

Note that $f(x|\theta) = f(x_{T(x)}|\theta) \frac{f(x|\theta)}{f(x_{T(x)}|\theta)}$

where $T(x_{T(\vec{x})}) = T(\vec{x})$. $f(x_{T(\vec{x})}|\theta) \triangleq g(T(x)|\theta)$

$\therefore f(x|\theta) / f(x_{T(\vec{x})}|\theta) = h(\vec{x})$. By Factorization Thm □

2') $T(\vec{x})$ is minimal:

By Factorization Thm: $\tilde{T}(\vec{x}) = \tilde{T}(\vec{\eta})$. for any other s.s. $\tilde{T}(x)$

$\Rightarrow \frac{f(x|\theta)}{f(\vec{\eta}|\theta)}$ is irrelevant with θ. $\therefore T(x) = T(\eta)$

Remark: i) $\forall$ s.s. satisfies ($\Leftarrow$). But only minimal s.s.

satisfies ($\Rightarrow$). "Minimal" is "necessary".

ii) To find minimal s.s. Firstly calculate

$f(x|\theta)/f(\eta|\theta)$. Find the condition that is

necessary and sufficient the equation doesn't

contain θ. (i.e. $T(\vec{x}) = T(\vec{\eta})$)

Thm. For family of density $\{f_i(x)\}_0^k$ has common

support. Then: i) $T(\vec{x}) = \left(\frac{f_1(x)}{f_0(x)}, \frac{f_2(x)}{f_0(x)} \cdots \frac{f_k(x)}{f_0(x)}\right)$

is minimal sufficient for the family. ii) If $T(x)$

is sufficient for the family. Then $T(X)$ is minimal.

Remark: The sufficient minimal statistic can be extend to nonparametric family:

· If $X_1, X_2 \cdots X_n \sim f$. unknown density. Then the order statistic is minimal sufficient.

Pf: If $f_i \sim \text{logistic}(\alpha_i, \beta_i) \Rightarrow$ By Thm: minimal S.S

$$T(x) = \left(\frac{\prod f_1(x_i)}{\prod f_0(x_i)} \cdots \frac{\prod f_p(x_i)}{\prod f_0(x_i)} \right) \xleftrightarrow{\text{one-to-one}} (X_{(1)} \cdots X_{(n)})$$

If f is nonparametric, by ii). $(X_{(1)} \cdots X_{(n)})$ is sufficient. So it's the minimal.

④ Ancillary Statistic:

· Def: A statistic $S(\vec{x})$ whose dist doesn't depend on θ. It's the ancillary statistic for θ.

Remark: It contains no information about θ.

e.g.
- i) For location family: $f(x-\theta)$.
  $R = X_{(n)} - X_{(1)}$ is ancillary. (Or $X-Y$)
- ii) For scale family: $f(\frac{x}{\sigma})$
  X/Y is ancillary statistic.

Remark: Minimal S.S. isn't indept with ancillary statistic.
It may contain ancillary statistic: $X \sim U(\theta, \theta+1)$
$X_{(n)}, X_{(1)}$ minimal S.S. But $X_{(n)} - X_{(1)}$ is ancillary.

③ Complete Statistic:

. Def: For statistic $T(\vec{X}) \sim f(x|\theta)$. It's called complete if $\forall g$. $E_\theta(g(T))=0$. $\forall\theta$. then $P_\theta(g(T)=0)=1$. $\forall \theta \in \Theta$.

Remark: i) It means any transformation of T won't contain ancillary statistic. If $E_\theta(g(T))$ is irrelevant with θ. Then $g(T) \equiv C$. w.p.1.

So the definition can be refined: $E_\theta(g(T)) \equiv C$. $\forall\theta$
$\Rightarrow P_\theta(g(T) \equiv C)=1$. $\forall\theta$.

Basu's Thm:

If $T(\vec{X})$ is complete sufficient statistic. (minimal)

Then $T(\vec{X})$ is indept with every ancillary statistic.

Pf: Suppose $S(\vec{X})$ is ancillary. indept with θ.

Prove: $P(S(\vec{X})=s \mid T(\vec{X})=t) = P(S(\vec{X})=s)$

Note that $P(S(\vec{X})=s) = \sum_t P(S(\vec{X})=s, T(\vec{X})=t)$

$= \sum P(S(\vec{X})=s \mid T(\vec{X})=t) P(T(\vec{X})=t)$ □

$g(T(X)) = P(S(\vec{X})=s \mid T(\vec{X})=t) - P(S(\vec{X})=s)$.

Remark: i) The converse is false.

ii) "sufficient" is necessary. Since there exists complete statistic not being sufficient:

$\{X_k\}_1^n \sim$ Poisson(λ). Let $T(\vec{x}) = X_1$. complete.

iii) The "minimal" can be omitted:

<u>Prop.</u> If minimal s.s exists. Then any complete sufficient statistic is minimal.

<u>For exponential family:</u>

$\cdot X_k \sim f(x|\vec{\theta}) = h(x) c(\vec{\theta}) e^{\sum_1^k w(\theta_i) t_i(x)}$, i.i.d.

$1 \le k \le n$, $\vec{\theta} = (\theta_1 \cdots \theta_k)$. Then $T(x) = (\sum_{i=1}^n t_1(X_i), \dots, \sum_1^n t_k(X_i))$

is complete if Θ contains an open set in $\mathbb{R}^k$.

<u>Pf:</u> By n-dimensional Laplace Transforming.

(2) <u>The Likelihood Principle:</u>

① $\cdot$ <u>Def:</u> Given $\vec{X} = \vec{x}$ observed. $L(\theta|\vec{x}) = f(\vec{x}|\theta)$

is called likelihood function.

<u>Remark:</u> It can used to summerize data.

<u>Likelihood Principle:</u>

If $\vec{x}$ and $\vec{y}$ are two sample points from $f(x|\theta)$.

$\forall\theta$, $L(\theta|\vec{x}) = C(x,y) L(\theta, \vec{y})$. Then the conclusion drawn from $\vec{x}$ and $\vec{y}$ concerning θ are same.

② The formal form:

· Def: Experiment E: $(\vec{X}, \theta, \{f(x|\theta)\})$, which means

$\vec{X}$ is random vector with pmf $f(x|\theta)$ for some fixed θ.

$\Rightarrow$ After the experiment was performed and having observed sample $\vec{X} = \vec{x}$, will make some conclusion $Ev(E, \vec{x})$ about θ.

Formal Sufficiency Principle:

For $E = (\vec{X}, \theta, \{f(x|\theta)\})$, $T(\vec{x})$ is s.s. for θ.

If $T(\vec{x}) = T(\vec{\eta})$, $\vec{x}, \vec{\eta}$ are 2 sample points.

Then $Ev(E, \vec{x}) = Ev(E, \vec{\eta})$

Remark: The Likelihood Principle can be used to derive it:

$$L(\theta|\vec{x}) = g(T(\vec{x})|\theta)h(\vec{x}) = g(T(\vec{\eta})|\theta)h(x) = \frac{h(x)}{h(\eta)} L(\theta|\vec{\eta})$$

Conditionality Principle:

· $E_1 = (X_1, \theta, \{f_1(x_1|\theta)\})$, $E_2 = (X_2, \theta, \{f_2(x_2|\theta)\})$, only the unknown θ are common. Consider mixed experiment: J is r.v. observed. indept of X_1, X_2, θ.

$P(J=1) = P(J=2) = \frac{1}{2}$. Perform $E_J = E^* = ((j, X_j), \theta, \{f^*(x^*|\theta)\})$

$x^* = (j, x_j)$, $f^*(x^*|\theta) = \frac{1}{2} f_j(x_j|\theta)$. $\Rightarrow Ev(E^*, x^*) = Ev(E_j, x_j)$

Remark: It means the conclusion only depends on which experiment is performed. indept with which one we choose.

⇒ Formal Likelihood Principle:

$E_1 = (X_1, \theta, \{f_1(\vec{x}_1|\theta)\})$, $E_2 = (X_2, \theta, \{f_2(x_2|\theta)\})$, only the unknown θ are common. x_1^*, x_2^* are sample point from E_1, E_2 resp. s.t. $L(\theta|x_1^*) = C\,L(\theta|x_2^*)$, $C = C(x_1^*, x_2^*)$

Then $Ev(E_1, x_1^*) = Ev(E_2, x_2^*)$

Cr. $E = (\vec{X}, \theta, \{f(x|\theta)\})$. is an experiment. Then: $Ev(E, \vec{x})$ only depends on $\vec{x}$ and E through $L(\theta|x)$

Thm. (Birnbaum's Thm)

The formal likelihood Principle ⟺ Formal Sufficient Principle and Conditionality Principle

Pf: (⟸) Def: $T(j, X_j) = \begin{cases} (1, X_1^*), & \text{if } j=1,\ X_1 = x_1^* \text{ or } j=2\ X_2 = x_2^* \\ (j, X_j), & \text{otherwise} \end{cases}$

The key: T ⟵ It's on sample space of E^* mixed experiment. (j, X_j).
takes on same value for sample point $(1, X_1^*)$, $(2, X_2^*)$ from each experiment

Claim: $T(j, X_j)$ is s.s for θ in $E^* \sim f^*$

$f^*((j, X_j)|\theta) = g(T(j, X_j)|\theta)$ $j \neq 2$.

$f^*((2, X_2^*)|\theta) = f(X_2^*|\theta)\,p(J=2) = C(x_1^*, x_2^*)\,f(x_1^*|\theta)\,p(J=2)$

$= f^*((1, X_1^*)|\theta)\,C(x_1^*, x_2^*) = g(T(2, X_2^*)|\theta)\,C(x_1^*, x_2^*)$

$\therefore T((j, x_j))$ is S.S. for $f^*((j, x_j)|\theta)$ about θ.

Since $T((1, x_1^*)) = T((2, x_2^*))$ By FSP:

$Ev(E^*, (1, x_1^*)) = Ev(E^*, (2, x_2^*))$

By CP: $Ev(E^*, (1, x_1^*)) = Ev(E_1, x_1^*) = Ev(E_2, x_2^*)$

($\Rightarrow$) Since $f^*((j, x_j)|\theta) = p(J=j) f(x_j|\theta)$ By FLP:

$\therefore \mathcal{L}(\theta|(j, x_j)) = C\mathcal{L}(\theta|x_j) \Rightarrow Ev(E^*, (j, x_j)) = Ev(E_j, x_j)$

If $T(\vec{x}) = T(\vec{\eta})$, we also have $\mathcal{L}(\theta|\vec{x}) = \mathcal{L}(\theta|\vec{\eta})$

<u>Remark</u>: i) FSP + CP $\Rightarrow$ Likelihood Principle. (By FLP $\Rightarrow$ LP)

Since $\mathcal{L}(\theta|\vec{x}) = c(\vec{x}, \vec{\eta})\mathcal{L}(\theta|\vec{\eta})$. Only one E in LP!

ii) <u>pf of Cor:</u>

For $E = E_1 = E_2$. $\mathcal{L}(\theta|\vec{x}) = \mathcal{L}(\theta|\vec{\eta})\, C(\vec{x}, \vec{\eta})$

$Ev(E_1, \vec{x}) = Ev(E, \vec{x}) = Ev(E, \vec{\eta}) = Ev(E_2, \vec{\eta})$

iii) Many common statistical model can't be applied in FLP. Since there may exist some information on θ doesn't base on sufficient statistic. (Violate FSP!)

iv) The pf of Thm isn't compelling. Since before CP coming, we should define S.S for each experiment (Then the "key" won't happen on separate S.S for each experiment: Sample space may be different for E_1, E_2!)

(3) The equivariance Principle:

· It states: If $T(\vec{x}) = T(\vec{\eta})$. Since $Ev(E, \vec{x})$ may be different from $Ev(E, \vec{\eta})$. But there's a certain relationship between the two inference

Type

i) Measurement Equivariance: the inference shouldn't depend on measurement scale.

ii) Formal Invariance: two inference problems has same formal structure of model (Dist's $f_1, f_2 \in \{f(x|\theta) | \theta \in \Theta\}$ and set of allowable or wrong inferences e.g. $\theta > 0$)

$\Rightarrow$ Then the same inference procedure sholud be used.

Equivariance Principle:

$Y = g(X)$ is change of measurement scale, s.t. Y has the same formal structure model as X.

$\Rightarrow$ Then the inference procedure should be measure equivariant and formally equivariant.

<u>Remark</u>: The transformations of scale measurement form a "Group". Denote $\mathcal{G}$. For $g \in \mathcal{G}$, Then

g satisfies

i) measurement equivariant: 若 $T(X)$ estimate θ, $Y = g(X)$
$T^*(Y)$ estimate $g(\theta)$

ii) Formal invariant: 若 $Y = g(X)$. Then we have
: $T(g(X)) = g(T(X))$
i.e. $T^*(X) = T(X)$

· <u>Def</u>: $\mathcal{F} = \{f(x|\theta) \mid \theta \in \Theta\}$. $\mathcal{G}$ is a Group of transform on sample space $\mathcal{X}$. Then $\mathcal{F}$ is invariant under $\mathcal{G}$ 若 $\forall \theta \in \Theta$, $g \in \mathcal{G}$, $\exists!\, \theta' \in \Theta$, s.t. $Y = g(X) \sim f(x|\theta')$ 若 $X \sim f(x|\theta)$, instally.