Chapter 4

Sufficient Statistics

4.1 Statistics and Sampling Distributions

Suppose that the data $Y_1, ..., Y_n$ is drawn from some population. The observed data is $Y_1 = y_1, ..., Y_n = y_n$ where $y_1, ..., y_n$ are numbers. Let $\boldsymbol{y} = (y_1, ..., y_n)$. Real valued functions $T(y_1, ..., y_n) = T(\boldsymbol{y})$ are of interest as are vector valued functions $\boldsymbol{T}(\boldsymbol{y}) = (T_1(\boldsymbol{y}), ..., T_k(\boldsymbol{y}))$. Sometimes the data $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are random vectors. Again interest is in functions of the data. Typically the data has a joint pdf or pmf $f(y_1, ..., y_n|\boldsymbol{\theta})$ where the vector of unknown parameters is $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$. (In the joint pdf or pmf, the $y_1, ..., y_n$ are dummy variables, not the observed data.)

Definition 4.1. A **statistic** is a function of the data that does not depend on any unknown parameters. The probability distribution of the statistic is called the **sampling distribution** of the statistic.

Let the data $\boldsymbol{Y} = (Y_1, ..., Y_n)$ where the Y_i are random variables. If $T(y_1, ..., y_n)$ is a real valued function whose domain includes the sample space $\mathcal{Y}$ of $\boldsymbol{Y}$, then $W = T(Y_1, ..., Y_n)$ is a statistic provided that T does not depend on any unknown parameters. The data comes from some probability distribution and the statistic is a random variable and hence also comes from some probability distribution. To avoid confusing the distribution of the statistic with the distribution of the data, the distribution of the statistic is called the sampling distribution of the statistic. If the observed data is $Y_1 = y_1, ..., Y_n = y_n$, then the observed value of the statistic is $W = w = T(y_1, ..., y_n)$. Similar remarks apply when the statistic $\boldsymbol{T}$ is vector valued and when the data $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are random vectors.

Often $Y_1, ..., Y_n$ will be iid and statistics of the form

$$\sum_{i=1}^{n} a_i Y_i \quad \text{and} \quad \sum_{i=1}^{n} t(Y_i)$$

are especially important. Chapter 10 and Theorems 2.17, 2.18, 3.6 and 3.7 are useful for finding the sampling distributions of some of these statistics when the Y_i are iid from a given brand name distribution that is usually an exponential family. The following example lists some important statistics.

Example 4.1. Let the $Y_1, ..., Y_n$ be the data.
a) The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{4.1}$$

b) The *sample variance*

$$S^2 \equiv S_n^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2}{n-1}. \tag{4.2}$$

c) The *sample standard deviation* $S \equiv S_n = \sqrt{S_n^2}$.
d) If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics.*
e) The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,} \tag{4.3}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \quad \text{if n is even.}$$

f) The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{4.4}$$

g) The *sample maximum*

$$\max(\text{n}) = \text{Y}_{(\text{n})} \tag{4.5}$$

and the observed max $y_{(n)}$ is the largest value of the observed data.

h) The *sample minimum*

$$\min(\mathrm{n}) = \mathrm{Y}_{(1)} \tag{4.6}$$

and the observed min $y_{(1)}$ is the smallest value of the observed data.

Example 4.2. Usually the term "observed" is dropped. Hence below "data" is "observed data", "observed order statistics" is "order statistics" and "observed value of MED(n)" is "MED(n)."
Let the data be $9, 2, 7, 4, 1, 6, 3, 8, 5$ (so $Y_1 = y_1 = 9, ..., Y_9 = y_9 = 5$). Then the order statistics are $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then MED($n$) $= 5$ and MAD(n) $= 2 =$ MED$\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

Example 4.3. Let the $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$. Then

$$T_n = \frac{\sum_{i=1}^n (Y_i - \mu)^2}{n}$$

is a statistic iff μ is known.

The following theorem is extremely important and the proof follows Rice (1988, p. 171-173) closely.

Theorem 4.1. Let the $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$.
a) The sample mean $\overline{Y} \sim N(\mu, \sigma^2/n)$.
b) $\overline{Y}$ and S^2 are independent.
c) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Hence $\sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2_{n-1}$.
d) $\frac{\sqrt{n}(\overline{Y}-\mu)}{S} = \frac{(\overline{Y}-\mu)}{S/\sqrt{n}} \sim t_{n-1}$.

Proof. a) This result follows from Theorem 2.17e.

b) The moment generating function of $(\overline{Y}, Y_1 - \overline{Y}, ..., Y_n - \overline{Y})$ is

$$m(s, t_1, ..., t_n) = E(\exp[s\overline{Y} + t_1(Y_1 - \overline{Y}) + \cdots + t_n(Y_n - \overline{Y})]).$$

By Theorem 2.22, $\overline{Y}$ and $(Y_1 - \overline{Y}, ..., Y_n - \overline{Y})$ are independent if

$$m(s, t_1, ..., t_n) = m_{\overline{Y}}(s)\ m(t_1, ..., t_n)$$

where $m_{\overline{Y}}(s)$ is the mgf of $\overline{Y}$ and $m(t_1, ..., t_n)$ is the mgf of $(Y_1 - \overline{Y}, ..., Y_n - \overline{Y})$.
Now

$$\sum_{i=1}^n t_i(Y_i - \overline{Y}) = \sum_{i=1}^n t_i Y_i - \overline{Y} n \overline{t} = \sum_{i=1}^n t_i Y_i - \sum_{i=1}^n \overline{t} Y_i$$

and thus

$$s\overline{Y} + \sum_{i=1}^{n} t_i(Y_i - \overline{Y}) = \sum_{i=1}^{n}[\frac{s}{n} + (t_i - \overline{t}\ )]Y_i = \sum_{i=1}^{n} a_i Y_i.$$

Now $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n}[\frac{s}{n} + (t_i - \overline{t}\ )] = s$ and

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n}[\frac{s^2}{n^2} + 2\frac{s}{n}(t_i - \overline{t}\ ) + (t_i - \overline{t}\ )^2] = \frac{s^2}{n} + \sum_{i=1}^{n}(t_i - \overline{t}\ )^2.$$

Hence

$$m(s, t_1, ..., t_n) = E(\exp[s\overline{Y} + \sum_{i=1}^{n} t_i(Y_i - \overline{Y})]) = E[\exp(\sum_{i=1}^{n} a_i Y_i)]$$

$$= m_{Y_1,...,Y_n}(a_1, ..., a_n) = \prod_{i=1}^{n} m_{Y_i}(a_i)$$

since the Y_i are independent. Now

$$\prod_{i=1}^{n} m_{Y_i}(a_i) = \prod_{i=1}^{n} \exp\left(\mu a_i + \frac{\sigma^2}{2}a_i^2\right) = \exp\left(\mu \sum_{i=1}^{n} a_i + \frac{\sigma^2}{2}\sum_{i=1}^{n} a_i^2\right)$$

$$= \exp\left[\mu s + \frac{\sigma^2}{2}\frac{s^2}{n} + \frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \overline{t}\ )^2\right]$$

$$= \exp\left[\mu s + \frac{\sigma^2}{2n}s^2\right]\ \exp\left[\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \overline{t}\ )^2\right].$$

Now the first factor is the mgf of $\overline{Y}$ and the second factor is $m(t_1, ..., t_n) = m(0, t_1, ..., t_n)$ since the mgf of the marginal is found from the mgf of the joint distribution by setting all terms not in the marginal to 0 (ie set $s = 0$ in $m(s, t_1, ..., t_n)$ to find $m(t_1, ..., t_n)$). Hence the mgf factors and

$$\overline{Y} \perp\!\!\!\perp (Y_1 - \overline{Y}, ..., Y_n - \overline{Y}).$$

Since S^2 is a function of $(Y_1 - \overline{Y}, ..., Y_n - \overline{Y})$, it is also true that $\overline{Y} \perp\!\!\!\perp S^2$.

c) $(Y_i - \mu)/\sigma \sim N(0,1)$ so $(Y_i - \mu)^2/\sigma^2 \sim \chi^2_1$ and

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 \sim \chi^2_n.$$

Now

$$\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n (Y_i - \overline{Y} + \overline{Y} - \mu)^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 + n(\overline{Y} - \mu)^2.$$

Hence

$$W = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \overline{Y})^2 + \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 = U + V.$$

Since $U \perp\!\!\!\perp V$ by b), $m_W(t) = m_U(t)\ m_V(t)$. Since $W \sim \chi^2_n$ and $V \sim \chi^2_1$,

$$m_U(t) = \frac{m_W(t)}{m_V(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$$

which is the mgf of a χ^2_{n-1} distribution.

d)

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

and

$$S^2/\sigma^2 = \frac{\frac{(n-1)S^2}{\sigma^2}}{n-1} \sim \frac{1}{n-1}\chi^2_{n-1}.$$

Suppose $Z \sim N(0,1)$, $X \sim \chi^2_{n-1}$ and $Z \perp\!\!\!\perp X$. Then $Z/\sqrt{X/(n-1)} \sim t_{n-1}$.
Hence

$$\frac{\sqrt{n}(\overline{Y} - \mu)}{S} = \frac{\sqrt{n}(\overline{Y} - \mu)/\sigma}{\sqrt{S^2/\sigma^2}} \sim t_{n-1}.$$

QED

Theorem 4.2. Let the $Y_1, ..., Y_n$ be iid with cdf F_Y and pdf f_Y.
a) The pdf of $T = Y_{(n)}$ is

$$f_{Y_{(n)}}(t) = n[F_Y(t)]^{n-1} f_Y(t).$$

b) The pdf of $T = Y_{(1)}$ is

$$f_{Y_{(1)}}(t) = n[1 - F_Y(t)]^{n-1} f_Y(t).$$

c) Let $2 \le r \le n$. Then the joint pdf of $Y_{(1)}, Y_{(2)}, ..., Y_{(r)}$ is

$$f_{Y_{(1)},...,Y_{(r)}}(t_1, ..., t_r) = \frac{n!}{(n-r)!}[1 - F_Y(t_r)]^{n-r} \prod_{i=1}^{r} f_Y(t_i).$$

Proof of a) and b). a) The cdf of $Y_{(n)}$ is

$$F_{Y_{(n)}}(t) = P(Y_{(n)} \le t) = P(Y_1 \le t, ..., Y_n \le t) = \prod_{i=1}^{n} P(Y_i \le t) = [F_Y(t)]^n.$$

Hence the pdf of $Y_{(n)}$ is

$$\frac{d}{dt} F_{Y_{(n)}}(t) = \frac{d}{dt}[F_Y(t)]^n = n[F_Y(t)]^{n-1} f_Y(t).$$

b) The cdf of $Y_{(1)}$ is

$$F_{Y_{(1)}}(t) = P(Y_{(1)} \le t) = 1 - P(Y_{(1)} > t) = 1 - P(Y_1 > t, ..., Y_n > t)$$

$$= 1 - \prod_{i=1}^{n} P(Y_i > t) = 1 - [1 - F_Y(t)]^n.$$

Hence the pdf of $Y_{(n)}$ is

$$\frac{d}{dt} F_{Y_{(n)}}(t) = \frac{d}{dt}(1 - [1 - F_Y(t)]^n) = n[1 - F_Y(t)]^{n-1} f_Y(t). \quad \text{QED}$$

To see that c) may be true, consider the following argument adapted from Mann, Schafer and Singpurwalla (1974, p. 93). Let Δt_i be a small positive number and notice that $P(E) \equiv$

$$P(t_1 < Y_{(1)} < t_1 + \Delta t_1, t_2 < Y_{(2)} < t_2 + \Delta t_2, ..., t_r < Y_{(r)} < t_r + \Delta t_r)$$

$$= \int_{t_r}^{t_r + \Delta t_r} \cdots \int_{t_1}^{t_1 + \Delta t_1} f_{Y_{(1)},...,Y_{(r)}}(w_1, ..., w_r) dw_1 \cdots dw_r$$

$$\approx f_{Y_{(1)},...,Y_{(r)}}(t_1, ..., t_r) \prod_{i=1}^{r} \Delta t_i.$$

Since the event E denotes the occurrence of no observations before t_i, exactly one occurrence between t_1 and $t_1+\Delta t_1$, no observations between $t_1+\Delta t_1$ and t_2 and so on, and finally the occurrence of $n-r$ observations after $t_r+\Delta t_r$, using the multinomial pmf shows that

$$P(E) = \frac{n!}{0!1!\cdots 0!1!(n-r)!}\rho_1^0\rho_2^1\rho_3^0\rho_4^1\cdots\rho_{2r-1}^0\rho_{2r}^1\rho_{2r+1}^{n-r}$$

where

$$\rho_{2i} = P(t_i < Y < t_i + \Delta t_i) \approx f(t_i)\Delta t_i$$

for $i = 1, ..., r$ and

$$\rho_{2r+1} = P(n-r\ Y's > t_r + \Delta t_r) \approx (1-F(t_r))^{n-r}.$$

Hence

$$P(E) \approx \frac{n!}{(n-r)!}(1-F(t_r))^{n-r}\prod_{i=1}^r f(t_i)\prod_{i=1}^r \Delta t_i$$

$$\approx f_{Y_{(1)},...,Y_{(r)}}(t_1,...,t_r)\prod_{i=1}^r \Delta t_i,$$

and result c) seems reasonable.

Example 4.4. Let $Y_1, ..., Y_n$ be iid from the following distributions.
a) Bernoulli(ρ): Then $Y_{(1)} \sim$ Bernoulli(ρ^n).
b) Geometric(ρ): Then $Y_{(1)} \sim$ Geometric($1-(1-\rho)^n$).
c) Burr(ϕ, λ): Then $Y_{(1)} \sim$ Burr($\phi, \lambda/n$).
d) EXP(λ): Then $Y_{(1)} \sim$ EXP(λ/n).
e) EXP(θ, λ): Then $Y_{(1)} \sim$ EXP($\theta, \lambda/n$).
f) Pareto PAR(σ, λ): Then $Y_{(1)} \sim$ PAR($\sigma, \lambda/n$).
g) Gompertz Gomp(θ, ν): Then $Y_{(1)} \sim$ Gomp($\theta, n\nu$).
h) Rayleigh $R(\mu, \sigma)$: Then $Y_{(1)} \sim R(\mu, \sigma/\sqrt{n})$.
i) Truncated Extreme Value TEV(λ) : Then $Y_{(1)} \sim$ TEV(λ/n).
j) Weibull $W(\phi, \lambda)$: Then $Y_{(1)} \sim W(\phi, \lambda/n)$.

Proof: a) $Y_i \in \{0,1\}$ so $Y_{(1)} \in \{0,1\}$. Hence $Y_{(1)}$ is Bernoulli, and $P(Y_{(1)} = 1) = P(Y_1 = 1, ..., Y_n = 1) = [P(Y_1 = 1)]^n = \rho^n$.

b) $P(Y_{(1)} \le y) = 1 - P(Y_{(1)} > y) = 1 - [1-F(y)]^n =$
$1 - [1-(1-(1-\rho)^{\lfloor y \rfloor +1}]^n = 1 - [(1-\rho)^n]^{\lfloor y \rfloor +1} = 1 - [1-(1-(1-\rho)^n)]^{\lfloor y \rfloor +1}$
for $y \ge 0$ which is the cdf of a Geometric($1-(1-\rho)^n$) random variable.

Parts c)-j) follow from Theorem 4.2b. For example, suppose $Y_1, ..., Y_n$ are iid EXP(λ) with cdf $F(y) = 1 - \exp(-y/\lambda)$ for $y > 0$. Then $F_{Y_{(1)}}(t) = 1 - [1 - (1 - \exp(-t/\lambda))]^n = 1 - [\exp(-t/\lambda)]^n = 1 - \exp[-t/(\lambda/n)]$ for $t > 0$. Hence $Y_{(1)} \sim$ EXP(λ/n).

4.2 Minimal Sufficient Statistics

For parametric inference, the pmf or pdf of a random variable Y is $f_{\boldsymbol{\theta}}(y)$ where $\boldsymbol{\theta} \in \Theta$ is unknown. Hence Y comes from a family of distributions indexed by $\boldsymbol{\theta}$, and quantities such as $E_{\boldsymbol{\theta}}(g(Y))$ depend on $\boldsymbol{\theta}$. Since the parametric distribution is completely specified by $\boldsymbol{\theta}$, an important goal of parametric inference is finding good estimators of $\boldsymbol{\theta}$. For example, if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\boldsymbol{\theta} = (\mu, \sigma)$ is fixed but unknown, $\boldsymbol{\theta} \in \Theta = (-\infty, \infty) \times (0, \infty)$ and $E_{\boldsymbol{\theta}}(\overline{Y}) \equiv E_{(\mu,\sigma)}(\overline{Y}) = \mu$. Since $V_{\boldsymbol{\theta}}(\overline{Y}) \equiv V_{(\mu,\sigma)}(\overline{Y}) = \sigma^2/n$, $\overline{Y}$ is a good estimator for μ if n is large. The notation $f_{\boldsymbol{\theta}}(\boldsymbol{y}) \equiv f(\boldsymbol{y}|\boldsymbol{\theta})$ is also used.

The basic idea of a sufficient statistic $\boldsymbol{T}(\boldsymbol{Y})$ for $\boldsymbol{\theta}$ is that all of the information needed for inference from the data $Y_1, ..., Y_n$ about the parameter $\boldsymbol{\theta}$ is contained in the statistic $\boldsymbol{T}(\boldsymbol{Y})$. For example, suppose that $Y_1, ..., Y_n$ are iid binomial$(1, \rho)$ random variables. Hence each observed Y_i is a 0 or a 1 and the observed data is an n–tuple of 0's and 1's, eg 0,0,1,...,0,0,1. It will turn out that $\sum_{i=1}^n Y_i$, the number of 1's in the n–tuple, is a sufficient statistic for ρ. From Theorem 2.17a, $\sum_{i=1}^n Y_i \sim$ BIN(n, ρ). The importance of a sufficient statistic is *dimension reduction*: the statistic $\sum_{i=1}^n Y_i$ has all of the information from the data needed to perform inference about ρ, and the statistic is one dimensional and thus much easier to understand than the n dimensional n–tuple of 0's and 1's. Also notice that all n–tuples with the same number of 1's have the same amount of information needed for inference about ρ: the n–tuples 1,1,1,0,0,0,0 and 0,1,0,0,1,0,1 both give $\sum_{i=1}^n Y_i = 3$.

Definition 4.2. Suppose that $(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ have a joint distribution that depends on a vector of parameters $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta$ where Θ is the parameter space. A statistic $\boldsymbol{T}(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ is a **sufficient statistic** for $\boldsymbol{\theta}$ if the conditional distribution of $(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n)$ given $\boldsymbol{T} = \boldsymbol{t}$ does not depend on $\boldsymbol{\theta}$ for any value of $\boldsymbol{t}$ in the support of $\boldsymbol{T}$.

Example 4.5. Suppose $T(\boldsymbol{y}) \equiv 7 \ \forall \boldsymbol{y}$. Then T is a constant and any constant is independent of a random vector $\boldsymbol{Y}$. Hence the conditional distribution $f_{\boldsymbol{\theta}}(\boldsymbol{y}|T) = f_{\boldsymbol{\theta}}(\boldsymbol{y})$ is not independent of $\boldsymbol{\theta}$. Thus T is not a sufficient statistic.

Often $\boldsymbol{T}$ and $\boldsymbol{Y}_i$ are real valued. Then $T(Y_1, ..., Y_n)$ is a sufficient statistic if the conditional distribution of $\boldsymbol{Y} = (Y_1, ..., Y_n)$ given $T = t$ does not depend on $\boldsymbol{\theta}$. The following theorem provides such an effective method for showing that a statistic is a sufficient statistic that the definition should rarely be used to prove that the statistic is a sufficient statistic.

Regularity Condition F.1: If $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a family of pmfs for $\boldsymbol{\theta} \in \Theta$, assume that there exists a set $\{\boldsymbol{y}_i\}_{i=1}^{\infty}$ that does not depend on $\boldsymbol{\theta} \in \Theta$ such that $\sum_{i=1}^{\infty} f(\boldsymbol{y}_i|\theta) = 1$ for all $\boldsymbol{\theta} \in \Theta$. (This condition is usually satisfied. For example, F.1 holds if the support $\mathcal{Y}$ is free of $\boldsymbol{\theta}$ or if $\boldsymbol{y} = (y_1, ..., y_n)$ and y_i takes on values on a lattice such as $y_i \in \{1, ..., \theta\}$ for $\theta \in \{1, 2, 3, ...\}$.)

Theorem 4.3: Factorization Theorem. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$ denote a family of pdfs or pmfs for $\boldsymbol{Y}$. For a family of pmfs, assume condition F.1 holds. A statistic $\boldsymbol{T}(\boldsymbol{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$ iff for all sample points $\boldsymbol{y}$ and for all $\boldsymbol{\theta}$ in the parameter space Θ,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})\ h(\boldsymbol{y})$$

where both g and h are nonnegative functions. The function h does not depend on $\boldsymbol{\theta}$ and the function g depends on $\boldsymbol{y}$ only through $\boldsymbol{T}(\boldsymbol{y})$.

Proof for pmfs. If $\boldsymbol{T}(\boldsymbol{Y})$ is a sufficient statistic, then the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t}$ does not depend on $\boldsymbol{\theta}$ for any $\boldsymbol{t}$ in the support of $\boldsymbol{T}$. Taking $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{y})$ gives

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})) \equiv P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y}))$$

for all $\boldsymbol{\theta}$ in the parameter space. Now

$$\{\boldsymbol{Y} = \boldsymbol{y}\} \subseteq \{\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})\} \tag{4.7}$$

and $P(A) = P(A \cap B)$ if $A \subseteq B$. Hence

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) = P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \text{ and } \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y}))$$

$$= P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y}))P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{y})) = g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})h(\boldsymbol{y}).$$

Now suppose

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})\ h(\boldsymbol{y})$$

for all $\boldsymbol{y}$ and for all $\boldsymbol{\theta} \in \Theta$. Now

$$P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t}) = \sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} f(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{t}|\boldsymbol{\theta}) \sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y}).$$

If $\boldsymbol{Y} = \boldsymbol{y}$ and $T(\boldsymbol{Y}) = \boldsymbol{t}$, then $T(\boldsymbol{y}) = \boldsymbol{t}$ and $\{\boldsymbol{Y} = \boldsymbol{y}\} \subseteq \{\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t}\}$. Thus

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t}) = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})}{P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})} = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})}{P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{t})}$$

$$= \frac{g(\boldsymbol{t}|\boldsymbol{\theta})\ h(\boldsymbol{y})}{g(\boldsymbol{t}|\boldsymbol{\theta}) \sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y})} = \frac{h(\boldsymbol{y})}{\sum_{\{\boldsymbol{y}:\boldsymbol{T}(\boldsymbol{y})=\boldsymbol{t}\}} h(\boldsymbol{y})}$$

which does not depend on $\boldsymbol{\theta}$ since the terms in the sum do not depend on $\boldsymbol{\theta}$ by condition F.1. Hence $\boldsymbol{T}$ is a sufficient statistic. QED

Remark 4.1. If no such factorization exists for $\boldsymbol{T}$, then $\boldsymbol{T}$ is not a sufficient statistic.

Example 4.6. To use factorization to show that the data $\boldsymbol{Y} = (Y_1, ..., Y_n)$ is a sufficient statistic, take $\boldsymbol{T}(\boldsymbol{Y}) = \boldsymbol{Y}$, $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})$, and $h(\boldsymbol{y}) = 1$ $\forall \boldsymbol{y}$.

Example 4.7. Let $X_1, ..., X_n$ be iid $N(\mu, \sigma^2)$. Then

$$f(x_1, ..., x_n) = \prod_{i=1}^n f(x_i) = \left[\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\mu^2}{2\sigma^2})\right]^n \exp(\frac{-1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i)$$

$$= g(T(\boldsymbol{x})|\boldsymbol{\theta})h(\boldsymbol{x})$$

where $\boldsymbol{\theta} = (\mu, \sigma)$ and $h(\boldsymbol{x}) = 1$. Hence $T(\boldsymbol{X}) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is a sufficient statistic for (μ, σ) or equivalently for (μ, σ^2) by the factorization theorem.

Example 4.8. Let $Y_1, ..., Y_n$ be iid binomial(k, ρ) with k known and pmf

$$f(y|\rho) = \binom{k}{y} \rho^y (1-\rho)^{k-y}\ I_{\{0,...,k\}}(y).$$

Then

$$f(\boldsymbol{y}|\rho) = \prod_{i=1}^n f(y_i|\rho) = \prod_{i=1}^n \left[\binom{k}{y_i} I_{\{0,...,k\}}(y_i)\right] (1-\rho)^{nk} \left(\frac{\rho}{1-\rho}\right)^{\sum_{i=1}^n y_i}.$$

Hence by the factorization theorem, $\sum_{i=1}^n Y_i$ is a sufficient statistic.

Example 4.9. Suppose $X_1, ..., X_n$ are iid uniform observations on the interval $(\theta, \theta+1)$, $-\infty < \theta < \infty$. Notice that

$$\prod_{i=1}^n I_A(x_i) = I(\text{all } \mathrm{x_i} \in \mathrm{A}) \quad \text{and} \quad \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{I_{A_n}(x)} = \mathrm{I_{\cap_{i=1}^n A_i}(x)}$$

where the latter holds since both terms are 1 if x is in all sets A_i for $i = 1, ..., n$ and both terms are 0 otherwise. Hence $f(\boldsymbol{x}|\theta) =$

$$\prod_{i=1}^n f(x_i|\theta) = \prod_{i=1}^n 1I(x_i \geq \theta)I(x_i \leq \theta+1) = 1I(\min(x_i) \geq \theta)I(\max(x_i) \leq \theta+1).$$

Then $h(\boldsymbol{x}) \equiv 1$ and $g(\boldsymbol{T}(\boldsymbol{x})|\theta) = I(\min(x_i) \geq \theta)I(\max(x_i) \leq \theta+1)$, and $\boldsymbol{T}(\boldsymbol{x}) = (X_{(1)}, X_{(n)})$ is a sufficient statistic by the factorization theorem.

Remark 4.2. i) Suppose $Y_1, ..., Y_n$ are iid from a distribution with support $\mathcal{Y}_i \equiv \mathcal{Y}^*$ and pdf or pmf $f(y|\boldsymbol{\theta}) = k(y|\boldsymbol{\theta})I(y \in \mathcal{Y}^*)$. Then $f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^n k(y_i|\boldsymbol{\theta}) \prod_{i=1}^n I(y_i \in \mathcal{Y}^*)$. Now the support of $\boldsymbol{Y}$ is the n-fold cross product $\mathcal{Y} = \mathcal{Y}^* \times \cdots \times \mathcal{Y}^*$, and $I(\boldsymbol{y} \in \mathcal{Y}) = \prod_{i=1}^n I(y_i \in \mathcal{Y}^*) =$
$I(\text{ all } y_i \in \mathcal{Y}^*)$. Thus $f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^n k(y_i|\boldsymbol{\theta}) I(\text{ all } y_i \in \mathcal{Y}^*)$.

ii) If $\mathcal{Y}^*$ does not depend on $\boldsymbol{\theta}$, then $I(\text{ all } y_i \in \mathcal{Y}^*)$ is part of $h(\boldsymbol{y})$. If $\mathcal{Y}^*$ does depend on unknown $\boldsymbol{\theta}$, then $I(\text{ all } y_i \in \mathcal{Y}^*)$ could be placed in $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$. Typically $\mathcal{Y}^*$ is an interval with endpoints a and b, not necessarily finite. For pdfs, $\prod_{i=1}^n I(y_i \in [a,b]) = I(a \leq y_{(1)} < y_{(n)} \leq b) = I[a \leq y_{(1)}]I[y_{(n)} \leq b]$. If both a and b are unknown parameters, put the middle term in $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$. If both a and b are known, put the middle term in $h(\boldsymbol{y})$. If a is an unknown parameter and b is known, put $I[a \leq y_{(1)}]$ in $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$ and $I[y_{(n)} \leq b]$ in $h(\boldsymbol{y})$.

iii) $\prod_{i=1}^n I(y_i \in (-\infty, b)) = I(y_{(n)} < b)$.
$\prod_{i=1}^n I(y_i \in [a, \infty)) = I(a \leq y_{(1)})$ et cetera.

iv) Another useful fact is that $\prod_{j=1}^k I(\boldsymbol{y} \in A_j) = I(\boldsymbol{y} \in \cap_{j=1} A_j)$.

Example 4.10. Try to place any part of $f(\boldsymbol{y}|\boldsymbol{\theta})$ that depends on $\boldsymbol{y}$ but not on $\boldsymbol{\theta}$ into $h(\boldsymbol{y})$. For example, if $Y_1, ..., Y_n$ are iid $U(\theta_1, \theta_2)$, then $f(\boldsymbol{y}|\boldsymbol{\theta}) =$

$$\prod_{i=1}^n f(y_i|\boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y_i \leq \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 \leq y_{(1)} < y_{(n)} \leq \theta_2).$$

Then $h(\boldsymbol{y}) \equiv 1$ and $\boldsymbol{T}(\boldsymbol{y}) = (Y_{(1)}, Y_{(n)})$ is a sufficient statistic for (θ_1, θ_2) by factorization.

If θ_1 or θ_2 is known, then the above factorization works, but it is better to make the dimension of the sufficient statistic as small as possible. If θ_1 is known, then

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{(\theta_2 - \theta_1)^n} I(y_{(n)} \leq \theta_2) I(\theta_1 \leq y_{(1)})$$

where the first two terms are $g(T(\boldsymbol{y})|\theta_2)$ and the third term is $h(\boldsymbol{y})$. Hence $T(\boldsymbol{Y}) = Y_{(n)}$ is a sufficient statistic for θ_2 by factorization. If θ_2 is known, then

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 \leq y_{(1)}) I(y_{(n)} \leq \theta_2)$$

where the first two terms are $g(T(\boldsymbol{y})|\theta_1)$ and the third term is $h(\boldsymbol{y})$. Hence $T(\boldsymbol{Y}) = Y_{(1)}$ is a sufficient statistic for θ_1 by factorization.

There are infinitely many sufficient statistics (see Theorem 4.8 below), but typically we want the dimension of the sufficient statistic to be as small as possible since lower dimensional statistics are easier to understand and to use for inference than higher dimensional statistics. Data reduction is extremely important and the following definition is useful.

Definition 4.3. Suppose that $Y_1, ..., Y_n$ have a joint distribution that depends on a vector of parameters $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta$ where Θ is the parameter space. A sufficient statistic $\boldsymbol{T}(\boldsymbol{Y})$ for $\boldsymbol{\theta}$ is a **minimal sufficient statistic** for $\boldsymbol{\theta}$ if $\boldsymbol{T}(\boldsymbol{Y})$ is a function of $\boldsymbol{S}(\boldsymbol{Y})$ for any other sufficient statistic $\boldsymbol{S}(\boldsymbol{Y})$ for $\boldsymbol{\theta}$.

Remark 4.3. A useful mnemonic is that $\boldsymbol{S} = \boldsymbol{Y}$ is a sufficient statistic, and $\boldsymbol{T} \equiv \boldsymbol{T}(\boldsymbol{Y})$ is a function of $\boldsymbol{S}$.

Warning: Complete sufficient statistics, defined below, are primarily used for the theory of uniformly minimum variance estimators, which are rarely used in applied work unless they are nearly identical to the corresponding maximum likelihood estimators.

Definition 4.4. Suppose that a *statistic* $\boldsymbol{T}(\boldsymbol{Y})$ has a pmf or pdf $f(\boldsymbol{t}|\boldsymbol{\theta})$. Then $\boldsymbol{T}(\boldsymbol{Y})$ is a *complete sufficient statistic* for $\boldsymbol{\theta}$ if $E_{\boldsymbol{\theta}}[g(\boldsymbol{T}(\boldsymbol{Y}))] = 0$ for all $\boldsymbol{\theta}$ implies that $P_{\boldsymbol{\theta}}[g(\boldsymbol{T}(\boldsymbol{Y})) = 0] = 1$ for all $\boldsymbol{\theta}$. The function g may not depend on any unknown parameters.

The following two theorems are useful for finding minimal sufficient statistics.

Theorem 4.4: Lehmann-Scheffé Theorem for Minimal Sufficient Statistics (LSM). Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pmf or pdf of an iid sample $\boldsymbol{Y}$. Let $c_{\boldsymbol{x},\boldsymbol{y}}$ be a constant. Suppose there exists a function $\boldsymbol{T}(\boldsymbol{y})$ such that for any two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $R_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})/f(\boldsymbol{y}|\boldsymbol{\theta}) = c_{\boldsymbol{x},\boldsymbol{y}}$ for all $\boldsymbol{\theta}$ in Θ iff $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$. Then $\boldsymbol{T}(\boldsymbol{Y})$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.

In the Lehmann-Scheffé Theorem, for R to be constant as a function of $\boldsymbol{\theta}$, define $0/0 = c_{\boldsymbol{x},\boldsymbol{y}}$. Alternatively, replace $R_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})/f(\boldsymbol{y}|\boldsymbol{\theta}) = c_{\boldsymbol{x},\boldsymbol{y}}$ by $f(\boldsymbol{x}|\boldsymbol{\theta}) = c_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{y}|\boldsymbol{\theta})$ in the above definition.

Finding sufficient, minimal sufficient, and complete sufficient statistics is often simple for a kP-REF (k-parameter regular exponential family). **If the family given by Equation (4.8) is a kP-REF, then the conditions for Theorem 4.5abcd are satisfied as are the conditions for e)** if $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$. In a), k does not need to be as small as possible. In Corollary 4.6 below, assume that both Equation (4.8) and (4.9) hold.

Note that any one to one function is onto its range. Hence if $\boldsymbol{\eta} = \tau(\boldsymbol{\theta})$ for any $\boldsymbol{\eta} \in \Omega$ where τ is a one to one function, then $\tau : \Theta \to \Omega$ is one to one and onto. Thus there is a one to one (and onto) inverse function τ^{-1} such that $\boldsymbol{\theta} = \tau^{-1}(\boldsymbol{\eta})$ for any $\boldsymbol{\theta} \in \Theta$.

Theorem 4.5: Sufficiency, Minimal Sufficiency, and Completeness of Exponential Families. Suppose that $Y_1, ..., Y_n$ are iid from an exponential family

$$f(y|\boldsymbol{\theta}) = h(y)c(\boldsymbol{\theta}) \exp\left[w_1(\boldsymbol{\theta})t_1(y) + \cdots + w_k(\boldsymbol{\theta})t_k(y)\right] \qquad (4.8)$$

with the natural parameterization

$$f(y|\boldsymbol{\eta}) = h(y)b(\boldsymbol{\eta}) \exp\left[\eta_1 t_1(y) + \cdots + \eta_k t_k(y)\right] \qquad (4.9)$$

so that the joint pdf or pmf is given by

$$f(y_1, ..., y_n|\boldsymbol{\eta}) = (\prod_{j=1}^{n} h(y_j))[b(\boldsymbol{\eta})]^n \exp[\eta_1 \sum_{j=1}^{n} t_1(y_j) + \cdots + \eta_k \sum_{j=1}^{n} t_k(y_j)]$$

which is a k-parameter exponential family. Then

$$\boldsymbol{T}(\boldsymbol{Y}) = (\sum_{j=1}^{n} t_1(Y_j), ..., \sum_{j=1}^{n} t_k(Y_j)) \text{ is}$$

a) a sufficient statistic for $\boldsymbol{\theta}$ and for $\boldsymbol{\eta}$,
b) a minimal sufficient statistic for $\boldsymbol{\eta}$ if $\eta_1, ..., \eta_k$ do not satisfy a linearity constraint,
c) a minimal sufficient statistic for $\boldsymbol{\theta}$ if $w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta})$ do not satisfy a linearity constraint,
d) a complete sufficient statistic for $\boldsymbol{\eta}$ if Ω contains a k–dimensional rectangle,
e) a complete sufficient statistic for $\boldsymbol{\theta}$ if $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$ and if Ω contains a k–dimensional rectangle.

Proof. a) Use the factorization theorem.
b) The proof expands on remarks given in Johanson (1979, p. 3) and Lehmann(1983, p. 44). The ratio

$$\frac{f(\boldsymbol{x}|\boldsymbol{\eta})}{f(\boldsymbol{y}|\boldsymbol{\eta})} = \frac{\prod_{j=1}^{n} h(x_j)}{\prod_{j=1}^{n} h(y_j)} \exp[\sum_{i=1}^{k} \eta_i (T_i(\boldsymbol{x}) - T_i(\boldsymbol{y}))]$$

is equal to a constant with respect to $\boldsymbol{\eta}$ iff

$$\sum_{i=1}^{k} \eta_i [T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})] = \sum_{i=1}^{k} \eta_i a_i = d$$

for all η_i where d is some constant and where $a_i = T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})$ and $T_i(\boldsymbol{x}) = \sum_{j=1}^{n} t_i(x_j)$. Since the η_i do not satisfy a linearity constraint, $\sum_{i=1}^{k} \eta_i a_i = d$ for all $\boldsymbol{\eta}$ iff all of the $a_i = 0$. Hence

$$\boldsymbol{T}(\boldsymbol{Y}) = (T_1(\boldsymbol{Y}), ..., T_k(\boldsymbol{Y}))$$

is a minimal sufficient statistic by the Lehmann-Scheffé LSM theorem.
c) Use almost the same proof as b) with $w_i(\boldsymbol{\theta})$ in the place of η_i and $\boldsymbol{\theta}$ in the place of $\boldsymbol{\eta}$. (In particular, the result holds if $\eta_i = w_i(\boldsymbol{\theta})$ for $i = 1, ..., k$ provided that the η_i do not satisfy a linearity constraint.)
d) See Lehmann (1986, p. 142).
e) If $\boldsymbol{\eta} = \tau(\boldsymbol{\theta})$ then $\boldsymbol{\theta} = \tau^{-1}(\boldsymbol{\eta})$ and the parameters have just been renamed.

Hence $E_{\boldsymbol{\theta}}[g(\boldsymbol{T})] = 0$ for all $\boldsymbol{\theta}$ implies that $E_{\boldsymbol{\eta}}[g(\boldsymbol{T})] = 0$ for all $\boldsymbol{\eta}$, and thus $P_{\boldsymbol{\eta}}[g(\boldsymbol{T}(\boldsymbol{Y})) = 0] = 1$ for all $\boldsymbol{\eta}$ since $\boldsymbol{T}$ is a complete sufficient statistic for $\boldsymbol{\eta}$ by d). Thus $P_{\boldsymbol{\theta}}[g(\boldsymbol{T}(\boldsymbol{Y})) = 0] = 1$ for all $\boldsymbol{\theta}$, and $\boldsymbol{T}$ is a complete sufficient statistic for $\boldsymbol{\theta}$.

Corollary 4.6: Completeness of a kP–REF. Suppose that $Y_1, ..., Y_n$ are iid from a kP–REF (k-parameter regular exponential family)

$$f(y|\boldsymbol{\theta}) = h(y)c(\boldsymbol{\theta}) \exp\left[w_1(\boldsymbol{\theta})t_1(y) + \cdots + w_k(\boldsymbol{\theta})t_k(y)\right]$$

with $\boldsymbol{\theta} \in \Theta$ and natural parameterization given by (4.9) with $\boldsymbol{\eta} \in \Omega$. Then

$$\boldsymbol{T}(\boldsymbol{Y}) = (\sum_{j=1}^{n} t_1(Y_j), ..., \sum_{j=1}^{n} t_k(Y_j)) \text{ is}$$

a) a minimal sufficient statistic for $\boldsymbol{\theta}$ and for $\boldsymbol{\eta}$,
b) a complete sufficient statistic for $\boldsymbol{\theta}$ and for $\boldsymbol{\eta}$ if $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$.

Proof. The result follows by Theorem 4.5 since for a kP–REF, the $w_i(\boldsymbol{\theta})$ and η_i do not satisfy a linearity constraint and Ω contains a k–dimensional rectangle. QED

Theorem 4.7: Bahadur's Theorem. A finite dimensional complete sufficient statistic is also minimal sufficient.

Theorem 4.8. A one to one function of a sufficient, minimal sufficient, or complete sufficient statistic is sufficient, minimal sufficient, or complete sufficient respectively.

Note that in a kP–REF, the statistic $\boldsymbol{T}$ is k–dimensional and thus $\boldsymbol{T}$ is minimal sufficient by Theorem 4.7 if $\boldsymbol{T}$ is complete sufficient. Corollary 4.6 is useful because often you know or can show that the given family is a REF. The theorem gives a particularly simple way to find complete sufficient statistics for one parameter exponential families and for any family that is known to be REF. If it is known that the distribution is regular, find the exponential family parameterization given by Equation (4.8) or (4.9). These parameterizations give $t_1(y), ..., t_k(y)$. Then $\boldsymbol{T}(\boldsymbol{Y}) = (\sum_{j=1}^{n} t_1(Y_j), ..., \sum_{j=1}^{n} t_k(Y_j))$.

Example 4.11. Let $X_1, ..., X_n$ be iid $N(\mu, \sigma^2)$. Then the $N(\mu, \sigma^2)$ pdf is

$$f(x|\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\mu^2}{2\sigma^2})}_{c(\mu,\sigma)\geq 0} \exp(\underbrace{\frac{-1}{2\sigma^2}}_{w_1(\boldsymbol{\theta})} \underbrace{x^2}_{t_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{w_2(\boldsymbol{\theta})} \underbrace{x}_{t_2(x)}) \underbrace{I_{\Re}(x)}_{h(x)\geq 0},$$

with $\eta_1 = -0.5/\sigma^2$ and $\eta_2 = \mu/\sigma^2$ if $\sigma > 0$. As shown in Example 3.1, this is a 2P–REF. By Corollary 4.6, $\boldsymbol{T} = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a complete sufficient statistic for (μ, σ^2). The one to one functions

$$\boldsymbol{T}_2 = (\overline{X}, S^2) \text{ and } \boldsymbol{T}_3 = (\overline{X}, S)$$

of $\boldsymbol{T}$ are also complete sufficient where $\overline{X}$ is the sample mean and S is the sample standard deviation. $\boldsymbol{T}, \boldsymbol{T}_2$ and $\boldsymbol{T}_3$ are minimal sufficient by Corollary 4.6 or by Theorem 4.7 since the statistics are 2 dimensional.

Example 4.12. Let $Y_1, ..., Y_n$ be iid binomial(k, ρ) with k known and pmf

$$f(y|\rho) = \binom{k}{y} \rho^y (1-\rho)^{k-y} \; I_{\{0,...,k\}}(y)$$

$$= \underbrace{\binom{k}{y} I_{\{0,...,k\}}(y)}_{h(y)\geq 0} \underbrace{(1-\rho)^k}_{c(\rho)\geq 0} \exp[\underbrace{\log(\frac{\rho}{1-\rho})}_{w(\rho)} \underbrace{y}_{t(y)}]$$

where $\Theta = (0, 1)$ and $\Omega = (-\infty, \infty)$. Notice that $\eta = \log(\frac{\rho}{1-\rho})$ is an increasing and hence one to one function of ρ. Since this family is a 1P–REF, $T_n = \sum_{i=1}^n t(Y_i) = \sum_{i=1}^n Y_i$ is complete sufficient statistic for ρ.

Compare Examples 4.7 and 4.8 with Examples 4.11 and 4.12. The exponential family theorem gives more powerful results than the factorization theorem, but often the factorization theorem is useful for suggesting a potential minimal sufficient statistic.

Example 4.13. In testing theory, a single sample is often created by combining two independent samples of iid data. Let $X_1, ..., X_n$ be iid exponential (θ) and $Y_1, ..., Y_m$ iid exponential $(\theta/2)$. If the two samples are independent, then the joint pdf $f(\boldsymbol{x}, \boldsymbol{y}|\theta)$ belongs to a regular one parameter exponential family with complete sufficient statistic $T = \sum_{i=1}^n X_i + 2 \sum_{i=1}^m Y_i$. (Let $W_i = 2Y_i$. Then the W_i and X_i are iid and Corollary 4.6 applies.)

Rule of thumb 4.1: A k–parameter minimal sufficient statistic for a d–dimensional parameter where $d < k$ will not be complete. In the following example $d = 1 < 2 = k$. (A rule of thumb is something that is frequently true but can not be used to rigorously prove something. Hence this rule of thumb can not be used to prove that the minimal sufficient statistic is not complete.)

Warning: Showing that a minimal sufficient statistic is not complete is of little applied interest since complete sufficient statistics are rarely used in applications; nevertheless, many qualifying exams in statistical inference contain such a problem.

Example 4.14, Cox and Hinckley (1974, p. 31). Let $X_1, ..., X_n$ be iid $N(\mu, \gamma_o^2 \mu^2)$ random variables where $\gamma_o^2 > 0$ is *known* and $\mu > 0$. Then this family has a one dimensional parameter μ, but

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\gamma_o^2\mu^2}} \exp\left(\frac{-1}{2\gamma_o^2}\right) \exp\left(\frac{-1}{2\gamma_o^2\mu^2}x^2 + \frac{1}{\gamma_o^2\mu}x\right)$$

is a two parameter exponential family with $\Theta = (0, \infty)$ (which contains a one dimensional rectangle), and $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a minimal sufficient statistic. (Theorem 4.5 applies since the functions $1/\mu$ and $1/\mu^2$ do not satisfy a linearity constraint.) However, since $E_\mu(X^2) = \gamma_o^2\mu^2 + \mu^2$ and $\sum_{i=1}^n X_i \sim N(n\mu, n\gamma_o^2\mu^2)$ implies that

$$E_\mu[(\sum_{i=1}^n X_i)^2] = n\gamma_o^2\mu^2 + n^2\mu^2,$$

we find that

$$E_\mu[\frac{n+\gamma_o^2}{1+\gamma_o^2}\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2] = \frac{n+\gamma_o^2}{1+\gamma_o^2} n\mu^2(1+\gamma_o^2) - (n\gamma_o^2\mu^2 + n^2\mu^2) = 0$$

for all μ so the minimal sufficient statistic is not complete. Notice that

$$\Omega = \{(\eta_1, \eta_2) : \eta_1 = \frac{-1}{2}\gamma_o^2\eta_2^2\}$$

and a plot of η_1 versus η_2 is a quadratic function which can not contain a 2–dimensional rectangle. Notice that (η_1, η_2) is a one to one function of μ, and thus this example illustrates that the rectangle needs to be contained in Ω rather than Θ.

Example 4.15. The theory does not say that any sufficient statistic from a REF is complete. Let Y be a random variable from a normal $N(0, \sigma^2)$ distribution with $\sigma^2 > 0$. This family is a REF with complete minimal sufficient statistic Y^2. The data Y is also a sufficient statistic, but Y is not a function of Y^2. Hence Y is not minimal sufficient and (by Bahadur's theorem) not complete. Alternatively $E_{\sigma^2}(Y) = 0$ but $P_{\sigma^2}(Y = 0) = 0 < 1$, so Y is not complete.

Theorem 4.9. Let $Y_1, ..., Y_n$ be iid.

a) If $Y_i \sim U(\theta_1, \theta_2)$, then $(Y_{(1)}, Y_{(n)})$ is a complete sufficient statistic for (θ_1, θ_2). See David (1981, p. 123.)

b) If $Y_i \sim U(\theta_1, \theta_2)$ with θ_1 known, then $Y_{(n)}$ is a complete sufficient statistic for θ_2.

c) If $Y_i \sim U(\theta_1, \theta_2)$ with θ_2 known, then $Y_{(1)}$ is a complete sufficient statistic for θ_1.

d) If $Y_i \sim U(-\theta, \theta)$, then $\max(|Y_i|)$ is a complete sufficient statistic for θ.

e) If $Y_i \sim EXP(\theta, \lambda)$, then $(Y_{(1)}, \overline{Y})$ is a complete sufficient statistic for (θ, λ). See David (1981, p. 153-154.)

f) If $Y_i \sim EXP(\theta, \lambda)$ with λ known, then $Y_{(1)}$ is a complete sufficient statistic for θ.

g) If $Y_i \sim$ Cauchy(μ, σ) with σ known, then the order statistics are minimal sufficient.

h) If $Y_i \sim$ Double Exponential(θ, λ) with λ known, then the order statistics $(Y_{(1)}, ..., Y_{(n)})$ are minimal sufficient.

i) If $Y_i \sim$ logistic(μ, σ), then the order statistics are minimal sufficient.

j) If $Y_i \sim$ Weibull(ϕ, λ), then the order statistics $(Y_{(1)}, ..., Y_{(n)})$ are minimal sufficient.

A **common midterm, final and qual question** takes $X_1, ..., X_n$ iid $U(h_l(\theta), h_u(\theta))$ where h_l and h_u are functions of θ such that $h_l(\theta) < h_u(\theta)$. The function h_l and h_u are chosen so that the min $= X_{(1)}$ and the max $= X_{(n)}$ form the 2-dimensional minimal sufficient statistic by the LSM theorem. Since θ is one dimensional, the rule of thumb suggests that the minimal sufficient statistic is not complete. State this fact, but if you have time find $E_\theta[X_{(1)}]$ and $E_\theta[X_{(n)}]$. Then show that $E_\theta[aX_{(1)} + bX_{(n)} + c] \equiv 0$ so that $\boldsymbol{T} = (X_{(1)}, X_{(n)})$ is not complete.

Example 4.16. The uniform distribution is tricky since usually $(X_{(1)}, X_{(n)})$ is minimal sufficient by the LSM theorem, since

$$f(\boldsymbol{x}) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 < x_{(1)} < x_{(n)} < \theta_2)$$

if $n > 1$. But occasionally θ_1 and θ_2 are functions of θ such that the indicator can be written as $I(\theta > T)$ or $I(\theta < T)$ where the minimal sufficient statistic T is a one dimensional function of $(X_{(1)}, X_{(n)})$. If $X \sim U(c_1 + d_1\theta, c_2 + d_2\theta)$ where $d_1 < 0$ and $d_2 > \theta$, then

$$T = \max(\frac{X_{(1)} - c_1}{d_1}, \frac{X_{(n)} - c_2}{d_2})$$

is minimal sufficient.

Let $X_1, ..., X_n$ be iid $U(1-\theta, 1+\theta)$ where $\theta > 0$ is unknown. Hence

$$f_X(x) = \frac{1}{2\theta} I(1 - \theta < x < 1 + \theta)$$

and

$$\frac{f(\boldsymbol{x})}{f(\boldsymbol{y})} = \frac{I(1 - \theta < x_{(1)} \leq x_{(n)} < 1 + \theta)}{I(1 - \theta < y_{(1)} \leq y_{(n)} < 1 + \theta)}.$$

This ratio may look to be constant for all $\theta > 0$ iff $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$, but it is not. To show that $\boldsymbol{T}_1 = (X_{(1)}, X_{(n)})$ is not a minimal sufficient statistic, note that

$$f_X(x) = \frac{1}{2\theta} I(\theta > 1 - x) I(\theta > x - 1).$$

Hence

$$\frac{f(\boldsymbol{x})}{f(\boldsymbol{y})} = \frac{I(\theta > \max(1 - x_{(1)}, x_{(n)} - 1))}{I(\theta > \max(1 - y_{(1)}, y_{(n)} - 1))}$$

which is constant for all $\theta > 0$ iff $T_2(\boldsymbol{x}) = T_2(\boldsymbol{y})$ where $T_2(\boldsymbol{x}) = \max(1 - x_{(1)}, x_{(n)} - 1)$. Hence $T_2 = T_2(\boldsymbol{X}) = \max(1 - X_{(1)}, X_{(n)} - 1)$ is minimal sufficient by the LSM theorem. Thus $\boldsymbol{T}_1$ is not a minimal sufficient statistic (and so not complete) since $\boldsymbol{T}_1$ is not a function of T_2.

To show that $\boldsymbol{T}_1$ is not complete using the definition of complete statistic, first find $E(\boldsymbol{T}_1)$. Now

$$F_X(t) = \int_{1-\theta}^{t} \frac{1}{2\theta} dx = \frac{t + \theta - 1}{2\theta}$$

for $1-\theta < t < 1+\theta$. Hence by Theorem 4.2a),

$$f_{X_{(n)}}(t) = \frac{n}{2\theta}\left(\frac{t+\theta-1}{2\theta}\right)^{n-1}$$

for $1-\theta < t < 1+\theta$ and

$$E_\theta(X_{(n)}) = \int x f_{X_{(n)}}(x)dx = \int_{1-\theta}^{1+\theta} x\frac{n}{2\theta}\left(\frac{x+\theta-1}{2\theta}\right)^{n-1} dx.$$

Use u–substitution with $u = (x+\theta-1)/2\theta$ and $x = 2\theta u + 1 - \theta$. Hence $x = 1+\theta$ implies $u = 1$, and $x = 1-\theta$ implies $u = 0$ and $dx = 2\theta du$. Thus

$$E_\theta(X_{(n)}) = n\int_0^1 \frac{2\theta u + 1 - \theta}{2\theta} u^{n-1} 2\theta du =$$

$$= n\int_0^1 [2\theta u + 1 - \theta]u^{n-1} du = 2\theta n\int_0^n u^n du + (n - n\theta)\int_0^1 u^{n-1} du =$$

$$2\theta n \frac{u^{n+1}}{n+1}\bigg|_0^1 + n(1-\theta)\frac{u^n}{n}\bigg|_0^1 =$$

$$2\theta\frac{n}{n+1} + \frac{n(1-\theta)}{n} = 1 - \theta + 2\theta\frac{n}{n+1}.$$

Note that $E_\theta(X_{(n)}) \approx 1+\theta$ as you should expect.

By Theorem 4.2b),

$$f_{X_{(1)}}(t) = \frac{n}{2\theta}\left(\frac{\theta-t+1}{2\theta}\right)^{n-1}$$

for $1-\theta < t < 1+\theta$ and thus

$$E_\theta(X_{(1)}) = \int_{1-\theta}^{1+\theta} x\frac{n}{2\theta}\left(\frac{\theta-x+1}{2\theta}\right)^{n-1} dx.$$

Use u–substitution with $u = (\theta - x + 1)/2\theta$ and $x = \theta + 1 - 2\theta u$. Hence $x = 1+\theta$ implies $u = 0$, and $x = 1-\theta$ implies $u = 1$ and $dx = -2\theta du$. Thus

$$E_\theta(X_{(1)}) = \int_1^0 \frac{n}{2\theta}(\theta + 1 - 2\theta u)u^{n-1}(-2\theta)du = n\int_0^1 (\theta + 1 - 2\theta u)u^{n-1} du =$$

$$n(\theta+1)\int_0^1 u^{n-1}du - 2\theta n\int_0^1 u^n du = (\theta+1)n/n - 2\theta n/(n+1) = \theta+1-2\theta\frac{n}{n+1}.$$

To show that $\boldsymbol{T}_1$ is not complete try showing $E_\theta(aX_{(1)} + bX_{(n)} + c) = 0$ for some constants a, b and c. Note that $a = b = 1$ and $c = -2$ works. Hence $E_\theta(X_{(1)} + X_{(n)} - 2) = 0$ for all $\theta > 0$ but $P_\theta(g(\boldsymbol{T}) = 0) = P_\theta(X_{(1)} + X_{(n)} - 2 = 0) = 0 < 1$ for all $\theta > 0$. Hence $\boldsymbol{T}_1$ is not complete.

Definition 4.5. Let $Y_1, ..., Y_n$ have pdf or pmf $f(\boldsymbol{y}|\theta)$. A statistic $\boldsymbol{W}(\boldsymbol{Y})$ whose distribution does not depend on θ is called an **ancillary statistic.**

Theorem 4.10, Basu's Theorem. Let $Y_1, ..., Y_n$ have pdf or pmf $f(\boldsymbol{y}|\theta)$. If $\boldsymbol{T}(\boldsymbol{Y})$ is a k-dimensional complete sufficient statistic, then $\boldsymbol{T}(\boldsymbol{Y})$ is independent of every ancillary statistic.

Remark 4.4. Basu's Theorem says that if $\boldsymbol{T}$ is minimal sufficient and complete, then $\boldsymbol{T} \perp\!\!\!\perp R$ if R is ancillary. Application: If $\boldsymbol{T}$ is minimal sufficient, R ancillary and R is a function of $\boldsymbol{T}$ (so $R = h(\boldsymbol{T})$ is not independent of $\boldsymbol{T}$), then $\boldsymbol{T}$ is not complete. Since θ is a scalar, usually need $k = 1$ for $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{Y}) = T(\boldsymbol{Y}) = T$ to be complete.

Example 4.17. Suppose $X_1, ..., X_n$ are iid uniform observations on the interval $(\theta, \theta + 1)$, $-\infty < \theta < \infty$. Let $X_{(1)} = \min(X_1, ..., X_n)$, $X_{(n)} = \max(X_1, ..., X_n)$ and $\boldsymbol{T}(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ be a minimal sufficient statistic. Then $R = X_{(n)} - X_{(1)}$ is ancillary since $R = max(X_1 - \theta, ..., X_n - \theta) + \theta - [min(X_1 - \theta, ..., X_n - \theta) + \theta] = U_{(n)} - U_{(1)}$ where $U_i = X_i - \theta \sim U(0, 1)$ has a distribution that does not depend on θ. R is not independent of $\boldsymbol{T}$, so $\boldsymbol{T}$ is not complete.

Example 4.18. Let $Y_1, ..., Y_n$ be iid from a location family with pdf $f_Y(y|\theta) = f_X(y - \theta)$ where $Y = X + \theta$ and $f_X(y)$ is the standard pdf for the location family (and thus the distribution of X does not depend on θ).
Claim: $\boldsymbol{W} = (Y_1 - \overline{Y}, ..., Y_n - \overline{Y})$ is ancillary.

Proof: Since $Y_i = X_i + \theta$,

$$\boldsymbol{W} = \left(X_1 + \theta - \frac{1}{n}\sum_{i=1}^n (X_i + \theta), ..., X_n + \theta - \frac{1}{n}\sum_{i=1}^n (X_i + \theta)\right)$$

$$= (X_1 - \overline{X}, ..., X_n - \overline{X})$$

and the distribution of the final vector is free of θ. QED

Application: Let $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$. For any fixed σ^2, this is a location family with $\theta = \mu$ and complete sufficient statistic $T(\boldsymbol{Y}) = \overline{Y}$. Thus $\overline{Y} \perp\!\!\!\perp \boldsymbol{W}$ by Basu's Theorem. Hence $\overline{Y} \perp\!\!\!\perp S^2$ for any known $\sigma^2 > 0$ since

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

is a function of $\boldsymbol{W}$. Thus $\overline{Y} \perp\!\!\!\perp S^2$ even if $\sigma^2 > 0$ is not known.

4.3 Summary

1) A statistic is a function of the data that does not depend on any unknown parameters.

2) For parametric inference, the data $Y_1, ..., Y_n$ comes from a family of parametric distributions $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$. Often the data are iid and $f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$. The parametric distribution is completely specified by the unknown parameters $\boldsymbol{\theta}$. The statistic is a random vector or random variable and hence also comes from some probability distribution. The distribution of the statistic is called the sampling distribution of the statistic.

3) For iid $N(\mu, \sigma^2)$ data, $\overline{Y} \perp\!\!\!\perp S^2$, $\overline{Y} \sim N(\mu, \sigma^2/n)$ and $\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sim \sigma^2 \chi^2_{n-1}$.

4) For iid data with cdf F_Y and pdf f_Y, $f_{Y_{(n)}}(t) = n[F_Y(t)]^{n-1} f_Y(t)$ and $f_{Y_{(1)}}(t) = n[1 - F_Y(t)]^{n-1} f_Y(t)$.

5) A statistic $\boldsymbol{T}(Y_1, ..., Y_n)$ is a *sufficient statistic* for $\boldsymbol{\theta}$ if the conditional distribution of $(Y_1, ..., Y_n)$ given $\boldsymbol{T}$ does not depend on $\boldsymbol{\theta}$.

6) A sufficient statistic $\boldsymbol{T}(\boldsymbol{Y})$ is a *minimal sufficient statistic* if for any other sufficient statistic $\boldsymbol{S}(\boldsymbol{Y})$, $\boldsymbol{T}(\boldsymbol{Y})$ is a function of $\boldsymbol{S}(\boldsymbol{Y})$.

7) Suppose that a *statistic* $\boldsymbol{T}(\boldsymbol{Y})$ has a pmf or pdf $f(\boldsymbol{t}|\boldsymbol{\theta})$. Then $\boldsymbol{T}(\boldsymbol{Y})$ is a *complete statistic* if $E_{\boldsymbol{\theta}}[g(\boldsymbol{T}(\boldsymbol{Y}))] = 0$ for all $\boldsymbol{\theta} \in \Theta$ implies that $P_{\boldsymbol{\theta}}[g(\boldsymbol{T}(\boldsymbol{Y})) = 0] = 1$ for all $\boldsymbol{\theta} \in \Theta$.

8) A one to one function of a sufficient, minimal sufficient, or complete sufficient statistic is sufficient, minimal sufficient, or complete sufficient respectively.

9) **Factorization Theorem.** Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ denote the pdf or pmf of a sample $\boldsymbol{Y}$. A statistic $\boldsymbol{T}(\boldsymbol{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$ iff for all sample

points $\boldsymbol{y}$ and for all $\boldsymbol{\theta}$ in the parameter space Θ,

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})\ h(\boldsymbol{y})$$

where both g and h are nonnegative functions.

Tips: i) for iid data with marginal support $\mathcal{Y}_i \equiv \mathcal{Y}^*$, $I_{\mathcal{Y}}(\boldsymbol{y}) = I(\text{all } y_i \in \mathcal{Y}^*)$. If $\mathcal{Y}^* = (a,b)$, then $I_{\mathcal{Y}}(\boldsymbol{y}) = I(a < y_{(1)} < y_{(n)} < b) = I(a < y_{(1)})I(y_{(n)} < b)$. Put $I(a < y_{(1)})$ in $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$ if a is an unknown parameter but put $I(a < y_{(1)})$ in $h(\boldsymbol{y})$ if a is known. If both a and b are unknown parameters, put $I(a < y_{(1)} < y_{(n)} < b)$ in $g(\boldsymbol{T}(\boldsymbol{y})|\boldsymbol{\theta})$. If $b = \infty$, then $I_{\mathcal{Y}}(\boldsymbol{y}) = I(a < y_{(1)})$. If $\mathcal{Y}^* = [a,b]$, then $I_{\mathcal{Y}}(\boldsymbol{y}) = I(a \le y_{(1)} < y_{(n)} \le b) = I(a \le y_{(1)})I(y_{(n)} \le b)$. ii) Try to make the dimension of $\boldsymbol{T}(\boldsymbol{y})$ as small as possible. Put anything that depends on $\boldsymbol{y}$ but not $\boldsymbol{\theta}$ into $h(\boldsymbol{y})$.

10) **Minimal and complete sufficient statistics for k-parameter exponential families:** Let $Y_1, ..., Y_n$ be iid from an exponential family $f(y|\boldsymbol{\theta}) = h(y)c(\boldsymbol{\theta})\exp[\sum_{j=1}^k w_j(\boldsymbol{\theta})t_j(y)]$ with the natural parameterization $f(y|\boldsymbol{\eta}) = h(y)b(\boldsymbol{\eta})\exp[\sum_{j=1}^k \eta_j t_j(y)]$. Then
$\boldsymbol{T}(\boldsymbol{Y}) = (\sum_{i=1}^n t_1(Y_i), ..., \sum_{i=1}^n t_k(Y_i))$ is

a) a minimal sufficient statistic for $\boldsymbol{\eta}$ if the η_j do not satisfy a linearity constraint and for $\boldsymbol{\theta}$ if the $w_j(\boldsymbol{\theta})$ do not satisfy a linearity constraint.

b) a complete sufficient statistic for $\boldsymbol{\theta}$ and for $\boldsymbol{\eta}$ if $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$ and if Ω contains a k–dimensional rectangle.

11) **Completeness of REFs:** Suppose that $Y_1, ..., Y_n$ are iid from a kP–REF

$$f(y|\boldsymbol{\theta}) = h(y)c(\boldsymbol{\theta})\exp\left[w_1(\boldsymbol{\theta})t_1(y) + \cdots + w_k(\boldsymbol{\theta})t_k(y)\right] \tag{4.10}$$

with $\boldsymbol{\theta} \in \Theta$ and natural parameter $\boldsymbol{\eta} \in \Omega$. Then

$$\boldsymbol{T}(\boldsymbol{Y}) = (\sum_{i=1}^n t_1(Y_i), ..., \sum_{i=1}^n t_k(Y_i)) \quad \text{is}$$

a) a minimal sufficient statistic for $\boldsymbol{\eta}$ and for $\boldsymbol{\theta}$,
b) a complete sufficient statistic for $\boldsymbol{\theta}$ and for $\boldsymbol{\eta}$ if $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$ and if Ω contains a k–dimensional rectangle.

12) For a 2-parameter exponential family ($k = 2$), η_1 and η_2 satisfy a linearity constraint if the plotted points fall on a line in a plot of η_1 versus η_2. If the plotted points fall on a nonlinear curve, then $\boldsymbol{T}$ is minimal sufficient but Ω does not contain a 2-dimensional rectangle.

13) **LSM Theorem:** Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pmf or pdf of a sample $\boldsymbol{Y}$. Let $c_{\boldsymbol{x},\boldsymbol{y}}$ be a constant. Suppose there exists a function $\boldsymbol{T}(\boldsymbol{y})$ such that for any two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $R_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})/f(\boldsymbol{y}|\boldsymbol{\theta}) = c_{\boldsymbol{x},\boldsymbol{y}}$ for all $\boldsymbol{\theta}$ in Θ iff $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$. Then $\boldsymbol{T}(\boldsymbol{Y})$ is a minimal sufficient statistic for $\boldsymbol{\theta}$. (Define $0/0 \equiv c_{\boldsymbol{x},\boldsymbol{y}}$.)

14) *Tips for finding sufficient, minimal sufficient and complete sufficient statistics.* a) Typically $Y_1, ..., Y_n$ are iid so the joint distribution $f(y_1, ..., y_n) = \prod_{i=1}^n f(y_i)$ where $f(y_i)$ is the marginal distribution. Use the **factorization theorem** to find the candidate sufficient statistic $\boldsymbol{T}$.
b) Use factorization to find candidates $\boldsymbol{T}$ that might be minimal sufficient statistics. Try to find $\boldsymbol{T}$ with as small a dimension k as possible. If the support of the random variable depends on θ, often $Y_{(1)}$ or $Y_{(n)}$ will be a component of the minimal sufficient statistic. To prove that $\boldsymbol{T}$ is minimal sufficient, use the **LSM theorem**. **Alternatively prove or recognize that Y comes from a regular exponential family.** $\boldsymbol{T}$ will be minimal sufficient for $\boldsymbol{\theta}$ if Y comes from an exponential family as long as the $w_i(\boldsymbol{\theta})$ do not satisfy a linearity constraint.
c) **To prove that the statistic is complete, prove or recognize that Y comes from a regular exponential family**. Check whether $\dim(\Theta) = k$. If $\dim(\Theta) < k$, then the family is usually not a kP–REF and Theorem 4.5 and Corollary 4.6 do not apply. The uniform distribution where one endpoint is known also has a complete sufficient statistic.
d) Let k be free of the sample size n. Then a $k-$dimensional complete sufficient statistic is also a minimal sufficient statistic (**Bahadur's theorem**).
e) To show that a statistic $\boldsymbol{T}$ is not a sufficient statistic, either show that factorization fails or find a minimal sufficient statistic $\boldsymbol{S}$ and show that $\boldsymbol{S}$ is not a function of $\boldsymbol{T}$.
f) To show that $\boldsymbol{T}$ is not minimal sufficient, first try to show that $\boldsymbol{T}$ is not a sufficient statistic. If $\boldsymbol{T}$ is sufficient, find a minimal sufficient statistic $\boldsymbol{S}$ and show that $\boldsymbol{T}$ is not a function of $\boldsymbol{S}$. (Of course $\boldsymbol{S}$ will be a function of $\boldsymbol{T}$.) The **Lehmann-Scheffé (LSM) theorem cannot be used to show that a statistic is not minimal sufficient.**
g) To show that a sufficient statistics $\boldsymbol{T}$ is not complete, find a function $g(\boldsymbol{T})$ such that $E_{\boldsymbol{\theta}}(g(\boldsymbol{T})) = 0$ for all $\boldsymbol{\theta}$ but $g(\boldsymbol{T})$ is not equal to the zero with probability one. Finding such a g is often hard, unless there are clues. For example, if $\boldsymbol{T} = (\overline{X}, \overline{Y}, ...)$ and $\mu_1 = \mu_2$, try $g(\boldsymbol{T}) = \overline{X} - \overline{Y}$. As a **rule of thumb**, a k–dimensional minimal sufficient statistic will generally not

be complete if $k > dim(\Theta)$. In particular, if $\boldsymbol{T}$ is k–dimensional and $\boldsymbol{\theta}$ is j–dimensional with $j < k$ (especially $j = 1 < 2 = k$) then $\boldsymbol{T}$ will **generally not be complete.** If you can show that a k–dimensional sufficient statistic $\boldsymbol{T}$ is not minimal sufficient (often hard), then $\boldsymbol{T}$ is not complete by Bahadur's Theorem. Basu's Theorem can sometimes be used to show that a minimal sufficient statistic is not complete. See Remark 4.3 and Example 4.17.

15) A **common** question takes $Y_1, ..., Y_n$ iid $U(h_l(\theta), h_u(\theta))$ where the min $= Y_{(1)}$ and the max $= Y_{(n)}$ form the 2-dimensional minimal sufficient statistic. Since θ is one dimensional, the minimal sufficient statistic is probably not complete. Find $E_\theta[Y_{(1)}]$ and $E_\theta[Y_{(n)}]$. Then show that $E_\theta[aY_{(1)} + bY_{(n)} + c] \equiv 0$ so that $\boldsymbol{T} = (Y_{(1)}, Y_{(n)})$ is not complete.

4.4 Complements

Some minimal sufficient statistics and complete sufficient statistics are given below for distributions that are not exponential families.

Stigler (1984) presents Kruskal's proof that $\overline{Y} \perp\!\!\!\perp S^2$ when the data are iid $N(\mu, \sigma^2)$, but Zehna (1991) states that there is a flaw in the proof.

The Factorization Theorem was developed with increasing generality by Fisher, Neyman and by Halmos and Savage (1949).

Bahadur's Theorem is due to Bahadur (1958) and Lehmann and Scheffé (1950).

Basu's Theorem is due to Basu (1959). Also see Koehn and Thomas (1975) and Boos and Hughes-Oliver (1998). An interesting alternative method for proving independence between two statistics that works for some important examples is given in Datta and Sarker (2008).

Some techniques for showing whether a statistic is minimal sufficient are illustrated in Sampson and Spencer (1976).

4.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

4.1. Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, which is an exponential family. Show that the sample space of (T_1, T_2) contains an open subset of $\mathcal{R}^2$, if $n \geq 2$ but not if $n = 1$.

Hint: Show that if $n \geq 2$, then $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i^2$. Then $T_2 = aT_1^2 + b(X_1, ..., X_n)$ for some constant a where $b(X_1, ..., X_n) = \sum_{i=1}^n (X_i - \overline{X})^2 \in (0, \infty)$. So range$(T_1, T_2) = \{\ (t_1, t_2) | t_2 \geq at_1^2\ \}$. Find a. If $n = 1$ then $b(X_1) \equiv 0$ and the curve can not contain an open 2-dimensional rectangle.

4.2. Let $X_1, ..., X_n$ be iid exponential(λ) random variables. Use the Factorization Theorem to show that $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for λ.

4.3. Let $X_1, ..., X_n$ be iid from a regular exponential family with pdf

$$f(x|\boldsymbol{\eta}) = h(x)b(\boldsymbol{\eta}) \exp[\sum_{i=1}^k \eta_i t_i(x)].$$

Let $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), ..., T_k(\boldsymbol{X}))$ where $T_i(\boldsymbol{X}) = \sum_{j=1}^n t_i(X_j)$.

a) Use the factorization theorem to show that $\boldsymbol{T}(\boldsymbol{X})$ is a k-dimensional sufficient statistic for $\boldsymbol{\eta}$.

b) Use the Lehmann Scheffé LSM theorem to show that $\boldsymbol{T}(\boldsymbol{X})$ is a minimal sufficient statistic for $\boldsymbol{\eta}$.
(Hint: in a regular exponential family, if $\sum_{i=1}^k a_i \eta_i = c$ for all $\boldsymbol{\eta}$ in the natural parameter space for some fixed constants $a_1, ..., a_k$ and c, then $a_1 = \cdots = a_k = 0$.)

4.4. Let $X_1, ..., X_n$ be iid $N(\mu, \gamma_o^2 \mu^2)$ random variables where $\gamma_o^2 > 0$ is **known** and $\mu > 0$.

a) Find a sufficient statistic for μ.

b) Show that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a minimal sufficient statistic.

c) Find $E_\mu[\sum_{i=1}^n X_i^2]$.

d) Find $E_\mu[(\sum_{i=1}^n X_i)^2]$.

e) Find

$$E_\mu[\frac{n + \gamma_o^2}{1 + \gamma_o^2} \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2].$$

(Hint: use c) and d).)

f) Is the minimal sufficient statistic given in b) complete? Explain.

4.5. If $X_1, ..., X_n$ are iid with $f(x|\theta) = \exp[-(x-\theta)]$for $x > \theta$, then the joint pdf can be written as

$$f(\boldsymbol{x}|\theta) = e^{n\theta} \exp(-\sum x_i) I[\theta < x_{(1)}].$$

By the factorization theorem, $\boldsymbol{T}(\boldsymbol{X}) = (\sum X_i, X_{(1)})$ is a sufficient statistic. Show that $R(\theta) = f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ can be constant even though $\boldsymbol{T}(\boldsymbol{x}) \neq \boldsymbol{T}(\boldsymbol{y})$. Hence the Lehmann-Scheffé theorem does not imply that $\boldsymbol{T}(\boldsymbol{X})$ is a minimal sufficient statistic.

4.6. Find a complete minimal sufficient statistic if $Y_1, ..., Y_n$ are iid from the following 1P-REFs.

a) $Y \sim$ binomial (k, ρ) with k known.
b) $Y \sim$ exponential (λ).
c) $Y \sim$ gamma (ν, λ) with ν known.
d) $Y \sim$ geometric (ρ).
e) $Y \sim$ negative binomial (r, ρ) with r known.
f) $Y \sim$ normal (μ, σ^2) with σ^2 known.
g) $Y \sim$ normal (μ, σ^2) with μ known.
h) $Y \sim$ Poisson (θ).

4.7. Find a complete minimal sufficient statistic if $Y_1, ..., Y_n$ are iid from the following 1P-REFs.

a) $Y \sim$ Burr Type XII (ϕ, λ) with ϕ known.
b) $Y \sim$ chi(p, σ) with p known
c) $Y \sim$ double exponential (θ, λ) with θ known.
d) $Y \sim$ two parameter exponential (θ, λ) with θ known.
e) $Y \sim$ generalized negative binomial (μ, κ) with κ known.
f) $Y \sim$ half normal (μ, σ^2) with μ known.
g) $Y \sim$ inverse Gaussian (θ, λ) with λ known.
h) $Y \sim$ inverted gamma (ν, λ) with ν known.
i) $Y \sim$ lognormal (μ, σ^2) with μ known.
j) $Y \sim$ lognormal (μ, σ^2) with σ^2 known.
k) $Y \sim$ Maxwell-Boltzmann (μ, σ) with μ known.
l) $Y \sim$ one sided stable (σ).
m) $Y \sim$ Pareto (σ, λ) with σ known.
n) $Y \sim$ power (λ).
o) $Y \sim$ Rayleigh (μ, σ) with μ known.
p) $Y \sim$ Topp-Leone (ν).
q) $Y \sim$ truncated extreme value (λ).
r) $Y \sim$ Weibull (ϕ, λ) with ϕ known.

4.8. Find a complete minimal sufficient statistic $\boldsymbol{T}$ if $Y_1, ..., Y_n$ are iid from the following 2P-REFs.

a) The beta (δ, ν) distribution.
b) The chi (p, σ) distribution.
c) The gamma (ν, λ) distribution.
d) The lognormal (μ, σ^2) distribution.
e) The normal (μ, σ^2) distribution.

4.9. i) Show that each of the following families is a 1P-REF. ii) Find a complete minimal sufficient statistic if $Y_1, ..., Y_n$ are iid from the 1P-REF.

a) Let

$$f(y) = \frac{\log(\theta)}{\theta - 1}\theta^y$$

where $0 < y < 1$ and $\theta > 1$.
Comment:

$$F(y) = \frac{\theta^y - 1}{\theta - 1}$$

for $0 < y < 1$, and the mgf

$$m(t) = \frac{\log(\theta)}{\theta - 1} \frac{e^{(t+\log(\theta))} - 1}{t + \log(\theta)}.$$

b) Y has an inverse Weibull distribution.
c) Y has a Zipf distribution.

4.10. Suppose Y has a log-gamma distribution, $Y \sim LG(\nu, \lambda)$.
i) Show the Y is a 2P-REF.
ii) If $Y_1, ..., Y_n$ are iid $LG(\nu, \lambda)$, find a complete minimal sufficient statistic.
iii) Show $W = e^Y \sim$ gamma (ν, λ).

Problems from old quizzes and exams.

4.11. Suppose that $X_1, ..., X_m; Y_1, ..., Y_n$ are iid N$(\mu, 1)$ random variables. Find a minimal sufficient statistic for μ.

4.12. Let $X_1, ..., X_n$ be iid from a uniform $U(\theta - 1, \theta + 2)$ distribution. Find a sufficient statistic for θ.

4.13. Let $Y_1, ..., Y_n$ be iid with a distribution that has pmf $P_\theta(X = x) = \theta(1 - \theta)^{x-1}$, $x = 1, 2, ...,$ where $0 < \theta < 1$. Find a minimal sufficient statistic for θ.

4.14. Let $Y_1, ..., Y_n$ be iid Poisson(λ) random variables. Find a minimal sufficient statistic for λ using the fact that the Poisson distribution is a regular exponential family (REF).

4.15. Suppose that $X_1, ..., X_n$ are iid from a REF with pdf (with respect to the natural parameterization)

$$f(x) = h(x) c^*(\boldsymbol{\eta}) \exp[\sum_{i=1}^{4} \eta_i t_i(x)].$$

Assume $\dim(\Theta) = 4$. Find a complete minimal sufficient statistic $\boldsymbol{T}(\boldsymbol{X})$ in terms of n, t_1, t_2, t_3, and t_4.

4.16. Let X be a uniform $U(-\theta, \theta)$ random variable (sample size $n = 1$). a) Find $E_\theta X$. b) Is $T(X) = X$ a complete sufficient statistic? c) Show that $|X| = \max(-X, X)$ is a minimal sufficient statistic.

4.17. A fact from mathematics is that if the polynomial $P(w) = a_n w^n + a_{n-1} w^{n-1} + \cdots + a_2 w^2 + a_1 w + a_0 \equiv 0$ for all w in a domain that includes an open interval, then $a_n = \cdots = a_1 = a_0 = 0$. Suppose that you are trying to use the Lehmann Scheffé (LSM) theorem to show that $(\sum X_i, \sum X_i^2)$ is a minimal sufficient statistic and that you have managed to show that

$$\frac{f(\boldsymbol{x}|\mu)}{f(\boldsymbol{y}|\mu)} \equiv c$$

iff

$$-\frac{1}{2\gamma_o^2 \mu^2}[\sum x_i^2 - \sum y_i^2] + \frac{1}{\gamma_o^2 \mu}[\sum x_i - \sum y_i] \equiv d \tag{4.11}$$

for all $\mu > 0$. Parts a) and b) give two different ways to proceed.

a) Let $w = 1/\mu$ and assume that γ_o is known. Identify a_2, a_1 and a_0 and show that $a_i = 0$ implies that $(\sum X_i, \sum X_i^2)$ is a minimal sufficient statistic.

b) Let $\eta_1 = 1/\mu^2$ and $\eta_2 = 1/\mu$. Since (4.11) is a polynomial in $1/\mu$, can η_1 and η_2 satisfy a linearity constraint? If not, why is $(\sum X_i, \sum X_i^2)$ a minimal sufficient statistic?

4.18 Let $X_1, ..., X_n$ be iid Exponential(λ) random variables and $Y_1, ..., Y_m$ iid Exponential($\lambda/2$) random variables. Assume that the Y_i's and X_j's are independent. Show that the statistic $(\sum_{i=1}^n X_i, \sum_{i=1}^m Y_i)$ is not a complete sufficient statistic.

4.19. Let $X_1, ..., X_n$ be iid gamma(ν, λ) random variables. Find a complete, minimal sufficient statistic $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}))$. (Hint: recall a theorem for exponential families. The gamma pdf is (for x > 0)

$$f(x) = \frac{x^{\nu-1} e^{-x/\lambda}}{\lambda^\nu \Gamma(\nu)}.)$$

4.20. Let $X_1, ..., X_n$ be iid uniform($\theta - 1, \theta + 1$) random variables. The following expectations may be useful:

$$E_\theta \overline{X} = \theta, \; E_\theta X_{(1)} = 1 + \theta - 2\theta \frac{n}{n+1}, \; E_\theta X_{(n)} = 1 - \theta + 2\theta \frac{n}{n+1}.$$

a) Find a minimal sufficient statistic for θ.

b) Show whether the minimal sufficient statistic is complete or not.

4.21. Let $X_1, ..., X_n$ be independent identically distributed random variables with pdf

$$f(x) = \sqrt{\frac{\sigma}{2\pi x^3}} \exp\left(-\frac{\sigma}{2x}\right)$$

where x and σ are both positive. Find a sufficient statistic $T(\boldsymbol{X})$ for σ.

4.22. Suppose that $X_1, ..., X_n$ are iid beta(δ, ν) random variables. Find a minimal sufficient statistic for (δ, ν). Hint: write as a 2 parameter REF.

4.23. Let $X_1, ..., X_n$ be iid from a distribution with pdf

$$f(x|\theta) = \theta x^{-2}, \quad 0 < \theta \le x < \infty.$$

Find a sufficient statistic for θ.

4.24. Let $X_1, ..., X_n$ be iid with a distribution that has pdf

$$f(x) = \frac{x}{\sigma^2} \exp\left(\frac{-x}{2\sigma^2}\right)$$

for $x > 0$ and $\sigma^2 > 0$. Find a minimal sufficient statistic for σ^2 using the Lehmann-Scheffé theorem.

4.25. Let $X_1, ..., X_n$ be iid exponential (λ) random variables. Find a minimal sufficient statistic for λ using the fact that the exponential distribution is a 1P–REF.

4.26. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$. Find a complete sufficient statistic for (μ, σ^2).

4.27. (Jan. 2003 Qual) Let X_1 and X_2 be iid Poisson (λ) random variables. Show that $T = X_1 + 2X_2$ is not a sufficient statistic for λ. (Hint: the Factorization Theorem uses the word *iff*. Alternatively, find a minimal sufficient statistic S and show that S is not a function of T.)

4.28. (Aug. 2002 Qual): Suppose that $X_1, ..., X_n$ are iid $N(\sigma, \sigma)$ where $\sigma > 0$.

a) Find a minimal sufficient statistic for σ.

b) Show that $(\overline{X}, S^2)$ is a sufficient statistic but is not a complete sufficient statistic for σ.

4.29. Let $X_1, ..., X_n$ be iid binomial$(k = 1, \theta)$ random variables and $Y_1, ..., Y_m$ iid binomial$(k = 1, \theta/2)$ random variables. Assume that the Y_i's and X_j's are independent. Show that the statistic $(\sum_{i=1}^n X_i, \sum_{i=1}^m Y_i)$ is not a complete sufficient statistic.

4.30. Suppose that $X_1, ..., X_n$ are iid Poisson(λ) where $\lambda > 0$. Show that $(\overline{X}, S^2)$ is not a complete sufficient statistic for λ.

4.31. (Aug. 2004 Qual): Let $X_1, ..., X_n$ be iid beta(θ, θ). (Hence $\delta = \nu = \theta$.)

a) Find a minimal sufficient statistic for θ.

b) Is the statistic found in a) complete? (prove or disprove)

4.32. (Sept. 2005 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables with probability mass function

$$f(x) = P(X = x) = \frac{1}{x^\nu \zeta(\nu)}$$

where $\nu > 1$ and $x = 1, 2, 3, ...$ Here the zeta function

$$\zeta(\nu) = \sum_{x=1}^{\infty} \frac{1}{x^\nu}$$

for $\nu > 1$.

a) Find a minimal sufficient statistic for ν.

b) Is the statistic found in a) complete? (prove or disprove)

c) Give an example of a sufficient statistic that is strictly not minimal.

4.33. Let $X_1, ..., X_n$ be a random sample from a half normal distribution with pdf

$$f(x) = \frac{2}{\sqrt{2\pi}\ \sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x > \mu$ and μ is real.

Find a sufficient statistic $\boldsymbol{T} = (T_1, T_2, ..., T_k)$ for (μ, σ) with dimension $k \leq 3$.

4.34. Let $X_{(1)} = \min_{1 \leq i \leq n} X_i$. If $X_1, ..., X_n$ are iid exponential(1) random variables, find $E(X_{(1)})$.

4.35. Let $X_{(n)} = \max_{1 \leq i \leq n} X_i$. If $X_1, ..., X_n$ are iid uniform(0,1) random variables, find $E(X_{(n)})$.

4.36. (Aug. 2009 Qual): Let $X_1, ..., X_n$ be iid uniform$(\theta, \theta + 1)$ random variables where θ is real.

a) Find a minimal sufficient statistic for θ.

b) Show whether the minimal sufficient statistic is complete or not.

4.37. (Similar to Sept. 2010 Qual): Suppose that $X_1, X_2, ..., X_n$ are independent identically distributed random variables from normal distribution with unknown mean μ and known variance σ^2. Consider the parametric function $g(\mu) = e^{2\mu}$.

a) Derive the uniformly minimum variance unbiased estimator (UMVUE) of $g(\mu)$.

b) Find the Cramer-Rao lower bound (CRLB) for the variance of an unbiased estimator of $g(\mu)$.

c) Is the CRLB attained by the variance of the UMVUE of $g(\mu)$?

4.38. (Sept. 2010 Qual): Let $Y_1, ..., Y_n$ be iid from a distribution with pdf

$$f(y) = \ 2 \ \tau \ y \ e^{-y^2} \ (1 - e^{-y^2})^{\tau - 1}$$

for $y > 0$ and $f(y) = 0$ for $y \leq 0$ where $\tau > 0$.

a) Find a minimal sufficient statistic for τ.

b) Is the statistic found in a) complete? Prove or disprove.