6.6 The EM (Expectation-Maximization) algorithm.

The EM algorithm is based on the idea of replacing a difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original problem. It is particularly suited to "missing data" problems.

Consider a sample of $n$ items, where $n_1$ of the items are observed while $n_2 = n-n_1$ items are not observable. Denote the observed items by $\mathbf{X} = (X_1, \ldots, X_{n_1})$, and the unobserved items by $\mathbf{Z} = (Z_1, \ldots, Z_{n_2})$.

Let $g(\mathbf{X} | \theta)$ denote the pdf of $\mathbf{X}$
Let $h(\mathbf{X}, \mathbf{Z} | \theta)$ denote the joint pdf of $\mathbf{X}$ and $\mathbf{Z}$.
Let $k(\mathbf{Z} | \theta, \mathbf{X})$ denote the conditional pdf of $\mathbf{Z}$ given $\mathbf{X}$.

$$k(\mathbf{Z} | \theta, \mathbf{X}) = \frac{h(\mathbf{X}, \mathbf{Z} | \theta)}{g(\mathbf{X} | \theta)}$$

The observed likelihood function is $L(\theta | \mathbf{X}) = g(\mathbf{X} | \theta)$
Complete likelihood function is $L_c(\theta | \mathbf{X}, \mathbf{Z}) = h(\mathbf{X}, \mathbf{Z} | \theta)$

The EM algorithm allows us to maximize $L(\theta | \mathbf{X})$ by working with $L_c(\theta | \mathbf{X}, \mathbf{Z})$.

We have $L(\theta | \mathbf{X}) = \frac{L_c(\theta | \mathbf{X}, \mathbf{Z})}{k(\mathbf{Z} | \theta, \mathbf{X})}$. 


Let $\theta^{(0)}$ be an arbitrary value of $\theta$.

\[
\ln L(\theta | x) = \int \ln L(\theta | z) \, k(\frac{z}{\theta^{(0)}}) \, dz
\]

\[
= \int \left[ \ln \frac{L^c(\theta | z, \frac{z}{\theta^{(0)}})}{k(\frac{z}{\theta}, \theta)} \right] \, k(\frac{z}{\theta^{(0)}}, \theta) \, dz
\]

\[
= \mathbb{E}_{\frac{z}{\theta^{(0)}} | x, \theta^{(0)}} \left[ \ln L^c(\theta | z, \frac{z}{\theta^{(0)}}) \right]
- \mathbb{E}_{\frac{z}{\theta^{(0)}} | x, \theta^{(0)}} \left[ \ln k(\frac{z}{\theta}, \theta) \right]
\]

Denote the first term by $Q(\theta | \theta^{(0)}, x)$.

\[
Q(\theta | \theta^{(0)}, x) = \mathbb{E}_{\frac{z}{\theta^{(0)}} | x, \theta^{(0)}} \left[ \ln L^c(\theta | z, \frac{z}{\theta^{(0)}}) \right]
\]

The expectation above, which defines $Q$, is called the E-step of the EM algorithm. The maximization of $Q$ is called the M-step.

Let $\hat{\theta}^{(1)}$ be the argument which maximizes $Q(\theta | \theta^{(0)}, x)$.

$\hat{\theta}^{(1)}$ is called the first step estimate of $\theta$.

Then compute,

\[
Q(\theta | \hat{\theta}^{(1)}, x) = \mathbb{E}_{\frac{z}{\hat{\theta}^{(1)}} | x, \hat{\theta}^{(1)}} \left[ \ln L^c(\theta | z, \frac{z}{\hat{\theta}^{(1)}}) \right]
\]

Let $\hat{\theta}^{(2)}$ be the argument which maximizes $Q(\theta | \hat{\theta}^{(1)}, x)$.

$\hat{\theta}^{(2)}$ is the second step estimate of $\theta$. 
EM algorithm creates a sequence $\hat{\theta}^{(m)}$ from an initial value $\theta^{(0)}$, according to

$$\hat{\theta}^{(m+1)} = \text{the value that maximizes } Q(\theta | \hat{\theta}^{(m)}, x)$$

$\hat{\theta}^{(m)}$ converges in probability to the maximum likelihood estimate as $m \to \infty$.

**Example.** Let $f(x)$ be any pdf. Then the family of pdf's $f(x-\theta)$ indexed by the parameter $\theta$ ($-\infty < \theta < \infty$) is called the location family with standard pdf $f(x)$ and $\theta$ is called the location parameter. As an example, if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $-\infty < x < \infty$ then $f(x-\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$, $-\infty < x < \infty$ be $X \sim N(\theta, 1)$.

Now suppose $X_1, \ldots, X_n$ is a random sample from a distribution with pdf $f(x-\theta)$ and cdf $F(x-\theta)$.

Let $Z_1, \ldots, Z_n$ denote the censored observations. We only know that $Z_j > a$ for some $a$ which is known and that $Z_j$'s are independent of $X_i$'s.
As an example, suppose we observe the survival times of $n$ patients for 10 years. Suppose $n_1$ of the patients died before the 10th year and $n_2$ patients were still alive, $n_1 + n_2 = n$, here $a = 10$.

$$L(\theta | x) = \prod_{i=1}^{n_1} P(Z_i > a) \prod_{i=1}^{n_2} f(x_i; -\theta)$$

$$= \left[1 - F(a; -\theta)\right]^{n_1} \prod_{i=1}^{n_2} f(x_i; -\theta)$$

$$L^c(\theta | x, z) = \prod_{i=1}^{n_1} f(z_i; -\theta) \prod_{i=1}^{n_2} f(x_i; -\theta)$$

$$k(z; \theta, z) = \frac{L^c(\theta | x, z)}{L(\theta | x)} = \frac{\prod_{i=1}^{n_1} f(z_i; -\theta)}{\left[1 - F(a; -\theta)\right]^{n_2}}$$
Thus, $Z_1, \ldots, Z_n$ is a random sample from a distribution with pdf \[ \frac{f(z-\theta)}{[1-F(a-\theta)]} \] for $z > a$.

Let $\theta^{(0)}$ be an arbitrary fixed value of $\theta$.

\[ \Theta(\theta | \theta^{(0)}, x) = E_x \left[ \frac{\sum_{i=1}^{n_1} \ln f(x_i-\theta) + \sum_{i=1}^{n_2} \ln f(z_i-\theta)}{\theta^{(0)}} \right] \]

\[ = \sum_{i=1}^{n_1} \ln f(x_i-\theta) + n_2 E_x \left[ \frac{\ln f(z-\theta)}{\theta^{(0)}} \right] \]

This is the $E$-step. For the $m$-step we need to maximize $\Theta(\theta | \theta^{(0)}, x)$ to obtain $\theta^{(1)}$.

\[ \frac{\partial \Theta}{\partial \theta} = - \left[ \sum_{i=1}^{n_1} \frac{f'(x_i-\theta)}{f(x_i-\theta)} + n_2 \int_0^\infty \frac{f'(z-\theta)}{f(z-\theta)} \frac{f(z-\theta^{(0)})}{[1-F(a-\theta^{(0)})]} dz \right] \]

Suppose $x \sim N(\theta, 1)$. Then $f(x) = \frac{1}{\sqrt{2\pi}^n} e^{-\frac{(x-\theta)^2}{2}}$ for $-\infty < x < \infty$.

\[ f'(x) = -\frac{(x-\theta)}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} \]

\[ \frac{f'(x)}{f(x)} = -\theta \]
\[ \frac{\partial \bar{\theta}}{\partial \theta} = \sum_{i=1}^{n_1} (x_i - \theta) + n_2 \theta^{(0)} + n_3 \int_{-\infty}^{\infty} \frac{\text{erf} \left( \frac{z - \theta^{(0)}}{2} \right)}{\sqrt{2\pi}} \, dz \]

where \( \Phi \) is the cdf of \( \mathcal{N}(0,1) \).

\[ \frac{\partial \bar{\theta}}{\partial \theta} = n_1 (\bar{x} - \theta) + n_2 \theta + n_3 \int_{\alpha}^{\infty} \frac{(z - \theta^{(0)})}{\sqrt{2\pi}} \, dz \]

\[ = n_1 (\bar{x} - \theta) - n_2 (\theta - \theta^{(0)}) + \frac{n_3}{\sqrt{2\pi}} \int_{\alpha - \theta^{(0)}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \, dw \]

\[ = n_1 (\bar{x} - \theta) - n_2 (\theta - \theta^{(0)}) + \frac{n_3}{\sqrt{2\pi}} \left[ \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \right]_{\alpha - \theta^{(0)}}^{\infty} \]

\[ = n_1 (\bar{x} - \theta) - n_2 (\theta - \theta^{(0)}) + \frac{n_3}{\sqrt{2\pi}} \left[ \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \right] - \frac{(\theta - \theta^{(0)})^2}{2} = 0 \]

\[ \hat{\theta}^{(1)} = \frac{n_1}{n} \bar{x} + \frac{n_2}{n} \theta^{(0)} + \frac{n_3}{n} \left( \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \right) \]

In general, given the \( m^{\text{th}} \) step EM estimate \( \hat{\theta}^{(m)} \), the \( (m+1)^{\text{th}} \) step EM estimate is

\[ \hat{\theta}^{(m+1)} = \frac{n_1}{n} \bar{x} + \frac{n_2}{n} \hat{\theta}^{(m)} + \frac{n_3}{n} \left( \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \right) \]
Theorem. The sequence of estimates $\hat{\theta}^{(m)}$ defined by the EM algorithm satisfies

1) $E_{\tilde{Z} | x, \hat{\theta}^{(m)}} \left[ \ell L^c (\hat{\theta}^{(m+1)} | x, \tilde{Z} ) \right]$

$\geq E_{\tilde{Z} | x, \hat{\theta}^{(m)}} \left[ \ell L^c (\hat{\theta}^{(m)} | x, \tilde{Z} ) \right]$

That is after each step $Q ( \theta | \hat{\theta}^{(m)}, x )$ increases.

2) $E_{\tilde{Z} | x, \hat{\theta}^{(m)}} \left[ \ell \ln k (\tilde{Z} | \hat{\theta}^{(m+1)} | x ) \right]$

$\leq E_{\tilde{Z} | x, \hat{\theta}^{(m)}} \left[ \ell \ln k (\tilde{Z} | \hat{\theta}^{(m)} | x ) \right]$

That is after each step $E_{\tilde{Z} | x, \hat{\theta}^{(m)}} \left[ \ell \ln k (\tilde{Z} | \hat{\theta}^{(m)} | x ) \right]$ decreases.

Thus we need only maximize $Q$ since the log of the observed likelihood is the difference between $Q$ and the expected value of log of the conditional pdf of $Z$ given $X$ evaluated when $\theta = \hat{\theta}^{(m)}$. 
Chapter 7. Sufficiency.

Let \( X_1, \ldots, X_n \) be a random sample from a distribution with pdf \( f(x|\theta) \). Let \( Y = U(X_1, \ldots, X_n) \) be a statistic on which we wish to base a point estimate of \( \theta \).

Definition: \( Y = U(X_1, \ldots, X_n) \) is called a minimum variance unbiased estimator (MVUE) of \( \theta \) if \( Y \) is unbiased and if the variance of \( Y \) is less than or equal to the variance of every other unbiased estimator of \( \theta \).

The loss function \( L(\theta, a) \) represents the loss incurred in taking action \( a \) when the parameter value is \( \theta \). In the problem of point estimation the action consists of computing a function of \( Y \), say \( \delta(Y) \), as the point estimate of \( \theta \). Thus if we use \( \delta(Y) \) to estimate \( \theta \), our loss function is \( L(\theta, \delta(Y)) \).

Examples of Loss Functions:

- \( L(\theta, \delta(Y)) = (\theta - \delta(Y))^2 \) is called Square-Error Loss Function.

- \( L(\theta, \delta(Y)) = |\theta - \delta(Y)| \) is called Absolute-Error Loss Function.

- \( L(\theta, \delta(Y)) = \begin{cases} 0 & \text{if } |\theta - \delta(Y)| \leq a \\ b & \text{if } |\theta - \delta(Y)| > a \end{cases} \)

is called Goal Post Loss Function.
The performance characteristic of an estimator $\delta (Y)$ is reasonably reflected by its Risk Function $R[\theta, \delta (Y)]$.

$$R[\theta, \delta (y)] = E[ L[\theta, \delta (y)] ] = \int_{-\infty}^{\infty} L[\theta, \delta (y)] f_Y(y|\theta) \, dy$$

where $f_Y(y|\theta)$ is the pdf of $Y$.

If for every $\delta (y)$, we have

$$R[\theta, \delta^* (y)] \leq R[\theta, \delta (y)] \quad \text{for all } \theta \in \Theta$$

then $\delta^* (y)$ is uniformly better estimator and so it is the one that we would like to use. But this is usually impossible because $\delta$ that minimizes $R[\theta, \delta]$ for one value of $\theta$ may not minimize it for another value of $\theta$, as demonstrated by the following example.

Example: Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, 1)$.

Let $Y = \bar{X}$, then

$$L[\theta, \delta (y)] = \left[ \theta - \delta (y) \right]^2$$

Consider the following estimators:

$$\delta_1 (Y) = Y, \quad \delta_2 (Y) = 0.$$ 

Then,

$$R[\theta, \delta_1 (y)] = E[ (\theta - Y)^2 ] = E[ (\theta - \bar{X})^2 ] = V(\bar{X}) = \frac{1}{n}$$

$$R[\theta, \delta_2 (y)] = E[ (\theta - 0)^2 ] = \theta^2.$$
\[ R[\theta, \delta(y)] < R[\theta, \delta'(y)] \] if \(- \frac{1}{\sqrt{n}} < \theta < \frac{1}{\sqrt{n}}\)
\[ R[\theta, \delta(y)] \geq R[\theta, \delta'(y)] \] otherwise.

That is, one of these estimators is better than the other for some values of \(\theta\), and the other estimator is better for other values of \(\theta\).

If we restrict our attention to class of unbiased estimators and if \( E[\theta, \delta(y)] = E[\theta - d(y)]^2 \) then
\[ R[\theta, \delta(y)] = E[(\theta - d(y))^2] = V[\delta(y)] \]. Then the problem reduces to finding MVU estimator.

**Minimax Rule:**

An estimator which minimizes the maximum of the risk function is called minimax estimator.

If for every \(\delta(y)\), we have
\[ \max_\theta R[\theta, \delta^*(y)] \leq \max_\theta R[\theta, \delta(y)] \]
then \(\delta^*(y)\) is called a minimax estimator of \(\theta\).

**Equalizer Principle:** The risk function of a minimax estimator is a constant, does not depend upon \(\theta\).
The Likelihood Principle:

Suppose two different sets of data from possibly two different random experiments lead to respective likelihood functions \( L_1(\theta) \) and \( L_2(\theta) \), that are proportional to each other. These two data sets provide the same information about \( \theta \) and we should make the same inference about \( \theta \) from either.

Example: Let \( \theta \): Probability of heads with a given coin.

\[ H_0: \theta = 0.5 \quad H_1: \theta < 0.5 \]

An experiment is conducted by flipping the coin in a series of independent trials, the result of which is observation of one head and 9 tails.

a) Suppose the experiment consists of a predetermined 10 flips.

Let \( X \): Number of heads in \( n=10 \) flips, then

\[ L_1(\theta) = \binom{10}{1} \theta^1 (1-\theta)^9 = 10 \theta (1-\theta)^9 \]

\[ P\text{-value}= P[ X \leq 1 | \theta=0.5 ] = \sum_{x=0}^{1} \binom{10}{x} 0.5^x 0.5^{10-x} = 0.0107 \]

[ Do not reject \( H_0 \) at \( \alpha = 0.01 \)]

b) Suppose the experiment consists of tossing the coin until the first head is observed.
Let $Z$: Number of tosses until the first head.

$$f_Z(z) = (1-\theta)^{z-1} \theta \quad \text{if } z = 1, 2, \ldots$$

$$E(Z) = \frac{1}{\theta}$$

$H_0: \frac{1}{\theta} = 2$

$H_1: \frac{1}{\theta} > 2$

$P\text{-value} = P[Z \geq 10 | \theta = \frac{1}{2}]$

$$= 1 - P[Z < 9 | \theta = \frac{1}{2}] = 1 - \sum_{z=1}^{9} (0.5)^z = 0.00197$$

[Reject $H_0$ at $\alpha = 0.01$]

If significance level of $\alpha = 0.01$ is used, two models would lead to different conclusions, in contradiction of the likelihood principle.

---

**Sufficient Statistic.**

Definition: Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a distribution with pdf $f_X(x | \theta)$, $\theta \in \Omega$. The statistic $Y_1 = Y_1(X_1, \ldots, X_n)$ is called a **sufficient statistic** if and only if the conditional distribution of $X_1, \ldots, X_n$ given $Y_1 = y$, does not depend upon the parameter $\theta$.

That is, $Y_1$ is a sufficient statistic for $\theta$ if and only if

$$\frac{f(x_1, \ldots, x_n | \theta)}{f_Y(y | \theta)}$$

does not depend upon $\theta$. 

where \( f(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f_x(x_i | \theta) \)
is the joint pdf of \( X_1, \ldots, X_n \) and \( f_{Y_1}(y_1 | \theta) \) is the
pdf of \( Y_1 \).

Above definition implies that, in a sense, \( Y_1 \)
exhausts all the information contained in the sample
about \( \theta \). If \( Y_1 \) is a sufficient statistic for \( \theta \),
then the conditional distribution of any statistic \( Y_2 \)
given \( Y_1 = y_1 \) does not depend upon \( \theta \). As a
consequence once \( Y_1 \) is given it is impossible to
use \( Y_2 \) to make a statistical inference about \( \theta \).

Example: Let \( X_1, \ldots, X_n \) be a random sample from
Bernoulli(\( \theta \)). Let \( Y_1 = \sum_{i=1}^{n} X_i \). \( Y_1 \sim \text{Binomial}(n, \theta) \)
\[
\frac{f(x_1, \ldots, x_n | \theta)}{f_{Y_1}(y_1 | \theta)} = \frac{\prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}}{(y_1)! \theta^{y_1} (1-\theta)^{n-y_1}} = \frac{1}{(y_1)!} \quad \text{does not depend upon } \theta .
\]
Hence \( Y_1 \) is a sufficient statistic
for \( \theta \).

Example: Let \( X_1, \ldots, X_n \) be a random sample
from a distribution with pdf
\[ f_X(x|\theta) = \begin{cases} \exp(-(x-\theta)) & \text{if } \theta < x < \infty \\ 0 & \text{elsewhere} \end{cases} \]

where \(-\infty < \theta < \infty\).

\[
\left[ \text{Indicator Function} \right] I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}
\]

Whenever the support of \(X\) depends on \(\theta\), write \(f_X(x|\theta)\) using the indicator function.

We write:
\[
f_X(x|\theta) = \exp(-(x-\theta)) I_{(\theta, \infty)}(x)
\]

Let \(Y_1 = \min(x_1, \ldots, x_n)\).

\[
f_{Y_1}(y_1) = n \exp(-n(y_1-\theta)) I_{(\theta, \infty)}(y_1)
\]

\[
f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \exp(-(x_i-\theta)) I_{(\theta, \infty)}(x_i)
\]

\[
f_{Y_1}(y_1|\theta) = \frac{n \exp(-n(y_1-\theta)) I_{(\theta, \infty)}(y_1)}{\prod_{i=1}^{n} \exp(-(x_i-\theta)) I_{(\theta, \infty)}(x_i)}
\]

Observe:
\[
\prod_{i=1}^{n} I_{(\theta, \infty)}(x_i) = I_{(\theta, \infty)}(y_1)
\]

Thus,
\[
\frac{f(x_1, \ldots, x_n|\theta)}{f_{Y_1}(y_1|\theta)} = \frac{\exp(-\sum_{i=1}^{n}(x_i-\theta)) I_{(\theta, \infty)}(y_1)}{n \exp(-n(y_1-\theta)) I_{(\theta, \infty)}(y_1)} = \frac{\exp((-\sum_{i=1}^{n}x_i)+n\theta)}{n \exp(-ny_1)}
\]

does not depend upon \(\theta\).

\(Y_1\) is a sufficient statistic for \(\theta\).
Factorization Theorem.

Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a distribution with pdf $f(x|\theta), \theta \in \Omega$. The statistic $Y_1 = U_1(X_1, \ldots, X_n)$ is a sufficient statistic for $\theta$ if and only if we can find two non-negative functions $k_1$ and $k_2$ such that

$$f(x_1, \ldots, x_n|\theta) = k_1(y_1, \theta) \cdot k_2(x_1, \ldots, x_n)$$

where for fixed value of $y_1 = y_{1*}$, $k_2(x_1, \ldots, x_n)$ does not depend upon $\theta$ and $k_1(y_1, \theta)$ is a function of $y_1$ and $\theta$.

Example: Let $X_1, \ldots, X_n$ be a random sample from Bernoulli$(\theta)$.

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \frac{\theta^{\sum x_i} (1-\theta)^{n-\sum x_i}}{k_1(\sum x_i, \theta) \cdot k_2(x_1, \ldots, x_n)}$$

$k_2(x_1, \ldots, x_n)$ does not depend upon $\theta$. Thus $\sum x_i$ is a sufficient statistic for $\theta$.

Example: Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

where $\theta > 0$. 
\[ f(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^\theta \left( \prod_{i=1}^{n} x_i \right)^{\theta-1} \cdot \left( \frac{1}{\prod_{i=1}^{n} x_i} \right) \cdot k_1(\prod_{i=1}^{n} x_i, \theta) \cdot k_2(x_1, \ldots, x_n) \]

\( \frac{k_1(\prod_{i=1}^{n} x_i, \theta)}{k_2(x_1, \ldots, x_n)} \) does not depend on \( \theta \).

\[ \therefore \prod_{i=1}^{n} x_i \text{ is a sufficient statistic for } \theta. \]

**Example:** Let \( x_1, \ldots, x_n \) be a random sample from a distribution with pdf \( f_x(x | \theta) = e^{-x(x-\theta)} \cdot I(\theta, \infty) \).

\[ f(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} e^{-x_i(x_i-\theta)} \cdot I(\theta, \infty) \]

\[ = e^{\theta \cdot \prod_{i=1}^{n} x_i} \cdot I(\theta, \infty) \cdot \prod_{i=1}^{n} e^{-x_i} \]

\[ = e^{\theta \cdot \prod_{i=1}^{n} x_i} \cdot I(\theta, \infty) \cdot \prod_{i=1}^{n} e^{-x_i} \]

\[ = e^{\theta \cdot \left( \min(x_1, \ldots, x_n) \right)} \cdot I(\theta, \infty) \cdot k_1 \left[ \min(x_1, \ldots, x_n), \theta \right] \cdot k_2(x_1, \ldots, x_n) \]

\[ \therefore \min(x_1, \ldots, x_n) \text{ is a sufficient statistic for } \theta. \]

**Theorem:** If a MLE \( \hat{\theta} \) of \( \theta \) exists uniquely then \( \hat{\theta} \) is a function of a sufficient statistic for \( \theta \).
Theorem. If $Y_i$ is a sufficient statistic for $\theta$ then $E = \mathcal{W}(Y_i)$ is also sufficient for $\theta$, if $\mathcal{W}$ is one-to-one.

Recall the following Theorem:

Theorem. For any random variables $Y_1$ and $Y_2$

a) $E(Y_2) = E[E(Y_2 | Y_1)]$

b) $V(Y_2) = E[V(Y_2 | Y_1)] + V[E(Y_2 | Y_1)]$

and thus $V(Y_2) \geq V[E(Y_2 | Y_1)]$.

Let $Y_1$ be a sufficient statistic for $\theta$ and let $Y_2$ be an unbiased statistic for $\theta$. $Y_2$ is not a function of $Y_1$ alone.

Then $E(Y_2 | Y_1) = \Psi(Y_1)$ is a function of $Y_1$ alone (since the conditional distribution of $Y_2$ given $Y_1$ does not depend upon $\theta$) and $\Psi(Y_1)$ is an unbiased statistic for $\theta$ (by a) and its variance is less than that of $Y_2$ (by b).

\[
E[\Psi(Y_1)] = \theta \quad \text{and} \quad V[\Psi(Y_1)] \leq V(Y_2)
\]
Completeness.

Definition: A family of pdf's \( \{ f(z; \theta), \theta \in \mathbb{R} \} \) is said to be complete if for every real-valued function \( U \), \( E[U(z)] = 0 \) for each \( \theta \in \mathbb{R} \) implies \( U(z) = 0 \) at each point \( z \) at which at least one member of the family of pdf's is positive.

Example. Let \( f(x; \theta) = \theta^x (1-\theta)^{1-x}, \ x > 0, 1 \)
\( \Omega = \{ \theta : 0 < \theta < 1 \} \). Show that \( \{ f(x; \theta), \theta \in \Omega \} \) is complete.

\[
E[U(x)] = \sum_{x=0}^{x=1} u(x) \theta^x (1-\theta)^{1-x} = u(0)(1-\theta) + u(1)\theta
\]
\[
= u(\theta) + \theta[u(1) - u(0)] = 0 \quad \forall \ 0 < \theta < 1
\]

implies \( u(0) = 0, u(1) - u(0) = 0 \) i.e. \( u(0) = u(1) = 0 \)

\( \therefore \ u(x) = 0, \ x > 0, 1 \) \[ If \ a \ linear \ function \]
\( a + b \theta = 0 \) for more than one value of \( \theta \) then \( a = 0 \)

Thus \( E[U(x)] = 0 \ \forall \ 0 < \theta < 1 \rightarrow U(x) = 0, x > 0, 1 \).

\( \therefore \ \{ f(x; \theta), 0 < \theta < 1 \} \) is complete.
Example: \( f(x; \alpha) = \begin{cases} \frac{1}{\alpha} & 0 < x < \alpha \\ 0 & \text{elsewhere} \end{cases} \)

\( \mathcal{F} = \{ \alpha : \alpha > 0 \} \). Show that \( \mathcal{F} \) is complete.

\[ E[U(X)] = \int_0^\theta u(x) \frac{1}{\theta} \, dx = 0 \quad \forall \ \theta > 0 \]

\[ \Rightarrow \quad \int_0^\theta u(x) \, dx = 0 \quad \forall \ \theta > 0 \]

\[ \Rightarrow \quad \frac{d}{d\theta} \int_0^\theta u(x) \, dx = 0 \quad \forall \ \theta > 0 \]

\[ \Rightarrow \quad u(\theta) = 0 \quad \forall \ \theta > 0 \]

\[ \therefore \quad u(x) = 0 \quad \forall x > 0 \]

Example: \( f(x; \alpha) = \begin{cases} \frac{1}{2\alpha} & -\alpha < x < \alpha \\ 0 & \text{elsewhere} \end{cases} \)

\( \mathcal{F} = \{ \alpha : \alpha > 0 \} \). \( \mathcal{F} \) is not complete.

Let \( U(X) = X \) then
\[ E[U(X)] = E(X) = \int_\theta^0 \frac{1}{2\alpha} \, d\alpha = 0 \]

\( U(X) = X \) is a non-zero function of \( X \) such that
\[ E[U(X)] = 0 \quad \Rightarrow \quad U(X) = 0 \]

\[ \therefore \quad \mathcal{F} \] is not a complete family.
Definition. A sufficient statistic \( Y \) with pdf \( f_{Y_i}(y_1 | \theta) \) is said to be a complete sufficient statistic if \( \int f_{Y_i}(y_1 | \theta) \, dy_1 \) is complete.

Theorem. If \( Y \) is a complete sufficient statistic for \( \theta \) and if there is a function of \( Y \) which is unbiased for \( \theta \), then this function of \( Y \) is the unique MVUE of \( \theta \).

Example. Let \( X_1, \ldots, X_n \) be a random sample from Uniform \((0, \theta)\).

\[
f_X(x | \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{elsewhere} \end{cases}
\]

Since the support depends on \( \theta \), write

\[
f_X(x | \theta) = \frac{1}{\theta} \cdot I((0, \theta))
\]

Let \( Y_n = \max(X_1, \ldots, X_n) \).

\[
f_{Y_n}(y_n | \theta) = \frac{n \cdot y_n^{n-1}}{\theta^n} \cdot I((0, \theta))
\]
\[
\frac{f(x_1, \ldots, x_n | \theta)}{f_n(y_1, \ldots, y_n | \theta)} = \frac{1/\theta^n}{n \prod_{i=1}^{n} \frac{\Gamma((y_i)/\theta)}{\Gamma(\theta)}} = \frac{1}{n \Gamma^{n-1}(\theta)} \text{ does not depend upon } \theta.
\]

Thus \( y_n \) is a sufficient statistic for \( \theta \).

We will now show that \( \mathcal{L}_n(y_n | \theta), \theta > 0 \) is complete.

\[
E[u(y_n)] = \int_{0}^{\theta} u(y_n) n \frac{y_n^{n-1}}{\theta^n} dy_n = 0 \quad \forall \theta > 0
\]

\[
\rightarrow \int_{0}^{\theta} u(y_n) y_n^{n-1} dy_n = 0 \quad \forall \theta > 0
\]

\[
\rightarrow \frac{d}{d\theta} \int_{0}^{\theta} u(y_n) y_n^{n-1} dy_n = 0 \quad \forall \theta > 0
\]

\[
\rightarrow u(\theta) \theta^{n-1} = 0 \quad \forall \theta > 0 \rightarrow u(\theta) = 0 \quad \forall \theta > 0
\]

\[
\therefore u(y_n) = 0 \quad \forall y_n > 0, \theta > 0.
\]

Thus \( y_n \) is a complete sufficient statistic for \( \theta \).

To find the MVUE of \( \theta \), we need to find \( u(y_n) \) such that

\[
E[u(y_n)] = \int_{0}^{\theta} u(y_n) n \frac{y_n^{n-1}}{\theta^n} dy_n = \theta
\]

We have \( \int_{0}^{\theta} u(y_n) y_n^{n-1} dy_n = \frac{1}{n} \theta^n \). Take the derivative of both sides w.r.t. \( \theta \) we get

\[
u(\theta) \theta^{n-1} = \frac{n+1}{n} \theta^n, \quad u(\theta) = (\frac{n+1}{n}) \theta
\]

Thus \( u(y_n) = (\frac{n+1}{n}) y_n \) is the MVUE of \( \theta \).