$$H = \sum_{k=1}^{M} p_k \log_2\left(\frac{1}{p_k}\right)$$

Here M = 5 and putting the values of probabilities in above equation,

$$H = 0.4 \log_2\left(\frac{1}{0.4}\right) + 0.19 \log_2\left(\frac{1}{0.19}\right) + 0.16 \log_2\left(\frac{1}{0.16}\right)$$

$$+ 0.15 \log_2\left(\frac{1}{0.15}\right) + 0.15 \log_2\left(\frac{1}{0.15}\right)$$

$$= 2.2281 \text{ bits/message}$$

The average number of bits per message $\bar{N}$ is given by equation 2.9.1 as,

$$\bar{N} = \sum_{k=0}^{L-1} p_k n_k$$

Here p_k is the probability of k^{th} message and n_k are number of bits assigned to it. Putting the values in above equation,

$$\bar{N} = 0.4(1) + 0.19(3) + 0.16(3) + 0.15(3) + 0.15(3)$$

$$= 2.35$$

The code efficiency is given by equation 2.9.4 i.e.

$$\text{code efficiency } \eta = \frac{H}{\bar{N}}$$

$$= \frac{2.2281}{2.35} = 0.948$$

ECE

ii) To obtain Huffman code :

Table 2.10.7 illustrates the Huffman coding. Huffman coding is explained in Table 2.10.3 with the help of an example. The coding shown in Table 2.10.7 below is based on this explanation.

Table 2.10.7 To obtain Huffman code

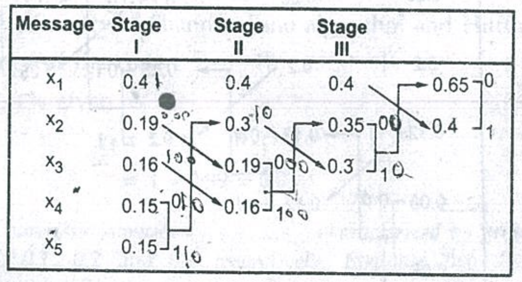


Table 2.10.4 shows how code words are obtained by tracing along the path. These code words are given in Table 2.10.8.

Table 2.10.8 Huffman coding

Message	Probability	Digits obtained by tracing $b_2\ b_1\ b_0$	Codeword $b_0\ b_1\ b_2$	Number of digits n_k
x_1	0.4	1	1	1
x_2	0.19	000	000	3
x_3	0.16	100	001	3
x_4	0.15	010	010	3
x_5	0.15	110	011	3

Now let us determine the average number of bits per message ($\bar{N}$). It is given as,

$$\bar{N} = \sum_{k=0}^{L-1} p_k n_k$$

Putting the values in above equation,

$$\bar{N} = 0.4(1) + 0.19(3) + 0.16(3) + 0.15(3) + 0.15(3)$$

$$= 2.35$$

Hence the code efficiency is,

$$\eta = \frac{H}{\bar{N}}$$

$$= \frac{2.2281}{2.35} = 0.948$$

Thus the code efficiency of Shannon-Fano code and Huffman code is same in this example.

Ex.2.10.3 *Compare the Huffman coding and Shannon-Fano coding algorithms for data compression. For a discrete memoryless source 'X' with six symbols $x_1, x_2, \ldots x_6$, find a compact code for every symbol if the probability distribution is as follows :*

$p(x_1) = 0.3 \quad p(x_2) = 0.25 \quad p(x_3) = 0.2$

$p(x_4) = 0.12 \quad p(x_5) = 0.08 \quad p(x_6) = 0.05$

Calculate entropy of the source, average length of the code, efficiency and redundancy of the code. **[May-2001, 10 Marks]**

Sol. : (I) Entropy of the source :

Entropy is given as,

$$H = \sum_{k=1}^{M} p_k \log_2\left(\frac{1}{p_k}\right)$$

For six messages above equation becomes,

$$= p_1 \log_2\left(\frac{1}{p_1}\right) + p_2 \log_2\left(\frac{1}{p_2}\right) + p_3 \log_2\left(\frac{1}{p_3}\right) + p_4 \log_2\left(\frac{1}{p_4}\right) + p_5 \log_2\left(\frac{1}{p_5}\right) + p_6 \log_2\left(\frac{1}{p_6}\right)$$

Putting values,

$$H = 0.3 \log_2\left(\frac{1}{0.3}\right) + 0.25 \log_2\left(\frac{1}{0.25}\right) + 0.2 \log_2\left(\frac{1}{0.2}\right) + 0.12 \log_2\left(\frac{1}{0.12}\right) + 0.08 \log_2\left(\frac{1}{0.08}\right) + 0.05 \log_2\left(\frac{1}{0.05}\right)$$

$$= 0.521 + 0.5 + 0.4643 + 0.367 + 0.2915 + 0.216$$

$$= 2.3568 \text{ bits of information/message.}$$

(i) To obtain codewords

Table 2.10.9 shows the procedure for Shannon-Fano coding. The partitioning is made as per the procedure discussed earlier.

Table 2.10.9 Shannon-Fano algorithm

Symbol	Probability	Stage-I	Stage-II	Stage-III	Stage-IV	Codeword	No. of bits per message n_k
x_1	0.3	0.45: 0	0			00	2
x_2	0.25	0	1			01	2
x_3	0.2	1	0.2: 0			10	2
x_4	0.12	0.45: 1	1	0.12: 0		110	3
x_5	0.08	1	0.25: 1	1	0	1110	4
x_6	0.05	1	1	0.13: 1	1	1111	4

(ii) To obtain average number of bits per message $(\overline{N})$:

$\overline{N}$ is given as, $\overline{N} = \sum_{k=0}^{L-1} p_k n_k$

Putting values in above equation from table 2.10.8,

$$\overline{N} = (0.3)(2) + (0.25)(2) + (0.2)(2) + (0.12)(3) + (0.08)(4) + (0.05)(4)$$

$$= 2.38$$

(iii) To obtain code efficiency :

Code efficiency is given as,

$$\eta = \frac{H}{\overline{N}} = \frac{2.3568}{2.38} = 0.99$$

(1) $H = \sum P_k \log\left(\frac{1}{P_k}\right)$

(2) $\bar{N} = \sum P_k N_k$

(iv) To obtain redundancy of the code :

Redundancy is given as,

$$\text{Redundancy } (\gamma) = 1 - \eta = 1 - 0.99 = 0.01$$

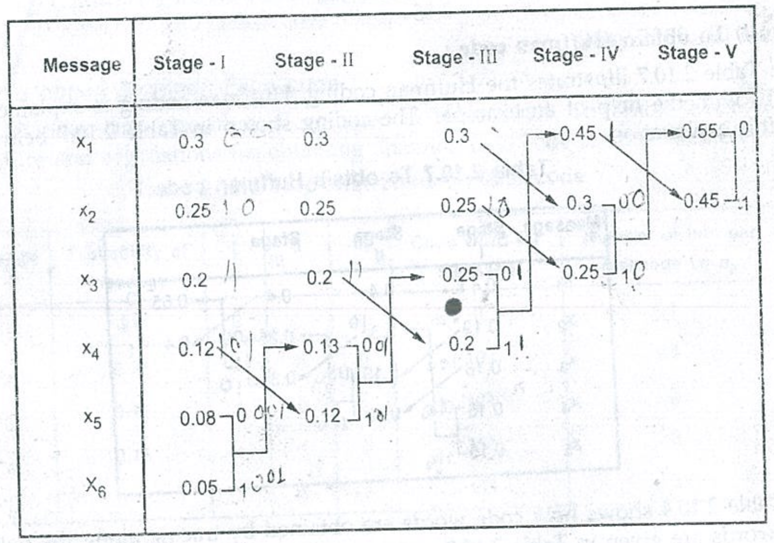
Here 0.01 indicates that there are 1% of redundant bits in the code.

(III) Huffman coding :

i) To obtain codewords

Table 2.10.10 lists the Huffman coding algorithm.

Table 2.10.10 : Huffman coding



Based on the above encoding arrangements following codes are generated.

Table 2.10.11 : Huffman codes

Message	Probability p_k	Digits obtained by tracing $b_3\, b_2\, b_1\, b_0$	Codeword $b_0\, b_1\, b_2\, b_3$	Number of digits n_k
x_1	0.3	00	00	2
x_2	0.25	10	01	2
x_3	0.2	11	11	2
x_4	0.12	101	101	3
x_5	0.08	0001	1000	4
x_6	0.05	1001	1001	4

(ii) To obtain average number of bits per message $(\overline{N})$:

$\overline{N}$ is given as,

$$\overline{N} = \sum_{k=0}^{L-1} p_k\, n_k$$

Putting values in above equation from table 2.10.11,

$$\overline{N} = (0.3)(2) + (0.25)(2) + (0.2)(2) + (0.12)(3) + (0.08)(4) + (0.05)(4)$$

$$= 2.38$$

(iii) To obtain code efficiency :

Code efficiency is given as,

$$\eta = \frac{H}{\overline{N}}$$

$$= \frac{2.3568}{2.38} = 0.99$$

Thus the code efficiency of Shannon-Fano algorithm and Huffman coding is same.

(iv) Redundancy of the code :

Redundancy (γ) is given as,

$$\gamma = 1 - \eta$$

$$= 1 - 0.99 = 0.01$$

Ex.2.10.4 *A DMS have five symbols $s_0, s_1, \ldots, s_4$, characterized by probability distribution as 0.4, 0.2, 0.1, 0.2 and 0.1 respectively. Evaluate two distinct variable length Huffman codes for the source to illustrate nonuniqueness of Huffman technique. Calculate the variance of the ensemble as defined by,*

$$\sigma^2 = \sum_{k=0}^{M-1} p_k \left[l_k - l_{avg} \right]^2$$

where p_k and l_k are probability and length of codeword respectively for symbol s_k and l_{avg} is average length. Conclude on result. **[May-2002, 10 Marks]**

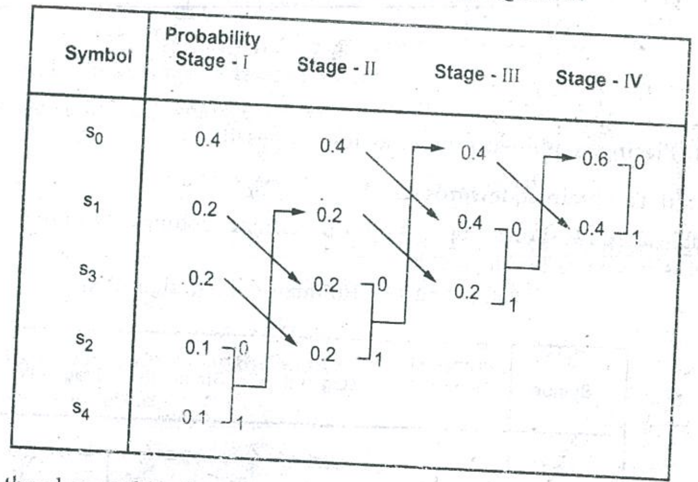
Sol. : Huffman coding can be implemented by placing the combined probability as high as possible or as low as possible. Let solve above example with these two techniques.

(I) Placing combined symbol as high as possible :

(i) To obtain codeword :

Table 2.10.12 lists the coding. The combined symbol is placed as high as possible.

Table 2.10.12 : Huffman coding algorithm



As per the above table, the codes are listed below :

Table 2.10.13

Symbol	Probability p_k	Digits obtained by tracing $b_2\, b_1\, b_0$	Codeword $b_0\, b_1\, b_2$	No. of bits per symbol n_k
s_0	0.4	0 0	0 0	2
s_1	0.2	0 1	1 0	2
s_2	0.1	0 1 0	0 1 0	3
s_3	0.2	1 1	1 1	2
s_4	0.1	1 1 0	0 1 1	3

(ii) To obtain average codeword length :

Average codeword length can be calcuiated as,

$$\overline{N} = \sum_{k=0}^{4} p_k n_k$$

$$= 0.4\ (2) + 0.2\ (2) + 0.1\ (3) + 0.2\ (2) + 0.1\ (3)$$

$$= 2.2$$

(iii) To obtain variance of code :

Variance can be calculated as,

$$\sigma^2 = \sum_{k=0}^{4} p_k \left[n_k - \overline{N}\right]^2$$

$$= 0.4[2-2.2]^2 + 0.2[2-2.2]^2 + 0.1[3-2.2]^2$$

$$+ 0.2[2-2.2]^2 + 0.1[3-2.2]^2$$

$$= 0.16$$

(II) Placing combined symbol as low as possible :

(i) To obtain codewords :

Table 2.10.14 shows the listing of huffman coding algorithm. The combined symbol is placed as low as possible.

Table 2.10.14 : Huffman coding algorithm

Symbol	Probability Stage - I	Stage - II	Stage - III	Stage - IV
s_0	0.4	0.4	0.4	0.6 0
s_1	0.2	0.2	0.4 0	0.4 1
s_3	0.2	0.2 0	0.2 1	
s_2	0.1 0	0.2 1		
s_4	0.1 1			

As per the above table, the codes are listed below :

Table 2.10.15

Symbol	Probability p_k	Digits obtained by tracing $b_2 b_1 b_0$	Codeword $b_0 b_1 b_2$	No. of bits per symbol n_k
s_0	0.4	1	1	1
s_1	0.2	1 0	0 1	2
s_2	0.1	0 1 0 0	0 0 1 0	4
s_3	0.2	0 0 0	0 0 0	3
s_4	0.1	1 1 0 0	0 0 1 1	4

(ii) To obtain average codeword length :

Average codeword length is given as,

$$\overline{N} = \sum_{k=0}^{4} p_k n_k$$

$$= 0.4\ (1) + 0.2\ (2) + 0.1\ (4) + 0.2\ (3) + 0.1\ (4)$$

$$= 2.2$$

(iii) To obtain variance of the code :

Variance can be calculated as,

$$\sigma^2 = \sum_{k=0}^{4} p_k \left[n_k - \overline{N}\right]^2$$

$$= 0.4[1-2.2]^2 + 0.2[2-2.2]^2 + 0.1[4-2.2]^2$$

$$+ 0.2[3-2.2]^2 + 0.1[4-2.2]^2$$

$$= 1.36$$

Results :

Sr. No.	Method	Average length	Variance
1	As high as possible.	2.2	0.16
2	As low as possible.	2.2	1.36

Above results show that average length of the codeword is same in both the methods. But minimum variance of Huffman code is obtained by moving the probability of a combined symbol as high as possible.

Ex.2.10.5 *A DMS has following alphabet with probability of occurrence as shown below :*

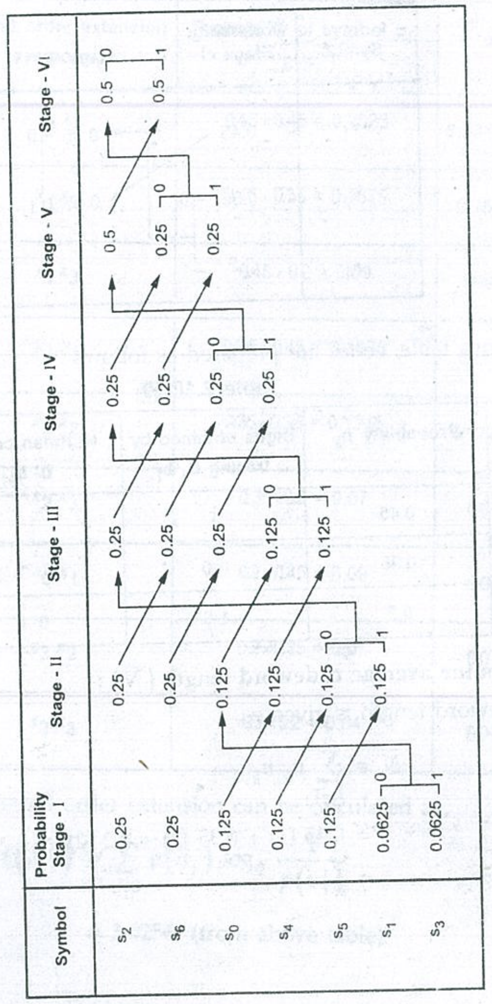
Symbol	s_0	s_1	s_2	s_3	s_4	s_5	s_6
Probability	0.125	0.0625	0.25	0.0625	0.125	0.125	0.25

Generate the Huffman code with minimum code variance. Determine the code variance and code efficiency. Comment on code efficiency. **[Dec-2002, 10 Marks]**

Sol. : (i) To obtain codewords :

Minimum code variance can be obtained in Huffman coding by putting the combined symbol probability as high as possible. Table 2.10.16 shows the coding based on this principle.

Table 2.10.16 : Huffman coding



Based on the above table, the codes are traced and generated as follows :

Table 2.10.17 : Codewords

Symbol	Probability	Digits obtained by tracing	Codeword	No. of bits per symbol n_k
s_0	0.125	1 0 0	0 0 1	3
s_1	0.0625	0 0 0 0	0 0 0 0	4
s_2	0.25	0 1	1 0	2
s_3	0.0625	1 0 0 0	0 0 0 1	4
s_4	0.125	0 1 0	0 1 0	3
s_5	0.125	1 1 0	0 1 1	3
s_6	0.25	1 1	1 1	2

(ii) To obtain average codeword length :

Average codeword length is given as,

$$\overline{N} = \sum_{k=0}^{6} p_k n_k$$

$$= 0.125\,(3) + 0.0625\,(4) + 0.25\,(2) + 0.0625\,(4) + 0.125\,(3) + 0.125\,(3) + 0.25\,(2)$$

$$= 2.625 \text{ bits/symbol.}$$

(iii) To obtain entropy of the source :

Entropy is given as,

$$H = \sum_{k=0}^{6} p_k \log_2\left(\frac{1}{p_k}\right)$$

$$= 0.125\log_2\left(\frac{1}{0.125}\right) + 0.0625\log_2\left(\frac{1}{0.0625}\right) + 0.25\log_2\left(\frac{1}{0.25}\right) + 0.0625\log_2\left(\frac{1}{0.0625}\right) + 0.125\log_2\left(\frac{1}{0.125}\right) + 0.125\log_2\left(\frac{1}{0.125}\right) + 0.25\log_2\left(\frac{1}{0.25}\right)$$

$$= 2.5 \text{ bits/symbol.}$$

(iv) To obtain code efficiency :

Code efficiency is given as,

$$\eta = \frac{H}{\overline{N}}$$

$$= \frac{25}{2.625} = 0.9523$$

Code efficiency is higher due to Huffman coding.

(v) To obtain variance :

$$\sigma^2 = \sum_{k=0}^{6} p_k \left(n_k - \overline{N}\right)^2$$

$$= 0.125(3-2.625)^2 + 0.0625(4-2.625)^2 + 0.25(2-2.625)^2$$

$$+ 0.0625(4-2.625)^2 + 0.125(3-2.625)^2$$

$$+ 0.125(3-2.625)^2 + 0.25(2-2.625)^2$$

$$= 0.4843$$

Ex.2.10.6 *A discrete memoryless source consists of three symbols x_1, x_2, x_3; with probabilities 0.45, 0.35 and 0.2 respectively. Determine the minimum variance Huffman codes for the source for following two alternatives :*
(i) Considering symbol by symbol occurrence.
(ii) Considering second order block extension of the source.
Determine the code efficiency for the two alternatives and comment on the efficiencies. **[May-2003, 10 Marks]**

Sol. : Following steps are required to solve this problem.

(I) Symbol by symbol occurrence

(i) Entropy of the source (H)
(ii) To determine Huffman code
(iii) To determine average codeword length $(\overline{N})$
(iv) To determine code efficiency (η)

(II) Second order extension

(i) To determine entropy of second order extension $H(X^2)$.
(ii) To determine Huffman code.
(iii) To determine average codeword length $(\overline{N})$
(iv) To determine code efficiency (η)

(I) Symbol by symbol occurrence :

(i) To determine entropy of the source :

Entropy is given as,

$$H = \sum_{k=1}^{3} p_k \log_2\left(\frac{1}{p_k}\right)$$

$$= 0.45 \log_2\left(\frac{1}{0.45}\right) + 0.35 \log_2\left(\frac{1}{0.35}\right) + 0.2 \log_2\left(\frac{1}{0.2}\right)$$

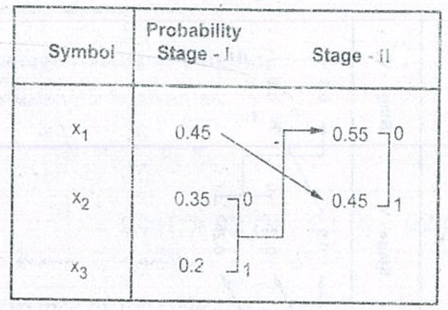
$$= 0.5184 + 0.5301 + 0.4643$$

$$= 1.5128$$

(ii) To determine Huffman code :

Table 2.10.18 lists the Huffman coding.

Table 2.10.18 : Huffman coding



Based on above table, codes are prepared as follows :

Table 2.10.19

Symbol	Probability p_k	Digits obtained by tracing b_0 b_1	Huffman code b_1 b_0	Number of bits / symbol n_k
x_1	0.45	1	1	1
x_2	0.35	0 0	0 0	2
x_3	0.2	1 0	0 1	2

(iii) To determine average codeword length $(\overline{N})$:

Average codeword length is given as,

$$\overline{N} = \sum_{k=1}^{3} p_k n_k$$

$$= 0.45\ (1) + 0.35\ (2) + 0.2\ (2)$$

$$= 1.55$$

(iv) To determine code efficiency :

Code efficiency is given as,

$$\eta = \frac{H}{\bar{N}}$$

$$= \frac{1.5128}{1.55} = 0.976$$

(II) Second order extension of the source :

(i) To determine entropy of second order extension :

In the second order extension, sequence of two symbols is considered. Since there are three symbols, second order extension will have nine symbols. Following table lists this second order extension.

Table 2.10.20 : Second order extension

i	Second order extension symbol σ_i	Probability of symbol σ_i $p(\sigma_i)$	$p(\sigma_i) \log_2 \frac{1}{p(\sigma_i)}$
1	$x_1 x_1$	$0.45 \times 0.45 = 0.2025$	$0.2025 \log_2 \frac{1}{0.2025} = 0.4665$
2	$x_1 x_2$	$0.45 \times 0.35 = 0.1575$	$0.1575 \log_2 \frac{1}{0.1575} = 0.42$
3	$x_1 x_3$	$0.45 \times 0.2 = 0.09$	$0.09 \log_2 \frac{1}{0.09} = 0.3126$
4	$x_2 x_1$	$0.35 \times 0.45 = 0.1575$	$0.1575 \log_2 \frac{1}{0.1575} = 0.42$
5	$x_2 x_2$	$0.35 \times 0.35 = 0.1225$	$0.1225 \log_2 \frac{1}{0.1225} = 0.371$
6	$x_2 x_3$	$0.35 \times 0.2 = 0.07$	$0.07 \log_2 \frac{1}{0.07} = 0.2685$
7	$x_3 x_1$	$0.2 \times 0.45 = 0.09$	$0.09 \log_2 \frac{1}{0.09} = 0.3126$
8	$x_3 x_2$	$0.2 \times 0.35 = 0.07$	$0.07 \log_2 \frac{1}{0.07} = 0.2685$
9	$x_3 x_3$	$0.2 \times 0.2 = 0.04$	$0.04 \log_2 \frac{1}{0.04} = 0.1857$

Entropy of the second order extension can be calculated as,

$$H(X^2) = \sum_{i=1}^{9} p(\sigma_i) \log_2 \frac{1}{p(\sigma_i)}$$

$$= 3.0254 \quad \text{(from above table)}$$

Above entropy can also be directly calculated as,

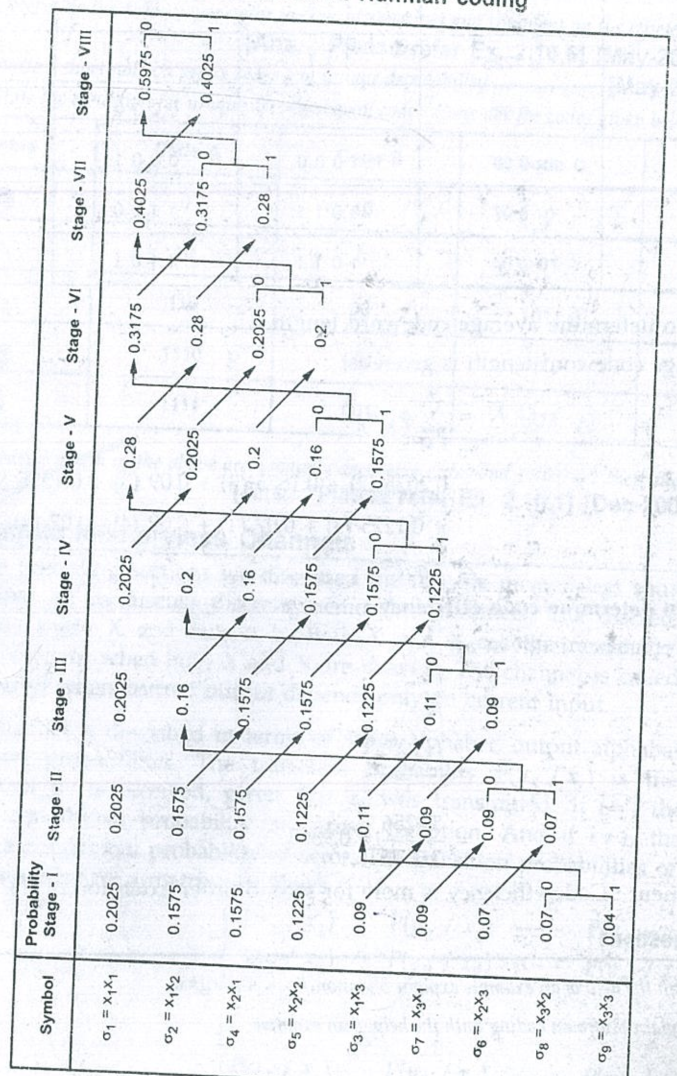
$$H(X^2) = 2H(X)$$

$$= 2 \times 15128 = 3.0256$$

(ii) To determine Huffman code :

Minimum variance Huffman code can be obtained by placing combined probabilities as high as possible. Table 2.10.21 lists the coding based on above principle.

Table 2.10.21 : Huffman coding



Based on above table, codes are prepared as follows :

Table 2.10.22

Symbol $p_k = \sigma_i$	Probability p_k	Digits obtained by tracing	Huffman code	Bits / message n_k
σ_1	0.2025	0 1	1 0	2
σ_2	0.1575	1 0 0	0 0 1	3
σ_3	0.09	0 1 0	0 1 0	3
σ_4	0.1575	1 1 0	0 1 1	3
σ_5	0.1225	1 1 1	1 1 1	3
σ_6	0.07	0 0 0 0	0 0 0 0	4
σ_7	0.09	1 0 0 0	0 0 0 1	4
σ_8	0.07	0 0 1 1	1 1 0 0	4
σ_9	0.04	1 0 1 1	1 1 0 1	4

(iii) To determine average codeword length :

Average codeword length is given as,

$$\overline{N} = \sum_{k=1}^{9} p_k n_k$$

$$= 0.2025\ (2) + 0.1575\ (3) + 0.09\ (3) + 0.1575\ (3) + 0.1225\ (3) + 0.07\ (4) + 0.09\ (4) + 0.07\ (4) + 0.04\ (4)$$

$$= 3.0675$$

(iv) To determine code efficiency :

Code efficiency is given as,

$$\eta = \frac{H}{\overline{N}} = \frac{H\left(X^2\right)}{\overline{N}} = \frac{3.0256}{3.0675} = 0.9863$$

Comment : Code efficiency is more for second order extension of the source.

Review Questions

1. *With the help of an example explain Shannon-Fano Algorithm.*
2. *Explain Huffman coding with the help of an example.*

Unsolved Examples

1. *A source generates 5 messages with probabilities of occurrence as shown below.*

Message	m_0	m_1	m_2	m_3	m_4
Probability	0.55	0.15	0.15	0.10	0.05

Apply Huffman coding algorithm and place the combined message as low as possible when its probability is equal to that of another message.

i) Calculate codeword for the messages.

ii) Calculate average codeword length (i.e. average number of binary digits per message)

[**Ans. :** (i) :

Message	Codeword
m_0	0
m_1	11
m_2	100
m_3	1010
m_4	1011

(ii) 1.9 binary digits / message]

2. *Apply Huffman coding algorithm to the messages in example 3. Place the combined message as high as possible when its probability is equal to that of another message.*

i) Calculate codeword for messages.

ii) Calculate average codeword length (i.e. average number of binary digits per message).

Hint : The stage-II for given data is shown below.

Message	Stage-I	Stage-II	
m_0	0.55	0.55	Combined probability is placed on higher side when they are equal.
m_1	0.15	0.15	
m_2	0.15	0.15	
m_3	0.10	0.15	
m_4	0.05		

In the above table observe that combined probability of m_3 and m_4 is equal to that of m_1 and m_2. The combined probability of m_3 and m_4 is placed on higher side in stage-II instead of bottom. This changes the codeword.

Ans : (i)