Problem 1.

**Erroneous answer.** We pick up a random bit by first picking a random source symbol $x_i$ with probability $p_i$, then picking a random bit from $c(x_i)$. If we define $f_i$ to be the fraction of the bits of $c(x_i)$ that are 1s, we find

$$\Pr(\text{bit is 1}) = \sum_{i=1}^{4} p_i f_i = 1/2 \cdot 0 + 1/4 \cdot 1/2 + 1/8 \cdot 2/3 + 1/8 \cdot 1 = 1/3.$$  

This answer is wrong. Do you see why?

The symbol $c$ and $d$ get encoded into longer-length binary strings than $a$, so when we pick up a bit from the compressed string at random we are more likely to land in a bit belonging to a $c$ or a $d$ than would be given by probability $p_i$. This means that all the probability need to be scaled up by $l_i$ and renormalized.

**Correct answer.** Every time symbol $x_i$ is encoded, $l_i$ bits are added to the binary string, of which $f_i l_i$ are 1s. The average number of 1s added per symbol is

$$\sum_{i=1}^{4} p_i f_i l_i,$$

and the average total number of bits added per symbol is

$$\sum_{i=1}^{4} p_i l_i.$$

So the fraction of 1s in the transmitted string is

$$\Pr(\text{bit is 1}) = \frac{\sum_{i=1}^{4} p_i f_i l_i}{\sum_{i=1}^{4} p_i l_i} = \frac{1/2 \cdot 0 + 1/4 \cdot 1 + 1/8 \cdot 2 + 1/8 \cdot 3}{7/4} = 1/2.$$  

Problem 2. The set of probabilities $\{p_1, p_2, p_3, p_4\} = \{1/6, 1/6, 1/3, 1/3\}$ gives rise to two different optimal sets of code-lengths, because at the second step of the Huffman coding we can choose any of the three possible pairings. We may either put them in a constant length code $\{00, 01, 10, 11\}$ or the code $\{000, 001, 01, 1\}$. Both codes have average length 2. Can you think of other solutions?

Problem 3.

There are 2 (FFFF,NNNN) World Series with 4 games. Each happens with probability $(1/2)^4$.

There are 8 = 2(4) World Series with 5 games. Each happens with probability $(1/2)^5$.

There are 20 = 2(5) World Series with 6 games. Each happens with probability $(1/2)^6$. 
There are $40 = 2^6$ World Series with 5 games. Each happens with probability $(1/2)^7$.

The entropy of $X$ is

$$H(X) = \sum p(x) \log \frac{1}{p(x)}$$

$$= 2(1/16 \log 16) + 8(1/32 \log 32) + 20(1/64 \log 64) + 40(1/128 \log 128)$$

$$= 5.81.$$  

The probability of a 4 game series ($Y = 4$) is $2(1/2)^4 = 1/8$.
The probability of a 5 game series ($Y = 5$) is $8(1/2)^5 = 1/4$.
The probability of a 6 game series ($Y = 6$) is $20(1/2)^6 = 5/16$.
The probability of a 7 game series ($Y = 7$) is $40(1/2)^7 = 5/16$.

The entropy of $Y$ is

$$H(Y) = \sum p(y) \log \frac{1}{p(y)}$$

$$= 1/8 \log 8 + 1/4 \log 4 + 5/16 \log 16/5 + 5/16 \log 16/5$$

$$= 1.92.$$  

Problem 4.

- The code \{0, 10, 11\} is a Huffman code for the distribution \{1/2, 1/4, 1/4\}.

- The code \{00, 01, 10, 110\} can be shortened to \{00, 01, 10, 11\} without losing its instantaneous property and therefore it is not optimal, so it cannot be a Huffman code.

- The code \{01, 10\} can be shortened to \{0, 1\}. Therefore it is not optimal and not a Huffman code.

Problem 5.

- The entropy of the source is 1.75.

- The corresponding Huffman code is \{0, 10, 110, 111\} and its average length is 1.75

- The corresponding Shannon-Fano code is as follows. In this case, the average codeword length is 2.75

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Probability</th>
<th>$F(x)$</th>
<th>$F(x)$ in binary</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1/2</td>
<td>0.25</td>
<td>0.001</td>
<td>001</td>
</tr>
<tr>
<td>b</td>
<td>1/4</td>
<td>0.75</td>
<td>0.5</td>
<td>10</td>
</tr>
<tr>
<td>c</td>
<td>1/8</td>
<td>0.875</td>
<td>0.1101</td>
<td>1101</td>
</tr>
<tr>
<td>d</td>
<td>1/8</td>
<td>1.00</td>
<td>0.1111</td>
<td>1111</td>
</tr>
</tbody>
</table>