Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
Department of Mechanical Engineering

6.050J/2.110J **Information and Entropy** **Spring 2003**

Issued: February 28, 2003 **Due: March 7, 2003**

Problem Set 4

Problem 1: Meet the Box People

A discrete character difference can often be determined by the difference in a single gene. We will refer to the outwardly detectable expression of the gene as its phenotype (red or blue for example). In higher organisms, genes are represented twice in each cell, as a gene pair. The genes we will be considering can come in one of two forms: either dominant or recessive. If a dominant gene is one of the genes of a pair, it will always be expressed. We will represent the dominant form of a gene with an uppercase letter and the recessive form with a lowercase letter.

In this problem, we consider the expression of two genes in an imaginary organism called a box person. We call them the r and c genes. The r gene determines the color of the box person, where red is the dominant phenotype and blue the recessive phenotype. The c gene determines the shape of the box person, where the circular shape is the dominant phenotype and the square shape the recessive phenotype. Here we will examine the progeny (offspring) of two box people. We will assume that both of the parents have RrCc gene combinations. Each child receives one of the mother's two color genes with equal probability (in this case either R or r) and one of the father's two color genes also with equal probability (again either R or r). For some reason the shape genes act differently: a child receives the dominant gene C only 30% of the time (rather than 50%) from a mother or a father that has one of each. We assume that for each child these four selections are independent. Below is a chart showing all sixteen possible results for a single child. This chart shows that the recessive gene is expressed in the shape or color of the child only if both of the genes are of the recessive type. (When this chart is printed without color, blue usually appears darker than red.)

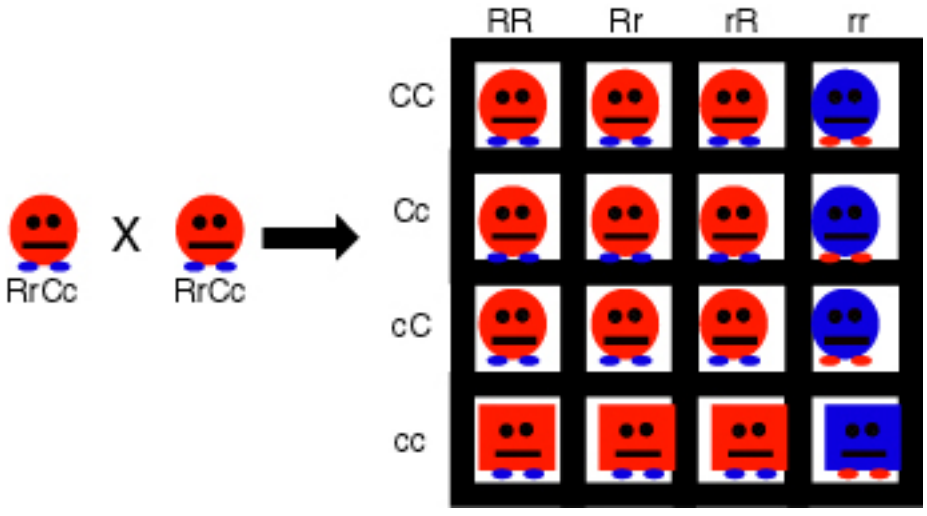

Figure 4–1: The Box People

a. What is the probability that one of the box people's offspring has...

   i. a circular phenotype?

   ii. a square phenotype?

   iii. a blue phenotype?

   iv. a red phenotype?

b. If you're told that one of their offspring has a circular phenotype, what's the probability that the offspring also has...

   i. a blue phenotype?

   ii. a red phenotype?

c. If instead of what you were told in part b., you're told that one of their offspring has a red phenotype, what's the probability that the offspring also has...

   i. a box phenotype?

   ii. a circular phenotype?

d. This question relates to another gene of the box people, unrelated to the r and c genes. Unfortunately, a mutant gene can turn box people into triangles late in life. A laboratory test has been developed which can spot the gene early so that the dreaded triangle transformation can be prevented by medications. This test is 95 percent accurate at spotting the gene when it is there. However, the test gives a "false positive" 0.4 percent of the time, falsely indicating that a healthy box person has the mutant gene. If 0.1 percent (be careful – that's one-tenth of one percent) of the box people have the mutant gene, what's the probability that a box person actually has the mutant gene if the test indicates that he or she does?

Problem 2: Huffman Coding

Note: It is not necessary to use MATLAB for this problem; however, you should feel free if you enjoy the challenge. If you decide to use MATLAB, please place your code in ps4p2.m.

You will make use of Huffman coding for this problem. You have been asked to encode a tounge-twister phrase compactly. This is the sequence of characters, you may recognize it from Problem Set 2:

peter piper picked a peck of pickled peppers

For your convenience here is the frequency distribution is listed in Table 4–1.

a. One way of coding this sequence would be to use a fixed-length code, with each code word long enough to encode fourteen different symbols (this is not a Huffman code). How many bits would be needed for this 44-character phrase using such a fixed-length code?

b. Determine the theoretical minimum number of bits required to encode the entire phrase (this is the information content of the phrase), assuming that each character is independent of the surrounding character. As reference, the equation from class for the average information of a single symbol is:

$$\sum_{i=1\ldots n} p_i \log_2\left(\frac{1}{p_i}\right) \tag{4–1}$$

Character	#	Frequency
p	9	20.46%
e	8	18.18%
space	7	15.91%
c	3	6.82%
i	3	6.82%
k	3	6.82%
r	3	6.82%
d	2	4.55%
a	1	2.27%
f	1	2.27%
l	1	2.27%
o	1	2.27%
s	1	2.27%
t	1	2.27%
Total	44	100.00%

Table 4–1: Frequency distribution of characters in "peter piper picked a peck of pickled peppers"

c. What is the theoretical contribution of each of the fourteen symbols to this average?

d. Derive a codebook for the fourteen symbols using Huffman coding. This codebook will, of course, depend on the frequencies of the symbols in the original phrase.

e. When the sequence is encoded using the codebook derived in part d. ...

   i. How many bits are needed?

   ii. How does this compare with the number of bits needed to use the fixed length code of part a?

   iii. How does this compare with the information content of the phrase as calculated in part b?

f. An alternate approach to producing a compact code would be to encode the notes as in part a. above and then use LZW lossless compaction. Compare the number of bits needed using LZW with the numbers derived above in parts a. and e. The LZW approach has the advantage that the dictionary does not have to be transmitted. Estimate how many bits are needed to transmit the codebook if Huffman coding is used.

Turning in Your Solutions

You should have 1 file, your diary (if you chose to use MATLAB for Problem 2, you may have an M-file as well). Name the diary `ps4diary`. If you are submitting an M-file for Problem 2, please name it `ps4p2.m`. You may turn in this problem set by e-mailing your M-files and diary. Do this either by attaching them to the e-mail as *text* files, or by pasting their content directly into the body of the e-mail (if you do the latter, please indicate clearly where each file begins and ends). The deadline for submission is the same no matter which option you choose.

Your solutions are due 5:00 PM on Friday, March 7, 2003. Later that day, solutions will be posted on the course website.

6.050J / 2.110J Information and Entropy
Spring 2008