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

6.050J/2.110J **Information and Entropy** **Spring 2005**

Issued: February 28, 2005 **Due: March 4, 2005**

Problem Set 5

Problem 1: Give Peas a Chance

The laws of heredity were first postulated in 1865 by Gregor Mendel (1822–1884) after a remarkable eight-year series of experiments on the common garden pea. He introduced probability to the study of heredity, and explained how offspring have a chance to exhibit traits not seen in the parents. He gave peas a chance.

Prior to Mendel, the accepted theory of inheritance was the blending theory, which stated that offspring would have traits intermediate between those of the parents. For example, a red flower crossed with a white flower would produce a pink flower. Observations to the contrary were considered "anomalies."

Mendel planned and executed his experiments well. It took him two years to decide to use the pea plant. It could be self-pollinated (from the same plant) or cross-pollinated (from different plants) easily, and protected against accidental pollination[6]. It produced offspring rapidly. It had several "binary" **characteristics**, each with two easily observed possible **traits** and no intermediate forms (he chose seven to study). He developed the skill to do the pollination. He kept records carefully. He knew statistics and probability, having studied science and mathematics at the University of Vienna. He belonged to the scientific community of his time and read the literature. This was a time of ferment in biology—blending theory was unable to explain Darwin's theory of evolution, which was published in 1859. Finally, Mendel was a monk at a monastery that provided him a garden to use, support, a lively intellectual environment, and encouragement to pursue his scientific studies.

The seven characterics studied by Mendel are:

1. Shape of ripe seeds, which could be smooth or wrinkled
2. Seed color, which could be green or yellow
3. Seed coat color, which could be white or grey
4. Ripe seed pod, which could be inflated or constricted
5. Unripe pod color, which could be green or yellow
6. Position of flower on stem, which could be axial or terminal
7. Length of stem, which could be long or short

Strains

The first thing Mendel did was produce "strains" of peas (plants whose offspring from self-pollination would consistently have the same traits). This took several generations of selection with careful record keeping, since it could not be known if a plant was a strain until several generations of its offspring were observed. If Mendel had wanted to study all seven characteristics together, he would have needed 128 strains, one for each combination of traits. However, he decided to study the characteristics separately, so he could use plants that were a strain for one characteristic but not necessarily others.

a. How many strains did he need?

[6] If you think of producing offspring as a flow of information, then accidental pollination is a source of errors of the sort discussed in Chapter 4.

Hybrids

Consider the seed-color characteristic, which is either green or yellow. According to blending theory, if a plant with yellow seeds is crossed with one with green seeds, the result should be a plant with light-green seeds, except for anomalies. Mendel crossed his strains, producing "hybrids." He observed that the seeds were **all yellow**. Today we would say yellow is the "dominant trait" and green the "recessive trait;" Mendel concluded as a result of his work that these words were appropriate and started using them.

b. Summarize his results and the predictions of blending theory in this table (you may assume that in blending theory anomalies happen 5% of the time):

	Blending Theory	**Mendel Observation**
yellow seeds		
light green seeds		
greenseeds		
Total	100%	100%

First Generation from the Hybrids

Mendel then took hybrid plants, all of which had yellow seeds, and self-pollinated them to produce a set of plants he called "first generation from the hybrids." He observed both yellow seeds and green seeds (never light green seeds). He did this experiment many times, using many different hybrid plants, ultimately making thousands of observations. Table 5–1 shows data from his paper about ten trials using ten different hybrid plants, each trial producing many offspring plants that survived to have their seed colors observed.

	Experiment 2	
	Color of Albumen	
Trial	Yellow	Green
1	25	11
2	32	7
3	14	5
4	70	27
5	50	14
6	20	6

Table 5–1: Second set of experiments: Self-pollination of the hybrids. Summary of 6 trials observing the color of the albumen. This table is extracted from the works published by Gregor Mendel

Mendel understood the difference between statistics and probability. Using these ten trials, he noted the observed ratio of yellow to green (a statistic). He wanted to model the observation of a characteristic as an outcome (in the sense of probability theory) and the possible traits as events. He wanted probability values $P(seed = yellow)$ and $P(seed = green)$ that were simple (a fraction between 0 and 1 involving only small integers).

c. Using the data from Table 5–1, add a row on the bottom for the total. Then add a column on the right for the percentage of green seeds for each trial (and for the total).

d. What do you propose for $P(seed = green)$ and $P(seed = yellow)$?

Mendel did many such experiments, for this characteristic and also six others. Some of his findings are in Table 5–2.

Experiment	Trait	count	Trait	count
Shape of ripe seeds	*smooth*	5474	*wrinkled*	1850
Seed color	*green*	2001	*yellow*	6022
Shape of ripe pod	*inflated*	882	*constricted*	299
Position of the flower	*axial*	651	*terminal*	207

Table 5–2: Summary of results for the first generation from the hybrids. This table is extracted from Gregor Mendel's original work.

e. For each of the four characteristics in Table 5–2, consider a similar probabilistic model and give the (simple) probabilities for the two events.

   All the characteristics have a dominant trait and a recessive trait (just like seed color). The dominant trait (the one displayed by hybrids) is the one most often seen in first generation plants.

f. Explain why the probability of a dominant trait in the first generation is not equal to the actual ratio of dominant trait count to total count in Table 5–2.

Second Generation from the Hybrids

Mendel then used self-pollination of the first-generation plants to obtain offspring which he called "second generation from the hybrids." He discovered that all such plants whose single parent exhibited the recessive trait (for example green seeds) also did so, but that first-generation plants exhibiting the dominant trait (for example, yellow seeds) produced offspring with the recessive trait in about 1/6 of the time and the dominant trait 5/6 of the time[7]. He modeled this in terms of probabilities, for example $P(r_2|D_1) = 1/6$. (The notation here is that r is a recessive trait and D a dominant trait, and the subscript refers to the generation from the hybrid.)

g. Find the four conditional probabilities $P(D_2|D_1)$, $P(D_2|r_1)$, $P(r_2|D_1)$, and $P(r_2|r_1)$.

h. Using Bayes' Theorem, find the joint probabilities $P(r_2, D_1)$ and $P(r_2, r_1)$. (Remember that for first-generation plants, $P(r_1) = 1/4$.)

i. Using the fact that D_1 and r_1 are a partition (mutually exclusive and exhaustive), find the unconditional probability $P(r_2)$.

j. Suppose you have a second-generation plant with recessive trait. What is the probability that its parent displayed the dominant trait?

Mendel went further, looking at third and later generations, and at plants produced by cross pollination, and developed a simple theory of heredity. His theory was not fully accepted for 35 years, but when it was, it provided the foundation for 20th century biology.

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 ps5p2.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 3:

[7] These numbers are valid going from first to second generation, but not for subsequent generations. Mendel explained them by saying that plants showing the dominant trait are of two types—those that are strains and those that are hybrids. The self-pollinated offspring of strains are all strains, while 2/3 of the self-pollinated offspring of hybrids that show the dominant trait are hybrids.

`de do do do de da da da`[8]

For your convenience the frequency distribution is listed in Table 5–1.

Character	#	Frequency
d	8	34.78%
space	7	30.43%
a	3	13.04%
o	3	13.04%
e	2	8.70%
Total	23	100.00%

Table 5–1: Frequency distribution of characters in "de do do do de da da da"

a. One way of coding this sequence would be to use a fixed-length code, with each code word long enough to encode five different symbols (this is not a Huffman code). How many bits would be needed for this 23-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:

$$\sum_{i=1\ldots n} p_i \log_2\left(\frac{1}{p_i}\right) \tag{5–1}$$

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

d. Derive a codebook for the five 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

Make sure you turn in your M-files and diary, if you used MATLAB for this assignment. You may turn in this problem set by e-mailing your M-files and diary to 6.050-submit@mit.edu. 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). Alternatively, you may turn in your solutions on paper in room 38-344. The deadline for submission is the same no matter which option you choose.

Your solutions are due 5:00 PM on Friday, March 4, 2005. Later that day, solutions will be posted on the course website.

[8] Lyrics adapted from The Police's song "De do do do, de da da da"