

Issued: **February 28, 2005**

Problem Set 5 Solutions

Due: **March 4, 2005**

Solution to **Problem 1: Give Peas a Chance**

Strains

Solution to Problem 1, part a.

Mendel needed 14 strains. If he had wanted to evaluate all possible cross interaction between all the traits he would have needed 2^7 instances (think of each characteristic as a bit taking any of two values, and there are 7 characteristics). However since he only wanted to evaluate each characteristic separately he only needed 14 strains.

Hybrids

Solution to Problem 1, part b.

The table below shows the predictions of Blending theory face to face with the observations of Mendel.

	Blending Theory	Mendel Observation
yellow seeds	2.5%	100%
light green seeds	95%	0%
greenseeds	2.5%	0%
Total	100%	100%

First Generation from the Hybrids

Trial	Experiment 2 Color of Albumen		Percentage of Yellow Seeds	Percentage of Green Seeds
	Yellow	Green		
1	25	11	69%	31%
2	32	7	82%	18%
3	14	5	74%	26%
4	70	27	72%	28%
5	50	14	78%	22%
6	20	6	77%	23%
TOTAL	211	70	75%	25%

Table 5-2: Table for the second set of experiments, extended with the total and the proportion of green and yellow seeds.

Solution to Problem 1, part c.

Table 5-2, shows the result of computing the proportion of green seeds.

Solution to Problem 1, part d.

Table 5-2 shows noticeable fluctuations in the proportion of green seeds around a value of 25%. Considering experimental errors, a good choice for the probabilities of each trait is: $P(\text{seed} = \text{green}) = 1/4$ and $P(\text{seed} = \text{yellow}) = 3/4$.

Mendel did many such experiments, for this characteristic and also six others. Some of his findings are in Table 5-3.

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

Table 5-3: Summary of results for the first generation from the hybrids. This table is extracted from Gregor Mendel's original work.

Solution to Problem 1, part e.

The proportion of 3 to 1 is still apparent from table 5-3.

Solution to Problem 1, part f.

This is the essential difference between statistics and probability. The more data we have the more we expect statistics to reproduce the probabilistic model. However nature samples from this probabilistic model randomly, and this random sampling will introduce certain departures from the expected ratio.

Second Generation from the Hybrids

Solution to Problem 1, part g.

the problem statement gives us already $P(r_2|D_1) = 1/6$. Consequently, since only two outcomes are possible: $P(D_2|D_1) = 1 - P(r_2|D_1) = 5/6$. We know that if the parent exhibited the recessive trait so will its offspring, therefore: $P(r_2|r_1) = 1$, and once again, since only two outcomes are possible: $P(D_2|r_1) = 0$.

Solution to Problem 1, part h.

Using Bayes' Theorem, $P(r_2, D_1) = P(r_2|D_1) \cdot P(D_1)$. In part c we determined that $P(D_1) = 3/4$, and with the numbers from part g:

$$P(r_2, D_1) = P(r_2|D_1) \cdot P(D_1) = 1/6 \cdot 3/4 = 1/8$$

. Similarly, for $P(r_2, r_1)$:

$$P(r_2, r_1) = P(r_2|r_1) \cdot P(r_1) = 1 \cdot 1/4 = 1/4$$

.

Solution to Problem 1, part i.

Since D_1 and r_1 are a partition $P(r_2) = P(r_2, D_1) + P(r_2, r_1) = 1/8 + 1/4 = 3/8$.

Solution to Problem 1, part j.

We are being asked to compute the probability: $P(D_1|r_2)$. Applying bayes rule in the other direction:

$$P(D_1|r_2) = P(D_1, r_2)/P(r_2) = \frac{1/8}{3/8} = 1/3$$

Solution to Problem 2: Huffman Coding**Solution to Problem 2, part a.**

to encode 5 characters we would need 3 bits. That is $2^3 = 8$ codewords, and a total of $3 \times 23 = 69$ bits to transmit the sentence.

Solution to Problem 2, part b.

Table 5-4 lists the calculation of the average information per symbol. Here we calculate an average of 2.1252 bits per symbol, or 49 bits.

Character	#	Frequency	$\log_2 \left(\frac{1}{p_i} \right)$	$p_i \log_2 \left(\frac{1}{p_i} \right)$
d	8	34.78%	1.5236	0.5299
space	7	30.43%	1.7162	0.5223
a	3	13.04%	2.9386	0.3833
o	3	13.04%	2.9386	0.3833
e	2	8.70%	3.5236	0.3064
Total	23	100.00%		2.1252

Table 5-4: Frequency distribution of characters in “de do do do de da da da”

Solution to Problem 2, part c.

See Table 5-4.

Solution to Problem 2, part d.

A possible code is derived below and listed in Table 5-5.

Start: (d='NA' $p = 0.3478$) (space='NA' $p = 0.3043$) (a='NA' $p = 0.1304$) (o='NA' $p = 0.1304$) (e='NA' $p = 0.0870$)

Next: (d='NA' $p = 0.3478$) (space='NA' $p = 0.3043$) (a='NA' $p = 0.1304$) (o='0' e='1' $p = 0.2174$)

Next: (d='NA' $p = 0.3478$) (space='NA' $p = 0.3043$) (a='0' o='10' e='11' $p = 0.3478$)

Next: (d='NA' $p = 0.3478$) (space='0' a='10' o='110' e='111' $p = 0.6521$)

Next: (d='0' space='10' a='110' o='1110' e='1111' $p = 1.0000$)

Character	Code
d	0
space	10
a	110
o	1110
e	1111

Table 5-5: Huffman code for “de do do do de da da da”

Solution to Problem 2, part e.

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

- i. See Table 5-6.

Character	# of Characters	Bits per Character	Bits Needed
d	8	1	8
space	7	2	14
a	3	3	9
o	3	4	12
e	2	4	8
Total	23		51

Table 5-6: Huffman code for “de do do do de da da da”

- ii. The fixed length code requires 69 bits, whereas Huffman coding requires 51 bits. So we find that the Huffman code does a better job than the fixed length code.
- iii. This number compares extremely well with the information content of 49 bits for the message as a whole.

Solution to Problem 2, part f.

The original message is 23 bytes long, and with LZW we know from Problem Set 3 we can encode the message using LZW in 18 bytes, with 15 additional entries in the dictionary. Thus we need 18+7=25 different dictionary entries (do not forget the 5 characters and the start and stop signals), for a total of 5 bits per byte. Thus we can compact the message down to $23 \times 5 = 115$ characters. Straight encoding needs 69 bits, and Huffman encoding needs 51 bits. Thus Huffman encoding does the best job of compacting the material (assuming we do not need to transmit the codebook).