Ling 236 Homework #1 answers

These are brief answers to the questions. Please talk to me or a fellow student if you can't make sense of the answers.

1. If a 5 letter word is formed at random (meaning that all sequences of 5 letters over the 26 letter Latin alphabet are equally likely), what is the probability that there is no letter that occurs more than once in the word?

   Answer: This is answered by a counting argument. There are $26 \times 26 \times 26 \times 26 \times 26$ five letter words, but if each letter can only be used once, there are $26 \times 25 \times 24 \times 23 \times 22$ possible five letter words (see Rice, ch. 1, Proposition A). So, the probability is the latter divided by the former, which is ≈ 0.664.

2. Combinatorics and probability:

   (a) Assume that the only 6 possible vowels in human languages are /a/, /e/, /i/, /o/, /u/, /ə/. With no other restrictions, how many 4 vowel systems are possible?
   Answer: This is $\binom{6}{4} = 15$ (Proposition B).

   (b) Assuming that each of those (4 vowel) inventories is equiprobable, what are the chances that a language will have /a/ in its inventory?
   Answer: If you have to choose /a/, then there are $\binom{5}{3} = 10$ ways you can fill out the inventory. So the probability is $\frac{10}{15} = \frac{2}{3}$.

   (c) What are the chances that a 4 vowel language will have /a/ and /u/ in its inventory?
   Answer: If you have to choose /a/ and /u/, then there are $\binom{4}{2} = 6$ ways you can fill out the inventory. So the chance is $\frac{6}{15} = \frac{2}{5}$.

   (d) Really, things aren't equiprobable (and some languages have more than 6 vowels!). According to the UCLA Phonological Segment Inventory Database sample, $P(\text{lg has /i/}) = 0.87$. Assuming that the other vowels are equiprobable, and that their occurrence is independent of /i/ and each other, reanswer questions (b) and (c) (for 4 vowel languages).
   Answer: By the above (symmetrically), 10 languages types have /i/ and /5/ don't. Of the ones that have /i/, there are $\binom{4}{2} = 6$ that have /a/. So their probability is $\frac{6}{10} \times 0.87 = 0.522$. Of the ones without /i/, there are $\binom{4}{3} = 4$ that have /a/. So, their probability is $\frac{4}{5} \times (1 - 0.87) = 0.104$. Thus the total probability of a language with /a/ is 0.626. It's a bit less than before (i.e., 2/3), because making /i/ more frequent, makes other things less frequent. The case for /a/ and /u/ is worked out similarly. The probability is $0.87 \times \binom{3}{1}/\binom{5}{3} + 0.13 \times \binom{3}{2}/\binom{5}{4} = 0.339$.

3. Suppose (in a certain genre of text) the probability that a word is a noun is 0.4, and the probability that a word is a verb is 0.2. Suppose also that the probability that the word is of Latin origin is 0.3.

   (a) Given just the above information what are the bounds on the minimum and maximum possible probability of a random word being a latinate noun.
   Answer: Since any interaction is possible, all we know is $P(\text{noun}, \text{latin}) \geq 0$, by axioms of probability, and $P(\text{noun}, \text{latin}) \leq 0.3$, since $P(A, B) \leq P(A)$, since $P(A) = P(A, B) + P(A - B)$, where $A - B$ is the set of things in A that are not in B. Therefore, $P(\text{noun}, \text{latin}) \leq \min(P(\text{noun}), P(\text{latin})) = 0.3$.

(b) Assuming that part of speech and latinate origin are independent, what is the probability that a random word is a latinate noun.

Answer: $0.4 \times 0.3 = 0.12$

(c) Suppose the probability of latinate nouns is actually 0.15. What is the probability that a random word is a noun not of Latin origin?

Answer: $P(\text{noun}, \overline{\text{latinate}}) = 0.25$, by the argument in (a).

4. In a multiple choice test, each question has 5 possible answers. The probability that Sue knows the correct answer to a question is 3/4. If she knows the correct answer, there is a 95% chance that she'll color in the right circle (sometimes she get gets confused...). If she doesn't she will guess randomly. Given that she gave the correct answer to question 10 on subjacency, what is the probability that she actually knows the answer?

Answer: This can be calculated directly, or by using Bayes' Rule, as I will do here. $P(K|A) = \frac{P(K)P(A|K)}{P(K)P(A|K)+P(\overline{K})P(A|\overline{K})} = \frac{0.75\times0.95}{0.75\times0.95+0.25\times0.2} = 0.9344$. So, she most likely knew the answer.

5. Which of these are true?

(a) $P(A|B,C) \leq P(A|C)$

Answer: False. Knowing B is true could make A more likely

(b) $P(A,B|C) \leq P(A|C)$

Answer: True. A and B being true cannot be more likely than A being true.

(c) $P(A \cup B) \geq P(A) + P(B) - 1$

Answer: True. $P(A \cup B) = P(A) + P(B) - P(A \cap B) \geq P(A) + P(B) - 1$

6. A very famous early example of probabilities done over text was A. A. Markov's counts of consonants and vowels and their sequencing. Here's the data from A. A. Markov's count of exactly 20000 letters of the first part of A. Pushkin's novel *Eugène Onégin* – in Russian. We assume that he just ignored spaces.

Markov defined two events, which he (unmnemonically) called:

E: a vowel occurred
F: a consonant occurred

He then did counts to get the following probabilities, all as relative frequency estimates:

p estimate of probability that the next letter is an event E
p_1 estimate of probability that the next letter is an event E, given that the preceding letter was an event E
p_2 estimate of probability that the next letter is an event E, given that the preceding letter was an event F
q estimate of probability that the next letter is an event F
q_1 estimate of probability that the next letter is an event F, given that the preceding letter was an event E
q_2 estimate of probability that the next letter is an event F, given that the preceding letter was an event F

In the text, Markov found 1104 instances of vowel-vowel sequences, while the total number of vowels in the text was 8638. Assume for convenience that we regard the text as circular, so that the very last letter counts as the preceding letter for the very first letter.

Using this information:

(a) Express the quantities p, p_1, p_2, q, q_1, and q_2 in a more mnemonic, intelligible notation (e.g., with events and conditional probabilities).

(b) What is the value of p, p_1, and p_2? By how much do p_1 and p_2 differ? What about for q, q_1 and q_2?

Answer: There are 8638 vowels, each of which is preceded by something. Since 1104 of them are preceded by vowels, 7534 of them were preceded by consonants. There are $20000 - 8638 = 11362$ consonants, and with circular text, there are as many VC pairs as CV pairs, so 7534 were preceded by vowels. The remaining 3828 were then preceded by consonants. Thus, the MLE estimates are:

$$\begin{aligned} P(L_n = V) = p &= 8638/20000 \approx 0.432 \\ P(L_n = V | L_{n-1} = V) = p_1 &= 1104/8638 \approx 0.128 \\ P(L_n = V | L_{n-1} = C) = p_2 &= 7534/11362 \approx 0.663 \\ \mathrm{p}(L_n = C) = q &= 11362/20000 \approx 0.568 \\ \mathrm{p}(L_n = C | L_{n-1} = V) = q_1 &= 7534/8638 \approx 0.872 \\ \mathrm{p}(L_n = C | L_{n-1} = C) = q_2 &= 3828/11362 \approx 0.337 \end{aligned}$$

The difference between p_1 and p_2 is 0.535, which is also the difference between q_1 and q_2 (obvious consequence of stochastic constraints).

(c) What simple fact of language does the differences between p and p_1 and p_2 (or between q, q_1, and q_2) capture?

Answer: Syllable structure is constrained; it's not the case that vowels and consonants occur independently in natural language. You are more likely to get a consonant after a vowel than another vowel, and similarly for after a consonant. (Here, to be precise, we are talking just in terms of the orthographic representation.)

(d) Suppose we see a short Russian text with CV structure:

CVCCVCCVCV

i. What is its probability in a unigraph model (using just p and q)?
Answer: $P_1(\text{CVCCVCCVCV}) = p^4 q^6 = 0.432^4 \times 0.568^6 \approx 0.00117$

ii. What is its probability in a digraph model (using p_1, p_2, q_1, and q_2)?
Answer: I didn't precisely say what to do with the first letter. You could either assume that this text was circular too, or use a unigram estimate for the first letter. I will do the latter here. $P_1(\text{CVCCVCCVCV}) = q p_2^4 q_1^3 q_2^2 \approx 0.00826$.

iii. Assuming that this is a representative text of Russian, which appears to be a better model of Russian CV structure?
Answer: The digraph model makes the text more than 7 times more likely. This means that it is capturing more of the structure of Russian, and hence is a better model of Russian. Strings that occur in the language are more likely in the model. In particular, both models get the relative frequency of vowels and consonants right, but the digraph model captures some of the sequence structure.