ANAGRAMS HOW MANY ARE THERE?

REX GOOCH
Letchworth, Herts, England

This short article was prompted by the formula for Prob(N) in Michael Keith's "What are the odds that X is an anagram of Y?" in *Word Ways*, August 2000. It seemed to me that multiplying the values of that function by the square of the number of words of a given length might be expected to give an estimate of the number of words of a given length that are permutations of another word.

I was unhappy that there seemed to be no reason for the form of the function chosen for curve-fitting (see below), and not clear how 10 or 4 or 7 related to words. As its predictions were also disappointing, I wondered what an admittedly crude estimate might yield. If we take a given word with l different letters, there are in all $l!$ (ie the product of all the integers from 1 to l) word strings that are permutations (anagrams) of it, including itself. For example, ABCD has 4x3x2x1 permutations, (but AAAA has just 1). There are potentially $26!/(26-l)!$ word strings of length l with all letters different, but a word list will only recognise say n of these as words (eg, there are 26x25x24x23 heterogrammatic word strings of length 4). So we reduce our number of permutations by scaling $l!$ according to our word list, ie by multiplying by n then dividing by 26!/(26-l)! This gives the expected number of permutations for a single word. Multiplying by n again gives the expected number of permutations for all words of length l. The assumptions made are fairly evident, so an exact answer is scarcely expected. Here are the results, expressed as the predicted number divided by the actual number of anagrams, so that 1 means perfect prediction (for example, my formula over-estimates the true number of 5-letter anagrams by 28%, and underestimates those of length 6 by 24%). "% anagrams" means the actual percentage of all words of the given length that, upon re-arrangement of letters, can produce at least one other word: the theoretical values for heterograms are calculated as $26!/(26-l)!$ divided by 26^l.

Word length, l	Prob(N) x n^2	$l!$ x $(26-l!)/26!$ x n^2	% anagrams	% Heterograms	
	both cols divided by number of anagrams			actual	theory
5	7.9	1.28	58	61	66
6	5.3	0.76	48	45	54
7	3.4	0.48	34	32	41
8	2.2	0.36	23	20	30
9	1.5	0.30	19	11.5	21
10	1.0	0.27	9.7	5.8	14
11	0.7	0.30	5.0	2.7	8
12	0.4	0.32	3.1	1.1	5
13	0.2	0.38	1.9	0.35	3
14	0.1	0.41	1.4	0.09	1

As Keith used capital N where I use l, Prob(N) is the reciprocal of 10 raised to the power

$(2.55l - 1.3)^{4/7}$

Given that vowels account for 40% of the letters in dictionary words, by length 13 a heterogram would be expected to use each one of the five vowels, which explains why the formula predicts too many heterograms at longer lengths (indeed, I found just one heterogrammatic anagram of length 14 or more: LYMPHADENOTICS, ENDOLYMPHATICS). If the 26!/(26-l)! is too big, then the formula for anagrams will under-estimate, more grieviously so for longer words. On the contrary, the *l*! over-estimates, as non-heterograms have fewer than *l*! distinct permutations. These two reasons for the mismatch between theory and actuality are not the full story, because of the rather poorer results for pure heterograms (not reported here).

Keith provided, for his formula, the ratio of predicted to actual anagrams for a much smaller word list (word lengths 5 to 11) were: 2.2, 1.6, 1.1, 0.75, 0.46, 0.31, 0.20. For this smaller wordlist, the results are better for short words but worse for longer words.

The percentage of heterograms is given to enable readers to see the magnitude of one assumption: a better theoretical formula should not be too difficult to develop.

Neither approach is especially accurate, though mine is a little more so, in addition to having a clear rationale: if accuracy be desired, perhaps someone would care to fit a curve to real data, with the number of anagrams depending both on the word length and on the size of the vocabulary.

I should like to thank Michael Keith for his helpful comments on a draft of this article.