ANAGRAMS AND THE BIRTHDAY PROBLEM

A. ROSS ECKLER
Morristown, New Jersey

In his landmark book *An Introduction to Probability Theory and its Applications* (Wiley, 1950), William Feller describes the birthday problem, showing how to calculate the probability that in a given group of people at least two share the same month-and-day birthday. (For 23 people, this probability exceeds one-half.) Instead of talking about people sharing birthdays, one can consider words sharing letter distributions; two words having the same distribution are anagrams. It is easy to mathematically derive the probability that two or more share a birthday because the 365 possible birthdays are equally likely to occur. The analogous anagram probability is difficult to calculate because the alternatives are not equally likely (two words are far more likely to share a distribution like ARTS than ZYQQ). Instead, one must evaluate the behavior of anagrams by observation—specifically, how many of a set of W words are anagrams?

On page ua of *Opperlans! Taal- & Letterkunde* (Querido, 2002), Battus (Hugo Brandt Corstius) evaluates how rapidly the number of anagrams A increase with W. Using a demographic argument in which a pestilence kills off half of a population which has both single people and married couples (anagrams), he shows that $A = W^2/N$, where N is a constant to be evaluated from the data. The square root of 2N yields the average number of words needed to contain a single pair of words that are anagrams of each other; it is the solution to $2 = W^2/N$. This is the linguistic equivalent of the Feller birthday problem. If one examines the commonest three-letter words as specified by Kucera & Francis, *Computational Analysis of Present-Day American English* (Brown University Press, 1967), one finds that only 30 words are needed to capture the anagram-pair who-how. A somewhat better estimate of the number needed can be derived by noting that there are 13 anagrams (she-he's, who-how, its-sit, was-saw, god-dog, now-own-won) among the commonest 123 three-letter words. A two-word anagram should on the average first appear when one has 123 words divided by the square root of 13/2, or 48.

The longer the word, the larger the value of N; one must observe more words to reach the first anagram pair. Using Kucera & Francis, the values of N associated with three-letter through seven-letter words are, respectively, 1200 (13 anagrams in 123 words), 4500 (24 anagrams in 329 words), 6700 (16 anagrams in 329 words), 9200 (16 anagrams in 383 words), and 18500 (14 anagrams in 510 words). For words of eight or more letters, tedious hand-processing of Kucera & Francis was replaced by tabulations from two small anagram dictionaries, R J Edwards *Eht Cdoorrssw Aaagmnr Acdiinorty* (1978) and B Wetterau *The Word Game Winning Dictionary* (1980). For eight-letter through twelve-letter words, Edwards yielded N values of 120000, 190000, 330000, 350000 and 1000000, and Wetterau yielded N values of 110000, 180000, 260000, 430000 and 730000. (The twelve-letter N values are quite uncertain, as they are based on only six and two anagrams, respectively; Webster's Second yielded 1062 twelve-letter anagrams and an N of 900000.)

Note that the Kucera & Francis N values do not track particularly well with the Edwards and Wetterau ones; the latter appear to be some four times as large as the former. Perhaps common words are much more likely to produce anagrams than rarer words are.

The $A = W^2/N$ law, it should be noted, applies only to sets of words for which all the anagrams come in pairs. This was (almost always) the case in the Kucera & Francis, the Edwards and the Wetterau data analyzed above. As more and more words are added, however, anagrams of three or more words appear, invalidating the square law. For example, there are 5159 four-letter words in Webster's Second, of which 3046 are anagrams, leading to an N of 8700. Similarly, there are 6879 anagrams among the 36402 eight-letter words in Webster's Second, leading to an N of 190000. This bias seems to introduce a factor of two in the value of N.

In *Opperslans!* Battus calculates an N of three million based on the unabridged Van Dale dictionary: 261000 words of which 22700 are anagrams. It is interesting to compare this with the above data. Let f(i) be the fraction of words of length i in Webster's Second; then

$$A = \Sigma A(i) = \Sigma W(i)^2/N(i) = [\Sigma f(i)^2/N(i)]W^2 = W^2/N$$

so that N is the inverse of $\Sigma f(i)^2/N(i)$. The table below details the calculation of N. For words of eight or more letters, the N(i) associated with Edwards and Wetterau have been reduced by a factor of 4 to make them more consistent with the Kucera & Francis data..

	f(i)	N(i)	$f(i)^2/N(i)$
3	.0052	1200	.00000002
4	.0200	4500	.00000009
5	.0400	6700	.00000024
6	.0705	9200	.00000054
7	.0971	18500	.00000051
8	.1238	29000	.00000053
9	.1357	46000	.00000040
10	.1311	74000	.00000023
11	.1125	98000	.00000013
12	.0894	225000	.00000004

Summing and inverting the last column, one obtains N equal to 370000, considerably less than the Battus value of 3 million. However, if one uses the Edwards and Wetterau values for N(i), correcting the Kucera & Francis values, N is increased to 1.5 million, still off by a factor of two. One should recall that the Battus statistics include many anagrams of three or more words, which has already been seen to increase the value of N by approximately a factor of two. So perhaps there is no disagreement between the Dutch and the English results!

The Commonest Anagrams (Kucera & Francis)

3 letters: who-how, now-own-won, was-saw, she-he's, god-dog, its-sit

4 letters: from-form, left-felt, name-mean, deal-lead, care-race, army-Mary, stop-post, life-file, note-tone, live-evil, more-Rome, send-ends

5 letters: night-thing, heart-earth, means-names, being-begin, quite-quiet, times-items, shape-phase, horse-shore

6 letters: center-recent, except-expect, course-source, direct-credit, danger-garden, master-stream, listen-silent, slight-lights

7 letters: courses-sources, leading-dealing, silence-license, medical-claimed, testing-setting, largely-gallery, leaders-dealers