WHAT ARE THE ODDS X IS AN ANAGRAM OF Y?

MICHAEL KEITH
Salem, Oregon

In a recent Word Ways article, "Celebrity One-Word Anagrams", the problem of finding famous (or even not-so-famous) people whose first and last names can be anagrammed into a single word was discussed. This raises questions of the form: how many people with N-letter combined names might we expect to encounter, in order to find one whose name is a single-word anagram? In this article I present some answers to this and other similar questions.

Let's first consider a more general problem. Suppose we have two N-letter strings, say S and T, constructed at random but with the important proviso that their letter distributions are, on average, the same as the text distribution for English. What is the probability that S and T are transposals (anagrams) of each other?

One way to proceed is to explicitly calculate the answer from the first principles of probability theory, but as this calculation seemed rather messy I decided instead to use a Monte Carlo method (i.e. the results of many random computer trials) to model the problem. For each value of N from 1 to 11 a computer program generated 100 million pairs of random strings with proper English-language letter distribution and checked each pair for inter-transposability, tabulating the results. (I ran the program several times in order to make sure the margin of error in the results was reasonable.) In the second column of the table below the value of 1/Prob(N) is given, which can be interpreted by saying that the odds of the strings S and T being anagrams of each other is 1/Prob(N). For example, the odds that two 3-digit strings are anagrams of each other is about one in 620.

1	14	14
2	110	139
3	620	751
4	2800	3072
5	10000	10653
6	34000	32948
7	99000	93571
8	290000	248510
9	640000	624989
10	1600000	1501864
11	3400000	3471681

To generalize this to all values of N, I did some numerical analysis and determined that the 1/Prob(N) values can be approximated fairly well by the function

1/Prob(N) = 10^[(2.55N - 1.3)^(4/7)]

where x^y means x raised to the power y. The values of this function are given in the third column on the preceding page; these compare favorably with the values in the second column.

We can now answer the question originally posed: what is the probability that a random N-letter string with English-letter frequencies is an anagram of a single English word? Suppose our dictionary has d(N) N-letter words. Then we have d(N) pairs of strings to examine, our target string paired with each of the d(N) dictionary words. Each of these pairs has Prob(N) of being an anagram, so the probability that some one of the pairs is an anagram is d(N)Prob(N). For example, suppose N=9 and our dictionary has 38,000 9-letter words. The above formula says 1/Prob(9) = 624989, or Prob(N) = 0.0000016. The total probability is thus 0.0000016x38000 = 0.061, which means that a random 9-letter string has about a 1-in-16 chance of being the anagram of an English word.

The fourth column of the table below gives the values of 1/d(N)Prob(N), which tells us the expected number of N-letter strings one needs to examine to find a one-word anagram, for N=7 to 19. The values of d(N) in the third column are based on a computer-readable word list containing Webster's Third Unabridged words plus inflected and derived forms. (If your dictionary has different values for d(N), replace the d(N) values below and recalculate the fourth column.)

7	93571	29000	3.2
8	248510	36000	6.9
9	624989	38000	16
10	1501864	36000	42
11	3471681	30000	116
12	7759876	23000	337
13	16840749	16000	1053
14	35604340	10000	3560
15	73530278	5900	12463
16	148676251	3300	45053
17	294894063	1800	163830
18	574717632	800	718397
19	1102107290	400	2755268

As one would expect, the values in the last column increase with larger N, but note that the values around N=7 are very small because 1/Prob(N) and d(N) are not that far apart. If one is handed random 7-letter strings with the proper letter distribution, roughly one in three should be the anagram of a single word! To test this I examined the 19796 distinct 7-letter surnames in the 1990 US Census database, and found 5676 (about 1 in 3.5) one-word anagrams.

This table predicts that about one in a thousand people with 13-letter names have a name that can be anagrammed into a single word. This agrees with the results of my "Celebrity One-Word Anagram" search. I employed a database having about three

thousand 13-letter celebrities, and found exactly three one-word anagrams (from Britney Spears, Eddie Charlton, Ingrid Steeger). The fact that the numbers in the last column are not astronomical, even for large N, suggested that I might have luck searching through a larger database of non-celebrity names, which yielded the 16-letter Carolyn McAlister MACROCRYSTALLINE and a couple of 17-letter near-misses. The table suggests that people whose name is an anagram of a dictionary-sanctioned 17-letter word (and maybe an even longer one) probably exist.

Note that the application of this analysis to people's names assumes that the letter frequency distribution for names is the same as for English text. This is probably not quite true, but based on my experience the error introduced by this assumption is not very significant. This could be remedied by redoing the Monte Carlo simulation, giving S the letter distribution for names and T the letter distribution for English text and deriving a new formula.

Our rough-and-ready formula for Prob(N) can be used to answer other questions related to accidental anagramming. For example, if I choose a random English novel from my shelf, what are the odds that the first L letters in the novel are an exact anagram of a line in one of Shakespeare's sonnets? Taking N=35 as the average length of a sonnet line, and recalling that there are 154 sonnets of 14 lines each, the odds are 1 in 1/(Prob(35)x154x14), or 1 in 3,800,000,000--not very good!

Thanks to Ross Eckler for posing the question that motivated this article.