Gleason Weights for Monicity

by

Lt. Alfred H. Clifford
(As told to Clifford by Lt. Andrew M. Gleason)

11 October 1944

Abstract

As applied to a frequency count, the word "monic" is short for "mono-alphabetic substitution on plain text". The standard I.C. test is a good measure of roughness (non-randomness), but these weights also take plain text distribution into account. This article gives their derivation. The British have long used these weights to score dotteries (Volume 3, Article I).

Source: ENIGMA Series Volume 9, Article 4
RIP 601, Box 169, 370 27/22/07
NARA, RG 38, Crane Collection

Editor: Frode Weierud, © May 2009

TOP SECRET ULTRA

FOREWORD TO ENIGMA SERIES

CRYPTANALYTICAL RESEARCH PAPERS

This series consists of original memoranda written by members of the cryptanalytical research section of the U.S. Naval Communications Intelligence Staff, and by others working with the research group. A brief description of the contents of each paper is given in the Index to each volume. While an effort toward completeness has been made, the reader is referred for greater detail to the various R.I.P's put out by the Atlantic Operations Department, especially R.I.P. 450. There he will also find polished techniques, which appear in this Series of their original form.

The name of the author and the date of the paper are also given in the Index, which lends an historical flavor to the Series. The Editor feels that there is considerable merit in an anthology for this sort, full of original ideas both good and bad, which supplements the finished publication. It should be further emphasized that R.I.P. 450 is concerned mainly with the techniques themselves, while this Series considers the cryptanalytical or mathematical theories which underlie the techniques. On the other hand, machine research (from an engineering point of view) is not covered in this Series.

Some of the papers in this Series are expository, but most represent original work. It must always be borne in mind that we owe to the British the basic solution of the Enigma, and many of the basic subsidiary techniques, together with the underlying mechanical and mathematical theories. Much of what we call "original" is only a retracing of steps previously taken by the British, and the Editor has striven to point this out in the Index. But there is also a great deal that extends or improves British methods, and some that strikes out in new directions.

It must be pointed out that the author of a paper may be entitled to credit only for his literary toil. Our group of eight or ten men worked as a team, and an assignment of "credit" would be as difficult as it is undesirable. In this line of endeavor, a chance remark may be worth a week's work.

GLEASON WEIGHTS FOR MONICITY

In order to tell whether or not a given frequency count is "monic", i.e. a simple substitution of plain text, the customary procedure is to find its I.C. This is a good measure of roughness, but does not take into account the type of roughness peculiar to the frequency of the language under consideration. An exact measure, using Bayes' Theorem, of the probability that a given frequency count is monic, involves a prohibitive calcualtion of symmetric functions of the P/L frequencies. The weights described below are an approximation to these. They have been in use for some years in England, but were discovered here independently by Lt. Gleason.

Suppose we make a large number of frequency counts, each on a sample of s objects which are of c different kinds. We can visualize making the frequency count by throwing the s objects, one by one, into one of c mutually exclusive cells.

If the probability of an object falling into the ith cell is f_i, then the probability that this cell will contain exactly k objects, when the count is complete, is

$$\binom{S}{K} f_i^{K} \left(1-f_i\right)^{S-K}$$

The expected number m_k of cells containing k objects is the sum of this over all of the c cells:

$$m_k = \sum_{i=1}^{c} \binom{S}{K} f_i^{K} \left(1-f_i\right)^{S-K} \qquad (1)$$

In the random case, $f_i = 1/c$, and the expected number n_k of cells containing k objects is

$$n_K = c\binom{s}{k}\left(\frac{1}{c}\right)^{K}\left(1-\frac{1}{c}\right)^{S-K} \qquad (2)$$

We refer to (1) as the "significant" and (2) as the "random" expectancy.

Now suppose we observe N_k cells with k objects for a particular sample of s objects of c kinds. The Bayes' factor in favor of this case being significant, by virtue of its having the value of N_k observed, is the ratio r_k of the probability that a significant sample will have N_k cells with k objects to that for a random sample. If we assume that the number of cells with k objects is distributed according to Poisson's law, then

$$r_K = \frac{\epsilon^{-m_K}\dfrac{(m_K)^{N_K}}{N_K!}}{\epsilon^{-n_K}\dfrac{(n_K)^{N_K}}{N_K!}} = \epsilon^{n_K - m_K}\left(\frac{m_K}{n_K}\right)^{N_K}.$$

We now make the second, and rather more daring approximation, and assume that the c probabilities of getting N_k cells with k objects, for k = 0, 1, 2, ...c, are independent. The Bayes' factor r in favor of the observed frequency count being significant, by virtue of its values of N_0, N_1, N_2, ..., N_c is then the product of the r_k over k = 0, 1, ..., c:

$$r = \prod_{K=0}^{c} e^{n_K - m_K}\left(\frac{m_K}{n_K}\right)^{N_K} = \prod_K \left(\frac{m_K}{n_K}\right)^{N_K}.$$

(The exponentials disappear, since $\sum n_K = \sum m_K = c$).

Taking the log of this, we get an additive weight:

$$w = \log r = \sum_K N_K \log \frac{m_K}{n_K} = \sum_K N_K w_K.$$

The numbers

$$w_K = \log \frac{m_K}{n_K}$$

are the Gleason weights. If each cell with k objects be given the weight w_k, and these are added up over all the cells, we get w. To evaluate them we use (1) and (2):

$$w_K = \log \frac{1}{c}\sum_{i=1}^{c} (cf_i)^K \left(\frac{c(1-f_i)}{c-1}\right)^{s-K}$$

$$= \log \frac{1}{c}\sum_{i=1}^{c} \left(\frac{(c-1)f_i}{1-f_i}\right)^K \left(\frac{c(1-f_i)}{c-1}\right)^{s}.$$

If we had a series expansion for w_k:

$$w_K = a_0 + a_1 K + a_2 K^2 + \ldots$$

and if we let k_i be the number of objects in the ith cell, then for any particular frequency count $\{K_i\}$

$$w = \sum_i w_{K_i} = a_0 c + a_1 s + a_2 \sum K_i^2 + \ldots\ldots$$

Since adding a constant to w, independent of the particular sample, i.e. the k_i, leaves just as good a weight, this shows that we can alter a_0 and a_1 in any way we like, i.e. make any linear transformation on w_k.