The Friedman Test

William Friedman (1891 – 1969) developed statistical methods for determining whether a cipher is monoalphabetic or polyalphabetic.

Index of coincidence

The probability of choosing two letters the same from ciphertext (i.e., two as or two bs or two cs or … or two zs) would be

$$I = \frac{n_a}{n} \times \frac{n_a - 1}{n-1} + \frac{n_b}{n} \times \frac{n_b - 1}{n-1} + \frac{n_c}{n} \times \frac{n_c - 1}{n-1} + \ldots + \frac{n_z}{n} \times \frac{n_z - 1}{n-1}$$

This number is denoted *I* and called the index of coincidence of the ciphertext.

Because Friedman denoted this number by the Greek letter kappa κ, it is sometimes called the Kappa Test.

English plaintext

The frequencies of the letters in English are:

Letter	a	b	c	d	e	f	g	h	i	j	k	l	m
Frequency	.082	.015	.028	.043	.127	.022	.020	.061	.070	.002	.008	.040	.024
Letter	n	o	p	q	r	s	t	u	v	w	x	y	z
Frequency	.067	.075	.019	.001	.060	.063	.091	.028	.010	.023	.001	.020	.001

Beker and Piper, *Cipher Systems: The Protection of Communications*, Wiley.

So, if a text were enciphered using a single alphabet, the probability of "drawing" two letters that are the same is:

aa	or	bb	or	cc	or	…	or	zz
$.082 \times .082$	+	$.015 \times .015$	+	$.028 \times .028$	+	…	+	$.001 \times .001$

This probability of "drawing" two letters that are the same – the index of coincidence -- is approximately $I \approx 0.0656010$.

Polyalphabetic ciphers

If more than one alphabet were used, the frequencies of the letters should be more nearly uniform. If they were uniform, the probability of "drawing" two letters that were the same would be:

$$I \approx \underbrace{\left(\frac{1}{26}\times\frac{1}{26}\right)+\left(\frac{1}{26}\times\frac{1}{26}\right)+\left(\frac{1}{26}\times\frac{1}{26}\right)+\ldots+\left(\frac{1}{26}\times\frac{1}{26}\right)}_{26 \text{ terms}}=\frac{1}{26}\approx 0.038.$$

Here is the idea of the test.

If the ciphertext were generated by a monoalphabetic cipher, we should determine *I* to be near 0.065 because a monoalphabetic cipher is just a permutation of the letters of a single alphabet. The frequencies of letters for the ciphertext alphabet should be nearly the same as for English – but in a different order.

If the cipher were generated by a polyalphabetic cipher, the frequencies of the letters would become more nearly uniform – more nearly the same for each letter. We should determine *I* to be near the $I = 0.038$.

We test the ciphertext by calculating *I* based on the ciphertext frequencies. The closer that *I* is to 0.065, the more likely it is that we have a monoalphabetic cipher. The closer that *I* is to 0.038, the more likely that we have a polyalphabetic cipher.

Recall that, using frequency analysis, peaks and valleys of frequencies suggest a monoalphabetic cipher and relatively uniform frequencies suggest a polyalphabetic cipher. The Friedman test is a statistical way of "looking for peaks and valleys versus uniform frequencies."