Breaking a Cæsar Cipher

Introduction

This presents two different methods of breaking a Cæsar cipher using statistics. The ciphertext is VRVRVHDPVWUHVV; the frequency counts are:

D	1	0.0714	R	2	0.1429	W	1	0.0714
H	2	0.1429	U	1	0.0714			
P	1	0.0714	V	6	0.4286			

As a preliminary, we note that the unicity distance of a Cæsar cipher is roughly 1.5, so 14 characters should be more than enough to find a unique key and decryption.

Method #1: frequency counts

The most frequent letters in English are ETAONRISH. The following table summarizes the overlap, assuming the key is the number indicated:

key	ct	letters	key	ct	letters	key	ct	letters	key	ct	letters
0	2	RH	7	4	VHPU	13	3	RVU	20	2	HU
1	2	PU	8	3	VPW	14	3	VHW	21	1	VD
2	3	VPW	9	2	RW	15	4	HDPW	22	3	DPW
3	**6**	**VRHDWU**	10	2	RD	16	3	HDU	23	1	P
4	3	VRW	11	2	DP	17	2	VR	24	2	RP
5	1	W	12	2	DU	18	2	WU	25	3	RHD
6	1	Y				19	2	WU			

The key value 3 leaps out. Notice that had we counted differences, the same thing would happen, as the differences are the same as the counts.

Method #2: correlations

For this, we need frequency counts of the letters in plaintext, and for this purpose (since we don't have the plaintext handy) we simply assume the plaintext follows the 1-gram, independent model. Note that this is usually a poor assumption since in English the letters are not independent, but we'll try it and see what happens.

The relevant equation is:

$$\Sigma_{0\le t<26} F_t(c)p(t-k) = 0.0714[p(3-k) + p(15-k) + p(20-k) + p(22-k)] + 0.1429[p(7-k) + p(17-k)] + 0.4286p(21-k)$$

and the correlations over the values of k come out to be:

key	r	key	r	key	r	key	r
0	0.0281	7	0.0590	13	0.0550	20	0.0252
1	0.0295	8	0.0425	14	0.0513	21	0.0500
2	0.0639	9	0.0331	15	0.0388	22	0.0253
3	**0.0762**	10	0.339	16	0.0400	23	0.0186
4	0.0514	11	0.0229	**17**	**0.0763**	24	0.0252
5	0.0160	12	0.0183	18	0.0330	25	0.0341
6	0.0214			19	0.0312		

The values 3 and 17 are just about equal, so both must be tried.

Answer

The key is 3; the message is `SOSOSEAMSTRESS`.