13 SINKOV's Test for the Period

We want to test a pretended period l whether it is the real period. We write the text in rows of width l and consider the columns.

- If l is the correct period, each column is monoalphabetically encrypted and has its coincidence index near the coincidence index of the plaintext language.
- Otherwise the columns are random garbage and have coincidence indices near the random value $\frac{1}{n}$. Or rather near the value for a polyalphabetic ciphertext of period (the true) l.

Maybe the columns are quite short, thus their coincidence indices are diffuse and give no clear impression. However we can put all the indices together without bothering about the different monoalphabets, and get a much more precise value, based on all the letters of the text.

Definition For a text $a \in \Sigma^*$ and $l \in \mathbb{N}_1$ the mean value

$$\bar{\varphi}_l(a) := \frac{1}{l} \cdot \sum_{i=0}^{l-1} \varphi(a_i a_{i+l} a_{i+2l} \ldots)$$

is called the SINKOV **statistic** of a of order l.

Note that $\bar{\varphi}_1 = \varphi$.

A Perl program, `phibar.pl`, is in `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Perl/`.

Example

Let us again examine the ciphertext from Section 9. We get the values:

$\bar{\varphi}_1(a)$	0.0442	$\bar{\varphi}_7(a)$	**0.0829**	$\bar{\varphi}_{13}(a)$	0.0444
$\bar{\varphi}_2(a)$	0.0439	$\bar{\varphi}_8(a)$	0.0443	$\bar{\varphi}_{14}(a)$	**0.0839**
$\bar{\varphi}_3(a)$	0.0440	$\bar{\varphi}_9(a)$	0.0427	$\bar{\varphi}_{15}(a)$	0.0432
$\bar{\varphi}_4(a)$	0.0438	$\bar{\varphi}_{10}(a)$	0.0421	$\bar{\varphi}_{16}(a)$	0.0439
$\bar{\varphi}_5(a)$	0.0430	$\bar{\varphi}_{11}(a)$	0.0426	$\bar{\varphi}_{17}(a)$	0.0444
$\bar{\varphi}_6(a)$	0.0435	$\bar{\varphi}_{12}(a)$	0.0432	$\bar{\varphi}_{18}(a)$	0.0419

The period 7 is overwhelmingly evident. The values other than at the multiples of 7 are in almost perfect compliance with a (German) ciphertext of period around 7.

A Short Ciphertext

Our example ciphertext was quite long, and it is no surprise that the statistical methods perform very well. To get a more realistic picture let us examine the following ciphertext of length 148:

```
MDJJL DSKQB GYMZC YKBYT ZVRYU PJTZN WPZXS KCHFG EFYFS ENVFW
KORMX ZQGYT KEDIQ WRVPM OYMQV DQWDN UBQQM XEQCA CXYLP VUOSG
EJYDS PYYNA XOREC YJAFA MFCOF DQKTA CBAHW FYJUI LXBYA DTT
```

The Kasiski test finds no reptitions of length 3 or more. It finds 16 repetitions of length 2 and no eye-catching pattern. The common factors 10 or 20 could be a hint at the correct period, but repetitions of length 2 are not overly convincing.

Repetition:	DS	SK	GY	YM	CY	BY	YT	TZ
Distance:	98	28	47	60	100	125	40	8
Repetition:	GE	FY	OR	MX	QW	DQ	AC	YJ
Distance:	60	94	60	31	12	50	40	21

The coincidence index of the text is 0.0386 and doesn't distinguish the ciphertext from random text. The first 40 values of the autocoincidence spectrum are

κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7	κ_8
0.0270	0.0203	0.0541	0.0405	0.0405	0.0338	0.0405	**0.0676**
κ_9	κ_{10}	κ_{11}	κ_{12}	κ_{13}	κ_{14}	κ_{15}	κ_{16}
0.0270	0.0473	0.0270	**0.0676**	0.0405	0.0473	0.0541	0.0541
κ_{17}	κ_{18}	κ_{19}	κ_{20}	κ_{21}	κ_{22}	κ_{23}	κ_{24}
0.0203	0.0203	**0.0608**	0.0473	0.0473	0.0135	0.0541	0.0270
κ_{25}	κ_{26}	κ_{27}	κ_{28}	κ_{29}	κ_{30}	κ_{31}	κ_{32}
0.0338	0.0405	0.0541	**0.0811**	0.0338	0.0338	0.0405	0.0203
κ_{33}	κ_{34}	κ_{35}	κ_{36}	κ_{37}	κ_{38}	κ_{39}	κ_{40}
0.0068	0.0473	0.0473	0.0270	0.0405	0.0066	0.0203	0.0473

Values above 0.06 occur for shifts of 8, 12, 19, 28, the latter being the largest one. This makes a diffuse picture, giving slight evidence for a period of 28. Finally let's try Sinkov's test. It gives as its first 40 values:

$\bar{\varphi}_1$ 0.0386	$\bar{\varphi}_2$ 0.0413	$\bar{\varphi}_3$ 0.0386	$\bar{\varphi}_4$ 0.0492	$\bar{\varphi}_5$ 0.0421	$\bar{\varphi}_6$ 0.0441	$\bar{\varphi}_7$ 0.0433	$\bar{\varphi}_8$ 0.0471
$\bar{\varphi}_9$ 0.0330	$\bar{\varphi}_{10}$ 0.0505	$\bar{\varphi}_{11}$ 0.0265	$\bar{\varphi}_{12}$ **0.0591**	$\bar{\varphi}_{13}$ 0.0333	$\bar{\varphi}_{14}$ 0.0486	$\bar{\varphi}_{15}$ 0.0444	$\bar{\varphi}_{16}$ 0.0410
$\bar{\varphi}_{17}$ 0.0280	$\bar{\varphi}_{18}$ 0.0395	$\bar{\varphi}_{19}$ 0.0439	$\bar{\varphi}_{20}$ **0.0589**	$\bar{\varphi}_{21}$ 0.0357	$\bar{\varphi}_{22}$ 0.0264	$\bar{\varphi}_{23}$ 0.0476	$\bar{\varphi}_{24}$ 0.0548
$\bar{\varphi}_{25}$ 0.0507	$\bar{\varphi}_{26}$ 0.0359	$\bar{\varphi}_{27}$ 0.0444	$\bar{\varphi}_{28}$ 0.0488	$\bar{\varphi}_{29}$ 0.0368	$\bar{\varphi}_{30}$ **0.0622**	$\bar{\varphi}_{31}$ 0.0312	$\bar{\varphi}_{32}$ 0.0323
$\bar{\varphi}_{33}$ 0.0091	$\bar{\varphi}_{34}$ 0.0294	$\bar{\varphi}_{35}$ 0.0429	$\bar{\varphi}_{36}$ **0.0611**	$\bar{\varphi}_{37}$ 0.0541	$\bar{\varphi}_{38}$ 0.0307	$\bar{\varphi}_{39}$ 0.0256	$\bar{\varphi}_{40}$ 0.0542

The values for 12, 20, 30, and 36 stand somewhat out, followed by the values for 24, 37, and 40, then 10 and 25—again there is no clear favorite. Let's discuss the candidate values for the period and rate each criterion as "good", "weak", or "prohibitive".

Period?	Pros and cons
8	$\varphi(c)$ should be slightly larger (weak). Only 3 repetition distances are multiples of 8 (weak). κ_8 and κ_{16} are good, κ_{40} is weak, κ_{24} and κ_{32} are prohibitive. $\bar{\varphi}_8$ is weak, $\bar{\varphi}_{16}$ and $\bar{\varphi}_{32}$ are prohibitive, $\bar{\varphi}_{24}$ and $\bar{\varphi}_{40}$ are good.
10	$\varphi(c)$ should be slightly larger (weak). 7 repetition distances are multiples of 10 (good). κ_{10}, κ_{20}, and κ_{40} are weak, κ_{30} is prohibitive. $\bar{\varphi}_{10}$, $\bar{\varphi}_{20}$, $\bar{\varphi}_{30}$, and $\bar{\varphi}_{40}$ are good.
12	$\varphi(c)$ should be slightly larger (weak). 4 repetition distances are multiples of 12 (good). κ_{12} is good, κ_{24} and κ_{36} are prohibitive. $\bar{\varphi}_{12}$, $\bar{\varphi}_{24}$, and $\bar{\varphi}_{36}$ are good.
19	0 repetition distances are multiples of 19 (prohibitive). κ_{19} is good, κ_{38} is prohibitive. $\bar{\varphi}_{19}$ and $\bar{\varphi}_{38}$ are prohibitive.
20	6 repetition distances are multiples of 20 (good). κ_{20} and κ_{40} are weak. $\bar{\varphi}_{20}$ and $\bar{\varphi}_{40}$ are good.
24	0 repetition distances are multiples of 24 (prohibitive). κ_{24} is prohibitive. $\bar{\varphi}_{24}$ is good.
28	Only 1 repetition distance is a multiple of 28 (weak). κ_{28} is good. $\bar{\varphi}_{28}$ is weak.
30	3 repetition distances are multiples of 30 (good). κ_{30} is prohibitive. $\bar{\varphi}_{30}$ is good.
36	0 repetition distances are multiples of 36 (prohibitive). κ_{36} is prohibitive. $\bar{\varphi}_{36}$ is good.
37	0 repetition distances are multiples of 37 (prohibitive). κ_{37} is prohibitive. $\bar{\varphi}_{37}$ is good.

To assess these findings let us score the values "good" as +1, "weak" as 0, and "prohibitive" as −1. Note that 3 repetitions for period 8 are weaker than 3 repetitions for period 30. The candidates 19, 24, 36, and 37 have negative weights, the candidates 8 and 28, zero weights. We skip them in the first round. Positive weights have 10 (3 of 9), 12 (3 of 8), 20 (3 of 5), and 30 (1 of 3). We rank them by their relative weights: 20 with score $0.6 = 3/5$, then 12 with score 0.375, then 10 and 30 with scores 0.333.

The most promising approach to further cryptanalysis starts from the hypothetical period 20, see Section 15.