STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

ZW 1953 - 014

Voordracht in de serie Actualiteiten

M. de Vries

31 oktober 1953

Statistische methoden in de crypto-analyse

1953

Voordracht door M.de Vries in de serie Actualiteiten
op 31 October 1953.

Statistische methoden in de crypto-analyse.

1. Inleiding.

De cryptologie is een toegepaste wetenschap waarin taal en wiskunde elkaar ontmoeten. Het probleem is allereerst van semantische aard. Informatie vervat in geschreven berichten moet verborgen worden. Om dit vraagstuk geschikt te maken voor een mathematische beschrijving ziet men (voorlopig) af van de semantische aspecten. Men stelt zich voor dat een imaginaire taalbron berichten letter voor letter voortbrengt volgens een bepaalde wh-verdeling, zodanig dat het optreden van een letter afhankelijk is van de voorafgaande letters. Een bericht kunnen we dus opvatten als een steekproef uit een stochastisch proces of een tijdreeks.

Als N het aantal letters is dan stellen we een bericht voor als volgt:

(1) $$t_1, t_2, t_3, \ldots, t_N.$$

De waarden die de stochastische variabelen t_i kunnen aannemen behoren tot de eindige verzameling

(2) $$A = (a_1, a_2, \ldots, a_n)$$

De verzameling A heet een normaal alphabet. De elementen a_i zijn de letters van de taal. Iedere permutatie van A

(3) $$A_i = (a_{i_1}, a_{i_2}, \ldots, a_{i_n})$$

wordt een alphabet genoemd.

Het proces wordt beschreven door de volgende waarschijnlijkheden:

(a) $p(i)$ is de waarschijnlijkheid dat een letter a_i voorkomt in een bericht.

(b) $p_i(j)$ is de voorwaardelijke waarschijnlijkheid dat een letter a_j volgt op een letter a_i. De whn $p_i(j)$ heten overgangswaarschijnlijkheden. Hieruit kunnen we afleiden:

(c) $p(ij)=p(i)p_i(j)$ is de wh [1] dat een groep van twee letters (bigrammen) $a_i\ a_j$ voorkomt in een bericht.

---

1) wh = waarschijnlijkheid
whn = waarschijnlijkheden.

Voor de meeste crypto-analytische problemen zijn de whn. gegeven door (a), (b) en (c) voldoende. Soms is het echter wenselijk de verzameling uit te breiden met:

(d) $p_{ij}(k)$, de wh. dat een letter a_k volgt op het bigram $a_i\ a_j$

(e) $p(ijk)$, de wh. van het voorkomen van een trigram $a_i a_j a_k$.

De whn., die het optreden van begin- en eindletters van woorden bepalen, zijn voor sommige problemen van groot belang.

Een zeer belangrijk statistisch kenmerk heeft betrekking op deelverzamelingen van letters uit een bericht. Voor iedere deelverzameling van uitgebreidheid N' die ontstaat door onafhankelijke trekkingen van letters uit een bericht geldt dat de frequentiequotiënten $\frac{n_i'}{N'}$ van de letters a_i naderen tot de whn. $p(i)$ van de taal.

(f) $$\lim \frac{n_i'}{N'} = p(i).$$

2. Crypto-transformaties.

Onder een crypto-transformatie verstaan we een transformatie die een bericht omzet in een reeks van letters die geen semantische correlatie vertoont met het oorspronkelijke bericht.

Men gaat uit van een streng parallelisme tussen de semantische en statistische kenmerken van de taal. Een transformatie van de statistische kenmerken zal een transformatie van de semantische eigenschappen en dus een versluiering van de informatie ten gevolge hebben.

Een crypto-transformatie wordt voorgesteld door T_s. De parameter s van de specifieke transformatie heet sleutel. Doorloopt s de verzameling S van alle mogelijke waarden die s kan aannemen, dan ontstaat een cryptografisch systeem T. Een cryptografisch systeem is dus een verzameling van crypto-transformaties. Na het toepassen van een crypto-transformatie ontstaat een cryptogram.

(4) $$T_s(t_1, t_2, \ldots, t_N) = c_1, c_2, c_3, \ldots, c_M.$$

Het toepassen van een crypto-transformatie noemt men gewoonlijk vercijferen.

3. Statistische methoden.

De statistische methoden die gebruikt worden bij het ontcijferen van geheimschriften zijn over het algemeen van zeer eenvoudige aard. De verdelingen van de statistische grootheden zijn

meestal discreet en de verzameling van waarden die de stochastische variabelen kunnen aannemen is eindig.

De gebruikte methoden munten ook niet altijd uit door overmaat van mathematische nauwkeurigheid.

Het voornaamste statistische hulpmiddel in de crypto-analyse zijn de whn. p(i) (fig. 1) van optreden van de letters a_i. In dit artikel zal geen gebruik worden gemaakt van de overgangswhn. $p_i(j)$ en bigramwhn. p(ij).

	A	B	C	D	E	F	G	H	I	J	K	L	M
n_i	65	17	13	61	200	6	37	30	66	17	23	35	22

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
n_i	108	57	13	0	63	35	58	18	24	18	0	0	14

$$n_i = 1000\, p_i .$$

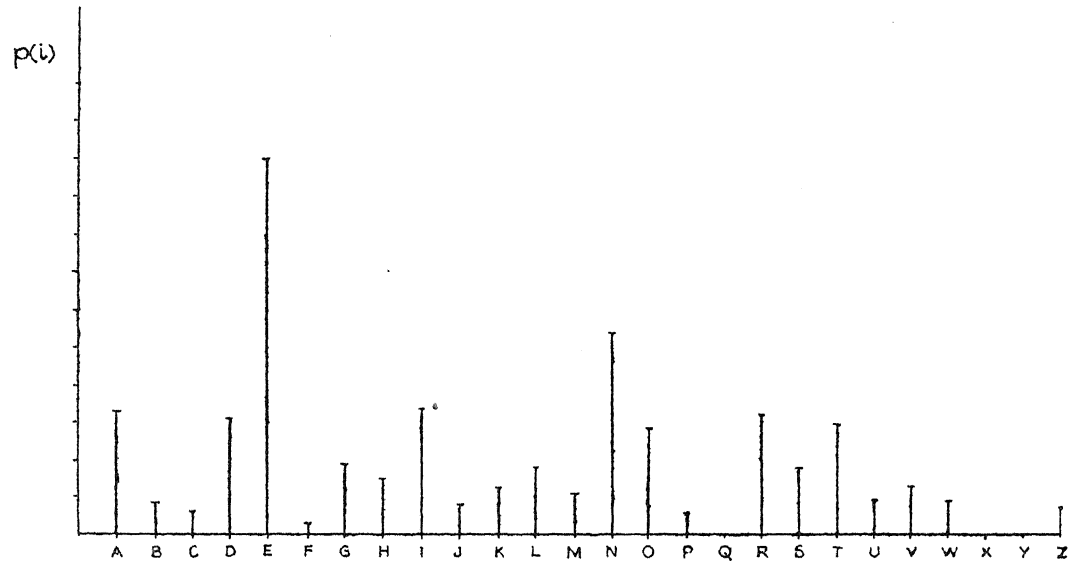

Fig. 1.

3.1. Homogeniteit.

Een cryptogram vertoont gewoonlijk een geheel andere frequentie-verdeling dan een normale tekst. De verdeling der letters nadert meer of minder tot de homogene verdeling. Het eerste werk van de crypto-analyst is een onderzoek van het cryptogram onder de hypothese dat dit ontstaan is door onafhankelijke trekkingen uit een homogeen universum met 26 kenmerken. Het lettermateriaal wordt doorzocht of er afwijkingen voorkomen die niet toevallig zijn ontstaan.

Laten wij aannemen dat we in het bezit zijn geraakt van 100 letters. Onder de nulhypothese is de verwachting voor een willekeurige letter 3,85. Sommige letters komen misschien niet voor in het cryptogram andere 6 of 7 maal.

Met behulp van de Poisson-verdeling [2] bepalen we dan de wh. dat b.v. een bepaalde letter niet zal voorkomen.

Volgens deze verdeling is de kans dat een bepaalde letter niet voorkomt 0,02. We verwerpen dan de hypothese dat deze afwijking bij toeval is ontstaan en daarmee de homogeniteit van de verdeling. De letters die in hun frequentie significant afwijken van de verwachting kunnen aanwijzingen geven over de gebruikte crypto-transformatie.

Dezelfde vragen kunnen beantwoord worden voor polygrammen van willekeurige lengte.

Het voorgaande was een inleiding tot een vraag die in de crypto-analyse zeer vaak beantwoord moet worden.

Gegeven een tekst van N elementen (onder elementen verstaan we enkele letters of polygrammen). Het aantal verschillende elementen is n.

Wat is, weer onder de hypothese dat de tekst is ontstaan door onafhankelijke trekkingen uit een homogeen verdeeld universum, de verwachting H_r van het aantal elementen dat r maal voorkomt?

Voor deze verwachting vinden we

$$H_r = \binom{N}{r} \frac{1}{n^{r-1}} \left(1 - \frac{1}{n}\right)^{N-r}. \tag{5}$$

Met behulp van deze formule zijn er nomogrammen vervaardigd waaruit, voor verschillende N, r en n, het antwoord is af te lezen. Wijkt het waargenomen aantal af van de verwachting dan wordt de Poisson-verdeling geraadpleegd om te zien of de afwijking significant is. Vooral het geval dat er voor een bepaalde r te veel herhalingen zijn is van groot belang. Hieruit kan soms de conclusie worden getrokken dat de crypto-transformatie is toegepast op groepen van r letters.

3.2. Coïncidentietoets A.

We gaan nu over tot het bespreken van een statistische grootheid van een ander karakter. Het doel is nog steeds te toetsen of het waarnemingsmateriaal afkomstig is uit een homogene verdeling.

We vormen uit een tekst alle $\binom{N}{2}$ combinaties van 2 letters. Bestaat een zo ontstaan paar uit twee gelijke letters, dan spreken wij van een coïncidentie.

De kans dat uit een homogene verdeling van 26 letters twee gelijke, maar overigens willekeurige letters worden getrokken,

---

2) E.C.Molina: Poisson's Experimential Binomial Limit, Table I.

is de som van de kansen voor de individuele letters:

(6) $$\kappa_h = \sum_1^{26} \left(\frac{1}{26}\right)^2 = 0{,}0385.$$

Dit is dus de kans op een coïncidentie. De grootheid κ_h heet de coïncidentie-constante voor homogeen verdeelde tekst.

De verwachting van het aantal coïncidenties is dus

(7) $$\kappa_h \frac{N(N-1)}{2}.$$

Het waargenomen aantal is:

(8) $$\sum_1^{26} \frac{n_i(n_i-1)}{2}$$

als n_i de frequentie van de letter a_i voorstelt.

Met behulp van de Poisson-verdeling worden de whn. bepaald van het voorkomen van 0,1,2,... coïncidenties.

Voor normale tekst is de kans op een coïncidentie

(9) $$\kappa_p = \sum_1^{26} p_i^2 = 0{,}0827$$

De grootheid κ_p heet de coïncidentie-constante voor normale tekst.

Er is geen reden de coïncidentietoets te beperken tot enkele letters. Als r de lengte is van het polygram, dan geldt voor homogeen verdeelde tekst

(10) $$\kappa_{hr} = \binom{N-r+1}{2} \sum_1^{26^r} \left(\frac{1}{26^r}\right)^2.$$

3.3. Coïncidentietoets B.

Een van de belangrijkste statistische technieken in de crypto-analyse staat bekend als coïncidentietoets B. Deze toets heeft als theoretische achtergrond de volgende overwegingen:

a) Als twee voldoende lange reeksen van letters worden gesuperponeerd dan zullen in een aantal gevallen gelijke letters boven elkaar staan. Deze paren gelijke letters noemen we weer coïncidenties.
b) Bestaan beide reeksen van letters uit homogeen verdeelde tekst, dan is de kans op een coïncidentie: $\kappa_h = 0{,}0385$.
c) Bestaan beide reeksen van letters uit normale tekst, dan is de kans op een coïncidentie: $\kappa_p = 0{,}0827$.

We stellen ons nu een crypto-transformatie voor T_{si} (s = constant) (i = 1,...,N), die afhangt van een tijdparameter i.

De transformatie bestaat dan uit een reeks van transformaties

(11) $$T_{S1}, T_{S2}, \ldots, T_{SN}$$

Als we de tekst voorstellen door:

(12) $$t_1, t_2, \ldots, t_N$$

dan is het cryptogram te schrijven als:

(13) $$T_{S1}\, t_1,\ T_{S2}\, t_2, \ldots,\ T_{SN}\, t_N .$$

Een andere normale tekst van dezelfde lengte:

(14) $$u_1, u_2, \ldots, u_N,$$

wordt met dezelfde transformatie vercijferd:

(15) $$T_{S1}\, u_1,\ T_{S2}\, u_2, \ldots,\ T_{SN}\, u_N .$$

Geldt nu: $t_k = u_k$

(16)

dan volgt: $T_{Sk}\, t_k = T_{Sk}\, u_k .$

Superponeren we de twee cryptogrammen zo dat de transformaties in de bovenste rij corresponderen met de transformaties in de onderste rij, dan is het resultaat gelijk aan een superpositie van twee normale teksten en de verwachting van het aantal coïncidenties is dan $\kappa_p\, N$.

Een onjuiste superpositie van twee cryptogrammen vercijferd met dezelfde sleutel zal ook aanleiding geven tot het optreden van coïncidenties, maar deze zijn dan geheel van toevallige aard. Het aantal zal dan niet ver liggen van $\kappa_h . N$.

In de crypto-analytische practijk komt het superponeren herhaaldelijk voor. Allereerst als methode om te weten te komen of meerdere cryptogrammen met dezelfde sleutel zijn vercijferd. Een ander voorbeeld dat terug te brengen is tot het vorige is het toepassen van een sleutel met een lange periode. Dit betekent dat het aantal transformaties T_{Si} korter is dan de tekst. Heeft men enige aanwijzingen omtrent de lengte van de sleutel en voldoene materiaal, dan verdeelt men de tekst in stukken en probeert de juiste superpositie te vinden. Met andere, niet altijd statistische, methoden worden de stukken dan simultaan ontcijferd.

3.4. Toets voor monoalphabeticiteit.

Het is zeer vaak van belang te weten of een letterverdeling afkomstig is uit een alphabet. Onder een alphabet verstaan we een willekeurige permutatie A_i van A. Bij grote aantallen let-

ters is het vergelijken van de frequentiequotiënten met de whn. p(i) meestal voldoende (fig. 1).

Bij kleine aantallen waarnemingen is het timmermansoog niet meer voldoende om tot een conclusie te komen. Men heeft daarom een methode bedacht die ook voor kleine aantallen aanwijst of een steekproef van letters afkomstig is uit een alphabet. Deze toets heet de φ-toets.

In paragraaf 3.2 is afgeleid dat de kans op een coïncidentie voor normale tekst gelijk is aan $\kappa_p = 0{,}0827$.

De verwachting van het aantal coïncidenties in een tekst van N letters is:

(17) $$\kappa_p \frac{N(N-1)}{2}$$

Als n_i de frequentie voorstelt van de letter a_i dan is het totaal aantal waargenomen coïncidenties gelijk aan

(18) $$\sum \frac{n_i(n_i-1)}{2}$$

Voor normale tekst moeten (17) en (18) aan elkaar gelijk zijn.

(19) $$\varphi = \sum n_i(n_i-1) = \kappa_p N(N-1).$$

Als $\frac{\varphi}{N(N-1)}$ niet te veel afwijkt van κ_p, dan mogen we aannemen dat de steekproef afkomstig is uit een alphabet.

3.5. K-toets.

We kunnen de φ-toets ook nog op een andere manier gebruiken. Niet slechts als aanwijzer of een steekproef van letters afkomstig is uit een alphabet of niet, maar ook om de mate aan te geven waarin de verdeling afwijkt van een normale tekst. Gewoonlijk gebruikt men een eenvoudige transformatie van φ.

(20) $$K = 26 \frac{\varphi}{N(N-1)}$$

Voor homogeen verdeelde tekst is dan:

(21) $$K = 26 . \kappa_h = 1$$

Voor normale tekst (Nederlands):

(22) $$K = 26 . \kappa_p = 2{,}15$$

K varieert dus tussen 1 en 2,15.

De K-toets wordt gebruikt als verschillende crypto-transformaties na elkaar zijn toegepast.

Bij de ontcijfering probeert men dan deze transformaties achtereenvolgens te elimineren, waarmee dan een stijging van K gepaard moet gaan. Meestal is de toepassing van de toets zeer moeilijk omdat de steekproef-fouten groter dan de stijging van K zijn.

3.6. Twee steekproeventoetsen.

In vele gevallen weten we van twee verdelingen dat het steekproeven zijn uit een normaal alphabet of een permutatie daarvan. Het is dan van belang te weten of beide alphabetten dezelfde permutatie van A voorstellen of niet. Bij grote aantallen letters is het vergelijken van de frequentie-verdelingen voldoende. Voor kleine aantallen waarnemingen heeft men verschillende methoden bedacht die aanwijzen of twee steekproeven afkomstig zijn uit dezelfde verdeling.

a) χ-toets.

In paragraaf 3.4 is afgeleid dat

(23) $$\Sigma(n_i^2 - n_i) = \kappa_p N^2 - \kappa_p N \ ;$$

(24) $$\Sigma n_i = N \ ;$$

(25) $$\Sigma n_i^2 = \kappa_p N^2 - \kappa_p N + N \ .$$

We veronderstellen nu dat twee monoalphabetische verdelingen tot hetzelfde alphabet behoren.

Worden de twee verdelingen gecombineerd, dan moet, onder de hypothese van gelijkheid, ook de nieuwe verdeling tot datzelfde alphabet behoren.

De letters van de eerste verdeling worden aangeduid met a_{i1} en hun frequenties met n_{i1}; voor de tweede verdeling resp. a_{i2} en n_{i2}. Het aantal letters in de eerste verdeling is N_1 en in de tweede verdeling N_2. Dan geldt:

(26) $$\Sigma(n_{i1}+n_{i2})^2 = \kappa_p(N_1+N_2)^2 - \kappa_p(N_1+N_2) + (N_1+N_2)$$

(27) $$\Sigma n_{i1}^2 + \Sigma n_{i2}^2 + 2\Sigma n_{i1}n_{i2} = \kappa_p(N_1^2+N_2^2+2N_1N_2) - \kappa_p(N_1+N_2) + (N_1+N_2)$$

uit (25) volgt:

(28) $$\Sigma n_{i1}^2 = \kappa_p N_1^2 - \kappa_p N_1 + N_1 \ ,$$

(29) $$\Sigma n_{i2}^2 = \kappa_p N_2^2 - \kappa_p N_2 + N_2 \ .$$

Na aftrekken links en rechts van het gelijkteken houden we over:

(30) $$2\Sigma n_{i1}n_{i2} = \kappa_p(2N_1N_2) \ ,$$

(31) $$\chi = \frac{\Sigma n_{i1}n_{i2}}{N_1N_2} = \kappa_p = 0{,}0827 \ .$$

De statistische grootheid χ stelt ons in staat te zien of twee steekproeven afkomstig zijn uit hetzelfde alphabet. Met

andere woorden of twee permutaties van het normale alphabet identiek zijn.

b) C-toets.

Deze toets wordt uitsluitend gebruikt als de aantallen letters in beide verdelingen gelijk zijn: $N_1 = N_2 = N$.

Men berekent:

$$(32) \qquad C = \frac{\Sigma \min(n_{i1}, n_{i2})}{2N} = \frac{\Sigma |n_{i1}+n_{i2}| - |n_{i1}-n_{i2}|}{4N} .$$

Voor twee identieke verdelingen is C = 0,5. De statistische grootheid C varieert tussen 0 en 0,5.

Voor C > 0,3 kan men met vrij grote zekerheid aannemen, dat de steekproeven afkomstig zijn uit dezelfde verdeling. Voor C ≦ 0,2 kan men besluiten tot ongelijkheid van de verdelingen. De toets is minder scherp dan de χ-toets.

4. Toepassingen.

In deze paragraaf zullen de beschreven statistische methoden toegepast worden bij de ontcijfering van een cryptogram.

We beginnen met de beschrijving van het gebruikte cryptografisch systeem.

De letters van het alphabet worden opgevat als de elementen van een eindige groep.

Een samenstellingsvoorschrift is gegeven dat we optelling noemen. Het eenvoudigst kunnen we dit bereiken door de letters (in een bepaalde volgorde) één-éénduidig af te beelden op de natuurlijke getallen van 0-25. De letters van het alphabet stellen dan de restklassen mod 26 voor. Met de letters kunnen we dan rekenen. In het voorbeeld zijn de letters in hun normale volgorde afgebeeld op de getallen 0-25.

A	B	C	D	E	F	G	H	I	J	K	L	M
0	1	2	3	4	5	6	7	8	9	10	11	12
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

Een eindige rij letters voorgesteld door

$$(33) \qquad s_1, s_2, \ldots, s_k$$

is gegeven als sleutel. Deze letters worden opgeteld bij de letters van de tekst op de volgende manier.

$$t_1, t_2, t_3, \ldots, t_k, t_{k+1}, \ldots, t_{2k}, t_{2k+1}, \ldots, t_{3k}, t_{3k+1} \ldots$$

$$(34) \quad s_1, s_2, s_3, \ldots, s_k, s_1, \ldots, s_k, s_1, \ldots, s_k, s_1 \ldots$$

$$c_1, c_2, c_3, \ldots, c_k, c_{k+1}, \ldots, c_{2k}, c_{2k+1}, \ldots, c_{3k}, c_{3k+1} \ldots$$

Deze transformatie is een voorbeeld van de transformatie be-

schreven in paragraaf 3.3, (11), (12) en (13). De parameters van de transformatie zijn de letters $s_1, \ldots, s_k$. De operatie is de optelling mod 26.

Om mnemotechnische redenen kiest men voor de eindige sleutel vaak een woord of een zin.

Het blijkt dat de gekozen sleutel cyclisch wordt gebruikt.

In de informatietheorie is het gebruikelijk de transformatie op te vatten als een storing, zoals ook voorkomt bij radio-uitzendingen en telefoongesprekken. Het vercijferen van een bericht met het beschreven cryptografische systeem kunnen we dus opvatten als een stochastisch proces dat systematisch gestoord wordt.

We mogen deze beschouwingswijze ook omkeren en het cryptogram opvatten als een stochastisch proces of tijdreeks ontstaan uit een cyclische component en een stochastische storing met bekende verdeling.

De analyse van een dergelijk cryptogram komt dus neer op het splitsen van een tijdreeks in een cyclische en een stochastische component.

Om een goede vercijfering tot stand te brengen moet de sleutel zeer lang zijn. Dit is te bereiken door het bericht achtereenvolgens te vercijferen met betrekkelijk korte sleutels. De lengte van de resulterende sleutel is dan hun kleinste gemene veelvoud. Het bewerkstelligen van een dergelijke vercijfering met behulp van potlood en papier is een zeer tijdrovende bezigheid die bovendien grote kans biedt op het maken van fouten. Verschillende typen apparaten waarmee berichten mechanisch of electrisch vercijferd worden berusten op het beginsel van een herhaalde toepassing van niet te lange sleutels.

De grote hoeveelheid tekst die nodig is om resultaten te verkrijgen verhindert een demonstratie van de statistische ontcijfering van een machinaal vercijferd cryptogram.

4.1. Voorbeeld.

```
      (10)       (20)       (30)       (40)       (50)
FATSM EEXVR WPZWK MXVKX TMTWE FLXTO WMGOP RXHRF ILBMS FQRML
      (60)       (70)       (80)       (90)       (100)
FDEGO WPKSG KMBLV XTZKZ EXUWK SMMWO CEFPI ETWDM YBUGO VSDTZ
      (110)      (120)      (130)      (140)      (150)
MTTHG ZMELG EUWPO ZDXGI DEXSK GUZSG SRUTH TIXHE KASZE GZWIL
      (160)      (170)      (180)      (190)      (200)
DETTB OWLBE DUSMP WOHXS WPTIJ DEGMF DCRXQ NOVSD GXALG SHOHR
      (210)      (220)      (230)      (240)      (250)
WMFWW SBQEG ZXACA IKVPW UAPKW UXAGB KPMVX DAGAX QYQSM IXANT
```

```
      (260)       (270)       (280)       (290)       (300)
VDOQR XLRHO NLAID MKDUG LAZME KDDMY EZSYW JOLVW GDXHW KMIUE
      (310)       (320)       (330)       (340)       (350)
EGPRM NWMJG KMWZE XVKXU EXFIL JLQOO MJGLR CHMLB NXZXY ZEGTI
      (360)       (370)       (380)       (390)       (400)
FDXWG TZTJH NDMDB ROOPE VYCBT HIFCS CKXBX ALNLI ODDWP IQIZM
      (410)       (420)       (430)       (440)       (450)
FUJLS IPXVN TVHOF ATSHU EBVNC YMSGT UTACD IXNIC JVKPM WLTHR
      (460)
KYOKU SMQIC.
```

Het totaal aantal letters in het cryptogram is 460.

De frequentieverdeling van de letters is:

	A	B	C	D	E	F	G	H	I	J	K	L	M
n_i	15	12	11	23	24	13	24	15	20	8	20	20	31
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
n_i	10	20	15	9	13	21	24	15	15	26	30	9	17

Al is deze verdeling veel vlakker dan voor normale tekst (fig. 1), de afwijkingen van de homogene verdeling zijn duidelijk. Het is dus niet nodig deze nog afzonderlijk aan te tonen.

We zullen de coïncidentie-toets A gebruiken om te onderzoeken of het aantal coïncidenties van tetragrammen significant afwijkt van de verwachting.

In de tekst vinden we de volgende coïncidenties:

FATS	op	de	plaatsen	1	en	415
XVKX	"	"	"	17	en	316
OVSD	"	"	"	95	en	187

De verwachting van het aantal tetragram-coïncidenties is

$$(35) \qquad \binom{N-z+1}{2} \sum_{1}^{26^4} \left(\frac{1}{26^4}\right)^2 = \binom{457}{2} \frac{1}{26^4} = 0{,}23.$$

De kans op het voorkomen van 3 coïncidenties in het gegeven cryptogram onder de hypothese dat het ontstaan is door onafhankelijke trekkingen uit een homogene verdeling is 0,0015.

De gevonden coïncidenties zijn dus met vrij grote zekerheid niet toevallig ontstaan.

Deze coïncidenties worden nu gebruikt om te onderzoeken of het cryptogram vercijferd is met een periodieke sleutel.

Als we inderdaad te maken hebben met een periodieke sleutel en de letters FATS zijn op beide plaatsen in het cryptogram ontstaan door vercijfering van gelijke letters uit de normale tekst met dezelfde sleutelletters, dan is de afstand tussen beide tetragrammen een veelvoud van de sleutellengte. We vinden:

FATS: 415-1 = 414 = 2.9.23
XVKX: 316-17 = 299 = 13.23
OVSD: 187-95 = 92 = 4.23

In alle drie gevallen vinden we dat de afstand tussen de coïncidenties een veelvoud is van 23. Voorlopig nemen we aan dat de tekst inderdaad vercijferd is met een sleutel van de lengte 23.

Er moet op gewezen worden dat dit een uiterst eenvoudig voorbeeld is. In de meeste gevallen is de afstand tussen de coïncidenties een functie van de sleutellengte. Deze functie is uit een klein aantal waarnemingen niet altijd éénduidig bepaald.

Het cryptogram wordt uitgeschreven in rijen van 23 letters (tabel 1)

F	A	T	S	M	E	E	X	V	R	W	P	Z	W	K	M	X	V	K	X	T	M	T
W	E	F	L	X	T	O	W	M	G	O	P	R	X	H	R	F	I	L	B	M	S	F
Q	R	M	L	F	D	E	G	O	W	P	K	S	G	K	M	B	L	V	X	T	Z	K
Z	E	X	U	W	K	S	M	M	W	O	C	F	F	P	I	E	T	W	D	M	Y	B
U	G	O	V	S	D	T	Z	M	T	T	H	G	Z	M	E	L	G	E	U	W	P	O
Z	D	X	G	I	D	E	X	S	K	G	U	Z	S	G	S	R	U	T	H	T	I	X
H	E	K	A	S	Z	E	G	Z	W	I	L	D	E	T	T	B	O	W	L	B	E	D
U	S	M	P	W	O	H	X	S	W	P	T	I	J	D	E	G	M	F	D	C	R	X
Q	N	O	V	S	D	G	X	A	L	G	S	H	O	H	R	W	M	F	W	W	S	B
Q	E	G	Z	X	A	C	A	I	K	V	P	W	U	A	P	K	W	U	X	A	G	B
K	P	M	V	X	D	A	G	A	X	Q	Y	Q	S	M	I	X	A	N	T	V	D	O
Q	R	X	L	R	H	O	N	L	A	I	D	M	K	D	U	G	L	A	Z	M	E	K
D	D	M	Y	E	Z	S	Y	W	J	O	L	V	W	G	D	X	H	W	K	M	I	U
E	E	G	P	R	M	N	W	M	J	G	K	M	W	Z	E	X	V	K	X	U	E	X
F	I	L	J	L	Q	O	O	M	J	G	L	R	C	H	M	L	B	N	X	Z	X	Y
Z	E	G	T	I	F	D	X	W	G	T	Z	T	J	H	N	D	M	D	B	R	O	O
P	E	V	Y	C	B	T	H	I	F	C	S	C	K	X	B	X	A	L	N	L	I	O
D	D	W	P	I	Q	I	Z	M	F	U	J	L	S	I	P	X	V	N	T	V	H	O
F	A	T	S	H	U	E	B	V	N	C	Y	M	S	G	T	U	T	A	C	D	I	X
N	I	C	J	V	K	P	M	W	L	T	H	R	K	Y	O	K	U	S	M	Q	I	C

Tabel 1

Wanneer een normale tekst geschreven wordt in rijen van 23 letters dan vormen de letters in de kolommen steekproeven uit normale tekst. In de inleiding (paragraaf 1 (f)) is vastgesteld dat voor een dergelijke steekproef geldt dat de frequentiequotiënten $\frac{n_i'}{N'}$ naderen tot de whn. p(i) van de taal.

Als onze veronderstelling juist is en de tekst is dus vercijferd met een sleutel van 23 letters, betekent dit dat in iedere kolom van tabel 1 letters staan die vercijferd zijn met eenzelfde (vooralsnog onbekende) sleutelletter.

Uit de algebraische eigenschappen van de transformatie volgt onmiddellijk dat de vercijfering van de letters uit een normaal alphabet een permutatie van dit alphabet tengevolge heeft. De frequentiequotiënten veranderen niet van waarde, maar

zijn toegevoegd aan andere letters. Men zegt meestal dat de letters "behoren" tot een bepaald alphabet.

We toetsen de vraag of de letters tot een alphabet behoren met de φ-toets. De frequenties van de letters voor ieder van de 23 kolommen vinden we in tabel 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A	-	2	-	1	-	1	1	1	2	1	-	-	-	-	1	-	-	2	2	-	1	-	-
B	-	-	-	-	-	1	-	1	-	-	-	-	-	-	-	1	2	1	-	2	1	-	3
C	-	-	1	-	1	-	1	-	-	-	2	1	1	1	2	-	-	-	-	1	1	-	1
D	2	3	-	-	-	5	1	-	-	-	-	1	1	-	-	1	1	-	1	2	1	1	1
E	1	7	-	-	1	1	5	-	-	-	-	-	1	1	-	3	1	-	1	-	-	3	-
F	3	-	1	-	1	1	-	-	-	2	-	-	-	1	-	-	1	-	2	-	-	-	1
G	-	1	3	1	-	-	1	3	-	2	4	-	1	1	3	-	2	1	-	-	-	1	-
H	1	-	-	-	1	1	1	1	-	-	-	2	1	-	4	-	-	1	-	1	-	1	-
I	-	2	-	-	3	-	-	-	2	-	2	-	1	-	1	2	-	1	-	-	-	5	-
J	-	-	-	2	-	-	-	-	-	3	-	1	-	2	-	-	-	-	-	-	-	-	-
K	1	-	1	-	-	2	-	-	-	2	-	2	-	3	2	-	2	-	2	1	-	-	2
L	-	-	1	3	1	-	-	-	1	2	-	3	1	-	-	-	2	2	2	1	1	-	-
M	-	-	4	-	1	1	-	2	6	-	-	-	3	-	2	3	-	3	-	1	4	1	-
N	1	1	-	-	-	-	1	1	-	1	-	-	-	-	-	1	-	-	3	1	-	-	-
O	-	-	2	-	-	1	3	1	1	-	3	-	-	1	-	1	-	1	-	-	-	1	5
P	1	1	-	3	-	-	1	-	-	-	2	3	-	-	1	2	-	-	-	-	-	1	-
Q	4	-	-	-	-	2	-	-	-	-	1	-	1	-	-	-	-	-	-	-	1	-	-
R	-	2	-	-	2	-	-	-	-	1	-	-	3	-	-	2	1	-	-	-	1	1	-
S	-	1	-	2	3	-	2	-	2	-	-	2	1	4	-	1	-	-	1	-	-	2	-
T	-	-	2	1	-	1	2	-	-	1	3	1	1	-	1	2	-	2	1	2	3	-	1
U	2	-	-	1	-	1	-	-	-	-	1	1	-	1	-	1	1	2	1	1	1	-	1
V	-	-	1	3	1	-	-	-	2	-	1	-	1	-	-	-	-	3	1	-	2	-	-
W	1	-	1	-	2	-	-	2	3	4	1	-	1	3	-	-	1	1	3	1	2	-	-
X	-	-	3	-	3	-	-	5	-	1	-	-	-	1	1	-	6	-	-	6	-	1	4
Y	-	-	-	2	-	-	-	1	-	-	-	2	-	-	1	-	-	-	-	-	-	1	1
Z	3	-	-	1	-	2	-	2	1	-	-	-	2	1	1	-	-	-	-	1	1	1	-

Tabel 2

Voor iedere kolom is N = 20. Het gemiddelde van $\frac{\varphi}{N(N-1)}$ voor alle 23 kolommen is

(36) $$\frac{1}{23}\sum_{1}^{23}\frac{\varphi_i}{N(N-1)} = 0{,}073.$$

Deze waarde is voldoende hoog om aan te nemen dat de verdelingen in alle kolommen bij een alphabet behoren en dat de lengte van de sleutel inderdaad gelijk is aan 23.

Hetzelfde resultaat had bereikt kunnen worden door toepassen van coïncidentietoets B.

Tabel 1 stelt dan de superposities voor van stukken cryptogram van 23 letters.

De coïncidenties worden dan geteld tussen alle combinaties 2 aan 2 van de zo gesuperponeerde rijen Het resultaat is natuurlijk precies gelijk aan de gemiddelde waarde van φ

De mogelijkheid bestaat dat verschillende kolommen vercijferd zijn met dezelfde sleutelletter. Dit betekent dat ze behoren tot hetzelfde alphabet. Om dit te onderzoeken wordt de χ-toets gebruikt. Alle alphabetten worden met elkaar vergeleken. De resultaten zijn te vinden in tabel 3. De onderstreepte getallen zijn significant en geven aan dat de twee kolommen hoogstwaarschijnlijk letterverdelingen bezitten uit hetzelfde alphabet. We vinden dat de volgende kolommen dezelfde verdeling bezitten.

a) 1=6 b) 2=7=16=22 16=22 7=22 c) 3=8=9=17=20 8=17=20 17=20 9=13=18=21 18=21 d) 4=12 e) 10=14=19

Na eliminatie van enkele twijfelachtige gevallen vinden we

1=6
2=7=16
3=8=17=20
4=12
10=14=19
18=21

De gelijke verdelingen worden gecombineerd in tabel 4:

	1+6	2+7+16	3+8+17+20	18+21	4+12	10+14+19
A	1	3	1	3	1	3
B	1	1	5	2	-	-
C	-	1	2	1	1	1
D	7	5	3	1	1	1
E	2	15	1	-	-	2
F	4	-	2	-	-	5
G	-	2	8	1	1	3
H	2	1	2	1	2	-
I	-	4	-	1	-	-
J	-	-	-	-	3	5
K	3	-	4	-	2	7
L	-	-	4	3	6	4
M	1	3	7	7	-	-
N	1	3	2	-	-	4
O	1	4	3	1	-	1
P	1	4	-	-	6	-
Q	6	-	-	1	-	-
R	-	4	1	1		1
S	-	4	-	-	4	5
T	1	4	4	5	2	-
U	3	1	2			
V	-	-	1	5	3	1
W	1	-	3	3	-	10
X	-	-	20	-	3	2
Y	-	-	-	-	2	-
Z	5	-	3	1	-	1

Tabel 3

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	38	13	20	18	83	25	25	15	33	23	28	38	33	25	25	28	15	48	33	33	25	23
2		8	20	50	60	135	15	25	18	25	20	50	30	15	90	35	18	45	18	18	90	9
3			28	45	28	35	90	80	53	70	20	53	40	65	45	75	68	30	75	70	40	70
4				40	13	28	18	40	40	43	75	40	38	25	25	23	40	38	18	38	28	10
5					10	33	55	68	43	28	40	55	60	35	48	65	35	38	60	35	78	35
6						42	25	28	20	23	33	42	40	35	40	35	40	40	53	48	33	50
7							23	23	15	58	33	33	45	33	65	20	28	33	23	23	65	48
8								57	40	42	10	40	42	65	23	100	40	28	100	40	38	65
9									40	40	18	75	48	42	63	13	90	48	28	95	55	13
10										38	35	40	73	33	13	63	35	75	42	38	10	28
11											40	35	25	57	45	25	53	20	25	42	48	53
12												25	45	50	30	30	30	38	33	23	30	25
13													30	48	60	25	50	25	33	70	53	8
14														33	23	50	18	57	38	23	38	45
15															30	40	45	18	48	28	45	30
16																23	48	28	33	57	75	30
17																	28	40	120	23	33	93
18																		45	30	83	28	30
19																			-	-	-	-
20																				-	-	-
21																					-	-
22																						-
23																						

Deze combinaties kunnen voor een deel al herkend worden als cyclische permutaties van het normale alphabet. Om dit vermoeden te bevestigen vergelijken we ze met een standaard alphabet (fig 1). Hiervoor geb uiken we weer de χ-toets. Het onderzochte alphabet wordt zo lang cyclisch gepermuteerd tot χ een waarde aanneemt die significant is.

Uit de algebraische eigenschappen van de transformatie blijkt dat het aantal plaatsen dat het alphabet cyclisch verschoven is gelijk is aan de waarde van de sleutelletter.

We vinden ten slotte:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

M A T H - M A T - S - H - S - A T I S T I - -

Het is een simpel werk deze letters aan te vullen tot een sleutelzin:

MATHEMATISCHE STATISTIEK.

Achteraf blijkt dat de combinatie van een paar ongelijke verdelingen de analyse niet heeft gehinderd.

De ontcijfering van het cryptogram is nu eenvoudig en wordt aan de lezer overgelaten.

Zonder enige veronderstellingen te maken over de inhoud van de oorspronkelijke tekst zijn we er in geslaagd met behulp van uitsluitend statistische methoden het cryptogram te ontcijferen.