Fall 2006
Chris Christensen
MAT/CSC 483

The Friedman Test

William Friedman (1891 – 1969) developed statistical methods for determining whether a cipher is monoalphabetic or polyalphabetic and for determining the length of the keyword if the cipher is polyalphabetic.

Friedman retired from the National Security Agency in 1955 after 35 years of service with U.S. cryptological activities. He transformed the methods and approaches of cryptology from the traditional into the modern by applying statistics to cryptology. His wife Elizebeth was also a cryptologist and served at one point with the Coast Guard, cryptanalyzing messages of the rumrunners.

Determining Whether We Have a Monoalphabetic Cipher or a Polyalphabetic Cipher

Friedman's method for determining whether a cipher is monoalphabetic or polyalphabetic is based upon the probability of randomly selecting two letters from an alphabet and having them be the same.

Let us first consider an example drawing two cards from a standard deck of 52. We want to know the probability that both cards are diamonds. There are 13 diamonds in the deck of 52; so, the probability that the first card selected is a diamond is $\frac{13}{52} = \frac{1}{4}$. Now only 12 diamonds remain among the remaining 51 cards; so, the probability that the second card selected is a diamond is $\frac{12}{51}$. The probability that both of these events; i.e., that two diamonds are drawn is $\frac{1}{4} \times \frac{12}{51} \approx 0.059$.

We will use the same reasoning to determine the probability of "drawing" two of the same letters from a ciphertext.

Consider the probability of randomly selecting two letters from a ciphertext alphabet and having them be the same. Let us say that there are n letters in our ciphertext and n_a `a`s in our ciphertext. Then the probability of selecting two `a`s would be $\frac{n_a}{n} \times \frac{n_a - 1}{n-1}$.

The probability of choosing two letters the same (i.e., two `a`s or two `b`s or two `c`s or … or two `z`s) would be

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

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

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

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

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 $I \approx 0.0656010$.

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:

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

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.

Typically we use both the Friedman test and frequency analysis to determine the kind of cipher we have.

Here is the Vigenère cipher that we cryptanalyzed in a previous section using the Kasiski test:

```
DBZMG   AOIYS   OPVFH   OWKBW   XZPJL   VVRFG   NBKIX
DVUIM   OPFQL   VVPUD   KPRVW   OARLW   DVLMW   AWINZ
DAKBW   MMRLW   QIICG   PAKYU   CVZKM   ZARPS   DTRVD
ZWEYG   ABYYE   YMGYF   YAFHL   CMWLW   LCVHL   MMGYL
DBZIF   JNCYL   OMIAJ   JCGMA   IBVRL   OPVFW   OBVLK
OPVUJ   ZDVLQ   XWDGG   IQEYF   BTZMZ   DVRMM   ANZWA
ZVKFQ   GWEAL   ZFKNZ   ZZVCK   VDVLQ   BWFXU   CIEWW
OPRMU   JZIYK   KWEXA   IOIYH   ZIKYV   GMKNW   MOIIM
KADUQ   WMWIM   ILZHL   CMTCH   CMINW   SBRHV   OPVSO
DTCMG   HMKCE   ZASYD   JKRNW   YIKCF   OMIPS   GAFZK
JUVGM   GBZJD   ZWWNZ   ZVLGT   ZZFZS   GXYUT   ZBJCF
PAVNZ   ZAVWS   IJVZG   PVUVQ   NKRHF   DVXNZ   ZKZJZ
ZZKYP   OIEXX   MWDNZ   ZQIMH   VKZHY   DVKYD   GQXYF
OOLYK   NMJGS   YMRML   JBYYF   PUSYJ   JNRFH   CISYL
N
```

We begin with the frequency analysis:

```
A    1111111111111111111                              19
B    11111111111111                                   14
C    1111111111111111                                 16
D    111111111111111111111                       20
E    11111111                                          8
F    1111111111111111111                              19
G    111111111111111111111                       20
H    111111111111                                     12
I    1111111111111111111111111                        26
J    1111111111111111                                 16
K    11111111111111111111111                     25
L    1111111111111111111111                           22
M    11111111111111111111111111111111                 32
N    1111111111111111                                 16
O    11111111111111111111                             18
P    1111111111111111                                 16
Q    1111111111                                  10
R    11111111111111                                   14
S    11111111111                                      11
T    111111                                            6
U    11111111111                                      11
V    1111111111111111111111111111111111               34
W    11111111111111111111111111111                    28
X    1111111111                                  10
Y    1111111111111111111111111111                     27
Z    111111111111111111111111111111111111111111       41
                                                     491
```

Notice that the relatively uniform frequencies suggest a polyalphabetic cipher. This should be confirmed by our calculation of *I*.

The calculation of *I* is easy.

$$
\begin{aligned}
I &= \left(\frac{19}{491}\times\frac{18}{490}\right)+\left(\frac{14}{491}\times\frac{13}{490}\right)+\left(\frac{16}{491}\times\frac{15}{490}\right)+\ldots+\left(\frac{41}{491}\times\frac{40}{490}\right)\\
&= \frac{(19\times 18)+(14\times 13)+(16\times 15)+\ldots+(41\times 40)}{491\times 490}\\
&\approx 0.044
\end{aligned}
$$

Because I is so near to 0.038 (random alphabet), we can reasonably assume that we have a polyalphabetic cipher. This confirms what we noticed in the frequency analysis.

Estimating the Length of the Keyword

Friedman also developed a method for estimating the length of the keyword. The statistics for estimating the length of the keyword are more complicated. (But, the calculations will not be hard.)

We are trying to estimate the length l of the keyword. We will develop an approximation formula for I, the index of coincidence; this formula will contain l and n, the number of letters in the ciphertext. Then, to get an approximation for the length l, we will solve for l in terms of I and n (we know n and can calculate I).

First, assume that we know l and arrange the ciphertext into l columns. Now each column corresponds to a Caesar cipher. Although the columns might not all have the same length, we will assume that the number of letters in the ciphertext is large enough so that we can assume that they each have length $\frac{n}{l}$; i.e., the error using this number for the length of each column is not large.

If we chose two letters from the ciphertext, what is the probability that they come from the same column and are the same letter? First, we select a letter from the ciphertext. This selection determines a column. The probability that the next letter chosen comes from the same column is $\frac{\frac{n}{l}-1}{n-1}$. Because both letters are selected from the same Caesar cipher alphabet, the probability that both are the same is approximately the same as for standard English 0.065. So, the probability that both letters are selected from the same column and are the same letter is approximately $\frac{\frac{n}{l}-1}{n-1} \times 0.065$.

The other possibility is that we select two letters from the ciphertext that come from different columns but are the same letter. What is that probability? First, we select a letter from the ciphertext. Again, this determines a column. The probability that the next letter comes from a

different column is $\frac{n-\frac{n}{l}}{n-1}$. Because the two letters are selected from different Caesar cipher alphabets, the probability that both are the same is approximately the same as for a random alphabet 0.038. So, the probability that both letters are selected from different columns and are the same letter is approximately $\frac{n-\frac{n}{l}}{n-1}\times 0.038$.

So we have two cases – the two letters are selected from the same column and are the same letter or the two letters are selected from different columns and are the same letter. To get an approximation of the index of coincidence *I*, the probability that the two letters selected are the same, we add these two probabilities:

$$I \approx \frac{\frac{n}{l}-1}{n-1}\times 0.065 + \frac{n-\frac{n}{l}}{n-1}\times 0.038.$$

Doing a bit of algebra to solve for *l*, we obtain:

$$I \approx \frac{\frac{n}{l}-1}{n-1}\times 0.065 + \frac{n-\frac{n}{l}}{n-1}\times 0.038$$

$$(n-1)I \approx \left(\frac{n}{l}-1\right)\times 0.065 + \left(n-\frac{n}{l}\right)\times 0.038$$

$$(n-1)I \approx \frac{n}{l}\times 0.065 - 0.065 + n\times 0.038 - \frac{n}{l}\times 0.038$$

$$(n-1)I + 0.065 - 0.038n \approx \frac{n}{l}\times(0.065-0.038)$$

$$(n-1)I + 0.065 - 0.038n \approx 0.027\frac{n}{l}$$

$$l \approx \frac{0.027n}{(n-1)I + 0.065 - 0.038n}$$

Now let us determine an approximation for the length of the keyword in the ciphertext given above.

We determined that n = 491, and we calculated above that $I \approx 0.044$; so,

$$l \approx \frac{0.027 \times 491}{(491-1) \times 0.044 + 0.065 - 0.038 \times 491} \approx 4.468 .$$

Recall that the Kasiski Test when applied to this ciphertext suggested that the length of the keyword is 5. In practice, we should consider the results of both tests.

The Columns are Monoalphabetic

Before ending, let us go back to the five alphabets of the ciphertext example given above and calculate I for alphabet one. I should be near 0.065, the monoalphabetic case of I.

Alphabet one:

$$I \approx \frac{4 \times 3 + 2 \times 1 + 6 \times 5 + 10 \times 9 + 6 \times 5 + 5 \times 4 + 7 \times 6 + 3 \times 2 + 4 \times 3 + 4 \times 3 + 13 \times 12 + 4 \times 3 + 3 \times 2 + 2 \times 1 + 4 \times 3 + 2 \times 1 + 4 \times 3 + 16 \times 15}{99 \times 98}$$

$I \approx 0.072$.

This confirms what we notice by looking at the frequencies of alphabet number one.

This suggests a brute force way to attack the length of the keyword. Assume that l = 1, 2, 3, …; for each of these values of l separate alphabets and calculate I; and determine for which value of l all the separated alphabets are monoalphabetic.

Consider the following ciphertext:

```
pukpz gmoqs ihzil ooiop xwtyf hxpfa epmng hhyfh
pelvy enzqo yatev vymxy oitsq nbnya fhohb uqbna
iimop iymqr xuflr durxk domvd stupd bsxyd uaxkl
oold ewate bufek umlpp digna fmoqs xatel voaes
qdkhu lphke gpsmt omsmo qsttq btzuc laduv agrxh
etaly avoun xbeew iktal uttsu agzno moafm oqsxc
qrlpa nlvrt alqnb nyafh ohbuq wapoh wppnh ataol
blnnn otyps ahpag lztkf pilja npouc aatex sqcmy
uctso ogamc mziek loogu qcmlp thate dlkbh hddbu
fhxvd dxycw xyfzn paalk rgaqw preov uuyl
```

If $l = 1$.

$I = .04408$

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

Both the frequencies and the calculation of I suggest that this has been encrypted with a polyalphabetic cipher.

If $l = 2$.

Alphabet number 1

$I = .04731$

```
A = *************
B = ********
C = *
D = *********
E = ****
F = *********
G = ********
H = ****
I = ***
J =
K = *********
L = ********
M = **************
N = ******
O = *************
P = ************
Q = ***************

R = ****
S = *
T = ***************
U = *********
V = ***
W = ****
X = **************
Y = ********
Z = ***
```

This does not appear to be monoalphabetic.

Alphabet number 2

I = .0521

```
A = *******************
B = ****
C = *******
D = *****
E = *************
F = ***
G = **
H = ***************
I = ********
J = *

K = **
L = ***************
M = ***
N = ***********
O = *****************
P = ***********
Q = *
R = ****
S = *************
T = *******
U = ************
V = *******
W = ***
X = *
Y = ********
Z = *****
```

This also does not appear to be monolalphabetic.

l = 2 does not seem to be the correct length.

If $l = 3$.

Alphabet number 1

$I = .40947$

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

This does not appear to be monoalphabetic.

Alphabet number 2

I = .04522

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

This does not appear to be monoalphabetic.

Alphabet number 3

I = .04347

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

This also does not appear to be monoalphabetic.

l = 3 does not seem to be the correct length.

If $l = 4$.

Alphabet number one

$I = .06367$

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

This appears to be monoalphabetic. It seems to correspond to a Caesar cipher. If this length is correct, what is the first letter of the keyword?

Frequency analysis and the Friedman test and the Kasiski test

Both frequency analysis and the Friedman test can be used to determine whether a cipher is monoalphabetic or polyalphabetic.

Both the calculation of l or the Kasiski test can be used to determine the length of the keyword.

If a polyalphabetic cipher is a Vigenère cipher and the Vigenère square consists of the 26 Caesar ciphers, then the keyword can be determined by separating alphabets and determining the shifts.

Alphabet number 2

I = .05286

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

This also appears to be a monoalphabtic cipher. It also seems to correspond to a Caesar cipher. If this length is correct, what is the second letter of the keyword?

Alphabet number 3

I = .05852

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

This also appears to be a monoalphabtic cipher. It also seems to correspond to a Caesar cipher. If this length is correct, what is the third letter of the keyword?

Alphabet number 4

I = .06831

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

It appears that the length of the keyword is l = 4. What is the keyword? What is the plaintext message?

Alphabet number one

```
A    1111
B    11
C    111111
D    1111111111
E
F
G    111111
H    1
I    11111
J    1111111
K    111
L    1
M    1111
N    1111
O    1111111111111
P    1111
Q    1
R
S    1
T
U
V    1111
W    1
X    11
Y    1111
Z    1111111111111111
```

The frequencies indicate that these letters correspond to a monoalphabetic cipher. In fact, they correspond to a Caesar cipher, and we can determine the first letter of the keyword of the Vigenère cipher (recall that it is V).

When we have the correct length of the keyword and have separated the alphabets correctly, each alphabet should have a frequency analysis that corresponds to a Caesar cipher and each alphabet should have a value of I near 0.065.

Exercises

1. Calculate *I* for the remaining four alphabets of the example. Compare these calculations with the frequency analyses that we did earlier.

2. Consider the following ciphertext:

```
wgixf irtnx amwpz gfcln bztef roozn maour tlrno
dsxjw xxdan zhdix nqtta hogcm rwrvj numyb gxavt
mgzdt ewlqs wvwtm lgblk nrins ozgif bgnlm fpsqn
xhvja ufgmj xyxum hqsxv vztea bzrpt lrijy ivnto
fywew uyfse beiaw vbimm igwhq ceytk ppien udmiq
nkmtw bnidy tgitm lcfqy fhegp ghewv viqbi pwsql
itmtp avlzk mdmao gxmsf bgxls sokdm eagyz azntg
zdhvx rameg qifre snood yeqts tlreg dirpt apirt
bnqfe zwaez mzukd meavz qlitz gytln bxqvi ntkpg
ieugz ywczu bowrl gxlmn viqwm gpsqx mpwgs amaaz
fhihv ofehf bgfew nvjih sfmju sgywy grium rbehc
zubep nukdi gnqtf oxumc mr
```

2a. Apply the Friedman test to the ciphertext: calculate *I* and approximate *l*.

2b. Do a frequency analysis of the ciphertext. Does this agree with the calculated value of *I* ?

2c. Do a Kasiski test to determine the length of the keyword. Does this agree with the calculated value of *l* ?

2d. Separate the alphabets corresponding to the length of the keyword. For each alphabet, calculate *I* and do a frequency analysis. For each alphabet, do these agree? Do they confirm that each alphabet corresponds to a monoalphabetic cipher?

3. In the calculation for *l*, we assumed that there were no repeated letters in the keyword. Where was that assumption used?

4. Consider the following ciphertext. $I = .04485$.

```
fewpo mfshg rygkm xyzvn yadcs icaqy asazk uyzbn
lygvy gdyez uvbmq ojrxx fejbx zvxxv xapbc oczbb
nuoca dyemy cdyqm ccrye wyfcu tomgr ycbyp smrzl
rguem cemoz cnnxw rcnuk nrxuo vyquh bhuan bvcwb
vfrka homgy ioduv xufez sswvo hgxoz lyeyz gryfo
frdnr bjnsl fklrx iguhb ghoen vdcfl yysyi oxgru
gnoes htdbr cjnxc frwvf cygue dbrsn nvcnx hnfsf
ygrdc zomgb uacgv dnrnm vqhnv mgrug ruqly rxyam
ccrye oxjsn udbrc uzoya sazkw bxzvq oeknv yhcbi
isxrn nuknn cospc psyad hhwvr bisdb rcyuk xooya
shgol pojgo xnxxv nyadc ssyqd briqb efqru iojey
pvnyq kmrdi swyfc utomv xxrzn uohnl fvxag rygkm
xshuk hqdio omhmw rcmse fyiwb wjyon rn
```

4a. Do a brute force attack on the length of the keyword by assuming that l = 1, 2, 3, …; for each of these values of l separating alphabets and calculating I; and determining for which value of l all the separated alphabets are monoalphabetic.

4b. Calculate l.

4c. Use the Kasiski test to determine the length of the keyword.

4d. Determine the keyword and the plaintext.