Fall 2006
Chris Christensen
MAT/CSC 483

Frequency Analysis

> *Cryptanalysis rests upon the fact that the letters of language have "personalities" of their own. … Though in a cryptogram they wear disguises, the cryptanalyst observes their actions and idiosyncrasies, and infers their identity from these traits. In ordinary monoalphabetic substitutions, [the cryptanalyst's] task is fairly simple because each letter's camouflage differs from every other letter's and the camouflage remains the same throughout the cryptogram.* David Kahn, The Codebreakers.

Frequency analysis is the basic tool of the cryptanalyst. It can be used to identify the type of cipher, and it can be used to identify plaintext/ciphertext correspondences.

Here are the plaintext frequencies:

```
a       1111111
b       1
c       111
d       1111
e       1111111111111
f       111
g       11
h       1111
i       1111111
j
k
l       1111
m       111
n       11111111
o       1111111
p       111
q
r       11111111
s       111111
t       111111111
u       111
v       1
w       11
x
y       11
z
```

Notice that there are peaks and valleys of frequencies corresponding to very frequent and very infrequent plaintext letters. Remember that Caesar ciphers are easy to spot by considering frequencies – the plaintext frequencies are just shifted by the numbers of places corresponding to the key. A more general simple substitution cipher will not have the frequencies shifted; the frequencies will be, perhaps, randomly rearranged corresponding to the permutation of the alphabet prescribed by the cipher. Here are frequencies that correspond to a simple substitution cipher with a randomly generated ciphertext alphabet.

```
A      111
B      11
C      11
D      11
E      1111111
F      1111
G      111
H      111
I      1
J
K      1
L      1111111111111
M      111
N
O      111
P      111111111
Q      1111
R
S      111111
T      1111
U      11111111
V      1111111
W      11111111
X
Y      11111111
Z
```

There are peaks and valleys of frequencies, but they are rather randomly arranged. Peaks and valleys of frequencies is a characteristic of a simple substitution cipher.

For the cipher having the frequencies above, it seems likely that ciphertext `L` corresponds to plaintext `e`, but it might be hard to identify just from the frequencies to what plaintext letters the other high frequency ciphertext letters correspond.

For the frequency patterns to be visible, the cryptanalyst needs a long message. Short messages are hard to attack. Try to cryptanalyze the following ciphertext message that was encrypted with a simple substitution cipher

`PNVAC`

Cryptanalysts pray for long messages; long messages are likely to have letter frequencies that correspond to what would be expected in standard English.

But, individual messages need not be long for the cryptanalyst to find a pattern; a collection of ciphertexts encrypted with the same key provides the same frequency information. Consider the ciphertexts given below. These messages are known to have been encrypted using the same method and key.

Ciphertext number one

```
VXANQ  NJM

A      1
B
C
D
E
F
G
H
I
J      1
K
L
M      1
N      11
O
P
Q      1
R
S
T
U
V      1
W
X      1
Y
Z
```

Ciphertext number two

TNWCD LTHBC JCN

A	
B	1
C	111
D	1
E	
F	
G	
H	1
I	
J	1
K	
L	1
M	
N	11
O	
P	
Q	
R	
S	
T	11
U	
V	
W	1
X	
Y	
Z	

Ciphertext number three

NJBCN AW

A	1
B	1
C	1
D	
E	
F	
G	
H	
I	
J	1
K	
L	
M	
N	11
O	
P	
Q	
R	
S	
T	
U	
V	
W	1
X	
Y	
Z	

Ciphertext number four

VDAAJ H

A	11
B	
C	
D	1
E	
F	
G	
H	1
I	
J	1
K	
L	
M	
N	
O	
P	
Q	
R	
S	
T	
U	
V	1
W	
X	
Y	
Z	

Ciphertext number five

WXACQ NAW

A	11
B	
C	1
D	
E	
F	
G	
H	
I	
J	
K	
L	
M	
N	1
O	
P	
Q	1
R	
S	
T	
U	
V	
W	11
X	1
Y	
Z	

Ciphertext number six

TNWCD LTH

A	
B	
C	1
D	1
E	
F	
G	
H	1
I	
J	
K	
L	1
M	
N	1
O	
P	
Q	
R	
S	
T	11
U	
V	
W	1
X	
Y	
Z	

Ciphertext number seven

UXDRB ERUUN

A	
B	1
C	
D	1
E	1
F	
G	
H	
I	
J	1
K	
L	
M	
N	1
O	
P	
Q	
R	11
S	
T	
U	111
V	
W	1
X	
Y	
Z	

Ciphertext number eight

UXDRB ERUUN

A	
B	1
C	
D	1
E	1
F	
G	
H	
I	
J	1
K	
L	
M	
N	1
O	
P	
Q	
R	11
S	
T	
U	111
V	
W	1
X	
Y	
Z	

Not much information is given by the individual ciphertext frequencies, but when they are collected, the pattern becomes clear.

A	111111
B	11111
C	1111111
D	1111
E	1
F	
G	
H	111
I	
J	1111
K	
L	11
M	1
N	11111111111
O	
P	
Q	11
R	11
S	
T	1111
U	111
V	11
W	1111111
X	11
Y	
Z	

Each message was encrypted with a Caesar cipher with key 9.

Unusual Frequencies

Unusual frequencies can occur when the plaintext is written in other than “standard” English (e.g., scientific papers) or when messages like this occur:

```
Zany Ezekiel, who studies zoology at the
Department of Zoology of the University of New
Zealand and plans to visit Zimbabwe stopped by
Zechariah’s Pizza, which is not in his zip code,
and picked up a pepperoni pizza.
```

A	111111111111
B	111
C	1111
D	1111111
E	1111111111111
F	1
G	11
H	1111111
I	11111111111111
J	
K	11
L	1111
M	11
N	1111111
O	1111111111111
P	111111111111
Q	
R	111
S	11111111
T	11111111
U	11
V	1
W	111
X	
Y	1111
Z	11111111111

Such messages when encrypted might be hard for the cryptanalyst to attack.

Creating a lipogram can be an interesting exercise for a writer.. A lipogram is writing that omits all words containing a particular letter. (*Lipogram* is derived from Greek words. *lipo* is derived from *leipein* which means to leave, lack, or be wanting; and *gram* is derived from *gramma* which means a letter.)

Examples of lipograms date back to Lasus of Achaia, a Greek poet of the 6th century B.C. The classical Greek writer Tryphiodorus composed an Odyssey of 24 volumes. Volume one leaves out the letter alpha, volume two leaves out the letter beta, etc.

Here are two examples of rather lengthy lipogramatic writings.

One of the most famous was written in 1969. Georges Perec (1936 – 1982) composed an 85,000-word novel *La disparation*. The novel is written without using the letter e. Remarkably, the English translation *A Void* by Gilbert Adair also does not include the letter e. Here is an excerpt from the English translation of *A Void*.

> Today, by radio, and also on giant hoardings, a rabbi, an admiral notorious for his links to masonry, a trio of cardinals, a trio, too, of significant politicians (bought and paid for by a rich an corrupt Anglo-Canadian banding corporation), inform us all of how our country now risks dying of starvation. A rumour, that's my initial thought as I switch off my radio, a rumour or possibly a hoax. Propaganda, I murmur anxiously – as though, just by saying so, I might allay my doubts – typical politician's propaganda. But public opinion gradually absorbs it as fact. Individuals start strutting around with stout clubs. 'Food, glorious food!' is a common cry (occasionally sung to Bart's music), with ordinary hard-working folk harassing officials, both local and national, and cursing capitalists and captains of industry. Cops shrink from going out on night shift. In Mâcon a mob storms a municipal building. In Rocadamour ruffians rob a hangar full of foodstuffs, pillaging tons of tuna fish, milk and cocoa, as also a vast quantity of corn – all of it, alas, totally unfit for human consumption. Without fuss or ado, and naturally without any sort of trial, an indignant crowd hangs 26 solicitors on hastily built scaffold in front of Nancy's law courts (this Nancy is a town, not a woman) and ransacks a local journal, a disgusting right-wing rag that is siding against it. Up and down this land of ours looting has brought docks, shops, and farms to a virtual standstill.
>
> The opening paragraph of *A Void* by Georges Perec translated by Gilbert Adair.
> First published in France as *La Disparition* by Editions Denöel in 1969. In Great Britain by Editions Denöel 1969. In English translation by Harvill 1994. Appears in *The Code Book* by Simon Singh.

The "26" that appears in this paragraph is a bit of cheating.

Prior to Perec's novel and Gilbert's translation, perhaps, the most famous lipogram was *Gadsby, A Story of Over 50,000 Words Without Using the Letter E* by Ernest Vincent Wright. It is a 267-page novel of "moderate" literary merit. The following is from David Kahn's *The Codebreakers*.

> Here is how the author summarizes his tale in his opening pages. The excerpt fairly illustrates the book's unique feature:
>
> > *Upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary "fill-ins" as "romantic moonlight casting murky shadows down a long, winding road." Nor will it say anything about twinklings lulling distant folds; robins caroling at twilight, nor any "warm glow of lamplight" from a cabin window. No. It is an account of up-and-doing activity; a vivid portrayal of Youth as it is today; and a practical discarding of that worn-out notion "a child don't know anything."*
>
> The author of *Gadsby*, a persevering, dauntless, white-haired old gentleman named Ernest Vincent Wright, enumerated some of the problems of his self-imposed task. He had to avoid most verbs in the past tense because they end in *–ed*. He could never use *the* or the pronouns *he she*, *they*, *we*, *me*, and *them*. *Gadsby* had to omit such seemingly indispensable verbs as *are*, *have*, *were*, *be*, and *very*. A purist, Wright refused to use numbers between 6 and 30, even as digits because an *e* was implied when they are spelled out. … Similarly he banned *Mr*. And *Mrs*. because of the *e* in their unabbreviated form. One of the most annoying problems would arise, when, near the end of a long paragraph, he could find no *e*-less word with which to complete the thought, and had to go back and rewrite the paragraph. So frequently did Wright find himself wanting to use a word containing *e* that he had to tie down the *e* typebar of his typewriter to make it impossible for one to slip in.

> "And many did try to do so," he says in his preface. "As I wrote along, in long-hand at first, a whole army of little *e*'s gathered around my desk, all eagerly expecting to be called upon. But gradually as they saw me writing on and on, without even noticing them, they grew uneasy; and with excited whisperings amongst themselves, began hopping up and riding on my pen, looking down constantly for a chance to drop off into some word; for all the world like seabirds perched, watching for a passing fish! But when they saw that I had covered 138 pages of typewriter size paper, they slid off onto the floor, walking sadly away, arm in arm; but shouting back: 'You certainly must have a hodge-podge of a yarn there with *Us*! Why, man! We are in every story ever written, *hundreds of thousands of times*! This is the first time we were ever shut out!'"

Lipogramatic writing could yield unusual frequencies, but it is not routinely done for cryptographic purposes.

History

Frequency analysis is not a new idea. It is the historically most basic technique of cryptanalysis. David Kahn claims that "cryptology was born of the Arabs." He points out that several cipher alphabets were included in a Ninth Century Arabic book of magic, and that a Fifteenth Century Arabic encyclopedia included a section on cryptology that includes:

> *Ibn ad-Duraihim has said: When you want to solve a message which you have received in code, begin first of all by counting the letters, and then count how many times each symbol is repeated and set down the totals individually. […] look which letters occur most frequently in the message and compare this with the pattern of letter-frequency previously mentioned. When you see that one letter occurs in the message more often than the rest, then assume that it is alif. …* David Kahn, *The Codebreakers*, quoting from Qalqashandi's *Subh al-a 'sha* [1412].

What about in the West? Cryptology seems to have developed much later.

> *The first Western instance of multiple cipher-representations occurs in a cipher … in 1401 … . Each of the plaintext vowels has several possible equivalents. This testifies silently that, by this time, the West*

> *knew cryptanalysis. There can be no other explanation for the appearance of these multiple substitutes … [this] indicates a knowledge of at least the outlines of frequency analysis.* David Kahn, *The Codebreakers*.

But, Kahn claims that there is no direct connection between the two developments.

> *Where did [the West's knowledge of frequency analysis] come from? It probably developed indigenously. Though it is true that contact with the Moslem and other civilizations during the Crusades triggered the cultural explosion of the Renaissance, and the Arabic works of science, mathematics, and philosophy poured into Europe from Moorish centers of scholarship in Spain, it seems unlikely that cryptanalysis emigrated from there.* David Kahn, *The Codebreakers*, p. 108.

Regardless of the origin of the idea of frequency analysis, it was known by the Fifteenth Century, and much of the subsequent history of cryptography develops from a struggle to nullify its effects. Later we will examine several methods to defeat the powers of frequency analysis. At the moment, it is the primary tool available to us for cryptanalysis.

For the moment, we will just "eyeball" frequency data collected by hand. Later we will apply the statistical techniques of William Friedman to draw inference from frequency data.

Exercises

1. Below are three ciphertext messages. One was encrypted with a Caesar cipher. One was encrypted with a simple substitution cipher that is not a Caesar cipher. One was encrypted by a cipher that is not a simple substitution cipher. Do a frequency analysis of each message and determine which is which.

Ciphertext number one

```
sxmux xzbat nvlgm ioznw yirnc qejfr qlbqj zxjlf
ervsp pobto szycj hlvfc nvwmt tvahm fgdcs qcmhk
rtktj wjnng lrvhs kgcwi owziw spmka msgyo elfav
prchp zzzxs gitpa xapbn vyitv phclo jsvjg admup
xecze hwsh
```

Ciphertext number two

```
odkbf axask eqdhq pmzaf tqdfd mufad ygotn qffqd
zayqd qyaza mxbtm nqfuo egnef ufgfu azera dmynu
fuage nqzqp uofmd zaxpf tqoad dqeba zpqzo qnqfi
qqzmd zaxpm zpvat zmzpd qimeo azpgo fqpuz eqhqd
mxfkb qearo apq
```

Ciphertext number three

```
lnoex vcwxt bwvod nxghp hixah dgceo yxgti xwing
cjznc jiinx gxkce jiocb vgpdi tbtep hohno uxgbt
ixwin xutho vgxth cbhxx ahicu xinti loini nxxmv
xdioc bcytb obygx fjxbi xdohc wxbcv gpdic zgtah
lxgxo bixgv xdixw
```

2. Below are three ciphertext messages. The plaintext message is the same for each cipher. One was encrypted with a Caesar cipher. One was encrypted with a simple substitution cipher that is not a Caesar cipher. One was encrypted with a cipher that is not a simple substitution cipher and is designed to destroy the value of frequency analysis by making frequencies relatively even. Do a frequency analysis of each message and determine which is which.

Ciphertext number one

```
jrfne hyhwf stsrl sesan zyeql mhjee qheeq lwlee
lrszm whypa hplqh blnlr szyhw tetls eqzap qtyhj
rfnez prhxe qlfcl hrkts patsl seqlj rfneh yhwfs
ezisl rblse qltrh jetzy shykt ktzsf yjrhs tlshy
ktyml rseql trtkl yetef mrzxe qlsle rhtes tyzrk
tyhrf xzyzh wnqhi letjs aiset eaetz yseql jrfne
hyhwf seseh svtsm htrwf stxnw liljh asllh jqwle
elrsj hxzam whplk tmmlr smrzx lblrf zeqlr wleel
rshyk eqljh xzamw hplrl xhtys eqlsh xleqr zapqz
aeeql jrfne zprhx
```

Ciphertext number two

```
kzgxb ivitg aqazm abacx wvbpm nikbb pibbp mtmbb
mzawn tivoc iompi dmxmz awvit qbqma bpwco pqvik
zgxbw oziub pmgem izlqa ocqam abpmk zgxbi vitga
bwjam zdmab pmqzi kbqwv aivlq lqwag vkzia qmaiv
lqvnm zabpm qzqlm vbqbg nzwub pmamb ziqba qvwzl
qvizg uwvwi txpij mbqka cjabq bcbqw vabpm kzgxb
ivitg ababi asqan iqztg aquxt mjmki cammi kptmb
bmzak iuwcn tioml qnnmz anzwu mdmzg wbpmz tmbbm
zaivl bpmki uwcnt iomzm uiqva bpmai umbpz wcopw
cbbpm kzgxb woziu
```

Ciphertext number three

```
dwgdu nmyrc zzhri xexlz fkuou ndtro ehame yrndv
mejno mszii pebqp idnjx ttvpk benvb mydul kzeyx
nhqae psxwv pqrpb uykyg vsedn tluew sdxhb azjew
afqsi wgjjd ybjyq kwzmi phuwf xoxca vpizb jwecb
sgkot errmk jwqsd genkh yidmy kvjco njjai rolbf
qztkv gvueo andlk flvcm vgahj gtros ujihi hcdgd
rkbso awjrv prplc uoava avdou ftqeo jqbnp pgjjz
ujzhb txtlo ervei lwwcm oosvc nwbba wfaxo flajs
gqwym uzrcu ntctm nfcxi trxnw ysjjk dbuoh wxxfj
raaab wxhzv wtizv
```

3. Consider the frequencies of the ciphertext letters in each of the following messages that have been encrypted using the same method and key. Then collect the frequencies. What can be said about the method? What can be said about the key? Using the frequency information obtained by collecting the messages, try to cryptanalyze the messages.

```
IURUT  KRY

ZNUXU  HXKJY

KGMRK  Y

XGIKX  Y

TUXYK

CORJI  GZY

IGXJO  TGRY

NORRZ  UVVKX  Y
```

4. Consider the frequencies of the ciphertext letters in each of the following messages that have been encrypted using the same method and key. Then collect the frequencies. What can be said about the method? What can be said about the key? Using the frequency information obtained by collecting the messages, try to cryptanalyze the messages.

```
XZNHV

PZSEE

LSGIY  H

ULMAK  U

LSILN

UKLAX  LN

QILYU

XYIRL  NA

HLEGZ  NALE
```

5. Of course, plaintext frequencies depend on language. Different languages have different patterns.

Here are frequencies for several languages taken from the American Cryptogram Association's *The ACA and You*.

English	etaonirshldcupfmwybgvkqxjz
German	enirsadtugholbmcwfkvzpjqxy
Latin	ieutamsnrodlvcpqbfgxhjkwyz
Spanish	eaosrnidlctumpgybqvhfz

If the cryptanalyst does not know what the plaintext language is, that must be quickly determined. Sometimes frequency analysis can aid in that determination.

Here is some frequency information taken from Classical Cryptography Course by Randy Nichols (LANAKI)

http://www.threaded.com/cryptography5.htm

```
%        12 10  8 8 7 7 7 6 5  4-3  2    1    <1
English  e/ t   a/o n i s r h /ldcu/pfmw/ybgv/kqxjz

          18 11 8  7   5     4    3    2    <1
German   e/ n/ i/ rs/ adtu/ gho/ lbm/ cw/ REMAINGING

         10 9  7    6    4    3    <2
Latin    i/ e/ uta/ srn/ om/ cpl/ REMAINING

          13  9  8  7    5   4    3   1    <1
Spanish  ea/ o/ s/ rni/ dl/ ctu/ mp/ gyb/ REMAINING
```

Notice that, in German, e is very much more frequent than other letters (typically in cryptography ä = ae, ö = oe, = ue), and n is also very frequent. In Spanish, a is as frequent as e, t is not as frequent as in English, and `osrn` are very close in frequency. In Latin, letter frequencies are somewhat even.

The following three ciphertext messages were encrypted with Caesar ciphers with different keys. One message is in English, one is in German, and one is in Spanish. Try to use frequency analysis to determine the plaintext languages.

Ciphertext number one

```
dpyrl oyluu bugby ylpul uthao lthap rgbyb ljrbu
kdluk lubuz klyhs nltlp uluao lvypl klymb urapv
ulurv twsle lycly hlukl yspjo lygbk pldlp alyll
uadpj rsbun bukmv lykly bunkp lzlzr lyuza bljrl
zbuzl ylolb apnli ylpul uthao lthap rclyk hurlu
dpypu lyzal yspup lgdlp klbaz jolun lsloy aluyp
lthuu bukdl plyza yhzz
```

Ciphertext number two

```
xassj xliqs wxfew mgviw ypxwm rryqf ivxli svcev
ijivq exwer hiypi vwxli sviqw svmkm reppc ehqmv
ihjsv xlimv xlisv ixmge pzepy ixlic leziq svivi
girxp ctvsz ihxsl ezimq tsvxe rxgvc txskv etlmg
ettpm gexms rw
```

Ciphertext number three

```
qnawj wlxac nbnaj dwljy rcjwn byjwx uzdno dnmnb
mnnby jwjjl dkjnw kdblj mnarz dnijb hjenw cdajn
wldkj anlrk rxwxc rlrjb mndwy jrbvd harlx mnbld
krnac xyxas djwmn parsj uejxc axjen wcdan axnby
jwxu
```

Run the alphabet on each and see whether you were correct.