Day 11 – Breaking Monoalphabetic Substitution

Last time - breaking rectangular transposition

The steps for breaking rectangular transposition:

1. Guess a length for the decrypting permutation, says $k$.
2. Arrange the ciphertext into $k$ columns and let $N$ be the height (i.e. number of rows) of the resulting rectangle.
3. For each pair $1 \leq i \neq j \leq k$, extract the columns $i$ and $j$ and count the number of occurrence of the pair of letters $\alpha\beta$ and call this $n^{(ij)}_{\alpha\beta}$.
4. For each pair $\alpha\beta$, let $p_{\alpha\beta}$ be the probability of the pair $\alpha\beta$ in the English language (obtain from the table of frequency for letter pairs). Compute

$$C_{ij} = \sum_{\alpha\beta} p_{\alpha\beta} \log n^{(ij)}_{\alpha\beta}.$$ 

So if we guessed the correct period then the matrix $[C_{ij}]_{1 \leq i \neq j \leq k}$ will have a substantially bigger number in each row, except one.

- If $C_{ij}$ is the substantially big number on row $i$ then $j$ follows $i$ in the decryption permutation.
- If row $k$ is the only row with no substantially big entry, then $k$ is the first entry in the decryption permutation.

If you hit "Break" on the page w/ the matrix...
If you hit "Break" on the page with the matrix
you will obtain another permutation.

This new perm. is the encrypting perm.
which should be the inverse of decrypting perm.

The following message was encrypted using rectangular transposition.
Decrypt it.

eone  edgo  yaoar  taalf  ciqal  sadga  biowy  enahki
snoor  rgeue  virnoo  ciht  whlbh  tobah  send  stpph
itaces  rfeeci  egc  xcnat  vrcoo  ahxwn  acfeo  rsepop
ecehe  rrrv  yeche  oagye  pceuo  vrildi  lweyg  brhke
rihrei  acef  wraog  hyula  ownee  terr  achrk  shano
cenjer  hlate  thsobl  sibehi  hrgpe  cehir  woggg  rofoo
skont  rfrw  ogld  tclmea  hisee  sitah  thlwo  gwqhr
slst  doocox  tmnhht  hache  rhleu  tgsr  tetia  etera
htisu  abct  ungen  euago  nrow  bhcnoe  eynie  oyshta
ogbbo  wcnt  gnocc  gpywr  zehay  rrvyog  eehuo  tagvey
cemnce  ridi  lnrhay  rhtehi  whria  eefin  roagh  yatao
wicet  tuvoo  chhrks  hronae  wegerb  katrl  subhi  iheih
tgpeac  elirc  cigg  snfoci  dmwwe  chewr  oewne  rashii
eswet  runnce  oand  hltun  enife  dlmwwe  ywlnt  bkalil
crhh  tehwe  accon  ynet  owoten  bibog  weeho  ektog
owanig  znacn  yhecew  yceuo  grooe  acwht  droob  yabel
reilk  threg  tlcman  swygg  oafhr  ewoew  hrtt  ekhto
srnuo  hryeb  thas  ovieck  edfr  bhegt  erhia  wqne

tofraz  way

To obtain the decrypting perm. from the matrix $[C_{ij}]_{1 \leq i,j \leq k}$.

Suppose $\pi = (a_1, a_2, ..., a_k)$ is the decrypting perm.

- $a_k$ is given by the row without any big entry.

- If $a_k = x$, now we look at column x to find the row of the big entry on this row. This location then gives $a_{k-1}$.

- In general, if $a_i = y$ and there is a big entry on row $z$ of column $y$, then $a_{i-1} = z$.

#Key = 26! $\approx 2^{88} \rightarrow \text{large!}$. 

Math 187 notes Page 2
Breaking Monoalphabetic substitution

In probability, any function of the outcome of our experiment can be referred to as a random variable.

If \( X \) is a random variable and \( \Omega = \{ x_1, x_2, \ldots, x_n \} \) is the set of all possible outcomes of \( X \), then the expectation/expected value of \( X \) is given by

\[
E(X) = x_1P(X = x_1) + \cdots + x_nP(X = x_n)
\]

**Example.** Find the expected number of boys in a family with two children

\[
\text{Expected } \# \text{ of boys} = (0) \left( \frac{1}{4} \right) + (1) \left( \frac{2}{4} \right) + (2) \left( \frac{1}{4} \right) = 1
\]

**Example.** Suppose we roll two fair 6-sided dice. Find the expected sum of the two faces.

\[
E = (2) \left( \frac{1}{36} \right) + (3) \left( \frac{2}{36} \right) + \cdots
\]
Example. A supposedly fair die rolled 1000 times produced the following result:

<table>
<thead>
<tr>
<th>Result</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>171</td>
<td>186</td>
<td>174</td>
<td>170</td>
<td>192</td>
<td>197</td>
</tr>
</tbody>
</table>

Do you believe this?

The Chi-square statistic shows the discrepancies observed frequencies are from their theoretical values.

- Compute the Chi-square statistic using the following formula:

\[
\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i \cdot p_i)^2}{n_i \cdot p_i}
\]

where
- \( k \) is the total number of entries. Here, \( k = 6 \)
- \( n_i \) is the observed frequency of the \( i^{th} \) entry.
- \( p_i \) is the theoretical probability of the \( i^{th} \) entry.
- \( n \) is the total number of observations.

- Compute the statistic above with the ones from the Chi-Square table to obtain the probability that the observed values differ from theoretical ones:

\[
\chi^2 = \sum_{i=1}^{6} \frac{(n_i - n_i \cdot \frac{1}{6})^2}{n_i \cdot \frac{1}{6}}
\]

\[
= \frac{(171 - 1000 \times \frac{1}{6})^2}{1000 \times \frac{1}{6}} + \frac{(186 - 1000 \times \frac{1}{6})^2}{1000 \times \frac{1}{6}} + \ldots = 27.95
\]
The probability that a fair die would produce a \( \chi^2 > 27.95 \) is less than 0.001%.
Known plaintext attack of monoalphabetic substitution.

Given the letter frequencies of a certain ciphertext as follows

<table>
<thead>
<tr>
<th>cipher-text</th>
<th>l</th>
<th>h</th>
<th>a</th>
<th>w</th>
<th>d</th>
<th>q</th>
<th>o</th>
<th>n</th>
<th>f</th>
<th>s</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>80</td>
<td>61</td>
<td>55</td>
<td>46</td>
<td>44</td>
<td>40</td>
<td>39</td>
<td>35</td>
<td>33</td>
<td>26</td>
<td>22</td>
</tr>
</tbody>
</table>

We know that the word “WHERE” was in the plain-text. We find in the ciphertext the two strings “HDFKP” and “PDLHL” that match the pattern of “WHERE”. Using the chi-square test, decide which of these two strings is the image of “WHERE”.

\[
\chi^2 = \sum_{i=1}^{26} \frac{(n_i - n_p_i)^2}{n_p_i}
\]

\[
P_W = P\left(W \mid \text{either } W \text{ or } H \text{ or } E \text{ or } R\right)
\]

\[
= \frac{0.02469}{0.02469 + 0.0625 + 0.1225 + 0.04065} = 0.0923
\]

\[
P_H = ... = 0.226
\]

\[
P_E = ... = 0.455
\]

\[
P_R = ... = 0.227
\]
<table>
<thead>
<tr>
<th>W</th>
<th>H</th>
<th>E</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_i$: 0.0923</td>
<td>0.226</td>
<td>0.455</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Solution (cont.) Test each candidate and compute each $\chi^2$ statistic.
The following message was encrypted using monoalphabetic substitution:

```
zitiq jajfe hoxax wawoz zrjogh eolfit soqga xwozo
zzjeg heofj tosje ioza ohsit fkgoj zibek tzzts
agjwh tsfjk qereu nhjng yAleC sqkoq gystt heoef
ikzst biktz zlazt ilzgs luhtt sqkkn jzett sttro
ypoez kqjgw sglzz yqzat qhjqs rizrta xozzz eqheo
fita zlthx juthq yeqhs gqzts olqgi dtztt soost
fkqht rwnoo fgsyq yqzil lmgys lbgjy klztt kttzst
stjnu tzuwts sfqsc tranjn bsgqy octrq gystt hzarrj
wqkav ioklz iktz szdj qeogk metax wawoz ztzeh
ngbts njegk zqjtg aoatz eqpqg wstql arqhr sqara
wawoz zrjogh eolfit soqga xwozo qzzt stztt syqtd
xthefz tznzi klzzt stzxx eqzak nxjfi gatzq Jadghk
izzts othba koomz gztj gatzq jgjbr afits zll znk
izzts vokkf sqqwe fswzf knzt tfits yaqzq egqur kkguz
iktz zltsz yuqts fkgzt rwnqg ngzzi stvzo yqzst
hzoq sqqzt oaziz tuteq hgzkg hstsp xazqg lztz
jgaze gjghg kltz saohe tzitk ilzzts egrhx gyooa
afstq ryqts atets qekieq sqqzt saqag ydkg kvgjg xjgs
qhrjg sfjqa aozkt qkzst hqozc toygq jyqik tzzts
xstf azkzo huwof tsqoz hzeqy kts stiz metax stswt
qogok egqgy eghec wrazaq eqzoq geofq itsae qhwer
luntr ypoez kqjgz bhekx uiqzy egqgy egtha oouou
ikhtr xzzoz ztyzr xtqy egqgy egfjz saqfd ywrqo
egzj htefz qegjy frzy tilts joqls yqygk klztt
zstq eizfjg ohszt zkizz lsggl uotyjg
```

Decrypt it, knowing that the plaintext contains the following words:

```
HOMOPHONIC SUBSTITUTION CHARACTERS LETTER
```