187 - Day 11

Friday, April 27, 2018 9:29 AM

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 $n_{\alpha\beta}^{(ij)}$.
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

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

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.

Col i	Col j
A	O
C	H
R	T
S	A
M	A
C	H

Compute $n_{\alpha\beta}^{(i,j)}$ where $\alpha\beta$ are all the pairs that occur in these cols

$n_{AO}^{(i,j)} = 1 = n_{RT}^{(i,j)} = n_{SA}^{(i,j)} = n_{MA}^{(i,j)}$

$n_{CH}^{(i,j)} = 2$

$C_{ij} = \sum_{\alpha\beta} p_{\alpha\beta} \cdot \log\left(n_{\alpha\beta}^{(ij)}\right) = p_{CH} \cdot \log(2)$

form a matrix

obtained in the table last time

count pairs that occur more than once

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

If you hit "Break" on the page w/ 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.

noeas	*cadpr*	*yaatr*	*atadf*	*eiynl*	*sadgo*	*liowy*	*enahi*
snoer	*ruaye*	*vinoo*	*citee*	*nhltb*	*todui*	*senot*	*stpph*
iates	*nfeei*	*ieyut*	*xonvt*	*nreoe*	*ahwwn*	*ocfea*	*rosyw*
eeehv	*rrrvo*	*yeehu*	*otagw*	*yeeuo*	*vridi*	*llowg*	*brthe*
iwhri	*aeeft*	*nroig*	*hyuta*	*owiee*	*ttvro*	*aehrk*	*shoma*
ewyer	*bkatr*	*lswbl*	*iihei*	*htrga*	*eetir*	*woigy*	*rnfoo*
ikout	*rfrna*	*ogltd*	*tleam*	*htsee*	*iitah*	*thtwo*	*guouh*
sllst	*daooe*	*iwmht*	*haehe*	*rhlue*	*tagsr*	*tetia*	*etera*
htisu	*acbtt*	*ngent*	*euayo*	*nroaw*	*htnoc*	*eyuie*	*osyta*
oybbo	*uemtv*	*gneoc*	*ggiwr*	*zehay*	*rrvoy*	*eehuo*	*tagwy*
eeuov	*ridil*	*lowgb*	*rthei*	*whria*	*eeftn*	*roigh*	*yutao*
wieet	*tvroa*	*ehrks*	*homae*	*wyerb*	*katrl*	*swbli*	*iheih*
trgae	*etirw*	*oigyr*	*nfooi*	*dnuww*	*ehewr*	*oucnv*	*rashi*
esivt	*nrmec*	*oanbt*	*htuii*	*enife*	*dhmio*	*ywttu*	*lkail*
cehah	*tehne*	*accou*	*nyest*	*oacty*	*bibay*	*uevoh*	*egtog*
omrig	*zancr*	*yhvew*	*eueyo*	*groew*	*aevht*	*droio*	*yubwl*
reilh	*ihreg*	*ttean*	*iwiyg*	*ouofr*	*ewaev*	*htart*	*ekito*
srmwo	*hryeb*	*thasr*	*awiek*	*eilir*	*lbegt*	*erhhi*	*wigna*
tofrx	*uoy*						

If you're using the applet, then you have to reverse the order of arrows to obtain the decrypting perm.

To obtain the decrypting perm. from the matrix $[C_{ij}]_{1 \le i,j \le k}$, we obtain the entries from Right to left.

- Suppose $\pi = (a_1, a_2, \ldots, a_k)$ 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 $a_{i-1} = z$

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

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

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 $\Omega = \{x_1, x_2, \ldots, x_k\}$ is the set of all possible outcomes of X, then the **expectation/expected value of** X is given by

$$\mathbb{E}(X) = x_1\mathbb{P}(X = x_1) + \cdots + x_k\mathbb{P}(X = x_k)$$

Ave. payoff in a game of chance

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

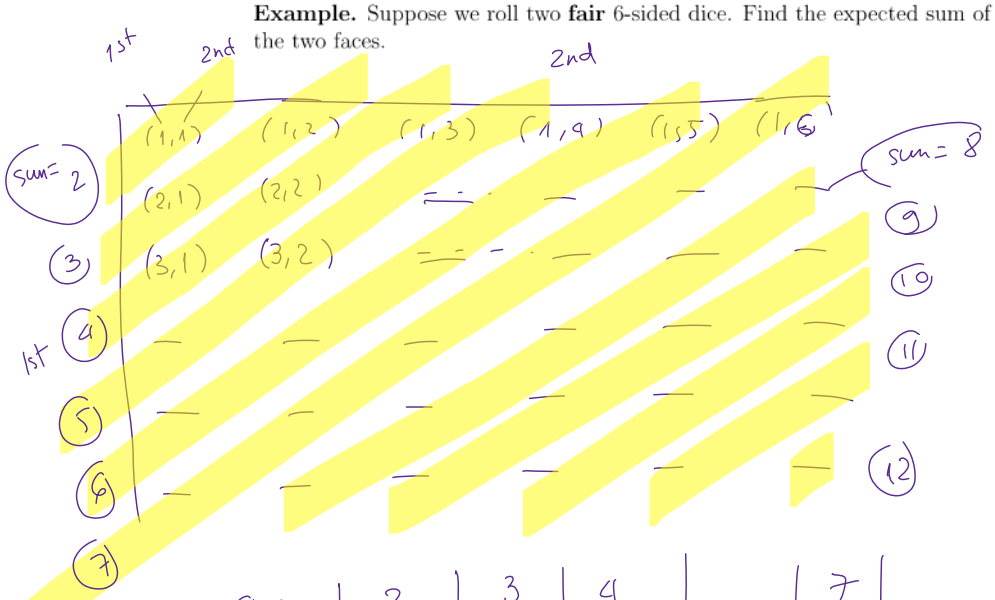
$\{GG, GB, BG, BB\}$

Expected # of boys $= (0)\left(\frac{1}{4}\right) + (1)\left(\frac{2}{4}\right) + (2)\left(\frac{1}{4}\right) = 1.$

$= \mathbb{P}(\text{win}) \cdot (\text{payoff for winning}) + \mathbb{P}(\text{lose}) \cdot (\text{payoff for losing})$

$= (3)\mathbb{P}(\text{win}) + (-1)\mathbb{P}(\text{lose})$ for Keno

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



sum	2	3	4	...	7	...
prob	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$		$\frac{6}{36}$	

$\mathbb{E} = (2)\left(\frac{1}{36}\right) + (3)\left(\frac{2}{36}\right) + \cdots$

(6-sided)

Example. A supposedly fair die rolled 1000 times produced the following result

Result	1	2	3	4	5	6
Frequency	171	186	174	170	192	107

n_1 n_2 n_3 ...

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

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

where

  - k is the total number of entries. Here, $k = 6$
  - n_i is the observed frequency of the i^{th} entry. n_i obtained from the result
  - p_i is the theoretical probability of the i^{th} entry. $\rightarrow p_i = \frac{1}{6}$ for all i
  - n is the total number of observations $\leftarrow n = \sum n_i = 1000$

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

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$$

$$= \frac{(171 - 1000 \times \frac{1}{6})^2}{1000 \times \frac{1}{6}} + \frac{(186 - 1000 \times \frac{1}{6})^2}{1000 \times \frac{1}{6}} + \dots = 27.95$$

$\chi^2 = 27.95$

$k = 6 \Rightarrow$ degree of freedom $= k - 1 = 5$

$\downarrow df \setminus p \longrightarrow$	10%	1%	0.01%	0.001%
2	4.60	9.21	13.81	18.42
3	6.25	11.34	16.26	21.10
4	7.77	13.27	18.46	23.51
5	9.23	15.08	20.51	25.74
6	10.64	16.81	22.45	27.85
7	12.01	18.47	24.32	29.87
8	13.36	20.09	26.12	31.82

27.95

deg. of freedom →

The probability that a pair 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

cipher-text	l	h	a	w	d	q	o	n	f	s	z
frequency	80	61	55	46	44	40	39	35	33	26	22

k	p	i	t	v	y	r	x	u	m	c	g	j	b	e
26	22	18	17	12	11	9	9	8	7	5	3	1	0	0

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

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i}$$

	W	H	E	R
prob.	0.02469	0.06025	0.1215	0.06063

these are NOT p_i!

$P_W = \mathbb{P}(W \mid \text{either W or H or E or R})$

$$= \frac{0.02469}{0.02469 + 0.06025 + 0.1215 + 0.06063} = 0.0923$$

$P_H = \ldots = 0.226$

$P_E = \ldots = 0.455$

$P_R = \ldots = 0.227.$

	W	H	E	R.
p_i	0.0923	0.226	0.455	0.227

Solution (cont.) Test each candidate and compute each χ^2 - statistic.

The following message was encrypted using monoalphabetic substitution:

zitig	*jgfig*	*hoeax*	*wazoz*	*xzogh*	*eofit*	*soaqa*	*xwazo*
zxzog	*heofi*	*tsohv*	*ioeia*	*ohukt*	*fkqoh*	*ztbzk*	*tzzts*
aeqhw	*tstfk*	*qetrw*	*nqhng*	*yatct*	*sqkro*	*yytst*	*hzeof*
itszt	*bzktz*	*ztsaz*	*itnqs*	*tutht*	*sqkkn*	*jxeij*	*gstro*
yyoex	*kzzgw*	*stqlz*	*iqhaz*	*qhrqs*	*raxwa*	*zozxz*	*ogheo*
fitsa	*zithx*	*jwtsg*	*yeiqs*	*qezts*	*atqei*	*ktzzt*	*soast*
fkqet	*rwnoa*	*fqszg*	*yzitl*	*tnygs*	*tbqjf*	*ktzit*	*ktzzt*
stjou	*izwts*	*tfkqe*	*trwnq*	*hngyy*	*octro*	*yytst*	*hzanj*
wgkav	*ioktz*	*itktz*	*ztsdj*	*qnghk*	*nwtax*	*wazoz*	*xztrw*
nghta	*njwgk*	*zittq*	*aotaz*	*vqnzg*	*wstql*	*azqhr*	*qsrax*
wazoz	*xzogh*	*eofit*	*saoaz*	*gkggl*	*qzzit*	*ktzzt*	*systd*
xtheo	*tazit*	*ktzzt*	*stoax*	*axqkk*	*nzitj*	*gazeg*	*jjghk*
tzzts	*ohthu*	*koaia*	*gzitj*	*gazeg*	*jjghe*	*ofits*	*ztbzk*
tzzts	*vokkf*	*sgwqw*	*knwtt*	*gsfts*	*iqfaz*	*oyvtq*	*kkgvz*
itktz	*ztstz*	*gwtst*	*fkqet*	*rwnqh*	*ngyzi*	*sttro*	*yytst*
hzeiq	*sqezt*	*sazit*	*hvteq*	*hhgkg*	*hutsp*	*xazzq*	*ltzit*
jgaze	*gjjgh*	*ktzzt*	*saohe*	*tzitk*	*tzzts*	*egxhz*	*gytoa*
afstq	*rgcts*	*atcts*	*qkeiq*	*sqezt*	*saqav*	*tqkkg*	*vjgst*
qhrjg	*stfga*	*aowkt*	*qkzts*	*hqzoc*	*taygs*	*tqeik*	*tzzts*
zitst	*axkzo*	*hueof*	*itseq*	*hwteg*	*jtcts*	*natex*	*stwst*
qlohu	*igjgf*	*ighoe*	*axwaz*	*ozxzo*	*gheof*	*itsae*	*qhwtc*
tsnro	*yyoex*	*kzoyz*	*ithxj*	*wtsgy*	*igjgf*	*ighta*	*oaiou*
iohqr	*rozog*	*hzgyo*	*hrohu*	*vioei*	*ktzzt*	*sajqf*	*zgvio*
eigzi	*tsavt*	*qkagh*	*ttrzg*	*rtzts*	*johti*	*gvjqh*	*nktzz*
tsatq	*eifkq*	*ohztb*	*zktzz*	*tseqh*	*wtegjt*		

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

HOMOPHONIC *SUBSTITUTION* *CHARACTERS* *LETTER*