Day 12 – Breaking Monoalphabetic Substitution

Known plaintext attack of monoalphabetic substitution.

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.
- 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

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**".

1

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

Step 1. Compute the theoretical probability of the letters W, H, E, R :

Letter	Relative frequency	Letter	Relative frequency
A	0.08399	N	0.06778
B	0.01442	O	0.07493
C	0.02527	P	0.01991
D	0.04800	Q	0.00077
E	0.12150	R	0.06063
F	0.02132	S	0.06319
G	0.02323	T	0.08999
H	0.06025	U	0.02783
I	0.06485	V	0.00996
J	0.00102	W	0.02464
K	0.00689	X	0.00204
L	0.04008	Y	0.02157
M	0.02566	Z	0.00025

These are NOT the theoretical p_i's!

$P_1 = P(W \mid \text{either } W \text{ or } H \text{ or } E \text{ or } R)$

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

$$= 0.0923$$

Similarly, $P_2 = P(H \mid W, H, E, R) = 0.226$

$P_3 = P(E \mid W, H, E, R) = 0.455$

$P_4 = P(R \mid W, H, E, R) = 0.227$

	W = H	H = D	E = F	R = K
p_i	0.0923	0.226	0.455	0.227
n_i	61	44	33	26

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$$

n_i ↓ freq. of each H, D, F, K in ciphertext

$n = \sum n_i = 164$

Step 2. Compute the χ^2 statistic for each candidate: "HDFKF"

$$\chi^2 = \frac{(61 - 164 \cdot 0.0923)^2}{164 \cdot 0.0923} + \frac{(44 - 164 \cdot 0.226)^2}{164 \cdot 0.226} + \frac{(33 - 164 \cdot 0.455)^2}{164 \cdot 0.455} + \frac{(26 - 164 \cdot 0.227)^2}{164 \cdot 0.227}$$

≈ 181.88

"PDLHL"

	W = P	H = D	E = L	R = H
p_i	0.0923	0.226	0.455	0.227
n_i	22	44	80	61

$n = 207$

$$\chi^2 = \frac{(22 - 207 \cdot 0.0923)^2}{207 \cdot 0.0923} + \frac{(44 - 207 \cdot 0.226)^2}{207 \cdot 0.226} + \dots \approx 6.59$$

Take the smaller χ^2 value to be the image of WHERE

(W → P ; H → D ; E → L ; R → H)

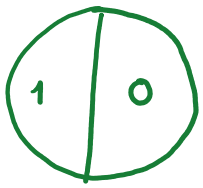
The following message was encrypted using monoalphabetic substitution:

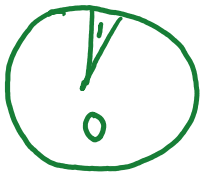
zitig	*jgfig*	*hoeax*	*wazoz*	*xzogh*	*eofit*	*soaqa*	*xwazo*
zxzog	*heofi*	*tsohv*	*ioeia*	*ohukt*	*fkqoh*	*ztbzk*	*tzzts*
aeqhw	*tstfk*	*qetrw*	*nqhng*	*yatct*	*sqkro*	*yytst*	*hzeof*
itszt	*bzktz*	*ztsaz*	*itnqs*	*tutht*	*sqkkn*	*jxeij*	*gstro*
yyoex	*kzzgw*	*stqlz*	*iqhaz*	*qhrqs*	*raxwa*	*zozxz*	*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*

put each plaintext letter into the applet and have it search for all possible ciphertext words. Then pick the one with the smallest χ^2 value that does not cause any conflict with the previous choices.





uncertainty

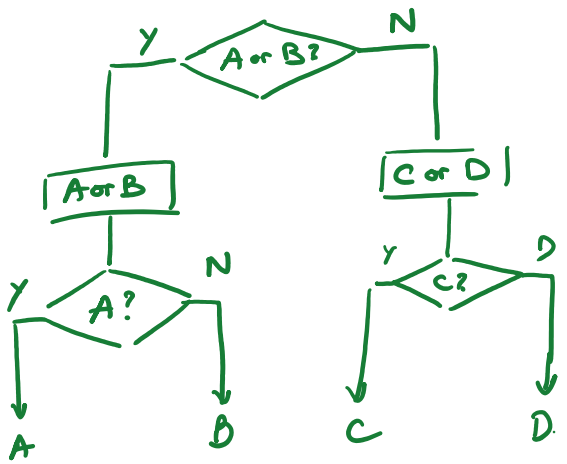
Elements of Information Theory

Consider two alphabets with letter probability

Letters	A	B	C	D
Alphabet 1 frequency	1/4	1/4	1/4	1/4
Alphabet 2 frequency	1/2	1/4	1/8	1/8

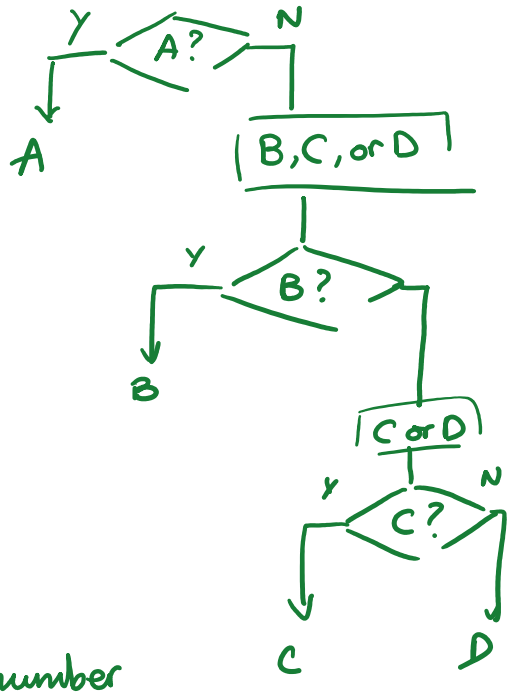
How many **yes/no questions** do we need to determine a letter in each alphabet?

Alphabet 1



2 Y/N questions to find 1 letter in this alphabet

Alphabet 2



The expected/average number of questions asked:

$$(1)\left(\frac{1}{2}\right) + (2)\left(\frac{1}{4}\right) + (3)\left(\frac{1}{4}\right) = 1.75$$

(A) (B) (C or D)

less uncertainty

Yes = 1 / No = 0

We can represent the outcome of a Yes/No question by a **single binary digit, call a bit**. Knowing the answer to a Yes/No question gains us **one bit of information.**

In general, if the experiment has N possible **equally likely** outcomes, then we need

$$\log_2(N)$$

bits of information to store a result of the experiment

* Roll a fair die: 6 equally likely outcomes 1 → 6

$\log_2(6) \approx 2.\ldots$ → round up to 3

one way to represent: 1 = 000 ; 3 = 010 ; 5 = 100
2 = 001 ; 4 = 011 ; 6 = 101

* 26 letters in the alphabet : $\log_2(26)$ → 5 bits

* ASCII: 128 symbols $\log_2(128) = 7$ bits

8 bits = 1 byte.

If the outcomes are **not equally likely** then to store the outcome $Z = a$ requires $\log_2\left(\frac{1}{p}\right)$ bits of information where $p = \mathbb{P}[Z = a]$.

So the **expected/average number of bits of information** required to store one spin of Z is

$$\sum_a p \cdot \log_2\left(\frac{1}{p}\right) = \text{entropy}.$$

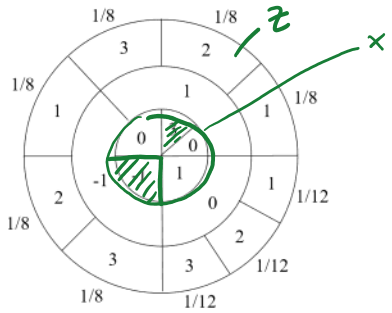
all pos. outcomes of Z

Definition 1. *The **entropy** of an event A is a measure of the uncertainty we feel about the occurrence of A.*

The entropy of a random variable X is given by

$$H(X) = \sum_a \mathbb{P}(X = a) \cdot \log_2 \left(\frac{1}{\mathbb{P}(X = a)} \right)$$

Example. Suppose that random variables X, Y, Z are obtained by spinning the wheel below, with X given by the innermost circle, Y given by the intermediate circle, and Z given by the outermost circle.



(a) Calculate $H(X)$.

a	-1	0	1	3
$\mathbb{P}(X=a)$	$1/4$	$3/8$	$1/4$	$1/8$

$$H(X) = \left(\frac{1}{4}\right) \cdot \log_2\left(\frac{1}{1/4}\right) + \left(\frac{3}{8}\right) \log_2\left(\frac{1}{3/8}\right) + \left(\frac{1}{4}\right) \log_2\left(\frac{1}{1/4}\right) + \left(\frac{1}{8}\right) \log_2\left(\frac{1}{1/8}\right) \approx 1.906.$$

(b) How many bits (of information) are required to store the results of 100,000 spins of Z?

(# of bits required to store the result of 1 spin) × 100,000

↓

entropy of Z : $H(Z)$

a	1	2	3
$\mathbb{P}(Z=a)$	$1/3$	$1/3$	$1/3$

$$H(Z) = \sum_a \mathbb{P}(Z=a) \cdot \log_2\left(\frac{1}{\mathbb{P}(Z=a)}\right)$$

or you can see that the outcomes of Z are equally likely.

$$H(Z) = \log_2(\#\text{outcomes}) = \log_2(3) \approx 1.585$$

So for 100,000 spins → 158,500.

Entropy = uncertainty about the random variable.
= amount of information (bits) that we gain by observing 1 result of the random variable.