187 - Day 08

Wednesday, April 18, 2018 8:30 AM

Day 08 – Breaking Vigenère cipher

Last time

Definition 1 (Permutation). *If an ordered list or **permutation** of k objects is to be formed by selecting from a collection of n objects (where $k \leq n$) then there are*

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+2) \cdot (n-k+1)$$

$(1,2,3) \neq (3,2,1)$

ways to do form this list. We denote this value by

$$P(n,k) := n \cdot (n-1) \cdot (n-2) \cdots (n-k+2) \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Definition 2 (Combination). *The number of unordered selections of **combinations** of n objects selected k at a time is given by*

$$C(n,k) = \frac{P(n,k)}{k!} = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

$(1,2,3) = (3,2,1)$

"double count" $k!$ times

This is called binomial coefficient, *read "n choose k."*

Definition 3 (Probability). *For an experiment where there are n different* equally likely *possible outcomes, then the **probability** of a result that can occur in k possible ways is given by $\frac{k}{n}$.*

Theorem 1 (Birthday problem). *Given that there are 365 days in a year (ignore leap year). If there are n people in a room, then the probability that **at least two have the same birthday** is given by*

$$1 - \frac{P(365,n)}{365^n} = 1 - \frac{365 \cdot 364 \cdots (365-n+1)}{365^n}$$

n	1	2	3	10	20	30	40	50
p	0	0.0027	0.0082	0.117	0.411	0.706	0.891	0.97

almost 1

hash collision.

Index of Coincidence

Example. Suppose there are 45 cards in a deck. Of those cards, 20 are labeled with "X", 15 are labeled with "Y", and 10 "Z". Suppose we pick a card at random, **put it back**, then shuffle the deck and pick another card at random. Find the probability that

a. the first letter is X and the second is Z

$$\mathbb{P}(X \text{ and then } Z) = \mathbb{P}(X) \cdot \mathbb{P}(Z) = \frac{20}{45} \cdot \frac{10}{45} = \frac{8}{81}$$

b. the two letters are X and Z $= (X \text{ and then } Z) \text{ or } (Z \text{ then } X)$

$$\mathbb{P}(X \text{ and } Z) = \mathbb{P}(X \text{ and then } Z) + \mathbb{P}(Z \text{ and then } X)$$

$$= \frac{8}{81} + \frac{8}{81} = \frac{16}{81}$$

c. the first letter is Y and the second letter is also Y

$$\mathbb{P}(Y \text{ and then } Y) = \mathbb{P}(Y) \cdot \mathbb{P}(Y) = \frac{15}{45} \cdot \frac{15}{45} = \frac{1}{9}.$$

same

d. the two letters are Y

$$\mathbb{P}(2Y\text{'s}) = \frac{\#\text{ of ways I can get } 2\ Y\text{'s}}{\#\text{ of ways of choosing 2 cards}} = \cancel{\frac{\binom{15}{2}}{\binom{45}{2}}}$$

w/o replacement.

$$= \frac{15 \cdot 15}{45 \cdot 45} = \frac{1}{9}$$

The following table give the relative frequency of the English alphabet letters in a 7834-letter sample of English writing.

Letter	Relative frequency (%)	Letter	Relative frequency (%)
A	8.399	N	6.778
B	1.442	O	7.493
C	2.527	P	1.991
D	4.800	Q	0.077
E	12.150	R	6.063
F	2.132	S	6.319
G	2.323	T	8.999
H	6.025	U	2.783
I	6.485	V	0.996
J	0.102	W	2.464
K	0.689	X	0.204
L	4.008	Y	2.157
M	2.566	Z	0.025

frequency $= \left(\dfrac{\text{\# of occurance for the letter}}{7834} \right) \times 100\%$

p_A = prob. of selecting an A.

Use the table to find the probability

a. of selecting two A's

$$= p_A^2 = (0.08399)^2 = 0.00705 = 0.705\%$$

b. of selecting two B's

$$= p_B^2 = (0.01442)^2 = 0.00021 = 0.021\%$$

c. that two randomly selected letters in English are identical

$$\mathbb{P}(2 \text{ identical letters}) = \mathbb{P}(2A\text{'s}) + \mathbb{P}(2B\text{'s}) + \dots + \mathbb{P}(2Z\text{'s})$$

$$= p_A^2 + p_B^2 + \dots + p_Z^2$$

3

$$= 0.065 = 6.5\%$$

plaintext	X	Y	Z	A
ciphertext	A	B	C	D

distribution of ciphertext letters & plaintext letter are the same for Caesar shift.

Example. Suppose we use the **Caesar cipher** where $A \to D$. What is the probability that a letter selected at random in the ciphertext

a. of selecting an A in the ciphertext?

$$\mathbb{P}(A \text{ in ciphertext}) = \mathbb{P}(X \text{ in plaintext}) = p_X = \boxed{0.204\%}$$
$$= 0.00204$$

b. of selecting a B in the ciphertext?

$$\mathbb{P}(B \text{ in ciphertext}) = \mathbb{P}(Y \text{ in plaintext}) = p_Y = \boxed{2.157\%}$$

$$* \; p_X = \frac{\# \text{ of } X}{\text{entire length}}$$

	X	A B C ...	N
odd	A	D	
even		N O	A

Example. Suppose we use the **Vigeneère cipher** with keyword DN. What is the probability

a. of selecting an A in the ciphertext?

$$\mathbb{P}(A \text{ in ciphertext}) = \mathbb{P}(X \text{ in p/t at odd position}) + \mathbb{P}(N \text{ in p/t at even position})$$

$$= \frac{p_X}{2} + \frac{p_N}{2} = \frac{1}{2}(0.209 + 6.778) = \boxed{3.491\%}$$

closer

b. of selecting a B in the ciphertext?

$$\mathbb{P}(B \text{ in ciphertext}) = \mathbb{P}(Y \text{ in p/t at odd position}) + \mathbb{P}(O \text{ in p/t at even position}) = \frac{1}{2}(p_Y + p_O) = \boxed{4.825\%}$$

distribution of ciphertext letters gets "flatten out".

Remark: in a Vigenère cipher with sufficiently long keyword, the probabilities of seeing any letter in the ciphertext will converge to

$$\frac{1}{26} = 0.0385 = 3.85\%$$

Definition 4 (Index of Coincidence). *The **index of coincidence** (for a ciphertext), denoted I, is the probability that two randomly selected letters in the ciphertext are identical.*

Remark:

- If $I \approx 0.065$ then the cipher is more likely to be mono-alphabetic substitution.
- For poly-alphabetic substitution, $0.0385 \leq I \leq 0.065$

Theorem 2. *Let $n_0, n_1, n_2, \ldots n_{24}, n_{25}$ be the respective counts of the letters $A, B, C, \ldots, Y, Z$. Let $n = \sum n_i$ be the total number of letters in the text then*

$$I = \frac{1}{n(n-1)} \sum_{i=0}^{25} n_i(n_i - 1).$$

Now if an English plaintext is encrypted using a Vigenère cipher with keyword of length k, then

$$I \approx \frac{0.0385 \cdot n(k-1) + 0.065(n-k)}{k(n-1)}, \text{ or equivalently,}$$

$$k \approx \frac{0.0265n}{(0.065 - I) + n(I - 0.0385)}.$$

$A =$ letter$_0$
$B =$ letter$_1$
$\vdots$
$Z =$ letter$_{25}$

Note: This is called the **Friedman Test**. It only gives an estimate for the keyword length k.

$I = \mathbb{P}($2 randomly selected letters in the ciphertext are the same$) = \sum_{i=0}^{25} \mathbb{P}($2 randomly selected letters are letter$_i)$

$$= \sum_{i=0}^{25} \frac{\binom{n_i}{2}}{\binom{n}{2}} = \sum_{i=0}^{25} \frac{\frac{n_i(n_i-1)}{2 \cdot 1}}{\frac{n(n-1)}{2 \cdot 1}}$$

Example. Suppose a ciphertext is encrypted with a Vigenère cipher with keyword of length k. The total number of letters in the ciphertext is $n = 337$ and the frequency count of the ciphertext is given by the table below. Estimate k.

Letter	Count	Letter	Count
A	13	N	11
B	18	O	17
C	12	P	21
D	15	Q	9
E	26	R	16
F	4	S	7
G	15	T	8
H	9	U	7
I	16	V	8
J	8	W	14
K	9	X	8
L	18	Y	20
M	22	Z	6

In MatLab

$$[n_0\ n_1\ \cdots\ n_{25}]\begin{bmatrix} n_0 & -1 \\ n_1 & -1 \\ \vdots & \vdots \\ n_{25} & -1 \end{bmatrix} = \left[\sum n_i^2 \ -\sum n_i\right]$$

$$I = \frac{1}{n(n-1)}\sum_{i=0}^{25} n_i(n_i-1) = \frac{1}{337\cdot 336}\left[13\cdot 12 + 18\cdot 17 + 12\cdot 11 + \cdots + 6\cdot 5\right]$$

$$= 0.0428$$

$$k \approx \frac{0.0265 \times 337}{(0.065 - 0.0428) + 337 \times (0.0428 - 0.0385)}$$

$$\approx 6.20$$

Kasiski Test

The **Kasiski Test** is another way of estimating the length of the keyword for Vigenère cipher. It obtains possible keyword lengths from the **gcd of the spacing between repeated letter groups** in the ciphertext.

Example. Consider the ciphertext

$$\underbrace{\text{I V E V Y G A R M L M Y}}_{12}\ \underbrace{\text{I V E K F D}}_{6}\ \underbrace{\text{I V E F R L}}_{6}$$

$k \approx \gcd(\text{spaces}) = \gcd(12, 6) = 6$

correct ans. for the example.

I V E
$\updownarrow$
T H E

You try to decrypt this!