CONDITIONAL PROBABILITY OF SUBSTITUTION CIPHER USING CAESAR CIPHER

Gabriel Lee

A00884904

INTRODUCTION

We use the English language in our daily lives and often don't think too much about it. It's interesting to talk about facts regarding language and it wouldn't surprise most people that the most common letters in the English alphabet are: *e*, *t*, *a*, *o*, *i*, and *n*. Although, this might be considered as a fun fact for most, for cryptanalysis, plays a bigger role.

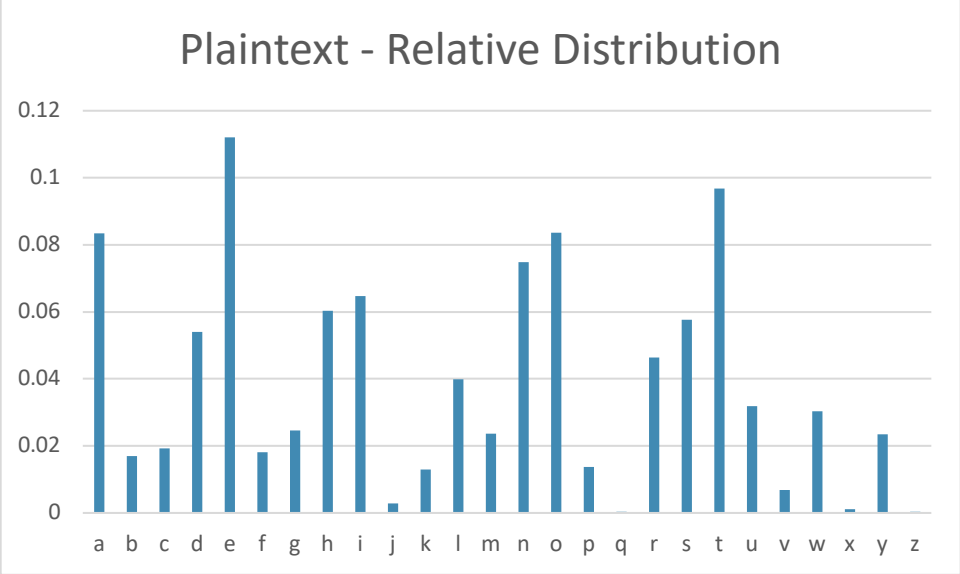
BACKGROUND

Possibly the simplest cipher that exists today is the *Caesar Cipher*. The *Caesar Cipher* is a simple substitution cipher. The encryption is done by shifting each character in the plaintext by a constant key. The most associated key is 3 for the *Caesar Cipher*. In which case, the plain text *banana* will become *ihuhuh*. The *Caesar Cipher* is not acceptable to use it by itself in our times, but it comes in handy studying substitution in cryptography.

TESTING

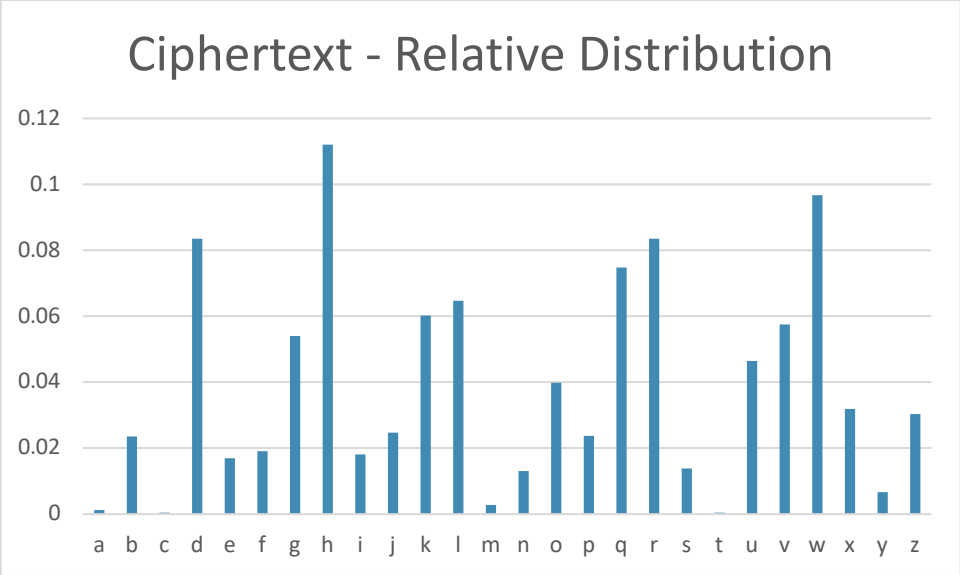
RELATIVE FREQUENCY DISTRIBUTION

Calculating the Relative Frequency Distribution (RFD) is quite simple. Looking at the *Adventures of Huckleberry Finn* by Mark Twain as an example, we can see the RFD of the English alphabet.



Looking at the graph above, we can conclude that the top 6 common letters are: *e*, *t*, *a*, *o*, *i*, and *n*.

If we encrypt, the *Adventures of Huckleberry Finn* by Mark Twain using the *Caesar Cipher* with the key of 3, we get the following graph.



With a quick look of the ciphertext RFD graph, we can quickly determine at the frequency shifted by 3. This effect is particularly easy for humans to identify because human minds are wired to identify shapes and patterns. In computing, however, that is not always the case.

Plaintext

letter	count	probability
a	36947	0.08345267
b	7509	0.01696068
c	8485	0.01916518
d	23906	0.05399679
e	49605	0.11204346
f	7992	0.01805163
g	10906	0.02463352
h	26660	0.06021729
i	28636	0.06468051
j	1238	0.00279629
k	5759	0.01300793
l	17637	0.03983692
m	10480	0.02367131
n	33119	0.07480632
o	37018	0.08361304
p	6111	0.013803
q	196	0.00044271
r	20554	0.04642559
s	25503	0.05760396
t	42825	0.09672938
u	14114	0.03187948

Ciphertext

letter	count	probability
a	528	0.0011926
b	10402	0.02349513
c	188	0.00042464
d	36947	0.08345267
e	7509	0.01696068
f	8485	0.01916518
g	23906	0.05399679
h	49605	0.11204346
i	7992	0.01805163
j	10906	0.02463352
k	26660	0.06021729
l	28636	0.06468051
m	1238	0.00279629
n	5759	0.01300793
o	17637	0.03983692
p	10480	0.02367131
q	33119	0.07480632
r	37018	0.08361304
s	6111	0.013803
t	196	0.00044271
u	20554	0.04642559

v	2993	0.00676033	**v**	25503	0.05760396
w	13419	0.03030967	**w**	42825	0.09672938
x	528	0.0011926	**x**	14114	0.03187948
y	10402	0.02349513	**y**	2993	0.00676033
z	188	0.00042464	**z**	13419	0.03030967

The human minds have its limitations to notice patterns, whereas computers can be near limitless as long as we can tell it how. Determining the top 6 most common letters in the table above will be time consuming and error prone if a human is to do it. However, using computing, getting the topmost characters and finding possible matches between the plain text and cipher text can be achieved in matter of seconds.

CONDITIONAL PROBABILITY

Human minds are great recognizing shape/patterns and is able to quickly determine the RFD of English alphabet especially when presented in a graph. However, computers are able to do it faster and at a much larger scale.

Using Conditional Probability, we can easily compute the probability and match the letter in the plaintext to the substituted letter in the ciphertext.

Letter: e

M	C	P(M=m)	P(C=c)	P(M=m|C=c)
e	a	0.11204346	0.0011926	93.9488636
e	b	0.11204346	0.02349513	4.76879446
e	c	0.11204346	0.00042464	263.856383
e	d	0.11204346	0.08345267	1.34259886
e	e	0.11204346	0.01696068	6.60607271
e	f	0.11204346	0.01916518	5.84619918
e	g	0.11204346	0.05399679	2.07500209
e	h	0.11204346	0.11204346	1
e	i	0.11204346	0.01805163	6.20683183
e	j	0.11204346	0.02463352	4.54841372
e	k	0.11204346	0.06021729	1.86065266
e	l	0.11204346	0.06468051	1.73226009
e	m	0.11204346	0.00279629	40.0686591
e	n	0.11204346	0.01300793	8.61347456
e	o	0.11204346	0.03983692	2.81255316
e	p	0.11204346	0.02367131	4.73330153
e	q	0.11204346	0.07480632	1.49778073
e	r	0.11204346	0.08361304	1.34002377
e	s	0.11204346	0.013803	8.11732941

e	t	0.11204346	0.00044271	253.086735
e	u	0.11204346	0.04642559	2.41339885
e	v	0.11204346	0.05760396	1.94506529
e	w	0.11204346	0.09672938	1.15831874
e	x	0.11204346	0.03187948	3.51459544
e	y	0.11204346	0.00676033	16.5736719
e	z	0.11204346	0.03030967	3.69662419

Character e is mapped to h with the distance of 3

Letter: t

M	C	P(M=m)	P(C=c)	P(M=m|C=c)
t	a	0.09672938	0.0011926	81.1079545
t	b	0.09672938	0.02349513	4.11699673
t	c	0.09672938	0.00042464	227.792553
t	d	0.09672938	0.08345267	1.15909275
t	e	0.09672938	0.01696068	5.70315621
t	f	0.09672938	0.01916518	5.04714202
t	g	0.09672938	0.05399679	1.79139128
t	h	0.09672938	0.11204346	0.86332023
t	i	0.09672938	0.01805163	5.35848348
t	j	0.09672938	0.02463352	3.92673758
t	k	0.09672938	0.06021729	1.60633908
t	l	0.09672938	0.06468051	1.49549518
t	m	0.09672938	0.00279629	34.592084
t	n	0.09672938	0.01300793	7.43618684
t	o	0.09672938	0.03983692	2.42813404
t	p	0.09672938	0.02367131	4.08635496
t	q	0.09672938	0.07480632	1.2930644
t	r	0.09672938	0.08361304	1.15686963
t	s	0.09672938	0.013803	7.00785469
t	t	0.09672938	0.00044271	218.494898
t	u	0.09672938	0.04642559	2.08353605
t	v	0.09672938	0.05760396	1.67921421
t	w	0.09672938	0.09672938	1
t	x	0.09672938	0.03187948	3.03422134
t	y	0.09672938	0.00676033	14.3083862
t	z	0.09672938	0.03030967	3.19137044

Character t is mapped to w with the distance of 3

Letter: a

M	C	P(M=m)	P(C=c)	P(M=m|C=c)
a	a	0.08345267	0.0011926	69.9753788
a	b	0.08345267	0.02349513	3.55191309
a	c	0.08345267	0.00042464	196.526596
a	d	0.08345267	0.08345267	1
a	e	0.08345267	0.01696068	4.92036223
a	f	0.08345267	0.01916518	4.3543901
a	g	0.08345267	0.05399679	1.54551159
a	h	0.08345267	0.11204346	0.74482411
a	i	0.08345267	0.01805163	4.622998
a	j	0.08345267	0.02463352	3.3877682
a	k	0.08345267	0.06021729	1.38585896
a	l	0.08345267	0.06468051	1.29022908
a	m	0.08345267	0.00279629	29.8441034
a	n	0.08345267	0.01300793	6.41552353
a	o	0.08345267	0.03983692	2.0948574
a	p	0.08345267	0.02367131	3.5254771
a	q	0.08345267	0.07480632	1.1155832
a	r	0.08345267	0.08361304	0.99808201
a	s	0.08345267	0.013803	6.04598265
a	t	0.08345267	0.00044271	188.505102
a	u	0.08345267	0.04642559	1.79755765
a	v	0.08345267	0.05760396	1.44873152
a	w	0.08345267	0.09672938	0.86274372
a	x	0.08345267	0.03187948	2.61775542
a	y	0.08345267	0.00676033	12.3444704
a	z	0.08345267	0.03030967	2.75333482

Character a is mapped to d with the distance of 3

Letter: o

M	C	P(M=m)	P(C=c)	P(M=m|C=c)
o	a	0.08361304	0.0011926	70.1098485
o	b	0.08361304	0.02349513	3.5587387
o	c	0.08361304	0.00042464	196.904255
o	d	0.08361304	0.08345267	1.00192167
o	e	0.08361304	0.01696068	4.92981755
o	f	0.08361304	0.01916518	4.36275781
o	g	0.08361304	0.05399679	1.54848155
o	h	0.08361304	0.11204346	0.74625542
o	i	0.08361304	0.01805163	4.63188188

o	j	0.08361304	0.02463352	3.39427838
o	k	0.08361304	0.06021729	1.38852213
o	l	0.08361304	0.06468051	1.29270848
o	m	0.08361304	0.00279629	29.901454
o	n	0.08361304	0.01300793	6.42785206
o	o	0.08361304	0.03983692	2.09888303
o	p	0.08361304	0.02367131	3.53225191
o	q	0.08361304	0.07480632	1.11772698
o	r	0.08361304	0.08361304	1
o	s	0.08361304	0.013803	6.05760105
o	t	0.08361304	0.00044271	188.867347
o	u	0.08361304	0.04642559	1.80101197
o	v	0.08361304	0.05760396	1.45151551
o	w	0.08361304	0.09672938	0.86440163
o	x	0.08361304	0.03187948	2.62278589
o	y	0.08361304	0.00676033	12.3681924
o	z	0.08361304	0.03030967	2.75862583

Character o is mapped to r with the distance of 3

Letter: i

M	C	P(M=m)	P(C=c)	P(M=m|C=c)
i	a	0.06468051	0.0011926	54.2348485
i	b	0.06468051	0.02349513	2.75293213
i	c	0.06468051	0.00042464	152.319149
i	d	0.06468051	0.08345267	0.77505616
i	e	0.06468051	0.01696068	3.81355706
i	f	0.06468051	0.01916518	3.37489688
i	g	0.06468051	0.05399679	1.19785828
i	h	0.06468051	0.11204346	0.57728052
i	i	0.06468051	0.01805163	3.58308308
i	j	0.06468051	0.02463352	2.62571062
i	k	0.06468051	0.06021729	1.07411853
i	l	0.06468051	0.06468051	1
i	m	0.06468051	0.00279629	23.1308562
i	n	0.06468051	0.01300793	4.97239104
i	o	0.06468051	0.03983692	1.62363214
i	p	0.06468051	0.02367131	2.73244275
i	q	0.06468051	0.07480632	0.86463963
i	r	0.06468051	0.08361304	0.77356961
i	s	0.06468051	0.013803	4.68597611

i	t	0.06468051	0.00044271	146.102041
i	u	0.06468051	0.04642559	1.39320813
i	v	0.06468051	0.05760396	1.12284829
i	w	0.06468051	0.09672938	0.66867484
i	x	0.06468051	0.03187948	2.02890747
i	y	0.06468051	0.00676033	9.56765787
i	z	0.06468051	0.03030967	2.13398912

Character i is mapped to l with the distance of 3

Letter: n

M	C	P(M=m)	P(C=c)	P(M=m|C=c)
n	a	0.07480632	0.0011926	62.7253788
n	b	0.07480632	0.02349513	3.18390694
n	c	0.07480632	0.00042464	176.164894
n	d	0.07480632	0.08345267	0.89639213
n	e	0.07480632	0.01696068	4.41057398
n	f	0.07480632	0.01916518	3.90324101
n	g	0.07480632	0.05399679	1.38538442
n	h	0.07480632	0.11204346	0.66765447
n	i	0.07480632	0.01805163	4.14401902
n	j	0.07480632	0.02463352	3.03676875
n	k	0.07480632	0.06021729	1.24227307
n	l	0.07480632	0.06468051	1.15655119
n	m	0.07480632	0.00279629	26.7520194
n	n	0.07480632	0.01300793	5.7508248
n	o	0.07480632	0.03983692	1.87781369
n	p	0.07480632	0.02367131	3.16020992
n	q	0.07480632	0.07480632	1
n	r	0.07480632	0.08361304	0.89467286
n	s	0.07480632	0.013803	5.41957126
n	t	0.07480632	0.00044271	168.97449
n	u	0.07480632	0.04642559	1.61131653
n	v	0.07480632	0.05760396	1.29863153
n	w	0.07480632	0.09672938	0.77335668
n	x	0.07480632	0.03187948	2.34653535
n	y	0.07480632	0.00676033	11.0654861
n	z	0.07480632	0.03030967	2.46806767

Character n is mapped to q with the distance of 3

The conditional probability is shown with the following notation: $P(M=m|C=c)$. This notates: given the possibility of *C*, what is the possibility of *M* happening. This is handy in substitution where we can determine the mapping of plaintext characters to the ciphertext equivalent.

By looking at the table above, we can determine that the key used in the *Caesar Cipher* is 3 based on calculating the conditional probability given M being *e*, *t*, *a*, *o*, *i*, and *n*.

CONCLUSION

Simply by looking at the conditional probabilities of *e*, *t*, *a*, *o*, *i*, and *n*, we are able to easily determine the plaintext/ciphertext pairing. Calculating the conditional probability is extremely handy in substitution ciphers. With modern ciphers that uses both substitution and diffusion, only using conditional probability will not do. However, it is still valuable concept that help us determine the substitution aspect of ciphers.