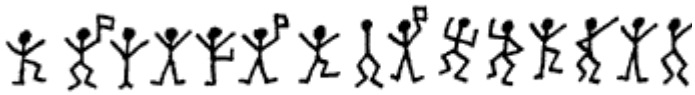


L'analyse des fréquences

1. Le principe

Un texte qui a été chiffré via une *substitution monoalphabétique* (i.e. une lettre correspond à une seule autre) présente une sécurité très limitée, pour ne pas dire quasi nulle si le message est suffisamment long. En effet, ce type de chiffrement est totalement vulnérable à un procédé appelé l'**analyse fréquentielle**.

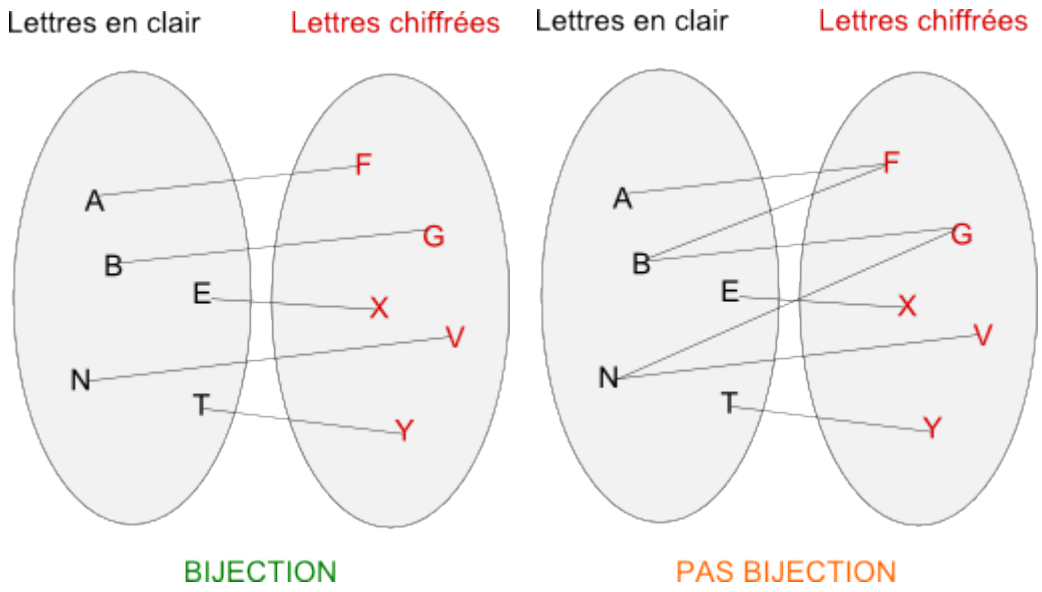
Exemple : le tableau suivant traduit une substitution mono-alphabétique (qui est un procédé bijectif) :

CLAIR	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CODE	O	M	X	P	Z	G	U	D	A	K	R	N	F	E	C	I	Q	T	L	J	V	W	Y	H	S	B

Le principe de déchiffrement est simple. Des études statistiques sur un ensemble de textes (de longueur et de niveau de langage moyens) permettent de mettre en évidence le *pourcentage moyen d'utilisation de chaque lettre*. Dès lors, il suffit d'établir un tableau des fréquences du texte chiffré, et de le comparer avec le tableau des fréquences pour un texte français normal. Si on constate par exemple que, dans le texte chiffré, la lettre Q apparaît énormément, il y a de très fortes chances qu'elle corresponde au E (lettre très courante en français). Il faut alors procéder ainsi pour chaque lettre et évaluer les possibilités.

Deux choses importantes à garder en tête :

- Le pourcentage varie selon les langues évidemment. Par exemple, en néerlandais, le Z est très courant, contrairement au français où il n'est quasiment jamais utilisé. Ce détail est **primordial** : si vous ne connaissez pas la langue dans laquelle le texte a été écrit, l'analyse des fréquences est inefficace voire impossible !!
- Si le procédé de chiffrement n'est pas bijectif, c'est-à-dire si vous avez plusieurs correspondances possibles pour chaque lettre, encore une fois, l'analyse fréquentielle est impossible... En effet, si tel était le cas, le tableau des fréquences serait totalement faussé !



Ces deux points limitent donc considérablement l'utilisation de l'analyse des fréquences...

2. Raisonnement pas à pas sur un exemple

Passons maintenant à la pratique. Voici un message chiffré (via substitution monoalphabétique donc) :

BWDDC YASWY WDTWD CTIHX DZWDS UHBWD NWYWI SVGWD UXSDZ WDXHC DIXID
WIUWU WZXMX TIXPW UXSDC GIYWB WYWTS WUWTI AWDIX GJZWU CTDWG JUWUW
DVGWB BWDDW BSMYW TICFF YXTIB NCUUX PWZWB WGYDA CYHDW IZWBW GYDXU
WDWID SBWTW DIXST DSXGT SMWXG ZWBXF XGIWH CGYVG CSTWT DWYXS ISBHX
DXSTD SXGTS MWXGZ GANXI SUWTI STFWY TXBTC GDXMC TDYWA WUUWT IXHHY SDXMW
AIYSD IWDDW VGGTA WYIXS TTCUT YWZWH WYDCT TWDZW DZWGJ DWJWD CGTBS
WGDWD ZWBWG YDXBG IWIAC TIYXS YWUWT IXBXF CSAXI NCBSV GWDWD CTIZC TTWDX
GJZWU CTDDC GDBXF CYUWD ZSTAG TWDWI ZWDGA AGTWD WIHXY BWGYD STAXT IXISC
TDDCY IDAYS UWDWI XAISC TDSTF XUXTI WDZWI YGSDW TIBWF YGSIZ WDWTI YXSBB
WDZWD FWUUW DZWDI YCGHW XGJZW ZSMWY DXTSU XGJFC TIUCG YSYBW TBWBX
MCSTW ZWASU WTIBW DYWAC BIWDT XHHCY IWTIV GWZCG BWGYD WIXFF BSAIS CTDWU
HWANW TIBWD NCUUW DZWHY CAYWW YWIBW DFWUU WDZWA CTAWM CSY

Attention : ce texte a été chiffré puis découpé en blocs pour rendre le déchiffrement plus difficile. Cela signifie qu'un mot peut-être à cheval sur plusieurs blocs, et qu'un bloc peut signifier plusieurs mots. Il est donc évident que le texte n'est pas constitué uniquement de mots de 5 lettres !

Etape 1 : préliminaires

Rendez-vous sur l'outil d'analyse des fréquences et saisissez le texte chiffré. Les deux tableaux côte à côte permettent une comparaison facile et quasiment directe, mais il ne faut pas s'y fier pour les lettres de milieu de tableau. Nous garderons ces deux tableaux sous les yeux pendant toute l'analyse.

Nous procéderons en 2 temps :

1. établir une correspondance entre quelques lettres
2. remplacer les lettres dans le texte suivant cette correspondance, et voir si ça pourrait marcher

Remarque : dès que le tableau des fréquences est établi, il est inutile de continuer le

travail sur le texte entier. Pour plus de facilité, nous n'allons considérer pour l'analyse que les 5 premiers blocs :

```
BWDDC YASWY WDTWD CTIHX DZWDS [...]
```

Etape 2 : l'analyse

Reprenons nos deux tableaux des fréquences. Le W apparaît à 17,85 % dans le texte, c'est une majorité écrasante et à n'en pas douter, le W correspond au E. On peut aussi raisonnablement supposer que le D correspond au S et le T au A. Ceci nous donne le tableau suivant :

CLAIR	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CODE	T				W														D							

En remplaçant, le texte devient :

```
BESSC YASEY ESAES CAIHX SZESS
```

Ce premier stade passé, c'est maintenant que commence la torture des méninges. Considérons le premier bloc : **BESSC**. Les lettres inconnues sont le **B** et le **C**. Il existe 2 possibilités pour ce bloc :

1. Soit c'est un mot entier de 5 lettres
2. Soit c'est plusieurs mots qui ont été collés lors du chiffrement

Première possibilité : connait-on un mot pas trop recherché de la forme "?ESS?" ? Il en existe bien sûr, comme "MESSE" par exemple. Mais cela pose un problème : le bloc que nous étudions est le **premier du texte**. On est donc sûr qu'il contient ou qu'il est le **premier mot de la première phrase**. Or, "Messe" pour débuter une phrase, c'est fort peu probable . En effet, une phrase n'a quasiment aucune chance de commencer par un nom commun (c'est en revanche très possible pour un nom propre, mais passons), mais plutôt par un déterminant ("le, "la", "les", "un", ...), un adverbe (hier, demain, ...), etc.

Ceci nous amène donc à retenir la deuxième possibilité : **BESSC** désigne plusieurs mots. De plus, en regard de ce qui a été dit une phrase plus tôt, il est très probable que la phrase commence par un déterminant, et avec une configuration telle que "BES..", on peut établir avec une grande certitude que le **B désigne le L**, pour donner "**LES** SC" ou bien "**LE** SSC". On peut donc ajouter une nouvelle lettre à notre tableau, le B :

CLAIR	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CODE	T				W							B							D							

Et le texte devient :

```
LESSC YASEY ESAES CAIHX SZESS
```

Maintenant que choisir ? "LES SC" ou "LE SSC" ? En y réfléchissant bien, c'est bien la première des deux qu'il faut conserver. En effet, si on prenait "LE SSC", cela voudrait dire que le deuxième mot commence par un double S, ce qui n'est pas envisageable. Va donc pour "LES SC".

```
LES SC YASEY ESAES CAIHX SZESS
```

Il est peu probable qu'une consonne suive un S en début de mot, on en déduit donc que le C désigne une voyelle. Mais quelle voyelle ? C'est maintenant qu'il faut ressortir nos deux tableaux des fréquences. Selon eux, le C correspondrait au L, ce qui est impossible puisqu'on a déjà établi la correspondance B-L. Voilà donc pourquoi il ne faut pas se fier aux lettres de milieu de tableau ! Pas de panique, ce n'est pas dramatique. On voit sur le tableau de gauche que le L est à 5.89%, ce qui est très proche du C de notre texte chiffré qui est à 5.6%. Comme le L n'est pas valable, regardons quelles autres lettres (et pas n'importe lesquelles, seulement les voyelles) sont proches de ce pourcentage. On trouve le **O (5.34%)** et le **U (6.05%)** ! C'est donc sans aucun doute une de ces voyelles que le C désigne.

Laquelle choisir ? Le O est plus proche : (5.6 - 5.34) est inférieur à (6.05 - 5.6) ... mais cette différence n'est pas significative. Jusqu'à présent, toutes nos suppositions étaient assez fiables et ne demandaient pas de "prise de risque", mais lors d'une analyse fréquentielle, il faudra bien quitter un moment ou à un autre cette sécurité. On va donc considérer, arbitrairement, que le **C désigne le O**. Si on découvre plus loin que ça mène à une impasse, on reviendra en arrière et on attribuera le C au U. On obtient :

LES SO YAS**E**Y **ESAES OA**IHX **S**Z**ES**S

Poursuivons avec le Y. Il suit un O, il peut donc être :

- une voyelle : I ou U (les autres sont peu probables)
- une consonne

Examinons les tableaux de fréquences. La lettre qui correspondrait au Y serait le R. Mais Y étant une lettre de milieu de tableau, il ne faut pas trop s'y fier. Examinons les lettres qui environnent le R : il y a le **I (7.23%) et le U (6.05%)**. Aie. Justement les deux lettres sur lesquelles on hésitait. Les tableaux ne peuvent donc pas nous aider plus que ça, il va falloir tester les possibilités.

Si Y = U :

LES SO UAS**EU ESAES OA**IHX **S**Z**ES**S

Si Y = I :

LES SO IAS**EI ESAES OA**IHX **S**Z**ES**S

Si Y = R :

LES SO RAS**ER ESAES OA**IHX **S**Z**ES**S

Première déduction : le deuxième mot commence par SOU, SOI, SOR, puisque "SO" n'existe pas dans la langue française. C'est cette piste que nous allons suivre : essayer de **deviner le deuxième mot de la phrase**. Trois possibilités donc (prenons 10 lettres pour être sûr de contenir tout le mot) :

SO UAS**EU ESA**
SO IAS**EI ESA**
SO RAS**ER ESA**

... Réflexion ...
... ...

... ...

SORCIERES ! Le mot SORCIERES apparaît si on considère la possibilité "R". Cela implique deux choses : **le A représente le C, et le S représente le I**. De plus, ceci confirmerait l'hypothèse qu'on avait faite plus haut, lorsqu'on a supposé que le C désigne le O. Allons vérifier sur les tableaux des fréquences :

C (3.32 %)	A (3.24%)
S (6.34%)	I (7.23%)

Les pourcentages sont très proches, et le mot existe dans la langue : on peut écarter l'hypothèse de la coincidence et mettre à jour notre tableau de substitution :

CLAIR	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CODE	T		A		W				S			B			C			Y	D							

Et le message :

LES SORCIERES AES OAIHX **S**Z**ESI**

Quelque chose cloche. Le troisième bloc, "AES", ne colle pas avec le début. Nous avons sans aucun doute commis une erreur plus haut. Cette erreur porte vraisemblablement sur le A, puisque le E et le S de AES sont confirmés par SORCIERES. Or, on avait convenu au tout début que le T désignait A, cette assertion est donc fausse : le T ne désigne pas A mais une autre lettre.

Reprenons les deux tableaux de fréquences : les pourcentages proches du A (7.68%) sont **le S (déjà attribué), le N (7.61%), et le T (7.30%)**. C'est donc entre ces deux dernières lettres qu'il faut trancher.

Si T représente T :

LES SORCIERES TES OTIHX **S**Z**ESI**

Si T représente N :

LES SORCIERES NES ONIHX **S**Z**ESI**

... Si on choisit T = N, on entrevoit la possibilité de faire apparaître le verbe de la phrase, **SONT**, précédé alors d'un **NE** de négation : "LES SORCIERES NE SONT". Optons donc pour ce choix, et une nouvelle déduction s'impose : **I désigne T** (hypothèse confirmée par les tableaux des fréquences). On a :

LES SORCIERES NE SONT HX **S**Z**ESI**

Continuons le travail de déduction. Qu'est-ce qui, généralement, vient directement après "ne sont" ? Réponse : **PAS** ! De plus, on dispose déjà d'un S confirmé, cette idée doit être exacte ! Dès lors, **H représente P, et X représente A**. Rapide coup d'oeil au tableau des fréquences : c'est vérifié !

LES SORCIERES NE SONT PAS Z**ESI**

Ouf ! Les 5 premiers blocs sont presque entièrement traduits. Récapitulatif du tableau :

CLAIR	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CODE	X		A		W				S			B		T	C	H		Y	D	I						

Ajoutons le reste de la ligne et remplaçons :

LES SORCIERES NE SONT PAS ZESI UPLES NERET IVGES UAISZ ESAPO STATS ETUEU EZAMA NTAPE UAISO GTREL ERENI EUENT CESTA GJZEU ONSEG

On trouve aisément la suite de la phrase : LES SORCIERES NE SONT PAS DE SIMPLES HERETIQUES MAIS [...]. Donc **Z vaut D, U vaut M, N vaut H, V vaut Q, G vaut U**. Le tableau :

CLAIR	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CODE	X		A	Z	W			N	S			B	U	T	C	H	V	Y	D	I	G					

Le texte :

LES SORCIERES NE SONT PAS DE SIMPLES HERETIQUES MAIS DES APOSTATS ET MEME DAMA NTAPE MAISO UTREL ERENI EMENT CESTA UJDEM ONSEU

Mais encore :

LES SORCIERES NE SONT PAS DE SIMPLES HERETIQUES MAIS DES APOSTATS ET MEME DAVANTAGE MAIS OUTRE LE RENIEMENT CEST AUX DEMONS EU

Le travail est presque terminé, le restant des lettres à déterminer est un jeu d'enfant en utilisant les quelques blocs suivants. Pour finir, la solution finale (avec la ponctuation en prime) :

Les sorcières ne sont pas de simples hérétiques mais des apostats et même davantage... mais outre le reniement, c'est aux démons eux-mêmes qu'elles se livrent, offrant l'hommage de leurs corps et de leurs âmes... Et s'il en est ainsi au niveau de la faute, pourquoi n'en serait-il pas ainsi au niveau du châtiment infernal ? -- Nous avons récemment appris avec tristesse ... qu'un certain nombre de personnes des deux sexes, oublieuses de leur salut et contrairement à la foi catholique, se sont donnés aux démons sous la formes d'incubes et de succubes ... et par leurs incantations, sorts, crimes et actions infamantes, détruisent le fruit des entrailles des femmes, des troupeaux, de divers animaux ; font mourir le blé, l'avoine, déciment les récoltes ... ; n'apportent que douleurs et afflictions ; empêchent les hommes de procréer et les femmes de concevoir ...

3. Récapitulatif

On a vu que l'analyse fréquentielle est une méthode longue, fastidieuse (parfois!), amusante (souvent!), et qui demande surtout de la logique. Cependant, elle est très limitée dans la pratique. Rappel des prérequis : il faut :

- Que le texte chiffré soit suffisamment long, sans quoi l'établissement du tableau des fréquences n'est pas fiable
- Savoir dans quelle langue le texte chiffré a été écrit
- Que le procédé de chiffrement soit bijectif
- De la persévérance

Retour au sommaire - #Remonter