Chapitre 3

Cryptanalyse

3.1 Point de vue bayésien

Faisons une petite expérience de pensée.

1. Quand on lance un dé, quelle est la chance que ce soit le chiffre 1 qui sorte ?
2. Maintenant je lance un dé dans un gobelet, quelle est la chance de faire 1 à présent ?
3. Je le dé mais vous ne la voyez pas, à combien estimer la probabilité que ce soit 1 ? Et si je vous affirme que c'est 1 ?
4. En supposant que la probabilité que j'essaie de vous induire en erreur est de 0.1 quelle est la probabilité que ce soit effectivement 1 sur le dé ?

Théorème de Bayes. Soit A et B deux évènements. La probabilité conditionnelle d'avoir A sachant B est donnée par la formule

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

De même, B sachant A est donné par

$$P(B|A) = \frac{P(A \cap B)}{P(A)},$$

donc en combinant les deux, nous obtenons l'expression suivant appelée théorème de Bayes,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Cette formule met en évidence un phénomène qui semble contre intuitif sur les tests à faux positifs. Supposons le cas suivant :

— Imaginons que l'on a un test pour une maladie qui touche 1 personne sur mille.

— Si un patient a contracté la maladie, le test le ressort positif, presque systématiquement, avec une probabilité 0,99 ;

— si un patient est sain, le test est correct, c'est-à-dire négatif avec une probabilité de $0,95$.
Désignons par A l'événement « Le patient a contracté la maladie » et par B l'événement « Le test est positif ». Le théorème de Bayes donne alors :

$$P(A|B) = \frac{0,99 \times 0,001}{0,99 \times 0,001 + 0,05 \times 0,999} \simeq 0,019.$$

Donc avec un résultat positif on a seulement $1,9\%$ de chance d'être touché par la maladie.

Bayes a utilisé cette formule en essayant d'estimer la distribution de probabilité d'un paramètre d'une variable aléatoire. C'est ça la grande innovation dans sa démarche de statisticien, c'est d'inférer de l'information sur un systèmes à partir d'observations statistiques.

Dans cette démarche, la probabilité exprime un degré de croyance en un événement. Le degré initial de croyance peut être basé sur des connaissances a priori, telles que les résultats d'expériences antérieures, ou sur des croyances personnelles concernant l'événement. La perspective bayésienne diffère d'un certain nombre d'autres interprétations de la probabilité, comme l'interprétation fréquentiste qui considère la probabilité comme la limite de la fréquence relative d'un événement après de nombreux essais.

> Le rapport du GIEC de 2007 conclut que :
> « l'activité humaine est fort probablement la cause du réchauffement global (à plus de 90 %). Le réchauffement global dans ce cas-ci réfère à une hausse de 0,75 degré de la température globale moyenne depuis les 100 dernières années. »

> What is the chance that an earthquake of magnitude 6.7 or greater will occur before the year 2030 in the San Francisco Bay Area ? The U.S. Geological Survey estimated the chance to be 0.7 ± 0.1 (USGS, 1999).

3.2 Chiffrements

La théorie de l'information permet de mesurer l'efficacité théorique d'un système de chiffrement. Claude Shannon travaillait d'ailleurs pour les services secrets américains pendant la seconde guerre mondiale avec notamment pour objectif de décrypter les codes ennemis. Certain de ses travaux on été déclassifiés seulement dans les années 80.

3.2.1 Le chiffrement de César

On appelle chiffrement de César la technique de chiffrement d'un texte qui consiste à décaler toutes les lettres d'un nombre constant dans l'alphabet. Elle était effectivement utilisée par Jules César dans ses correspondances secrètes.

Supposons qu'un chiffrement de César ait été utilisé sur un texte et que nous interceptons une certaine quantité, N lettres, du texte chiffré. Pour des valeurs relativement grandes de N, disons 50 lettres, il existe presque toujours une solution unique au chiffrement ; c'est-à-dire, une seule séquence de mots français qui se transforme en le message intercepté par une substitution simple. Cependant, avec une valeur plus petite de N, la chance d'avoir plus d'une solution est plus grande : avec 15 lettres, il y aura généralement un certain nombre de fragments de texte possibles qui pourraient

correspondre, tandis qu'avec $N = 8$ une grande proportion de toutes les séquences raisonnables en français de cette longueur sont possibles ($\frac{1}{8}$ selon Shannon), car il y a rarement plus d'une lettre répétée dans le texte intercepté. Avec $N = 1$, n'importe quelle lettre est clairement possible et a la même probabilité a posteriori que sa probabilité a priori. Pour une seule lettre, le système est parfait.

Cela est généralement visible des systèmes de chiffrement résolubles. Avant qu'une partie du message ne soit interceptée, nous pouvons deviner les probabilités a priori associées aux différents messages possibles, ainsi qu'aux différentes clés. Au fur et à mesure que le message est intercepté, le cryptanalyste calcule les probabilités a posteriori, et à mesure que N augmente, les probabilités de certains messages augmentent, et pour la plupart, diminuent, jusqu'à ce qu'il n'en reste finalement qu'un seul, qui a une probabilité presque égale à 1, tandis que la probabilité totale de tous les autres est presque nulle.

Decipherments	$N = 1$	$N = 2$	$N = 3$	$N = 4$	$N = 5$
C R E A S	.028	.0377	.1111	.3673	1
D S F B T	.038	.0314			
E T G C U	.131	.0881			
F U H D V	.029	.0189			
G V I E W	.020				
H W J F X	.053	.0063			
I X K G Y	.063	.0126			
J Y L H Z	.001				
K Z M I A	.004				
L A N J B	.034	.1321	.2500		
M B O K C	.025		.0222		
N C P L D	.071	.1195			
O D Q M E	.080	.0377			
P E R N F	.020	.0818	.4389	.6327	
Q F S O G	.001				
R G T P H	.068	.0126			
S H U Q I	.061	.0881	.0056		
T I V R J	.105	.2830	.1667		
U J W S K	.025				
V K X T L	.009				
W L Y U M	.015		.0056		
X M Z V N	.002				
Y N A W O	.020				
Z O B X P	.001				
A P C Y Q	.082	.0503			
B Q D Z R	.014				
H(decimal digits)	1.2425	.9686	.6034	.285	0

FIGURE 3.1 – Exemple de probabilité a posteriori pour un chiffrement de César par Shannon.

La Figure 3.1 montre les probabilités a posteriori pour un chiffrement de César appliqué à un texte en anglais, avec la clé choisie au hasard parmi les 26 possibilités. Pour permettre l'utilisation de tables de fréquences standard pour les lettres, digrammes et trigrammes, le texte a été commencé à un point aléatoire (en ouvrant un livre et en posant un crayon au hasard sur la page). Le message

sélectionné de cette manière commence par "creases to..." à l'intérieur du mot "increases". Si l'on savait que le message commence une phrase, un ensemble différent de probabilités pouvait être utilisé, correspondant aux fréquences des lettres, digrammes, etc., au début des phrases.

Le système de César avec une clé aléatoire est un chiffrement dont la clé choisie n'affecte pas les probabilités a posteriori puisqu'il permute les texte dans un ensemble de 26 textes. Il suffit alors de répertorier les déchiffrements possibles pour toutes les clés et de calculer leurs probabilités a priori. Les probabilités a posteriori sont obtenues en divisant ces probabilités a priori par leur somme. Ces déchiffrements possibles sont trouvés par le processus standard de "défilement de l'alphabet" à partir du message et sont répertoriés à gauche. Ils forment la classe de résidu pour le message. Pour une seule lettre interceptée, les probabilités a posteriori sont égales aux probabilités a priori pour les lettres et sont affichées dans la colonne intitulée $N = 1$. Pour deux lettres interceptées, les probabilités sont celles des digrammes ajustées pour sommer à 1 et sont affichées dans la colonne $N = 2$. On fait de même pour les suites des 3, 4 puis 5 lettres. Et on obtient une distribution d'entropie quasiment nulle !

3.2.2 Mesure de la connaissance

La question qui se pose est de savoir dans quelle mesure la connaissance d'un message chiffré donne de l'information sur le message clair, en supposant que le système utilisé est connu. Afin d'étudier les qualités du système cryptographique étudié, on s'imagine que l'envoi de messages est une expérience aléatoire, qui fait intervenir trois variables aléatoires M, K, et C. La première M donne le message à envoyer, puis la valeur de K correspond à la clé d'encodage choisie, et enfin C donne le message codé résultant. La situation est donc la suivante. On a :

1. Le résultat de la variable aléatoire M est l'un des messages clairs possibles qui est choisi par l'envoyeur avec une certaine probabilité.
2. La variable aléatoire K donne l'une des clés possibles selon le système choisi tirée selon un autre probabilité supposée indépendante.
3. Enfin, C donne le message codé résultant.

La situation décrite dans la partie précédente correspond au cas où l'entropie de la distribution de la clé K connaissant C est nulle. La faille théorique d'un système de cryptographie est donc atteinte quand

$$H(K|C) \approx 0.$$

Supposons que l'on intercepte C_N une partie de longueur N du code. Une question importante pour un-e cryptanalyste est de savoir quelle est la longueur de N minimal pour pouvoir espérer déchiffrer le message.

Proposition 3.2.1. *L'incertitude sur la clé $H(K|C_N)$ est décroissante en N.*

Démonstration. $H(K|C_N) = H(K|C_{N-1}, C_N)$. □

Dans la suite on omet l'indice N quand il n'est pas nécessaire. Lorsqu'on connaît la clé, on peut récupérer le message clair à partir du message codé. En conséquence, la quantité d'information globalement obtenue lorsqu'on connaît la valeur des deux variables aléatoires indépendantes M et K est donc égale à la somme :

$$H(K, C) = H(K) + H(M).$$

Ainsi,

$$\begin{aligned} H(K|C) &= H(K,C) - H(C) \\ &= H(K) + H(M) - H(C) \\ &= H(K) - [H(C) - H(M)] \\ &\geq H(K) - [\log G - H(M)]. \end{aligned}$$

Où G est le nombre total de messages possible de longueur N. Notre système de chiffrement étant déterministe, on a aussi au plus G codes C et donc $H(C) \leq \log G$. La quantité

$$D_N = \log G - H(M)$$

est appelée la redondance du message, vocabulaire qui est motivé par le résultat sur la compression.

Ainsi pour pouvoir espérer casser un code secret, il faut que la redondance du message soit supérieure ou égale à l'entropie de la clé. Si on choisit une clé aléatoire aussi longue que le message le code est théoriquement incassable, c'est le cas du code de Vernam utilisé par Che Guevara pour correspondre avec Fidel Castro.

Exemples du codage de Vigenère et Vernam Le chiffre de Vigenère est un algorithme de chiffrement polyalphabétique inventé par le cryptologue français Blaise de Vigenère au XVIe siècle. Il est basé sur un chiffrement par décalage auquel est ajouté l'utilisation d'un mot-clé qui modifie le décalage à chaque étape. Le chiffre de Vernam suis le même principe mais au lieu d'un mot-clé on prend une phrase qui doit avoir un nombre de lettres identique voire supérieur au nombre de caractères du message clair.

D'après l'estimation de Shannon, la redondance d'un texte anglais est d'environ $\log 26/2$ et donc le chiffre de Vernam ou celui de Vigenère (avec un motif plus court que l'on répète) sont théoriquement résolubles.

En fait son estimation de la redondance est obtenue de cette manière. Il observe que ces systèmes de chiffrement sont en générale résolubles et donc que la redondance de l'anglais est d'au moins 50%. Autrement dit, si je vous donne un texte coupé à un endroit quelconque, vous avez en moyenne une chance sur deux de deviner la prochaine lettre.

Cas du code de César Dénommons X_i la variable aléatoire qui donne la i-ème lettre du message clair. D'après Shannon,

$$H(X_{n+1}|X_1 \cdot X_2 \cdot \ldots \cdot X_n) \leq 1,6,$$

et la redondance d'un message de longueur n est d'au moins $1,6 \cdot n$.

Pour un codage de César, il y a 26 clés possibles tandis que pour un codage par substitution mono-alphabétique (c'est à dire en adoptant un alphabet secret), il y a :

$$26! \approx 4,0 \cdot 10^{26}$$

clés possibles. En supposant que le choix d'une clé est équiprobable, on trouve dans chacun des cas, $H(K) = \log 26 \approx 3,3$ et $H(K) = \log 26! \approx 26,6$. Donc pour pouvoir espérer déchiffrer le code dans le premier cas, il faut intercepter au moins 3 lettres et 17 dans le second cas.

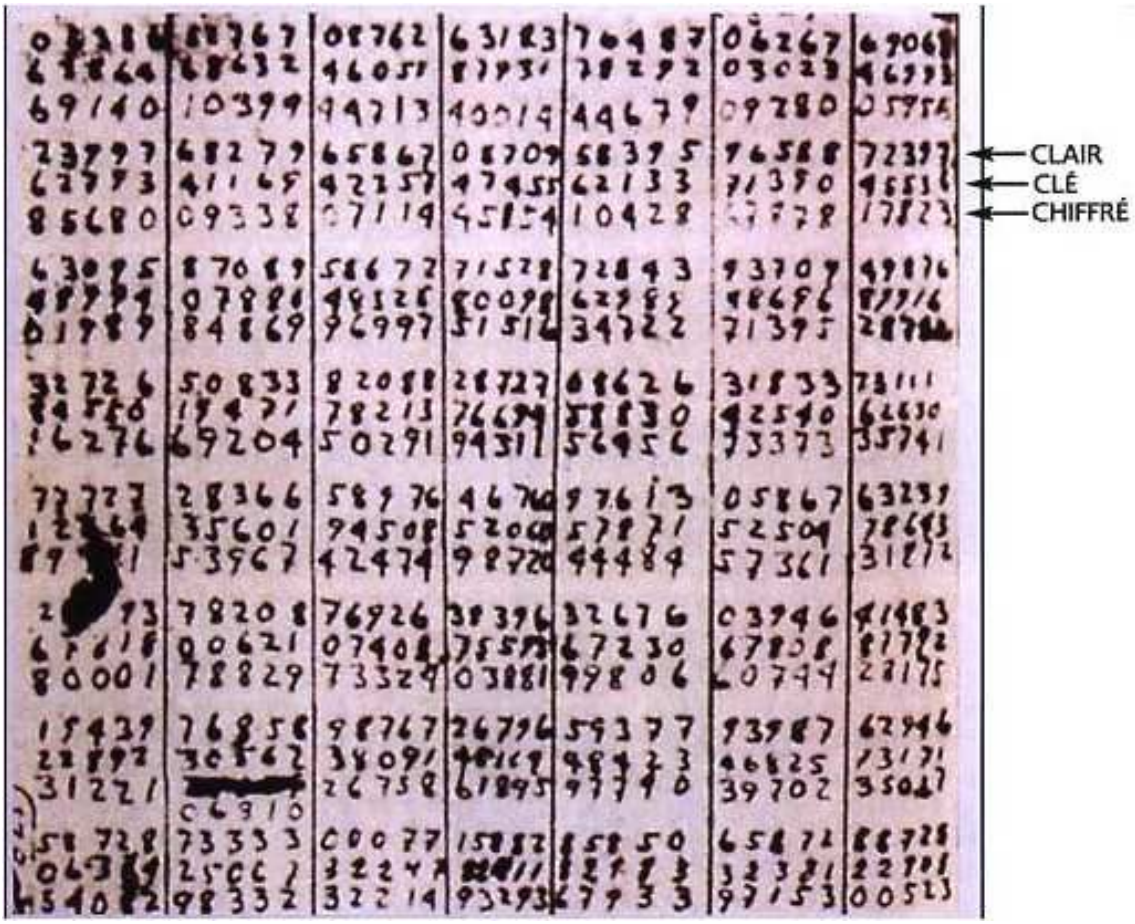

FIGURE 3.2 – message chiffré retrouvé sur che guevara le jour de son exécution en bolivie.

3.2.3 Banburisme

Pendant la deuxième guerre mondiale, une équipe de cryptanalystes sous la supervision d'Alan Turing a réussi l'exploit de déchiffrer quotidiennement la clef secrète des armées allemandes dans les laboratoires secrets de Bletchley Park. Comme la clef changeait chaque jour parmi un grand nombre de possibilité il fallait extraire des messages codés du jour environ 129 décibans, le ban étant l'unité d'information en base 10 et 10 décibans équivaut à un ban. Le mot *ban* vient d'ailleurs du nom de la ville Banbury à environ 30 kilomètres des laboratoires où des les lettres coïncidant entre deux textes étaient recherchées par un système de cartes perforées.

Introduction du poids d'une hypothèse Lorsque nous comparons des hypothèses les unes avec les autres à la lumière des données, il est souvent pratique de comparer le logarithme de la

probabilité des données sous les hypothèses alternatives,

$$\text{‘vraisemblance logarithmique pour } H_i\text{’} = \log P(D|H_i), \tag{3.1}$$

ou, dans le cas où seules deux hypothèses sont comparées, nous évaluons les ‘cotes logarithmes’,

$$\log \frac{P(D|H_1)}{P(D|H_2)}, \tag{3.2}$$

qui est appelé le ‘poids en faveur de H_1’. La vraisemblance logarithmique d'une hypothèse, $\log P(D|H_i)$, est le négatif de la quantité d'information des données D : si les données ont une quantité d'information importante, étant donné une hypothèse, elles sont surprenantes pour cette hypothèse ; si sous une autre hypothèse nous ne sommes pas aussi surpris par les données, alors cette hypothèse devient plus probable.

Toutes ces quantités sont des logarithmes de probabilités, ou des sommes pondérées de logarithmes de probabilités, donc elles peuvent toutes être mesurées dans les mêmes unités.

Comment détecter que deux messages proviennent de machines avec une séquence d'état commune Dans le cas d'Enigma, deux machines avec les mêmes rotors et câblages vont coder un texte en se déplaçant de façon déterministe dans un ensemble de 26^3 permutations de l'alphabet mais en commençant à une point aléatoire du chemin. Comme un grand nombre de messages étaient interceptés, il y avait une forte probabilité de trouver deux morceaux de messages avec la même séquence de permutations. Le but était alors de trouver de tels messages.

Les hypothèses considérés sont alors l'hypothèse nulle, H_0, qui stipule que les machines sont dans des états différents et que les deux messages en clair ne sont pas liés ; et l'hypothèse de ‘correspondance’, H_1, qui dit que les machines sont dans le même état, et que les deux messages en clair ne sont pas liés. Aucune tentative n'est faite ici pour déduire l'état de l'une ou l'autre des machines. Les données fournies sont les deux textes chiffrés x et y ; supposons qu'ils ont tous deux une longueur T et que la taille de l'alphabet est A (26 pour Enigma). Quelle est la probabilité des données, étant donné les deux hypothèses ?

Tout d'abord, l'hypothèse nulle. Cette hypothèse affirme que les deux textes chiffrés sont donnés par

$$x = x_1x_2x_3... = c_1(u_1)c_2(u_2)c_3(u_3)... \tag{3.3}$$

et

$$y = y_1y_2y_3... = c'_1(v_1)c'_2(v_2)c'_3(v_3)... \tag{3.4}$$

où les codes c_t et c'_t sont deux permutations indépendantes variant dans le temps de l'alphabet, et $u_1u_2u_3...$ et $v_1v_2v_3...$ sont les messages en clair. Un calcul exact de la probabilité des données (x, y) dépendrait d'un modèle de langage du texte en clair, et d'un modèle de l'Enigma, mais si l'on suppose que chaque machine Enigma est une permutation aléatoire idéale variant dans le temps, alors la distribution de probabilité des deux textes chiffrés est uniforme. Tous les textes chiffrés ont la même probabilité :

$$P(x, y|H_0) = \frac{1}{A^{2T}} \tag{3.5}$$

Et pour H_1 ? Cette hypothèse affirme qu'une seule permutation variant dans le temps, c_t, est à la base à la fois de

$$x = x_1x_2x_3... = c_1(u_1)c_2(u_2)c_3(u_3)... \tag{3.6}$$

et

$$y = y_1 y_2 y_3 ... = c_1(v_1)c_2(v_2)c_3(v_3)... \tag{3.7}$$

Quelle est la probabilité des données (x, y) ? Nous devons faire quelques hypothèses sur le langage du texte en clair. Si le texte en clair était complètement aléatoire, alors la probabilité de $u_1u_2u_3...$ et $v_1v_2v_3...$ serait uniforme, de même que celle de x et y, donc la probabilité $P(x, y|H_1)$ serait égale à $P(x, y|H_0)$, et les deux hypothèses H_0 et H_1 seraient indiscernables.

Nous progressons en supposant que le texte en clair n'est pas complètement aléatoire. Les deux textes en clair sont écrits dans un langage, et ce langage comporte des redondances. Supposons par exemple que certaines lettres du texte en clair sont plus souvent utilisées que d'autres. Donc, même si les deux messages en clair sont sans rapport, ils sont légèrement plus susceptibles d'utiliser les mêmes lettres ; si H_1 est vrai, deux lettres synchronisées des deux textes chiffrés sont légèrement plus susceptibles d'être identiques. De même, si un langage utilise fréquemment certaines paires ou triplets de lettres, alors les deux messages en clair peuvent occasionnellement contenir les mêmes paires ou triplets de lettres en même temps, donnant ainsi, si H_1 est vrai, une petite série de 2 ou 3 lettres identiques. La Table 3.1 montre une telle coïncidence dans deux messages en clair qui sont sans rapport, sauf qu'ils sont tous deux écrits en anglais. Les cryptanalystes ont recherché parmi les bouts de messages des paires qui étaient anormalement similaires les unes aux autres, en comptant le nombre de monogrammes, de bigrammes, de trigrammes, etc., identiques.

u	LITTLE-JACK-HORNER-SAT-IN-THE-CORNER-EATING-A-CHRISTMAS-PIE--HE-PUT-IN-H
v	RIDE-A-COCK-HORSE-TO-BANBURY-CROSS-TO-SEE-A-FINE-LADY-UPON-A-WHITE-HORSE
matches:	.*....*..******.*.............*...........*.............*..........

TABLE 3.1 – Deux morceaux alignés de texte en anglais, u et v, avec des correspondances marquées par *. Remarquez qu'il y a douze correspondances, incluant une séquence de six, tandis que le nombre attendu de correspondances dans deux chaînes totalement aléatoires de longueur T = 74 serait d'environ 3. Les deux textes chiffrés correspondants provenant de deux machines dans des états identiques auraient également douze correspondances.

Prenons le cas simple d'un modèle de langage par monogramme et estimons quelle longueur de message est nécessaire pour pouvoir décider si deux machines sont dans le même état. Supposons que les lettres successives sont tirées de manière indépendante et identiquement distribuées (i.i.d.) à partir de la distribution d'une probabilité $\{p_i\}$.

La probabilité de x et y n'est pas uniforme : considérons deux caractères simples, $x_t = c_t(u_t)$ et $y_t = c_t(v_t)$; la probabilité qu'ils soient identiques est donnée par :

$$P(u_t = v_t) = \sum_i p_i^2 =: m$$

Nous appelons cette quantité m, pour 'probabilité de correspondance' ; pour l'anglais et l'allemand, m est d'environ $2/26$ plutôt que $1/26$ (la valeur qui serait valable pour une langue complètement aléatoire).

En supposant que c_t est une permutation aléatoire idéale, la probabilité de x_t et y_t est, par symétrie :

$$P(x_t, y_t|H_1) = \begin{cases} \frac{m}{A} & \text{si } x_t = y_t \\ \frac{(1-m)}{A(A-1)} & \text{pour } x_t \neq y_t \end{cases}$$

Étant donné une paire de textes chiffrés x et y de longueur T qui correspondent en M endroits et ne correspondent pas en N endroits, la poids logarithmique en faveur de H_1 est alors donnée par :

$$\begin{aligned}\log \frac{P(x,y|H_1)}{P(x,y|H_0)} &= M \log \frac{m/A}{1/A^2} + N \log \frac{\frac{(1-m)}{A(A-1)}}{1/A^2} \\ &= M \log mA + N \log \frac{(1-m)A}{A-1}.\end{aligned}$$

Chaque correspondance contribue de $\log(mA)$ en faveur de H_1 ; chaque non-correspondance contribue de $\log((1-m)A)$ en faveur de H_0.

TABLE 3.2 – Probabilité de correspondance pour le monogramme en anglais

Probabilité de correspondance en allemand	m	0.076
Probabilité de correspondance pour une loi uniforme	$1/A$	0.037
Preuve logarithmique pour H_1 par correspondance	$10 \log_{10} mA$	3.1 db
Preuve logarithmique pour H_1 par non-correspondance	$10 \log_{10} \frac{(1-m)A}{(A-1)}$	-0.18 db

Si, par exemple, il y avait $M = 4$ correspondances et $N = 47$ non-correspondances dans une paire de longueur $T = 51$, le poids en faveur de H_1 serait de +4 décibans, soit un rapport de vraisemblance de 2,5 pour 1.

Le poids attendu d'une ligne de texte composé de $T = 20$ caractères est l'espérance de l'expression ci-dessus, qui dépend de la véracité de H_1 ou de H_0. Si H_1 est vraie, alors les correspondances sont attendues à une fréquence m, et le poids attendu est alors de 1.4 décibans par 20 caractères. Si H_0 est vrai, alors les correspondances fortuites sont attendues à une fréquence de $1/A$, et le poids attendu est de -1.1 décibans par 20 caractères. En général, environ 400 caractères doivent être inspectés pour avoir un poids de preuve supérieur à cent pour un (20 décibans) en faveur de l'une ou l'autre hypothèse.

Ainsi, deux textes en clair en allemand ont plus de correspondances que deux chaînes aléatoires. De plus, parce que les caractères consécutifs en anglais ne sont pas indépendants, les statistiques de bigrammes et de trigrammes de l'anglais ne sont pas uniformes et les correspondances ont tendance à se produire par rafales de correspondances consécutives. En utilisant de meilleurs modèles de langage, les poids logarithmiques donnés par les séries de correspondances ont été calculées de manière plus précise. Un tel système de notation a été élaboré par Turing et affiné par Good. Les résultats positifs ont été transmis aux services de cryptanalyse automatisés et humains. Selon Good, le plus long faux-positif qui est apparu dans ce travail était une série de 8 correspondances consécutives entre deux machines qui étaient en réalité dans des états non liés.