Análise de Freqüências de Línguas

Bruno da Rocha Braga
Ravel / COPPE / UFRJ
brunorb@ravel.ufrj.br
http://www.ravel.ufrj.br/

24 de Março, 2003

Resumo

Para construção de ferramentas de cripto-análise é necessária a codificação de rotinas de análise de freqüência de letras, digramas ou mesmo n-gramas do texto. Uma vez obtidas estas facilidades de software, também é de grande valia para o cripto-analista o cálculo dessas medidas estatísticas para o idioma no qual suspeita-se estar o texto cifrado. Neste trabalho, calculamos estas medidas para a língua portuguesa e inglesa, bem como comentamos o processo usado na obtenção dessas medidas.

1. Introdução

Um conjunto P^n de símbolos encontrados em textos de uma língua L, que podem ser palavras, letras, digramas, trigramas ou qualquer n-grama, ocorrem com freqüências individuais que tendem sempre a um mesmo valor quanto maior for o texto de amostra avaliado; formando o que chamamos *histograma de freqüências* para um conjunto de símbolos desta língua. Estatisticamente falando, podemos dizer que P^n é uma variável aleatória que assume valores segundo uma freqüência teórica bem definida.

Uma *Variável Aleatória* é um termo estatístico para uma variável que assume valores de um Espaço Amostral, ou seja, o conjunto dos valores que são possíveis ocorrer num experimento. Uma distribuição de probabilidades é uma função cujo domínio são pontos do Espaço Amostral (EA) e a imagem consiste das probabilidades de ocorrência (freqüências) associadas a cada um desses pontos, sendo denotada por $P(X)$, onde X é a variável aleatória para este EA. Exemplo:

Experimento:	lançamento de três moedas
Variável Aleatória X:	número de "caras" obtidas; X = {0, 1, 2, 3}
Espaço Amostral EA (n(S)=8):	S = {ccc $^{X=3}$, cck $^{X=2}$, ckc $^{X=2}$, ckk $^{X=1}$, kcc $^{X=2}$, kck $^{X=1}$, kkc $^{X=1}$, kkk $^{X=0}$}
Probabilidades associadas P(X):	P(X=0)=1/8, P(X=1)=3/8, P(X=2)=3/8, P(X=3)=1/8.

Fazendo uma analogia com o estudo da língua, o histograma de freqüências de letras (a distribuição da variável aleatória P^1 obtida experimentalmente) mostra claramente que dada uma letra qualquer de um texto em português (um experimento), é mais provável que ela seja *a* do que *z*, pois $P(P^1=a) > P(P^1=z)$. Em diferentes idiomas, algumas letras (ou outro símbolo qualquer) são mais usados que outros, nos permitindo dizer que as

freqüências do histograma descrevem indivíduos de uma população de textos de um mesmo idioma, pois enunciam uma característica comum entre os indivíduos desta população.

1.1. Cálculo do Histograma de Freqüências

Para determinar o histograma de freqüências da língua portuguesa, executamos o seguinte experimento: selecionamos aleatoriamente textos de autores brasileiros e os dividimos em blocos de 100 KBytes de texto, de um total de 1,1 MByte. O gráfico a seguir mostra a variação da freqüência de ocorrência de cada letra para cada experimento realizado, que consistia em processar um bloco de texto a mais que no anterior, chegando a 11 no final. Note que esta variação no valor das freqüências calculadas é inferior à 10^{-3} (1%) antes mesmo do terceiro experimento (com 3 blocos processados), indicando empiricamente que há uma quantidade mínima de texto necessária (tamanho da amostra) para estimar, com uma desejada precisão, o histograma de freqüências do idioma.

Português do Brasil

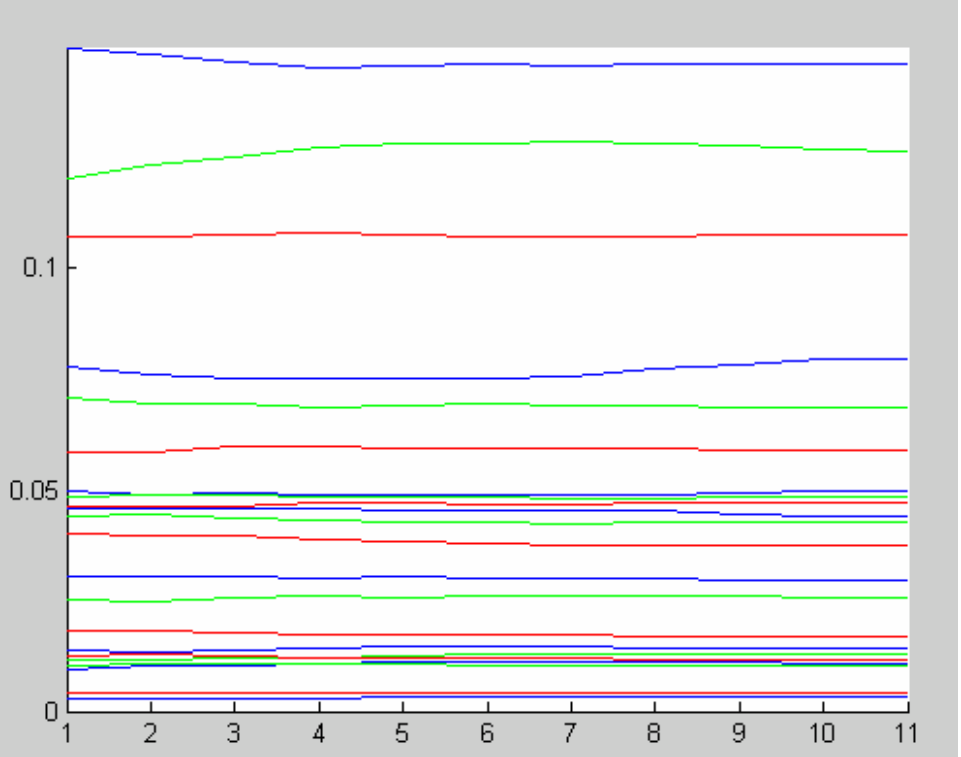

Figura 1.1.1 - Flutuação das freqüências de cada letra após 11 experimentos no Português

```
a 0.1496 0.1485 0.1466 0.1452 0.1459 0.1461 0.1460 0.1462 0.1462 0.1464 0.1464
e 0.1205 0.1236 0.1251 0.1274 0.1283 0.1285 0.1288 0.1284 0.1277 0.1270 0.1267
o 0.1073 0.1074 0.1075 0.1081 0.1075 0.1074 0.1071 0.1072 0.1078 0.1078 0.1078
s 0.0778 0.0762 0.0751 0.0751 0.0752 0.0753 0.0758 0.0774 0.0784 0.0797 0.0797
r 0.0707 0.0694 0.0694 0.0686 0.0689 0.0693 0.0691 0.0689 0.0688 0.0688 0.0687
i 0.0584 0.0587 0.0599 0.0598 0.0596 0.0594 0.0596 0.0593 0.0590 0.0590 0.0591
d 0.0497 0.0488 0.0493 0.0487 0.0487 0.0487 0.0489 0.0490 0.0493 0.0497 0.0498
n 0.0483 0.0487 0.0487 0.0486 0.0484 0.0483 0.0482 0.0481 0.0483 0.0485 0.0485
m 0.0462 0.0464 0.0463 0.0473 0.0471 0.0468 0.0468 0.0471 0.0472 0.0471 0.0470
u 0.0460 0.0460 0.0457 0.0459 0.0455 0.0455 0.0453 0.0452 0.0447 0.0442 0.0441
```

t	0.0442	0.0443	0.0437	0.0433	0.0427	0.0427	0.0425	0.0426	0.0427	0.0426	0.0427
c	0.0401	0.0396	0.0398	0.0389	0.0382	0.0379	0.0377	0.0375	0.0376	0.0376	0.0376
l	0.0305	0.0304	0.0304	0.0302	0.0304	0.0301	0.0301	0.0299	0.0297	0.0295	0.0295
p	0.0251	0.0248	0.0256	0.0259	0.0258	0.0259	0.0261	0.0259	0.0259	0.0258	0.0258
v	0.0183	0.0181	0.0176	0.0173	0.0174	0.0173	0.0173	0.0170	0.0169	0.0168	0.0167
h	0.0138	0.0135	0.0138	0.0143	0.0145	0.0146	0.0145	0.0143	0.0142	0.0142	0.0141
g	0.0115	0.0117	0.0120	0.0121	0.0125	0.0128	0.0129	0.0128	0.0129	0.0129	0.0129
b	0.0125	0.0131	0.0125	0.0119	0.0120	0.0120	0.0119	0.0118	0.0118	0.0116	0.0117
q	0.0093	0.0102	0.0103	0.0109	0.0110	0.0112	0.0112	0.0114	0.0111	0.0109	0.0109
f	0.0105	0.0108	0.0109	0.0107	0.0106	0.0104	0.0104	0.0103	0.0102	0.0102	0.0101
z	0.0041	0.0042	0.0043	0.0043	0.0042	0.0042	0.0041	0.0042	0.0042	0.0042	0.0042
j	0.0029	0.0028	0.0029	0.0030	0.0031	0.0032	0.0031	0.0031	0.0032	0.0032	0.0032
x	0.0025	0.0025	0.0025	0.0024	0.0024	0.0024	0.0024	0.0023	0.0023	0.0023	0.0023
k	0.0001	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001
y	0.0001	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
w	0.0001	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001

Tabela 1.1.1 - Flutuação das freqüências de cada letra após 11 experimentos no Português

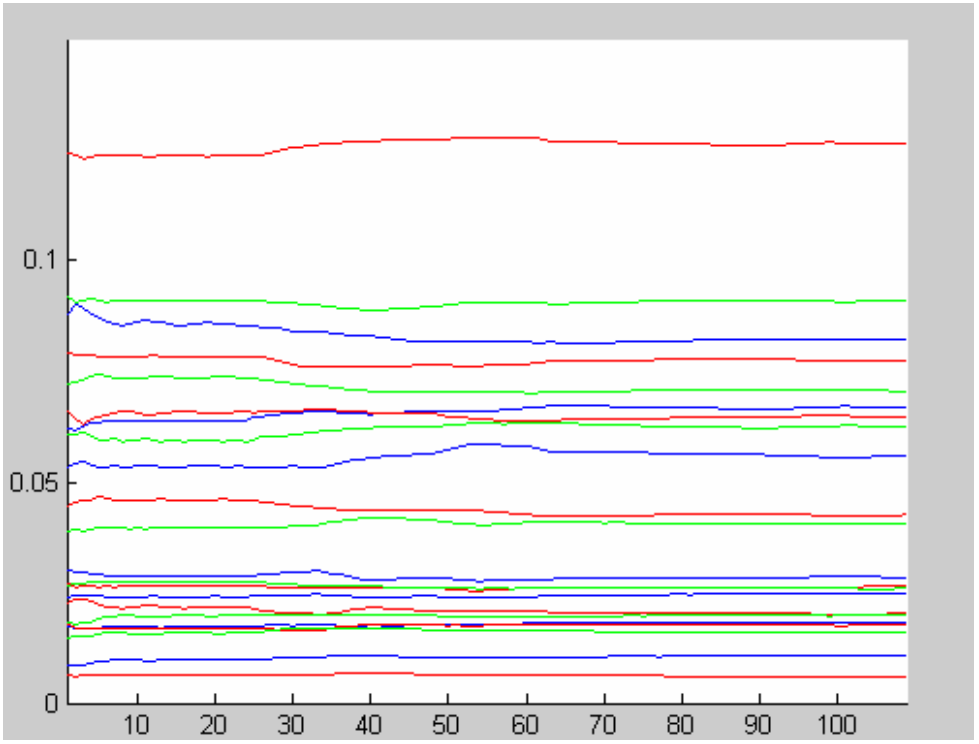

Figura 1.1.2 - Flutuação das freqüências de cada letra após 120 experimentos no Inglês

Observe ainda que o software desenvolvido para esta tarefa desconsidera acentos, bem como interpreta o Ç como a letra C; objetivando que o histograma calculado pudesse ser comparável com os histogramas de outras línguas baseadas no alfabeto latino. Portanto, nosso conjunto de valores para P^1 possui 26 elementos, as letras de A a Z. Para o problema real de cripto-análise tal simplificação não seria válida, pois lidamos com bytes e não letras.

1.2. Histogramas de Freqüências

Português do Brasil

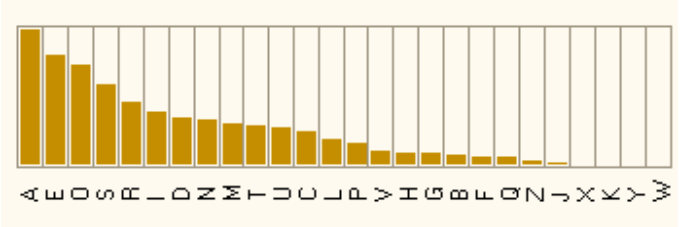

Figura 1.2.1 – Histograma de freqüências para o Português

```
1. A 14.64% E 12.70% O 10.78%
2. S 07.97% R 06.88% I 05.90%
3. D 04.97% N 04.85% M 04.71% U 04.42% T 04.26% C 03.76%
4. L 02.95% P 02.58%
5. V 01.68% H 01.42% G 01.29% B 01.16% Q 01.09% F 01.02%
6. Z 00.42% J 00.32% X 00.23%
7. K 00.01% Y 00.01% W 00.01%
```

Tabela 1.2.1 – Agrupamento das letras por similaridade das freqüências para o Português

Inglês Americano

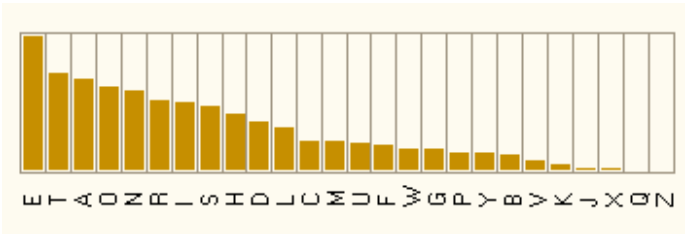

Figura 1.2.2 – Histograma de freqüências para o Inglês

```
1. e 12.66%
2. t 09.08% a 08.21% o 07.74% n 07.05%
3. i 06.69% r 06.47% s 06.27% h 05.61%
4. d 04.25% l 04.07%
5. c 02.84% u 02.65% m 02.59% f 02.49%
6. w 02.02% g 01.98% p 01.80% y 01.78% b 01.60%
7. v 01.06% k 00.59% x 00.19% j 00.15% q 00.10% z 00.07%
```

Tabela 1.2.2 – Agrupamento das letras por similaridade das freqüências para o Inglês

1.3. Interpretação dos Dados Coletados

Tais dados podem ser interpretados de diferentes maneiras, dependendo do problema ao qual eles se aplicam. No entanto, nos interessa ter certeza de que esses dados são confiáveis e que ocorrem nos textos de mesmo idioma, bem como saber qual o tamanho mínimo e características de um texto tal que seja possível reconhecer em que idioma ele está escrito usando o seu histograma.

A princípio, problema se resumiu a obter uma amostra de textos pertencentes a uma mesma população de textos de língua portuguesa. Pela hipótese, eles devem possuir um histograma suficientemente similar se pertencentes ao mesmo idioma. Textos técnicos, que possuem muitas palavras estrangeiras (geralmente do inglês) vão possuir uma variação maior na freqüência de algumas letras (principalmente o K, Y e W), mas não o suficiente para descaracterizar totalmente o histograma pois a grande maioria das palavras ainda pertencem ao português. Uma variação menor ocorreria para textos em Galego e outra ainda menor para textos do Português de Portugal.

Assim, podemos também dizer que os indivíduos dessa população possuem 26 variáveis aleatórias correlacionadas (P^A a P^Z), que os destingüem de textos de outros idiomas; isso somente se houver uma média de freqüência e uma variação suficientemente pequena para cada uma destas variáveis aleatórias.

Ao tomarmos uma amostra de uma população várias vezes (*n*), e medirmos a média da amostra M a cada vez, temos como montar a distribuição empírica da média (variável M), obtendo um E(M). E é sabido que essa média das amostras M é uma estimativa não tendenciosa da média da população (pois os textos foram selecionados) e possui aproximadamente distribuição normal. Não é preciso fazer isso para a o cálculo da freqüência média em si pois já o fizemos utilizando o método da seção anterior que mostrou graficamente a tendência para o valor médio, mas se nos interessa calcularmos a variância da freqüência amostral, vamos precisar analisar os histogramas das amostras de texto calculados isoladamente.

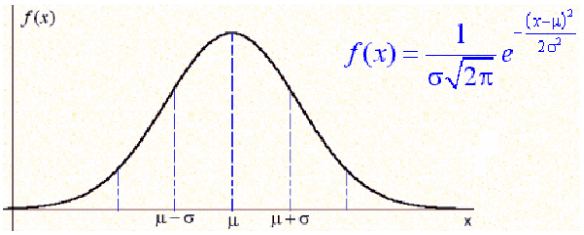

Figura 1.3.1 – Curva Normal

2. Aplicações do Histograma da Língua

2.1. Determinação do Idioma de um Texto

Uma vez calculado o histograma de freqüências da língua, podemos utilizar uma técnica estatística denominada distribuição Qui-quadrado para determinar se um texto está escrito neste idioma.

A distribuição Qui-quadrado permite que testemos a igualdade de mais do que duas médias (ou proporções) simultaneamente. *Exemplo*: se lançarmos um dado 60 vezes, sabemos que cada face deveria ter uma média de 10 aparecimentos; nossa hipótese é: a

distribuição de freqüências da variável aleatória de cada face é a mesma. Se não o for, o dado é viciado.

	X_1	X_2	X_3	X_4	X_5	X_6
o_i	15	7	4	11	6	17
e_i	10	10	10	10	10	10

H_0: $p_1 = ... = p_6 = 1/6$ ($p=e_i/n$)
10:10:10:10:10:10

Medida Qui-Quadrado
$$X^2 = \sum_{i=1..k} (o_i - e_i)^2/e_i$$

A tabela da distribuição Qui-quadrado pode ser consultada pelo grau de confiança (em geral, 0,05) e o grau de liberdade (v = k - 1). Todas as medidas e_i devem exceder 5.

Analogamente, podemos considerar a freqüência de cada letra e compará-las com as freqüências do histograma da língua usando o mesmo procedimento. Note também que se avaliarmos um texto sabidamente em português e o mesmo não for aprovado neste teste, significa que o texto não possui tamanho suficiente e seu histograma é, certamente, deturpado.

2.2. Cripto-análise

Considerando as cifras simétricas simples, ou seja, as de transposição e as de substituição, um ataque de texto conhecido quebra a cifra com apenas a comparação do histograma de freqüências da mensagem com o da mensagem cifrada, que terá os mesmos valores. Logo, os dados calculados neste artigo são apenas necessários em ataques de texto desconhecido contra cifras deste tipo.

Assim, processando o histograma de uma mensagem cifrada cujo conteúdo seja desconhecido, e ordenando-o por freqüência (da mesma forma como o histograma da língua foi processado), podemos determinar a correspondência entre os símbolos e descobrir a cifra, seja ela de transposição ou substituição. As cifras de transposição são mais simples de quebrar, e ainda por cima são um caso especial das cifras de substituição. Há 26! cifras de substituição possíveis, o que justifica aplicação de técnicas mais elaboradas que a força bruta.

Sendo de substituição, pode haver uma confusão na decisão entre letras que possuem naturalmente a freqüência parecida, principalmente as do agrupamento 3, 4, 5 e 6 da tabela 2. Uma solução estatística para esse problema seria também dispor do histograma de digramas da língua e do texto, e fazer paralelamente a correspondência entre eles. Quanto menor a mensagem cifrada, mais histogramas de n-gramas poderiam tornar-se necessários na quebra da cifra de substituição.

Note que para cifras como a Viginère, que usa um número *n* fixo de cifras de substituição diferentes repetidamente, somente a abordagem por histograma de letras é

viável. E antes, este valor *n* deve ser descoberto por força-bruta, o que não é tão difícil pois uma vez tentado o valor certo de *n*, os histogramas do texto cifrado por cada letra vão coincidir consideravelmente.

3. Código-Fonte

Gerador de Histograma em Perl com saída para MatLab

```
# freq.pl
# Levantamento de estatisticas sobre textos.
# Considera apenas as letras do alfabeto norte-americano (A-Z), case insensitive.

%casting = ("á", "a",  "à", "a",  "ä", "a",  "â", "a", "ã", "a",
               "é", "e",  "è", "e",  "ë", "e",  "ê", "e",
               "í", "i",  "ì", "i",  "ï", "i",  "î", "i",
               "ó", "o",  "ò", "o",  "ö", "o",  "ô", "o", "õ", "o",
               "ú", "u",  "ù", "u",  "ü", "u",  "û", "u",
               "ç", "c",  "ñ", "n",
               "Á", "a",  "À", "a",  "Ä", "a",  "Â", "a", "Ã", "a",
               "É", "e",  "È", "e",  "Ë", "e",  "Ê", "e",
               "Í", "i",  "Ì", "i",  "Ï", "i",  "Î", "i",
               "Ó", "o",  "Ò", "o",  "Ö", "o",  "Ô", "o", "Õ", "o",
               "Ú", "u",  "Ù", "u",  "Ü", "u",  "Û", "u",
               "Ç", "c",  "Ñ", "n");

$count=0;
$count_dg=0;
$end = 0; $loop_count = 0;

while(<STDIN>){
    chop ;
    tr/A-Z/a-z/;
    @x=split(//,$_);
    foreach $y(@x) {
      $y = ($casting{$y} eq "") ? $y : $casting{$y};
        $y =~ tr/a-z//cd; #if ($y eq "") { print "$yu ($casting{$yu}, $ye)"; }
    }
    $lastchar="" ;

   foreach $y(@x) {
     if ($y ne "") {
         # Contagem de Letras
        $cc{ $y } += 1 ;
        $count += 1 ;

        # Contagem de Digramas
        $dg=$lastchar.$y ;
        $dgcount{ $dg } += 1 ;
        $count_dg += 1 ;
        $lastchar=$y ;

        if (($count % (100*1024)) == 0) { &showresults; }
    }
   }
}

sub byletterfrequency { $cc{$b} <=> $cc{$a}; }
sub bydgfrequency { $dgcount{$b} <=> $dgcount{$a}; }

$end = 1;
&showresults;

sub showresults
{
 @colors = ( "b", "r", "g" );
```

```
@sortedbyfrequency=sort byletterfrequency a..z;
@digramsbyfrequency=sort bydgfrequency keys(%dgcount);

$sum=0; $tt=1; $loop_count++;
print "Frequencia de Letras ($count caracteres avaliados):\n";

if ($end != 0) {
        print "hold on\n";

        print "x = [";
        for ($i=1; $i<=$loop_count;$i++) {
          print " $i ";
        }

        print "];\n";
        print "axis( [1 $loop_count 0 0.15]);\n";
}

while($letter=shift(@sortedbyfrequency))
{
   $ct++;
   $matrix{$letter}[$loop_count] = $cc{$letter}/$count;

   if ($end == 0) {
       printf("%s %2.4f", $letter,$cc{$letter}/$count);
       printf(" %d", $cc{$letter});
       printf("\t");
       if ($tt++ % 5 == 0) {print "\n" ;}
   } else {
       $i=1;
       printf("$letter = [%2.4f ", $matrix{$letter}[$i]);
       for ($i=2; $i<=$loop_count; $i++) {
              printf("%2.4f ", $matrix{$letter}[$i]);
       }
       print("];\n");
       print("plot(x, $letter, \'$colors[$ct%3]\');\n");
   }
   $sum += $cc{$letter}/$count;
}

print "\nTotal: $sum\n\n";

print "Frequencia de Digrafos:\n";

$ct=0 ;
$sum=0; $tt=1;

while($dg=shift(@digramsbyfrequency))
{
  $matrix{$dg}[$loop_count] += $dgcount{$dg}/$count_dg;

   if ($end == 0) {
       printf("%2s %.2f\t", $dg,$dgcount{$dg}/$count_dg);
       printf(" %d", $dgcount{$dg});
       printf("\t");
       if ($tt++ % 5 == 0) {print "\n" ;}
   } else {
       $i=1;
       printf("$letter = [%2.4f ", $matrix{$dg}[$i]);
       for ($i=2; $i<=$loop_count; $i++) {
              printf("%2.4f ", $matrix{$dg}[$i]);
       }
       print("];\n");
       print("plot(x, $letter, \'$colors[$loop_count%3]\');\n");
   }

  $sum += $cc{$dg}/$count_dg;
}
print "\n$sum\n" ;
```

```
}
```

4. Referências

HOEL, P. *Estatística Elementar*, Editora Atlas, 1976.
SHANNON, C. *Communication Theory of Secrecy Systems*, 1949.