The Strength of the Bayes Score

The signage of the Bayes score is derived under the assumption that the score is normally distributed in right and wrong cases. Asymptotically there is a constant scoring rate per bit, and that rate is determined. Textlengths needed to attain certain signages for common attacks are calculated. The authors verify the accuracy of these textlength calculations (given the validity of the underlying mathematical model).

1. INTRODUCTION

It is well known that the signage of the approximate Bayes score for a regularly stepping machine (number of standard deviations between right and wrong case means) is equal to $\sqrt{\frac{T}{r}}$ (r the expected value of the square of the putative bulls, T the textlength).

---

*Originally submitted No. 284 of 8 September 1970; this paper won First Prize in the 1971 Crypto-Mathematics Institute Essay Contest.

NSAL-S 188,795

SECRET
(1) We prove for a lower bound model that the expected value of the factor in the right case has the form $CS^T$. Our model is essentially an exact model, although the score on $T$ bits is weaker than the Bayes score on 772 bits.

(2) We are able to determine $C$ and the value of $T$ is important. Were it substantially larger or smaller than $1$, the asymptotic effect takes place, a better approximation to the significance is given by

(3) We are able to do something else which

and this is to have some control over the accuracy of our estimates.

We are then asked to choose between two conflicting hypotheses, $H_1$ and $H_2$, where $H_1$ is the hypothesis that $p(X) = 0$ for all $1 \leq t \leq T$. $H_2$ is a bit more complicated. We define a probability function $P$ on the set $S^T$ (all $T$-long

(4) Given the usual assumptions about scores being normally distributed, we have been able to calculate the significance of the Bayes score. Before the asymptotic effect takes place, a better approximation to the significance is given by

(5) We are able to do something else which

and this is to have some control over the accuracy of our estimates.

We have not attempted to describe a specific COMSEC situation to which these results apply for the following reason, which, in fact, substantially limits the practical value of our findings:

We think of $[K]$ as the observations from a sequence of independent Bernoulli random variables $[X]$. We are then asked to choose between two conflicting hypotheses, $H_1$ and $H_2$, where $H_1$ is the hypothesis that $p(X) = 0$ for all $1 \leq t \leq T$. $H_2$ is a bit more complicated. We define a probability function $P$ on the set $S^T$ (all $T$-long

The important question left unanswered by this paper, and one which, it is hoped, will be the subject of a future one, is the extent to which the Bayes scores

For example, it turns out that for the Bayes score one

We summarize the asymptotic results:

It is well-known that the Bayes score will always require less textlength than the

score. Our results indicate that the significance obtained from a textlength of $T$ with a score can be attained with a textlength of approximately

This result is especially important for primary attacks where $n$ is small and the required textlength for scoring might not be available. The Bayes score will never be cheaper to compute than the score, for the work involved in calculating the factor on a textlength of $T$ is on the order of rather than $T$. A measure of the "efficiency" of over Bayes is thus given by $U/T$, which our results show to be the constant

this factor times the work needed to attain a certain significance gives the work needed to attain a certain Bayes significance.

We have not attempted to describe a specific COMSEC situation to which these results apply for the following reason, which, in fact, substantially limits the practical value of our findings:
sequences from $S$) so that if $Y = |Y|, \delta^T$

To test which of the two hypotheses is true, a scoring function $S$ (from $S'$ to the reals) is proposed and a threshold $U$ is set so that $H_0$ is accepted if $S(K) > U$ and $H_1$ is accepted if $S(K) < U$. There are, of course, two possibilities for error: we may accept $H_0$ when $H_1$ is true (type II error) or accept $H_1$ when $H_0$ is true (type I error). The celebrated Neyman-Pearson Lemma (see, e.g., [2], p. 66) suggests that the "best" score is

$$S(K) = \frac{\text{Prob}[H_0|K]}{\text{Prob}[H_1|K]}$$

in the following sense: if $S_1$ is some other scoring function, and thresholds $U_1$ and $U_2$ are chosen for the respective scores so that the probabilities of type I error are $a$, then the probability of a type II error using $S_1$ is less than or equal to that using $S$. What these thresholds are, and how small the probabilities of type II error then become, depend on knowledge of the distribution of the scoring function.

The scoring function $S_1$ is commonly called the Bayes factor, since in order to obtain a posteriori odds in favor of $H_1$ from the a priori odds, one multiplies by $S_1(K)$ (in particular,

$$\frac{\text{Prob}[H_1|K]}{\text{Prob}[H_0|K]} = S_1(K) \cdot \frac{\text{Prob}[H_1]}{\text{Prob}[H_0]}$$

A unique "best" scoring function does not exist. In fact, it is easy to see that if $f$ is monotonic increasing, the composite of $f$ with $S_1$ is as good as $S_1$ in the sense described above. In particular, a common scoring function, as good as the Bayes factor, is $S_1(K) = \log \left( \frac{\text{Prob}[H_1|K]}{\text{Prob}[H_0|K]} \right)$, commonly called the Bayes score. For the all-important example we described in our first paragraph, $S_1(K) = 2 \sum P(Y) \cdot \text{Prob} [\text{K}],$...
number is also the number of standard deviations between

the means, and from this it is easy to derive the probability of a type II error for a given type I error (because the scores are normally distributed).

III. THE UNESCRIBED BAYES FACTOR

Our intention was to calculate mean and standard deviation of the Bayes factor, but this is complicated by the peculiar nature of the probability function $P$. It is possible that the techniques we employ in sections IV and V could be modified to apply to the exact Bayes factor, but the calculations would certainly be more cumbersome.

For this reason we propose a fourth and fifth score, which we shall define shortly. We do not expect that these scores should ever be calculated for an actual key stream and hypothesis $H_1$, for the work would be comparable to the work in calculating the exact Bayes factor, and we know with certainty that the exact Bayes factor is the better score; however, the scores which we will introduce are closer in spirit to the Bayes score than is the score, and we will be at least partially successful in calculating their means and variances.

Our notation is awkward in that the second usage of $m$ depends on the $y$ currently being summed over, but the typist has already been overworked and the meaning of $m$ here and later should always be clear from context. This suggests that we define a new score $S$, by

which we call the Bayes factor, and $S_{(K)} = \log S_{(K)}$.

the Bayes score $S_{(K)}$ is intuitively a weaker scoring function because it ignores the effect of the Neyman Pearson Lemma.

IV. EXPECTED VALUE OF THE BAYES FACTOR

From this point on the terms Bayes factor and Bayes score will refer to the factor and score, and we set $N = \frac{T}{2}$. If $H_1$ is true, then for each $1 \leq t \leq N$, $\text{Prob}(X, = 0) = \frac{1}{2}$; so

and $E[S_{(K)}] = 1$.

Next suppose that $H_1$ is true and let

$|X| \sim S$. For $1 \leq t \leq N$ we say that the pair of sequences $Y$ and $Z$ match at $t$ if

$|Y| = |Z|$ and $|Z|$ match at $t$.
BAYES SCORE

If there is a match at \( t \), then \( \text{Prob}(\text{match at } t) = \frac{1}{2^t} \) and if there is no match at \( t \), then

If we are willing to assume that \( \mathbb{E}[t] = 0 \), we get that in most cases we can set \( \mathbb{b} = 1 + \mathbb{E}[t] \). It follows that \( \mathbb{E}[1 + (-1)^{t+2} Z^t \omega] = \mathbf{s}^t \), where \( k \) is the number of matches between \( \mathbb{Y} \) and \( \mathbb{X} \). If we let \( \mathbb{p}(N, k) \) be the probability that a pair of randomly selected elements of \( \mathcal{S}^N \) will have \( k \) matches, then

We devote the rest of this section to the calculation of the right hand expression, which we denote by \( \mathbb{L}(N, \mathbf{b}) \).

Let \( (U, V) \), \( S^N \times S^N \), with \( U = (U_1, \ldots, U_N) \) and \( V = (V_1, \ldots, V_N) \).

Let \( Q(N, k) \) - number of elements \((U, V)\) with exactly \( k \) matches, so

\[
Q(N, k) = \mathbb{L}(N-1, k) \quad \text{for} \quad 1 \leq k \leq N - 1.
\]

Because \( \mathbb{L}(N, k, N) = 0 \) for \( k \geq 1 \), this together with (1) yields

We next define a triangular array of integers, \( M(i, j) \), for \( 0 \leq i \leq l \), as follows:

\[
M(0, 0) = 1 \\
M(0, i) = 0 \\
M(i, j) = \sum_{\ell=0}^{j-1} M(i, \ell).
\]

Some values of \( M(i, j) \) are

<table>
<thead>
<tr>
<th>( i )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>15</td>
<td>19</td>
<td>25</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>14</td>
<td>14</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>42</td>
<td>28</td>
<td>14</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>132</td>
<td>90</td>
<td>48</td>
<td>20</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>429</td>
<td>297</td>
<td>165</td>
<td>75</td>
<td>27</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1490</td>
<td>1001</td>
<td>572</td>
<td>275</td>
<td>110</td>
<td>35</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>4862</td>
<td>3432</td>
<td>2002</td>
<td>1001</td>
<td>429</td>
<td>154</td>
<td>44</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The importance of the \( M(i, j) \) is that

We prove this by induction on \( N \). The result is trivial for \( N = 0 \). For \( N = 1 \) it becomes

Since the result clearly holds for \( l = N \) and \( l = 0 \), we may assume \( 1 \leq l < N - 1 \) and argue that

Let us count the \( Q(N, k) \) pairs of \( N \) sequences with \( k \) matches in terms of their initial bits. There are

Adding these together and applying (4), we get for \( k \geq 1 \)
or translating into "p" notation

This last formula is also valid only for \( k \geq 1 \). If we give \( p(0, k) \) its natural interpretation, we can trivially generalize the above formula to all \( k \) by writing

\[
1 = \sum p(N, k)
\]

In fact, since

\[
1 = \sum \, p(N, k)
\]

it follows that

\[
1 = \sum \, p(N, k)
\]

We next find a closed form expression for \( M(1, i) \), which we for notational simplicity refer to as \( M(i) \). The \( M(i) \) possess a convolution property

\[
M(i) = \sum_{j=0}^{i} M(j) M(i-j).
\]

To prove (6), we observe from (3) that

\[
M(0) = M(0,0) + M(1,0) \quad 1 \leq i \leq i \geq 2.
\]

Using this, we can prove by induction that

\[
M(i+1) = \sum_{j=0}^{i+1} M(i+1, j+1) \quad i \geq 1, i \geq 1.
\]

Now, by repeated use of this equation, we get

\[
M(i) = M(i-1) + \sum_{j=0}^{i-1} M(j, i-1)
\]

\[
= M(1) M(i-1) + \sum_{j=1}^{i-1} M(j+1, i-1) M(j+1, i-1)
\]

(since \( M(j, i) = 1 \))

By factoring out \( M(i-1) \) from (5), we get

\[
M(i) = M(i-1) + \sum_{j=1}^{i} M(j+1, i-1) M(j+1, i-1)
\]

(since \( M(j, i) = 1 \))

By factoring out \( M(i-1) \) from (6), we get

\[
M(i) = M(i-1) + \sum_{j=1}^{i} M(j+1, i-1) M(j+1, i-1)
\]

(since \( M(j, i) = 1 \))

Formula (6) having been verified, it follows that the power series

\[
g(x) = 1 + \sum_{j=0}^{\infty} \frac{-M(j)}{2 \cdot 4^{j}} x^j
\]

which converges in some open neighborhood about the origin, has the property that \( g(x) = 1 - x \), so it follows from the generalized binomial theorem, since \( g(x) = (1-x)^{-1/2} \), that

\[
M(i) = (1-x)^{-1/2} \quad \text{(8)}
\]

Recalling that the \( x \) appearing in (5) had the property

\[
M(i) = (1-x)^{-1/2} \quad \text{(8)}
\]
perform by induction that

For \( N = 1 \), (10) is clear. Assume it holds for \( N - 1 \). Then

(by (9))

EO 1.4. (c)
EO 1.4. (d)
(by induction hypothesis)
(by (9))

Combining (5), (8) and (10) we get that, for \( N \geq 0, k \geq 0 \)

We now let \( A(x,y) \) be the power series in two indeterminants defined by 

\[ h(x,y) = \sum p(N,k) x^N y^k \]

which converges in the region \( |x| < |y|^{-1} \). It follows from (11) and the fact that \( \sum \binom{1/2}{i} (-1)^i x^i = 1 - x^{1/2} \) that

\[ g(x,y) = h(x,y) - 1/4 xyh(x,y) + (1-x)^{1/2} (h(x,y) - 1) + 1/4 x h(x,y) \]

has the property that

\[ g_{N,k} = -2N \binom{1/2}{N} (-1)^N, N \neq 0. \]

\[ g_{N,0} = 0, k \neq 0. \]

\[ g_{N,0} = -2N \binom{1/2}{N} (-1)^N, N \neq 0. \]

\[ g_{N,0} = 0, k \neq 0. \]

\[ g_{N,0} = -2N \binom{1/2}{N} (-1)^N, N \neq 0. \]

Even more important, \( h(x,0) = \sum E(N,k) x^k \).

Let

\[ h(x,y) - x h(x,y) - x^2 \left( \frac{y-1}{4} \right)^2 h(x,y) = 1 - \frac{x(y-1)}{4 \sqrt{1-x}}. \]
Since \((1-a)^{-1} = \sum_{n=0}^\infty a^n\), comparing coefficients yields

\[
L = \sum_{n=0}^\infty \frac{(-1)^n}{n+1}.
\]

Let us look at the factor in the third term.

We can rewrite our previous equation, getting

\[
\text{From (12) it is easy to see that in the sense that the percentage error goes to 0 as } N \text{ gets large. More precisely:}
\]

\[
(13)
\]

(In section 7 we will prove the stronger result

\[
(14)
\]

able to estimate the accuracy of the

We here prove (13). It is obvious that

\[
\text{vorge of } \sum_{n=0}^\infty \left(\frac{-1}{i} \right) (-y)^i \text{ that } \lim_{n \to \infty} \sum_{i=1}^n \left(\frac{-1}{i} \right) (-y)^i = 0.
\]

Finally, pick some \(e\) with \(y < e < 6\). Then the series on the right hand side converges, and since \(y < 6\), \(\lim_{n \to \infty} \text{right hand side is 0.}

1.4. (c) VARIANCE OF THE BAYES FACTOR

Recalling the definition of \(S_i^2(K)\), we see that \(S^2(K)\) is equal to

We first take up the case where \(H_i\) is true. Then if \(Z\) and \(Z'\) match at

and otherwise the expected value is 1/4. It follows that the value of

\[
\text{E}[S^2(K)] \text{ is given by}
\]
We have illustrated by this argument that the variance of the score when $H_1$ is true equals the mean of the score when $H_0$ is true. This is actually true in general for any score of the form $\frac{\text{Prob}[K|H_1]}{\text{Prob}[K|H_0]}$ and can be proved by an elementary argument (see, e.g., [9]).

Next we do the case where $H_1$ is true. It is an easy exercise to show that the value of

$$p$$

where $k$ is the number of matches among $Y$, $Z$ and $Z'$ (if the three sequences match at $i$ this counts as three matches), and thus

$$p(N, k)$$

is the probability that three $N$-long sequences chosen at random will have exactly $k$ matches. The authors have been unable to calculate the density function $q$; the problem seems similar to that of calculating the function $p$ of chapter IV, but involves three-dimensional arrays rather than the two-dimensional array $M(i,j)$.

Without doing any hard work we can get a lower bound on the variance. Let $A$ be the random variable representing the number of matches between the first two sequences, $B$ the number of matches between the first and third, $D$ the number of matches between the second and third. $A$, $B$ and $D$ are identically distributed but patently not independent.

The authors have been unable to come up with a useful upper bound.

Summarizing the results of sections 4 and 5 (the reader should check that the $C$ and $C'$ corresponding to $b'$ are approximately the same as those corresponding to $b$), we get

$H_1$ true

$H_0$ true

mean of $S$: 1

variance of $S$: (upper bound)

since $E [S(K)]$ dominates $(E [S(K)])^2$.

VI. SIGMA OF THE BAYES SCORE

It has become part of the COMSEC folklore that log factors tend to be normally distributed, especially if scoring rates per bit are constant and low. For a further discussion of this principle, see [9]. Throughout the rest of this section, we assume that $S$ is normally distributed when either $H_0$ or $H_1$ is true. We seek the mean and standard deviation in the two cases.

103
$\mu$ and $\sigma'$ be the mean and standard deviation of $S$. It follows from the form of the moment generating function of a normal distribution (again, see [9]) that

$$\ln \sigma' = \mu + \frac{\sigma'}{2}$$
$$\ln(\sigma)^2 = 2\mu + 2\sigma^2.$$  

We have seen that $H_1$ is true, $\mu' = 1$ and whereas when $H_0$ is true, $\mu = 0$ and $\sigma = \sigma'$. Solving the simultaneous equations, we get:

\[
\begin{align*}
  H_1 \text{ true} & : \\
  H_0 \text{ true} & :
\end{align*}
\]

For very large $T$ this gives us the signages referred to in section 1. The effect of $\ln \sigma'$ being a positive number, is to keep the signage greater than.

Since there is a constant scoring rate per bit asymptotically, there is little advantage when $T$ is large.

We have calculated mean and variance for the score in the second case. We can approximate the first case by saying that the score we are deriving.

To conclude this section we present tables which give for common values of $\alpha$ the text length needed to attain certain signages using the formulas.

The reader should understand the following reservations before using these tables:

1. The text length needed to attain a certain signage of the Bayes score is greater than that needed for the true Bayes score, but not more than twice as much.

2. The values we have listed in the Bayes column are only lower bounds for the text length needed within the mathematical model we have set up, due to the independence assumption which we made at the end of section V.

3. The values listed in the Bayes column are accurate only to the extent to which the error term discussed in the next section is small.

4. We have assumed that the Bayes score is normally distributed.
In section VII it will be shown that all entries in our tables are accurate to $\pm 0.1 \sigma$ (that is, the Bayes textlength which we purport to give rise to a signagme of $n$ does indeed give rise to a signagme of $n \pm 0.1$). Note that were an entry for $1\sigma$ in the Bayes column computed, it would have been a negative textlength; this absurdity can be explained by the fact that for the textlength $T$ which has the property that

VII. ACCURACY

We have seen in section IV that

and in fact get an upper bound for $P(N)$. We refer to equation (12) in section IV. Since $|a| < 1$, it is clear that

We next estimate
By comparing
\[
\left( \frac{m}{n} \right) = \left( -1 \right)^{\left\lfloor \frac{n}{m+1} \right\rfloor} \cdot \frac{1}{n} \left( 1 - \frac{m+1}{n} \right)
\]
(15)
with Weierstrass's infinite product expansion for the reciprocal of the gamma function
\[
\frac{1}{\Gamma(x)} = x e^{\gamma} \prod_{n=1}^{\infty} \left( 1 + \frac{x}{n} \right)^{\frac{1}{e^{n}}} \left( 1 - \frac{1}{n} \right)
\]
(1 = Euler's constant)
and taking \( x = m - 1 \) it is easily seen that
\[
\lim_{n \to \infty} \left( \frac{m}{n} \right) N^{n-1} \left( -1 \right)^x = \frac{1}{\Gamma(-m)} .
\]
In particular
\[
\lim_{N \to \infty} \left( \frac{m}{n} \right) N^{n-1} \left( -1 \right)^x = \frac{1}{\Gamma(-m)} .
\]
(16)
It is clear that \( \left( \frac{m}{n} \right) \) decreases as \( i \) becomes larger, so

The sequence \( \sqrt{N} \left( 1 - \frac{1}{2N} \right) \) is increasing, since
\[
\sqrt{N} \left( 1 - \frac{1}{2N} \right) > \sqrt{N-1} .
\]
Hence,
\[
\sqrt{N} \left( 1 - \frac{1}{2N} \right) \leq \frac{1}{\sqrt{N}}
\]
by (16),

and
\[
\sqrt{N} \left( 1 - \frac{1}{2N} \right) \leq \frac{1}{\sqrt{N}}
\]
Next we estimate

Although the series does not converge, by Taylor's theorem with a remainder we can write
\[
R_N = \frac{\left( -\frac{1}{2} \right) \left( -y^{x+1} \right) \left( N+1 \right)}{(1-y)^{x+1}}
\]
for some \( y \) with \( 0 < y < 1 \). Since

we need only estimate \( R_N \). We have
\[
|R_N| \leq \frac{\left( -\frac{1}{2} \right) \left( -y^{x+1} \right) \left( N+1 \right)}{(1-y)^{x+1}}
\]
(16)
and thus that:

Finally, This proves that:

and thus that:
\[
F(N) = O \left( \frac{1}{\sqrt{N}} \right)
\]
By using formula (12) and our previous estimates we can get a more precise upper bound for \( F(N) \), namely:
For small $a$...