1 Recognizing Plaintext: FRIEDMAN's Most-Frequent-Letters Test

We begin with the first question: *Does a given text belong to a certain language?* FRIEDMAN gave a quite simple procedure for distinguishing valid text from random noise that works surprisingly well, even for short texts. Besides it makes a smooth introduction to statistical test theory.

FRIEDMAN's **Procedure**

Assume we are given a string of letters and want to decide whether it is a part of a meaningful text (in a given language, say English), or whether it is random gibberish. Our first contact with this problem was the exhaustion attack against the simple shift cipher that produced 26 strings, exactly one of which represented the correct solution. Cherry-picking it was easy by visual inspection. But for automating this decision procedure we would prefer a quantitative criterion.

Such a criterion was proposed by FRIEDMAN in Riverbank Publication No. 16 from 1918 [3]. The procedure is

1. Identify a set of most frequent letters from the target language. For English take `ETOANIRSHD` that make up 73.9% of an average English text but only $10/26 \approx 38.5\%$ of a random text.

2. Count the cumulative frequencies of these most-frequent letters for each of the candidate strings.

3. Pick the string with the highest score. If this doesn't work, also consider the next highest scores.

Example. For the CAESAR example in Section 1.3 the scores are in Table 1. We immediately see that the correct solution `CAESAR` has the highest score (even if this is not a genuine English word).

The example shows that FRIEDMAN's procedure seems to work well even for quite short strings. To confirm this observation we analyze the distribution of the Most-Frequent-Letters scores—in short **MFL scores**—for strings of natural languages and for random strings. First we consider this task from a theoretic viewpoint, then we also perform some empirical evaluations.

The distribution of MFL Scores

Consider strings of length r over an alphabet Σ whose letters are independently drawn with certain probabilities, the letter $s \in \Sigma$ with probability p_s. Let $\mathcal{M} \subseteq \Sigma$ be a subset and $p = \sum_{s \in \mathcal{M}} p_s$ be the cumulative probability

Table 1: Friedman *scores for the exhausion of a shift cipher*

FDHVDU	3		OMQEMD	3		XVZNVM	1	
GEIWEV	3		PNRFNE	4	<---	YWAOWN	3	
HFJXFW	1		QOSGOF	3		ZXBPXO	1	
IGKYGX	1		RPTHPG	3		AYCQYP	1	
JHLZHY	2		SQUIQH	3		BZDRZQ	2	
KIMAIZ	3		TRVJRI	4	<---	CAESAR	5	<===
LJNBJA	2		USWKSJ	2		DBFTBS	3	
MKOCKB	1		VTXLTK	2		ECGUCT	2	
NLPDLC	2		WUYMUL	0				

of the letters in $\mathcal{M}$. The **MFL score** of a string $a = (a_1, \ldots, a_r) \in \Sigma^r$ with respect to $\mathcal{M}$ is

$$N_{\mathcal{M}}(a) = \#\{i \mid a_i \in \mathcal{M}\}.$$

To make the scores for different lengths comparable we also introduce the **MFL rate**

$$\nu_{\mathcal{M}}(a) = \frac{N_{\mathcal{M}}(a)}{r}.$$

The MFL rate defines a function

$$\nu_{\mathcal{M}} \colon \Sigma^* \longrightarrow \mathbb{Q}.$$

(Set $\nu_{\mathcal{M}}(\emptyset) = 0$ for the empty string $\emptyset$ of length 0.)

The distribution of scores is binomial, that is the probability that a string $a \in \Sigma^r$ contains exactly k letters from $\mathcal{M}$ is given by the binomial distribution

$$P(a \in \Sigma^r \mid N_{\mathcal{M}}(a) = k) = B_{r,p}(k) = \binom{r}{k} \cdot p^k \cdot (1-p)^{r-k}.$$

Random strings. We take the 26 letter alphabet `A...Z` and pick a subset $\mathcal{M}$ of 10 elements. Then $p = 10/26 \approx 0.385$, and this is also the expected value of the MFL rate $\nu_{\mathcal{M}}(a)$ for $a \in \Sigma^*$. For strings of length 10 we get the two middle columns of Table 2.

English strings. Assuming that the letters of an English string are independent is certainly only a rough approximation to the truth, but the best we can do for the moment, and, as it turns out, not too bad. Then we take $\mathcal{M} = \{$`ETOANIRSHD`$\}$ and $p = 0.739$ and get the rightmost two columns of Table 2.

Table 2: *Binomial distribution for $r = 10$. The columns headed "Total" contain the accumulated probabilities.*

		$p = 0.385$ **(Random)**		$p = 0.739$ **(English)**	
Score	Coefficient	Probability	Total	Probability	Total
0	$B_{10,p}(0)$	0.008	0.008	0.000	0.000
1	$B_{10,p}(1)$	0.049	0.056	0.000	0.000
2	$B_{10,p}(2)$	0.137	0.193	0.001	0.001
3	$B_{10,p}(3)$	0.228	0.422	0.004	0.005
4	$B_{10,p}(4)$	0.250	**0.671**	0.020	**0.024**
5	$B_{10,p}(5)$	0.187	0.858	0.067	0.092
6	$B_{10,p}(6)$	0.097	0.956	0.159	0.250
7	$B_{10,p}(7)$	0.035	0.991	0.257	0.507
8	$B_{10,p}(8)$	0.008	0.999	0.273	0.780
9	$B_{10,p}(9)$	0.001	1.000	0.172	0.951
10	$B_{10,p}(10)$	0.000	1.000	0.049	1.000

A Statistical Decision Procedure

What does this table tell us? Let us interpret the cryptanalytic task as a decision problem: We set a threshold value T and decide:

- A string with score $\leq T$ is probably random. We discard it.
- A string with score $> T$ could be true plaintext. We keep it for further examination.

There are two kinds of possible errors in this decision:

1. A true plaintext has a low score. We miss it.
2. A random string has a high score. We keep it.

Example. Looking at Table 2 we are tempted to set the threshold value at $T = 4$. Then (in the long run) we'll miss 2.4% of all true plaintexts because the probability for an English 10 letter text string having an MFL score ≤ 4 is 0.024. On the other hand we'll discard only 67.1% of all random strings and erroneously keep 32.9% of them.

The lower the threshold T, the more unwanted random strings will be selected. But the higher the threshold, the more true plaintext strings will be missed. Because the distributions of the MFL scores for "Random" and "English" overlap there is no clear cutpoint that always gives the correct decision.

This is a typical situation for statistical decision problems (or **tests**). The statistician usually bounds one of the two errors by a fixed amount, usually 5% or 1%, and calls this the **error of the first kind**, denoted by α. (The complementary value $1 - \alpha$ is called the sensitivity of the test.) Then she tries to minimize the other error, the **error of the second kind**, denoted by β. The complementary value $1 - \beta$ is called the **power** (or specifity) of the test. FRIEDMAN's MFL-method, interpreted as a statistical test (for the "null hypothesis" of English text against the "alternative hypothesis" of random text), has a power of $\approx 67\%$ for English textstrings of length 10 and $\alpha = 2.4\%$. This α-value was chosen because it is the largest one below 5% that really occurs in the sixth column of Table 2.

To set up a test the statistician faces two choices. First she has to choose between "first" and "second" kind depending on the severity of the errors in the actual context. In our case she wants to bound the number of missed true plaintexts at a very low level—a missed plaintext renders the complete cryptanalysis obsolete. On the other hand keeping too many random strings increases the effort of the analysis, but this of somewhat less concern.

The second choice is the error level α. By these two choices the statistician adapts the test to the context of the decision problem.

Remark. We won't discuss the trick of raising the power by exhausting the α-level, randomizing the decision at the threshold value.

Note. There is another ("BAYESian") way to look at the decision problem. The **predictive values** give the probabilities that texts are actually what we decide them to be. If we decide "random" for texts with MFL score ≤ 4, we'll be correct for about 671 of 1000 random texts and err for 24 of 1000 English texts. This makes 695 decisions for random of which 671 are correct. The predictive value of our "random" decision is $96.5\% \approx 671/695$. The decision "English" for an MFL score > 4 will be correct for 976 of 1000 English texts and false for 329 of 1000 random texts. Hence the predictive value of the decision "English" is about $75\% \approx 976/1305$. That means that if we pick up texts (of length 10) with a score of at least 5, then (in the long run) one out of four selected texts will be random.

Other Languages: German and French

German: The ten most frequent letters are `ENIRSATDHU`. They make up 75.1% of an average German text.

French: The ten most frequent letters are `EASNTIRULO`. They make up 79.1% of an average French text.

With these values we supplement Table 2 by Table 3.

As before for English we get as conclusions for textstrings of length 10:

Table 3: *Distribution of MFL scores for* $r = 10$

	$p = 0.751$ **(German)**		$p = 0.791$ **(French)**	
Score	Probability	Total	Probability	Total
0	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000
3	0.003	0.003	0.001	0.001
4	0.016	**0.019**	0.007	0.008
5	0.058	0.077	0.031	**0.039**
6	0.145	0.222	0.098	0.137
7	0.250	0.471	0.212	0.350
8	0.282	0.754	0.301	0.651
9	0.189	0.943	0.253	0.904
10	0.057	1.000	0.096	1.000

German: With a threshold of $T = 4$ and $\alpha = 1.9\%$ the MFL-test has a power of 67%. The predictive value for "German" is $75\% \approx 981/1310$.

French: With a threshold of $T = 5$ and $\alpha = 3.9\%$ the MFL-test has a power of 86%. The predictive value for "French" is $87\% \approx 961/1103$.

Textstrings of length 20

The distribution is given in Table 4. We conclude:

English: With a threshold of $T = 10$ and $\alpha = 1.9\%$ the MFL-test has a power of 90% and a predictive value of $91\% \approx 981/1081$.

German: With a threshold of $T = 11$ and $\alpha = 4.0\%$ the MFL-test has a power of 96% and a predictive value of $96\% \approx 960/1002$.

French: With a threshold of $T = 12$ and $\alpha = 4.1\%$ the MFL-test has a power of 98.5% and a predictive value of $98.5\% \approx 959/974$.

Table 4: *Distribution of MFL scores for $r = 20$*

	Random		**English**		**German**		**French**	
Score	Prob	Total	Prob	Total	Prob	Total	Prob	Total
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2	0.005	0.005	0.000	0.000	0.000	0.000	0.000	0.000
3	0.017	0.022	0.000	0.000	0.000	0.000	0.000	0.000
4	0.045	0.067	0.000	0.000	0.000	0.000	0.000	0.000
5	0.090	0.157	0.000	0.000	0.000	0.000	0.000	0.000
6	0.140	0.297	0.000	0.000	0.000	0.000	0.000	0.000
7	0.175	0.472	0.000	0.000	0.000	0.000	0.000	0.000
8	0.178	0.650	0.001	0.001	0.001	0.001	0.000	0.000
9	0.148	0.798	0.004	0.006	0.003	0.004	0.001	0.001
10	0.102	**0.900**	0.013	**0.019**	0.010	0.013	0.003	0.004
11	0.058	**0.958**	0.034	0.053	0.026	**0.040**	0.010	0.013
12	0.027	**0.985**	0.072	0.125	0.060	0.100	0.028	**0.041**
13	0.010	0.996	0.125	0.250	0.111	0.211	0.064	0.105
14	0.003	0.999	0.178	0.428	0.168	0.379	0.121	0.226
15	0.001	1.000	0.201	0.629	0.202	0.581	0.184	0.410
16	0.000	1.000	0.178	0.807	0.191	0.772	0.217	0.627
17	0.000	1.000	0.119	0.925	0.135	0.907	0.193	0.820
18	0.000	1.000	0.056	0.981	0.068	0.975	0.122	0.942
19	0.000	1.000	0.017	0.998	0.022	0.997	0.049	0.991
20	0.000	1.000	0.002	1.000	0.003	1.000	0.009	1.000