Statistics and Information Theory

Yan Huang
Objective

• Histograms

• Chi-Square Tests

• Index of Coincidence

• Information Entropy
Histogram
Histogram

• Letter Frequency Table
import Data.List
import Data.Ratio

countif :: Char -> [Char] -> Int
countif c = length . filter (== c)

countif :: [Char] -> [(Char, Int)]

histogram m = map accum ['A'..'Z']
  where accum c = (c, countif c m)
Chi-Square Test
Chi-Square ($\chi^2$) Test

Let $A, B$ be two distributions over the same set $S$.

$$\chi^2(A, B) = \sum_{i \in S} \frac{(A(i) - B(i))^2}{B(i)}$$

Applications

Automatically identify likely-correct decryptions
Implementing $\chi^2$-test
Index of Coincidence
Vigenère Cipher.

Plaintext: THE EMPIRICAL JAPAN

Key: HELLO; HELLO; HELLO; HELLO; HELLO

Ciphertext: ALP P A W I C T O S N L A O U

Caesar: 7

We can use Index of Coincidence to predict the length of the key.
Index of Coincidence

• The index of coincidence of a given message is the probability that two randomly picked letters from the message happen to be the same.

• For a uniform distribution U over the English alphabet, $\text{IoC}(U) = 1/26$

\[
\frac{26}{\sum_{i=1}^{26} \frac{1}{26} - \frac{1}{26}} = \frac{1}{26} \approx 0.038
\]
Application of Index of Coincidence

• Predicting the key length of a Vigenère cipher

\[ \frac{I_{OC_{text}}}{I_{OC_{english}}} = \frac{f_{k-1}}{K} \approx \frac{I_{OC_{random}}} {0.06} \]

So you can derive \( K \) from the three \( I_{OC} \) scores.