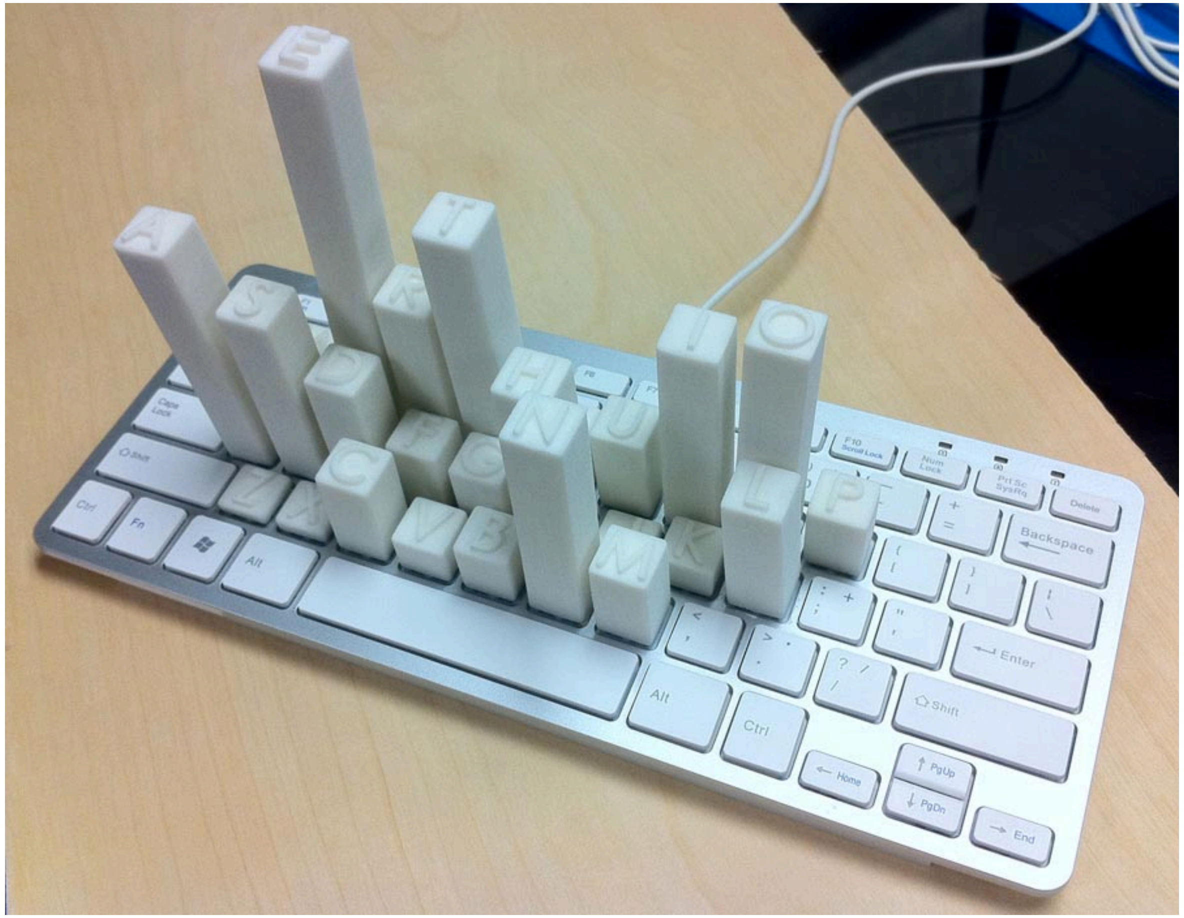
Statistics and Information Theory

Yan Huang

Objective

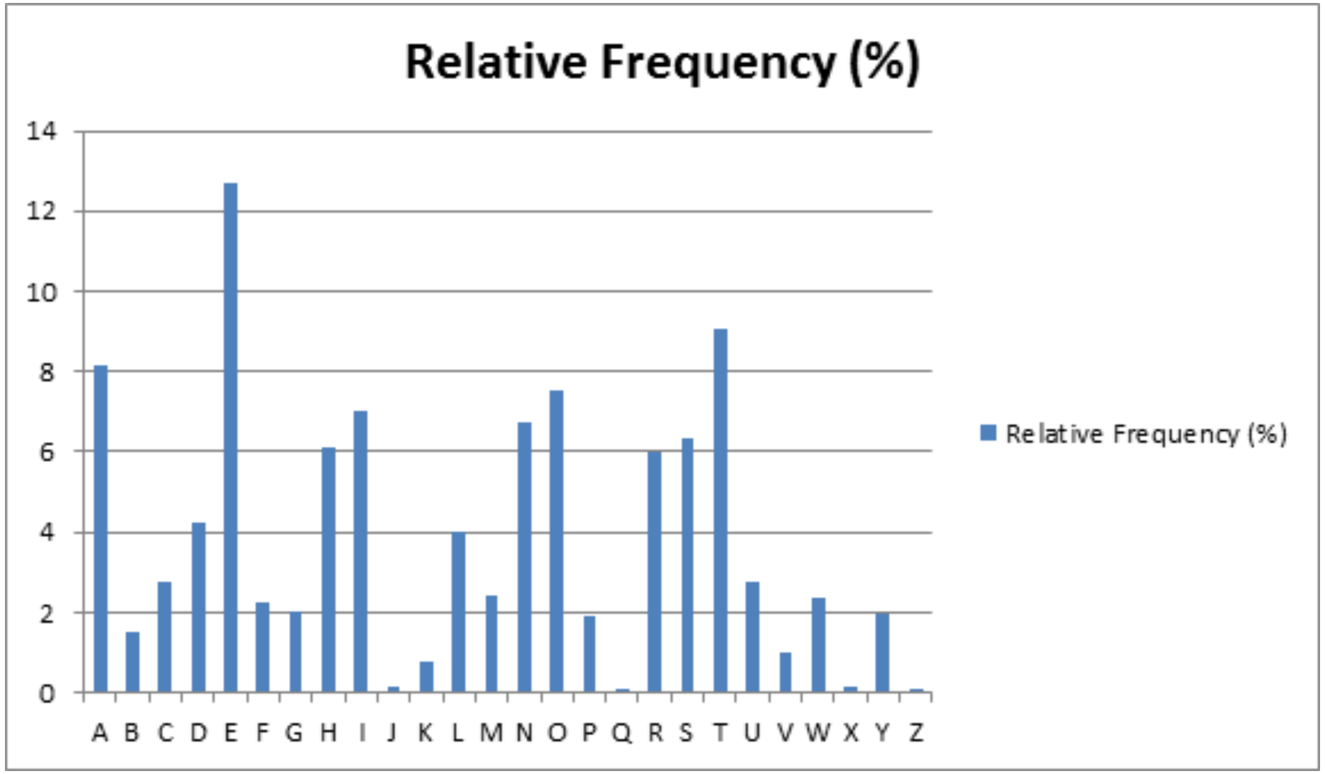
- Histograms
- Chi-Square Tests
- Index of Coincidence
- Information Entropy

Histogram



Histogram

- Letter Frequency Table



Compute the Histogram

```
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 (χ^2) Test

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

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

Applications

Automatically identify likely-correct decryptions

Implementing χ^2-test

Index of Coincidence

Vigenère Ciper.

Plaintext: THE EMPERIAL JAPAN....

Key: HELLO HELLO HELLO H

7 4

Ciphertext: ALPPAWICTOSNLAOU

caeserEnc 7 'T'

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

Index of Coincidence

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

"draw without replacement"

C_i — the number of occurances of letter 'i', then $IoC = \sum_{i='A'}^{'Z'} \frac{C_i(C_i-1)}{n^2}$

- For a *uniform* distribution U over the English alphabet,

  IoC(U) = 1/26 $\sum_{i=1}^{26} \frac{1}{26} \cdot \frac{1}{26} = \frac{1}{26} \approx 0.038$

Application of Index of Coincidence

- **Predicting the key length of a Vigenère cipher**

Predicting the Key length of Vigenère Cipher.

$$IoC_{\text{ciphertext}} = \frac{1}{K}\, IoC_{\text{English}} + \frac{K-1}{K}\, IoC_{\text{Uniform Random}}$$

IoC_{English}: 0.06...

$IoC_{\text{Uniform Random}}$: $\frac{1}{26}$

So you can derive K from the three IoC scores.