Cryptanalysis of the Vigenère Cipher Using the Kasiski Examination and Statistical Validation Techniques

Bruno Bavaresco Zaffari
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Email: brunobzaffari@hotmail.com

Professor: Iaçanã Ianiski Weber

Course: Software Reliability and Security

Abstract

This paper presents a structured approach to decrypting a Vigenère cipher through classical cryptanalysis techniques. Initially, the Kasiski examination is employed to estimate the key length by detecting repeated n-grams and calculating the greatest common divisors (GCD) of the distances between them. To validate the hypothesized key lengths, the Index of Coincidence (IC) is applied. Finally, the Chi-squared (χ^2) test is used to reconstruct the cipher key. Results confirm the success of the method, revealing the plaintext as an excerpt from "Papéis Avulsos" by Machado de Assis.

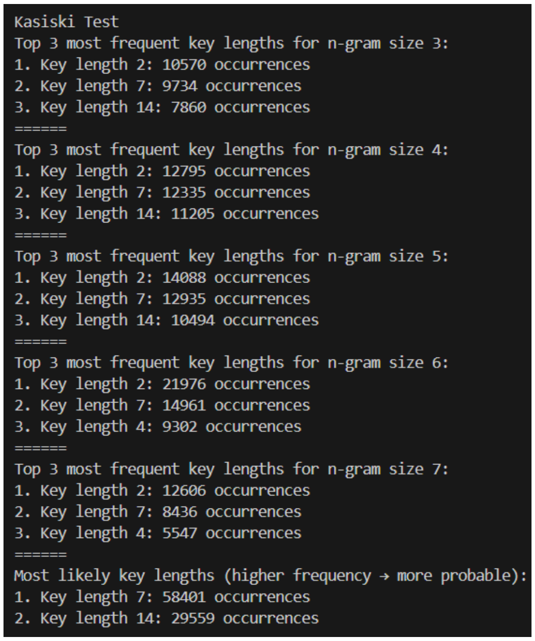
1. Introduction

The Vigenère cipher, a classical polyalphabetic encryption technique, poses considerable resistance to naive frequency analysis. However, it remains vulnerable to statistical and pattern-based attacks when the key length is sufficiently short. This paper documents a step-by-step cryptanalysis using Kasiski examination, IC, and χ^2 testing.

2. Estimating the Key Length Using Kasiski Examination

The first phase involves identifying repeated sequences (n-grams) within the ciphertext. These repetitions are potential markers of key length periodicity, assuming repeated plaintext segments were encrypted with the same key segment.

A sliding window approach was implemented to extract repeated n-grams ranging from 3 to 7 characters. To reduce noise, frequency thresholds were adapted to n-gram size: shorter sequences required more repetitions to be considered. This approach is justified statistically shorter n-grams tend to recur more frequently by chance and thus need stricter filtering, whereas longer n-grams are inherently more significant even with fewer occurrences. After identifying repeated n-grams, the positions were recorded, and distances between occurrences were computed. The GCD of these distances was calculated, as the true key length should be a divisor of these intervals.



```
Kasiski Test
Top 3 most frequent key lengths for n-gram size 3:
1. Key length 2: 10570 occurrences
2. Key length 7: 9734 occurrences
3. Key length 14: 7860 occurrences
======
Top 3 most frequent key lengths for n-gram size 4:
1. Key length 2: 12795 occurrences
2. Key length 7: 12335 occurrences
3. Key length 14: 11205 occurrences
======
Top 3 most frequent key lengths for n-gram size 5:
1. Key length 2: 14088 occurrences
2. Key length 7: 12935 occurrences
3. Key length 14: 10494 occurrences
======
Top 3 most frequent key lengths for n-gram size 6:
1. Key length 2: 21976 occurrences
2. Key length 7: 14961 occurrences
3. Key length 4: 9302 occurrences
======
Top 3 most frequent key lengths for n-gram size 7:
1. Key length 2: 12606 occurrences
2. Key length 7: 8436 occurrences
3. Key length 4: 5547 occurrences
======
Most likely key lengths (higher frequency → more probable):
1. Key length 7: 58401 occurrences
2. Key length 14: 29559 occurrences
```

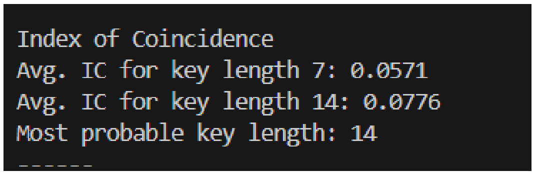
3. Validation with Index of Coincidence (IC)

To assess the plausibility of each candidate key length, the IC was computed for character groups formed by deinterleaving the ciphertext according to each key length.

A correct key length produces groups each encrypted with a Caesar cipher, hence retaining the original language frequency characteristics. The IC for Portuguese hovers around 0.076, whereas random text tends toward 0.0385.

Three key length candidates emerged: 2, 7, and 14. Length 2 was discarded due to overly regular patterns. Groupings with lengths 7 and 14 showed more promising IC averages, with 14 providing the closest match.

For instance, with a key length of 7, the ciphertext was partitioned into seven interleaved groups, each corresponding to the characters encrypted with the same letter from the key. This segmentation allows the application of monographic analysis to each group independently.



```
Index of Coincidence
Avg. IC for key length 7: 0.0571
Avg. IC for key length 14: 0.0776
Most probable key length: 14
------
```

4. Key Reconstruction Using Chi-Squared Test

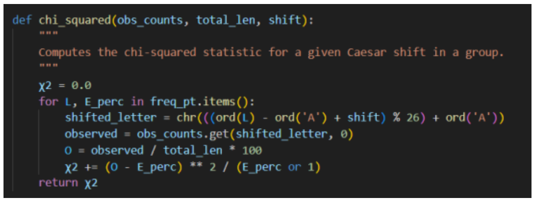
Once the key length was determined (14), the next step was deducing the actual key. Each of the 14 groups, corresponding to one key character, was tested against all 26 possible Caesar shifts.

The Chi-squared (χ^2) test was applied to measure the similarity between the letter frequency of the shifted ciphertext group and the standard Portuguese language frequency.

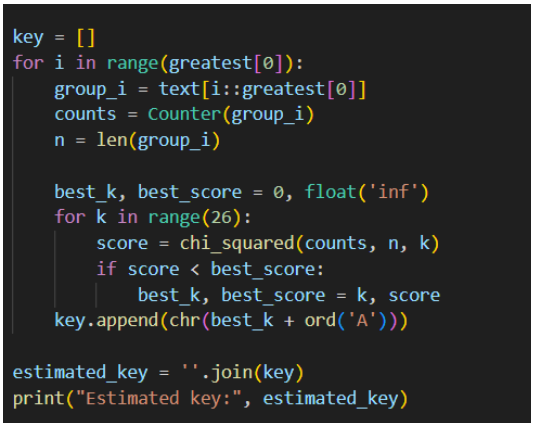
$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The shift with the lowest χ^2 score in each group was selected as the best approximation of the corresponding key letter. This process yielded the 14-character key.

Although alternative metrics like the mean absolute deviation could be employed, the χ^2 method alone provided sufficiently robust statistical grounding for key character estimation, given its sensitivity to frequency profile deviations.



```
def chi_squared(obs_counts, total_len, shift):
    """
    Computes the chi-squared statistic for a given Caesar shift in a group.
    """
    χ2 = 0.0
    for L, E_perc in freq_pt.items():
        shifted_letter = chr(((ord(L) - ord('A') + shift) % 26) + ord('A'))
        observed = obs_counts.get(shifted_letter, 0)
        O = observed / total_len * 100
        χ2 += (O - E_perc) ** 2 / (E_perc or 1)
    return χ2
```



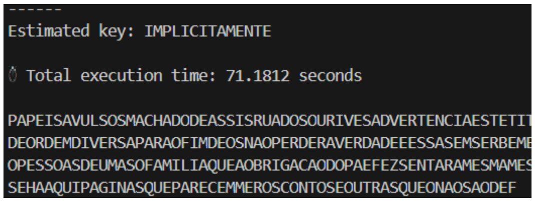
```
key = []
for i in range(greatest[0]):
    group_i = text[i::greatest[0]]
    counts = Counter(group_i)
    n = len(group_i)

    best_k, best_score = 0, float('inf')
    for k in range(26):
        score = chi_squared(counts, n, k)
        if score < best_score:
            best_k, best_score = k, score
    key.append(chr(best_k + ord('A')))

estimated_key = ''.join(key)
print("Estimated key:", estimated_key)
```

5. Results, Limitations, and Conclusion

The analysis yielded the key "IMPLICITAMENTE" with a length of 14. Upon decryption, the plaintext was identified as a passage from Machado de Assis's "Papéis Avulsos".



```
------
Estimated key: IMPLICITAMENTE

⏱ Total execution time: 71.1812 seconds

PAPEISAVULSOSMACHADODEASSISRUADOSOURIVESADVERTENCIAESTETIT
DEORDEMDIVERSAPARAOFIMDEOSNAOPERDERAVERDADEEESSASEMSERBEME
OPESSOASDEUMASOFAMILIAQUEAOBRIGACAODOPAEFEZSENTARAMESMAMES
SEHAAQUIPAGINASQUEPARECEMMEROSCONTOSEOUTRASQUEONAOSAODEF
```

This exercise illustrates the efficacy of classical cryptanalytic methods when paired with modern computational implementation. Even traditional ciphers like Vigenère can be broken effectively through statistical rigor and structured methodology.

It is worth noting that the effectiveness of Kasiski’s method diminishes significantly when the key length approaches the length of the ciphertext. In such cases, the statistical signal becomes diluted, and the Index of Coincidence tends to values indistinguishable from random noise, complicating analysis.

The decrypted passage clearly referenced "Papéis Avulsos" by Machado de Assis, confirming the textual origin and concluding the cryptanalytic effort with both statistical and literary validation.

References

1. [1] Kasiski, F. W. "Die Geheimschriften und die Dechiffrir-Kunst." Berlin: E. S. Mittler und Sohn, 1863.
2. [2] Sinkov, A. "Elementary Cryptanalysis: A Mathematical Approach." The Mathematical Association of America, 1966.
3. [3] Salomaa, A. "Public-Key Cryptography." Springer, 1990.
4. [4] Beker, H. and Piper, F. "Cipher Systems: The Protection of Communications." Wiley, 1982.