Sapienza
Università di Roma

Statistical Methods in Cybersecurity: Foundations, Applications, and Future Perspectives

Statistics a.y. 24-25

Scarselli Ilaria, 1918975 - scarselli.1918975@studenti.uniroma1.it

December, 2024

Contents

1 Introduction

Cybersecurity is a major challenge in today's digital world as cyber threats grow in complexity and frequency. Effective tools are needed to detect, analyze, and mitigate these threats, and **statistical methods** play a key role in this effort. These methods help identify patterns, assess risks, and make informed decisions under uncertainty.

This thesis focuses on classical statistical methods such as **probability theory**, **sampling techniques**, and foundational concepts like the **Law of Large Numbers (LLN)** and the **Central Limit Theorem (CLT)**. These tools help systems detect anomalies, evaluate cryptographic strength, and analyze large datasets, which are central to cybersecurity operations like intrusion detection and risk assessment.

The work highlights practical applications in **anomaly detection**, **cryptanalysis**, and **cyber threat modeling**, emphasizing how statistical methods optimize decision-making and computational efficiency. It also explores the limitations and future prospects of these techniques in the face of emerging technologies and evolving threats.

2 Probability and Frequency Analysis in Cybersecurity

Understanding probability and frequency is essential for predicting threats, assessing risks, and detecting unusual behavior in cybersecurity.

2.1 Probability Theory in Cybersecurity

Probability helps estimate the likelihood of attacks or vulnerabilities being exploited. Two main approaches are:

- **Frequentist Probability**: Defines probability based on long-term frequency. For example, if phishing emails occur 3 times per week, the daily probability can be estimated.

- **Bayesian Probability**: Updates probabilities using prior knowledge. For instance, if IP addresses linked to malware are detected, Bayesian methods adjust the likelihood of future attacks. It is fundamentally about updating beliefs with new evidence.

Application: In Intrusion Detection Systems (IDS), Bayesian methods quantify the likelihood that repeated access attempts to restricted areas signal malicious activity.

2.2 Frequency Analysis

Frequency analysis studies how often events occur, revealing patterns that may indicate attacks.

- **Cryptographic Analysis**: Historically, frequency analysis broke weak ciphers like the Caesar Cipher by analyzing letter frequencies in the various languages (changing the alphabet offered basically no help). Modern systems like AES resist this by ensuring ciphertext appears random, assuring that information about the plaintext isn't easily extracted observing the ciphertext.
- **Anomaly Detection**: Abnormal event frequencies, such as sudden spikes in login attempts, may signal brute-force attacks. Tools like histograms and probability distributions help identify these deviations.

2.3 Combining Probability and Frequency Analysis

Combining these methods strengthens threat detection:

Probability models estimate risks based on historical data.

Frequency analysis monitors current events for unusual patterns.

Example: In Security Information and Event Management (SIEM) systems, logged events are analyzed to identify risks and frequent anomalies. These are then presented as summarized reports and graphs of various types to operators, providing them with a clearer view without being overwhelmed by the constant incoming data.

The next section will explore how sampling techniques handle large datasets, a core aspect of modern security operations.

3 Statistical Sampling in Large Datasets for Cyber Threat Detection

The exponential growth of digital systems and network activity generates vast amounts of data, making it impractical to analyze every record individually. **Statistical sampling** provides a way to extract useful insights from large datasets while optimizing computational resources. This section explores key sampling techniques and their applications in cyber threat detection.

3.1 The Role of Sampling in Cybersecurity

In cybersecurity, sampling techniques allow analysts to work with manageable subsets of data without losing the ability to identify meaningful patterns. Sampling reduces the time and computational power needed to process massive datasets, which is critical for real-time threat detection.

Example: A network administrator analyzing millions of connection logs can apply sampling to extract a representative subset of records. This subset can then be analyzed for anomalies, such as unusual access patterns or suspicious traffic.

3.2 Key Sampling Methods

3.2.1 Simple Random Sampling

Simple random sampling involves selecting a subset of data where every element has an equal probability of being chosen. This method ensures an unbiased representation of the entire dataset. Multistage Sampling uses a combination of two or more SRS in multiple stages.

Application: In analyzing login records, random sampling can help detect brute-force attacks or password-guessing attempts by focusing on a representative subset of the logs. For example, a particular period can be selected as a first stage and then refined to focus on a specific part of the system (as a specific part of the organization network infrastructure, or a few machines).

3.2.2 Stratified Sampling

Stratified sampling divides the dataset into distinct groups, or *strata*, based on certain attributes. Samples are then drawn from each group to ensure proportional representation: in each stratum we take an SRS, that will be then combined to obtain the full sample.

Application: In intrusion detection systems, network traffic can be divided into strata based on protocols (e.g., HTTP, FTP, SSH). Sampling from each protocol ensures that anomalies are not overlooked within any specific category.

3.2.3 Systematic Sampling

In systematic sampling, data points are selected at regular intervals. This method is efficient for analyzing sequential records, such as time-stamped logs.

Application: A security team analyzing hourly network traffic can select every 100th log entry to identify deviations or trends.

3.3 Advantages and Limitations of Sampling

Advantages:

- Reduces computational costs for analyzing large datasets.
- Enables real-time threat detection by working on manageable subsets.
- Provides a scalable approach for monitoring massive systems.

Limitations:

- Sampling may overlook rare events or attacks if they do not appear in the subset.
- Requires careful design to ensure the sample is representative of the full dataset and is not biased, so carefully-tailored studies need to be conducted for every different context, to best optimize resource usage and correctly respond to cyber risks.

3.4 Practical Applications in Cyber Threat Detection

Statistical sampling is widely applied in cybersecurity to improve efficiency and accuracy. Key examples include:

- **Intrusion Detection**: Sampling network traffic logs to identify anomalies without analyzing every packet.
- **Malware Analysis**: Sampling large datasets of files to detect malicious patterns or signatures.
- **Risk Assessment**: Analyzing sampled user activity to detect unusual behaviors indicating compromised accounts.

4 Law of Large Numbers and Central Limit Theorem in Anomaly Detection

In the field of cybersecurity, statistical concepts such as the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT) play a critical role in anomaly detection, providing a mathematical framework to distinguish between normal and abnormal behaviors in network traffic. These statistical principles are especially useful in establishing baselines and thresholds, which are fundamental to identifying potential security threats in both a manual and an automated fashion (as with the SIEM discussed in earlier paragraphs).

4.1 Law of Large Numbers (LLN)

The Law of Large Numbers (LLN) states that as the sample size increases, the sample mean will converge to the expected value of the population. In the context of cybersecurity, LLN helps stabilize averages in cyber data, such as network traffic patterns, over time. For instance, when analyzing data from a network, the LLN suggests that the average behavior of network traffic will become more predictable and consistent as more data points are observed. This concept forms the foundation for establishing baseline measurements in anomaly detection.

Anomaly detection systems leverage this property to monitor network behavior, allowing the system to "learn" what constitutes typical activity. As more traffic data is gathered, the system refines its understanding of normal traffic patterns and less errors are made (alarms are raised in a more consistent way and less false positive are generated). Significant deviations from this baseline — such as a sudden spike in traffic or unusual access patterns — are flagged as potential anomalies.

4.2 Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) is another key concept in statistical analysis, which states that the distribution of the sample means will approach a normal distribution (a symmetric, bell-shaped probability distribution where most data points cluster around the mean, with decreasing frequency as you move farther away) as the sample size increases, regardless of the original data distribution. In anomaly detection, the CLT is applied to the analysis of large sets of network data, such as packet delays. As network traffic increases, the sample mean of packet delays will approximate a normal distribution.

For example, in the context of traffic analysis, CLT can be used to detect packet delays that deviate significantly from normal behavior. If a network's typical packet delay is usually distributed within a certain range, but a sudden deviation from this distribution occurs, this could indicate a potential security breach, such as a Distributed Denial-of-Service (DDoS) attack or a network performance issue caused by malicious activity (or more generally by a problem in the network infrastructure, at least).

4.3 Applications in Anomaly Detection

Both the Law of Large Numbers and the Central Limit Theorem contribute to the development of statistical thresholds that are vital for anomaly detection systems. LLN assists in establishing the baseline of normal network behavior, while CLT aids in defining the expected distribution of data over time. By combining these two concepts, cybersecurity systems can develop robust models for flagging suspicious activities, such as unexpected traffic surges, abnormal delays, or irregular access patterns, in an automatic manner to ensure a quick reaction to mitigate risks. These statistical tools are essential for detecting emerging threats in real-time, thereby improving the security and reliability of network infrastructure.

5 Statistical Properties of Cryptographic Techniques

Cryptographic systems depend on randomness for secure key generation and encryption. Weaknesses in randomness can expose systems to statistical attacks, making randomness a critical factor in cryptographic security.

5.1 Entropy and Randomness

Randomness is essential in cryptography, especially for generating secure keys in systems like AES. The strength of a key is often measured by its entropy, which quantifies its unpredictability. Shannon entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

, where:

- $H(X)$ is the entropy of the random variable X,
- $p(x_i)$ is the probability of the occurrence of each possible outcome x_i of X,
- The summation runs over all possible outcomes $x_1, x_2, \ldots, x_n$ of the random variable.

This formula calculates the average amount of "information" produced by the random variable. The term $p(x_i) \log_2 p(x_i)$ represents the contribution of each outcome x_i to the overall uncertainty, weighted by its probability. The negative sign ensures that entropy is a positive quantity.

5.1.1 Interpretation

- **High Entropy**: If each possible outcome x_i has an equal probability, i.e., $p(x_i) = \frac{1}{n}$ for all i, the entropy is maximized. This represents a system with maximum uncertainty, such as a fair coin flip.

- **Low Entropy**: If some outcomes have a much higher probability than others, for example, $p(x_1) = 1$ and $p(x_i) = 0$ for all $i > 1$, the entropy is low, indicating a predictable system.

Consider a simple random process where a random variable X represents the outcome of a die roll. The possible outcomes are $X = \{1, 2, 3, 4, 5, 6\}$, each with an equal probability of $\frac{1}{6}$. The probability of each outcome is $p(x_i) = \frac{1}{6}$, for $i = 1, 2, \ldots, 6$.

Now, we can calculate the entropy for this system:

$$H(X) = -\sum_{i=1}^{6} p(x_i) \log_2 p(x_i)$$

Since each outcome is equally likely, $p(x_i) = \frac{1}{6}$ for all i. Substituting into the formula:

$$H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6}$$

Simplifying further, since $\log_2 \frac{1}{6} = -\log_2 6$, we get:

$$H(X) = -\sum_{i=1}^{6} \frac{1}{6}(-\log_2 6) = \log_2 6$$

Thus, the entropy of this system is:

$$H(X) = \log_2 6 \approx 2.585$$

This indicates a moderate level of uncertainty in the outcome of the die roll.

Example with Lower Entropy:

Consider a biased coin where the probability of heads is $p(\text{Heads}) = 0.9$ and the probability of tails is $p(\text{Tails}) = 0.1$. In this case, the entropy is calculated as:

$$H(X) = -\left(p(\text{Heads}) \log_2 p(\text{Heads}) + p(\text{Tails}) \log_2 p(\text{Tails})\right)$$

Substituting the probabilities:

$$H(X) = -\left(0.9 \log_2 0.9 + 0.1 \log_2 0.1\right)$$

Evaluating the logarithms:

$$H(X) = -\left(0.9 \times (-0.137) + 0.1 \times (-3.322)\right) \approx 0.532$$

The entropy here is much lower than the die example, indicating that the outcome is much more predictable (with heads occurring 90% of the time).

In cryptography, the entropy of a key determines its strength. A key with higher entropy is more random and harder to predict, thus providing greater security. A higher entropy value indicates a stronger, more unpredictable key. For example, a 256-bit AES key has high entropy, ensuring strong encryption. Ensuring high entropy in key generation is crucial for resisting attacks.

5.2 Vulnerabilities

Weaknesses in Random Number Generators (RNGs) can compromise cryptographic security. If an RNG is poorly seeded or deterministic, it may produce predictable sequences, making it easier for attackers to guess cryptographic keys. Vulnerable systems are often based on pseudo-random number generators (PRNGs) that exhibit statistical patterns, which can be exploited if the internal state of the RNG is predictable.

5.3 Statistical Attacks

Cryptographic systems are vulnerable to statistical attacks that exploit patterns in encryption. Two major types are ***Linear Cryptanalysis*** and ***Differential Cryptanalysis***.

5.3.1 Linear Cryptanalysis

Linear cryptanalysis finds linear relationships between plaintext, ciphertext, and keys. By analyzing statistical correlations, attackers can reveal bits of the key with fewer known plaintext-ciphertext pairs than brute-force methods require.

5.3.2 Differential Cryptanalysis

Differential cryptanalysis analyzes how changes in plaintext affect ciphertext. This attack uses statistical methods to trace key relationships, particularly effective against block ciphers, and can reduce the effort needed to break the encryption.

5.4 Case Study: Weaknesses in Poorly Generated Keys

A notable real-world case involved the OpenSSL library (CVE-2008-0166), where weak private keys were generated due to flawed RNGs. Attackers detected these weak keys by analyzing the statistical properties of the public keys, revealing low entropy in the key generation. This led to improvements in RNG design and cryptographic protocols to enhance security.

6 Quantum Probability Theory and Cybersecurity

6.1 Classical vs. Quantum Probability

Classical probability is based on deterministic systems with well-defined states and outcomes. In contrast, **quantum probability** arises from the principles of quantum mechanics, where systems can exist in a *superposition* of states until measured. Unlike classical randomness, quantum randomness is inherently unpredictable and non-deterministic, providing stronger guarantees for cryptographic systems.

6.2 Quantum Cryptography

Quantum cryptography leverages quantum probability to enable secure communication. A notable example is **Quantum Key Distribution (QKD)**. In QKD:

- Quantum states (e.g., photons) encode cryptographic keys.
- Any attempt to measure these states (eavesdropping) disturbs the system, due to the *Heisenberg uncertainty principle* (there is a limit to the precision with which certain pairs of physical properties can be simultaneously known).
- Probabilistic measurements detect such disturbances, ensuring security.

6.3 The Future of Quantum Cybersecurity

Quantum probability plays a foundational role in developing cryptographic systems that are theoretically unbreakable. With advancements in quantum computing and quantum communication, quantum cybersecurity promises:

- Unconditional security through quantum mechanics.
- Robust systems resistant to both classical and quantum attacks.

While practical implementation remains a challenge, quantum cryptography represents a paradigm shift in secure communications.

7 Integration of Statistical Models in Modern Cybersecurity Systems

By combining statistical methods and probability models, it is possible to greatly enhance the implementation of secure and robust systems against external threats, while also enabling timely intervention through faster problem detection via partially automated alert systems.

7.1 Practical Example: Bayesian Malware Detection

Consider malware detection using Bayesian probability:

- Statistical baselines for system behavior (e.g., process execution times or network packet rates) are established using LLN.
- As new activity is observed, Bayesian inference calculates the probability that the activity is malicious given prior baselines.
- If the probability exceeds a threshold, an alert is triggered.

7.2 Quantum Probability Advancements

Quantum probability offers future improvements for cybersecurity. By leveraging quantum randomness, quantum encryption systems provide enhanced key security. Additionally, quantum probability models can refine intrusion detection, allowing systems to analyze patterns with greater precision and resilience against sophisticated attacks. The study of this field is essential because, with the advent of modern supercomputers and quantum cryptography, problems previously considered difficult will become decidable. It is therefore necessary to continue strengthening this area to keep pace with modern challenges.

8 Conclusions

Classical and modern statistical methods, along with quantum probability theory, are crucial tools in cybersecurity. These methods enable effective anomaly detection, enhance cryptographic strength, and optimize cyber data analysis. As the role of statistics continues to grow, integrating quantum models into real-world systems will be essential for advancing cybersecurity, addressing emerging threats, and ensuring robust protection against future challenges.

A Appendix: Overview of AES, RSA, Caesar Cipher, and DES

A.1 AES (Advanced Encryption Standard)

AES is a symmetric encryption algorithm using block sizes of 128 bits and key sizes of 128, 192, or 256 bits. It provides strong security by employing multiple rounds of substitution and permutation, making it resistant to brute-force and cryptanalysis attacks.

Why More Secure: AES has a large key space and complex transformations, making it much harder to break compared to older algorithms, with the use of non-linear transformations.

A.2 RSA (Rivest-Shamir-Adleman)

RSA is an asymmetric encryption algorithm based on the difficulty of factoring large primes. It uses a public key for encryption and a private key for decryption, commonly used for secure communication.

Why More Secure: RSA's security is based on the computational difficulty of factoring large numbers, which is infeasible with current technology.

A.3 Caesar Cipher

The Caesar cipher shifts letters by a fixed amount and is easily broken with brute-force or frequency analysis.

Why Less Secure: With only 25 possible shifts, it can be quickly deciphered by attackers.

A.4 DES (Data Encryption Standard)

DES is a symmetric cipher that uses a 56-bit key and is vulnerable to brute-force attacks.

Why Less Secure: The small 56-bit key size makes DES susceptible to modern computational power and easy to break.

A.5 Comparison

- **Caesar Cipher** is weak due to a small key space.
- **DES** is insecure because of its small key size (56 bits).
- **AES** and **RSA** are much more secure due to larger key sizes and complex encryption methods.