1. Theoretical Background

1.1 Population, Statistical Units, and Distribution

In statistics, the **population** is the entire set of individuals, objects, or observations under investigation. For example, when studying exam scores in a classroom, the population consists of all students in that class, with size **N**.

A **statistical unit** is the smallest element of the population about which data is collected. In this case, each student is a statistical unit.

A **distribution** describes how the values of a variable are spread among statistical units. For example, the distribution of exam scores shows how many students fall within ranges such as 0–50, 51–75, and 76–100. Distributions can be represented by tables, histograms, or probability models.

Types of distributions:

- **Frequency distribution:** shows how often values occur in observed data.
- **Probability distribution:** a theoretical model of the likelihood of outcomes.
- **Univariate distribution:** involves a single variable.
- **Bivariate distribution:** involves two variables (e.g., height vs weight).
- **Multivariate distribution:** involves three or more variables.

Distributions are quantified using **frequencies**:

- **Absolute frequency (f):** the raw count of occurrences. Example: if 20 out of 100 students score 85, then f = 20.
- **Relative frequency (f/N):** the proportion of the population, here 20/100 = 0.2.
- **Percentage frequency:** relative frequency × 100 = 20%.

In the **penetration simulation**, the **population** is the set of attackers, the **statistical units** are individual attackers, and the **distribution** describes how many penetrations each attacker achieves.

1.2 Notion of the Arithmetic Average

The **average** summarizes a dataset with a typical or central value.

- **Mean (arithmetic average):**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Example: [10, 20, 30, 40, 50] → mean = 30.

Derivation: the mean is the value that minimizes the sum of squared deviations:

$$S = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **Median:** the middle value after ordering data. For [3, 5, 7], median = 5. For [3, 5, 7, 9], median = (5+7)/2 = 6.
- **Mode:** the most frequent value. For [4, 4, 5, 6, 4, 7], mode = 4.

These measures behave differently depending on the distribution: the mean is influenced by outliers, while the median is more robust.

1.3 Floating-Point Representation and Computational Issues

Computers represent real numbers in **floating-point arithmetic**, which has finite precision. This leads to:

- **Rounding errors:** some numbers (e.g., 0.1) cannot be represented exactly in binary, leading to small inaccuracies.
- **Catastrophic cancellation:** subtracting nearly equal numbers causes loss of significant digits.
- **Overflow/Underflow:** extremely large numbers exceed representable limits (overflow), while extremely small numbers become 0 (underflow).

Example: 1.23456789 − 1.23456780 should be 0.00000009, but floating-point cancellation can distort the result.

1.4 Numerical Stability and Knuth's Contributions

To reduce errors, **Donald Knuth** and others developed stable numerical algorithms, such as:

- **Compensated summation (Kahan/Knuth algorithm):** maintains a correction term to reduce rounding error in summing many numbers.
- **Adaptive algorithms:** choose computational formulas that minimize error depending on the situation.
- **Interval arithmetic:** keeps track of error bounds.
- **Error analysis techniques:** estimate and control rounding error growth.

These approaches ensure accurate computations in scientific and statistical applications.

2. Applications / Practice

2.1 Problem Description

We simulate **m attackers** attempting to penetrate **n servers**. Each penetration attempt succeeds with probability **p**.

- Each attacker's progress is represented as a **line**:
  - flat if penetration fails,
  - a jump upward if penetration succeeds.
- After n attempts, we calculate the **distribution of successes among attackers**:
  - Some attackers may have 0 penetrations, some 1, some 2, etc.
  - This distribution is visualized with a **histogram**, where:
    - the x-axis = number of penetrations (0…n),
    - the y-axis = number of attackers achieving that count.