Research Homework 2

In statistics, **frequencies** are used to describe the distribution of a dataset, providing information on how often different values occur within the data. The two main types of frequency are **absolute frequency** and **relative frequency**.

- **Absolute Frequency**: of a value (or class of values) is simply the number of times that value appears in a dataset.

  f_i = \text{number&nbsp;of&nbsp;occurrences&nbsp;of&nbsp;value&nbsp;i}

- **Relative Frequency**: shows the proportion or percentage with which a certain value occurs compared to the total number of observations. It is calculated by dividing the absolute frequency of a value by the total number of observations.

$$f_r = \frac{n}{f_i}$$

⚠ **Difference**

- **Absolute Frequency**: This is a count. It represents how many times a value or event occurs in a dataset.
- **Relative Frequency**: This is a proportion or percentage. It shows how significant a value is in relation to the entire dataset.

Definitions of μ and σ^2

In statistics, **μ (mu)** and **σ² (sigma squared)** are fundamental parameters used to describe the characteristics of a probability distribution, particularly in the context of the **normal distribution** (or Gaussian distribution). These two values help summarize the central tendency and variability of data.

μ (Mu) - The Mean

μ represents the **mean** or **expected value** of a probability distribution. It indicates the central point or the average around which the data values are clustered.

Formula:

For a population:

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

Where:

- x_i are the individual data points,
- n is the total number of data points in the population.

> μ is the average value of all the data points in a dataset. It tells you where the center of the data is located.

σ^2 (Sigma Squared) - The Variance

σ² represents the **variance** of a distribution, which measures the degree of spread or dispersion of data points from the mean. It tells us how much the values deviate from the mean.\

> Variance is the average of the squares of the deviations of each value from the average.

Formula:

For a population:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Where:

- x_i are the individual data points,
- μ is the mean of the population,
- n is the total number of data points.

> Variance is essentially the **average of the squared differences** from the mean. A higher variance indicates that the data points are more spread out, while a lower variance means they are closer to the mean.

If you are calculating the variance for a sample (rather than for the entire population), you use a slightly modified formula known as the sample variance. This formula adjusts the estimate by dividing by $n - 1$ instead of n:

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

ℹ Related Concept

Standard Deviation (σ): This is the square root of the variance. While variance is in squared units, standard deviation provides a more intuitive measure of dispersion by returning to the original units of the data. The formula is:

$$\sigma = \sqrt{\sigma^2}$$

Standard deviation is commonly used in many statistical contexts to describe variability.

Interpretation

- Low Variance: indicates that the data are low dispersed and close to the average.
- High Variance: indicates that the data are highly dispersed and far from the average

Relationship between the two distributions

The two graphs, generated by the `drawChartAbsolute` and `drawChartRelative` functions, depict the same simulation of cyber attacks but use two different methods to represent the data: **absolute frequency** and **relative frequency**.

Absolute Frequency

This graph shows the total number of successful attacks for each level of compromised systems (from -systems to +systems), using a histogram that represents the actual counts. The graph traces lines indicating the penetration score of each attacker and draws a horizontal histogram representing the number of attackers reaching each level. The height of each bar is proportional to the absolute number of successes.

Relative Frequency

The graph shows the percentage of successes compared to the total number of attacks for each level of compromised systems. The height of the bars in the final chart represents the relative frequency of successful attacks. This graph provides a clearer view of the effectiveness of the attacks in relation to the number of attackers by showing the frequency with which each level was successful relative to the total number of attacks. Note how each frequency is always between 1 and minus 1.

Relationship

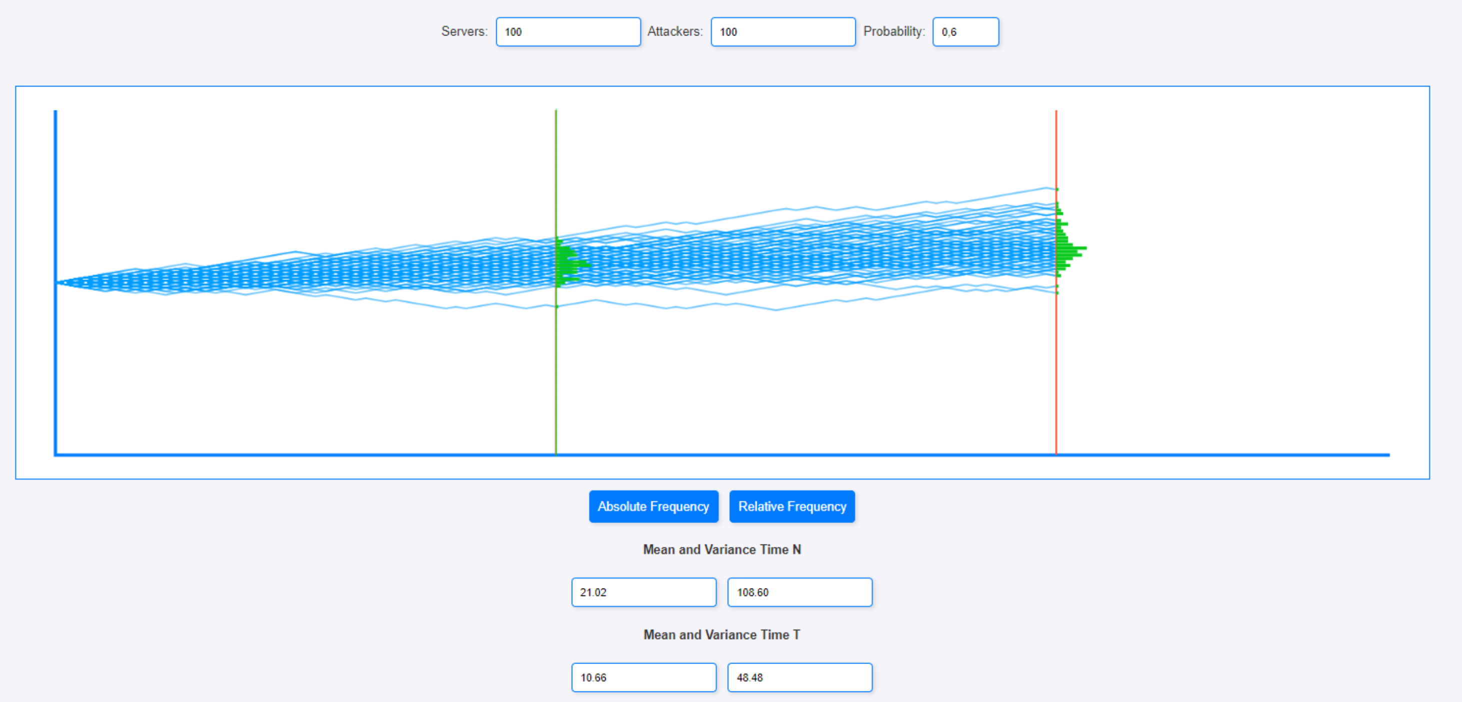
- **Complementarity**: The graphs complement each other. The absolute frequency graph provides an overview of the attackers' performance, while the relative frequency graph offers an understanding of their efficiency.
- **Data Interpretation**: If the absolute frequency graph shows a high number of successes at a specific level, the relative frequency graph may indicate whether that level was reached frequently relative to the total number of attempts.
- **Statistical Analysis**: Both graphs provide information on mean and variance, but the context changes. The mean of absolute frequencies is useful for evaluating overall impact, while the mean of relative frequencies indicates the consistency of attackers' performance.

Difference

The difference between absolute and relative frequency in data representation lies in how each conveys information about the outcomes of a particular event, such as successful attacks in a simulation. Absolute frequency focuses on the raw counts of occurrences, illustrating the total number of attackers that successfully reached each level of compromised systems. This method provides a clear view of the impact and scale of the attacks, making it easy to grasp the total success rate at a glance. In contrast, relative frequency expresses these successes as a proportion of the total number of attempts, offering insights into the effectiveness of the attacks relative to the overall attacker population. By normalizing the data, relative frequency allows for comparisons across different scenarios, highlighting patterns in performance that might be obscured by sheer count alone.

Observation about mean and variance

Absolute



Servers: 100 Attackers: 100 Probability: 0,6

Mean and Variance Time N

21.02 108.60

Mean and Variance Time T

10.66 48.48

Mean Comparison:

- **At Time T:** The mean of **10.66** indicates that, on average, attackers have compromised approximately 11 systems. This suggests a low to moderate level of success in the early phase of the simulation.
- **At Final Time (N):** The mean increases to **21.02**, reflecting improved performance and effectiveness of the attackers over time. This increase indicates that as attackers gain experience, their average success in compromising systems also increases.

Variance Comparison:

- **At Time T:** The variance of **48.48** indicates a relatively low spread of data points around the mean. This suggests that the attackers' performances are somewhat consistent, with most attackers achieving results close to the mean. Lower variance often indicates that there are fewer outliers or extreme values in the data.
- **At Final Time (N):** The variance increases to **108.60**, indicating a greater spread in the performance of attackers. This higher variance suggests that while some attackers have become very successful, others have struggled significantly. The presence of higher variance implies a

more diverse range of outcomes, indicating that the behaviors or strategies of attackers are less uniform than at time T.

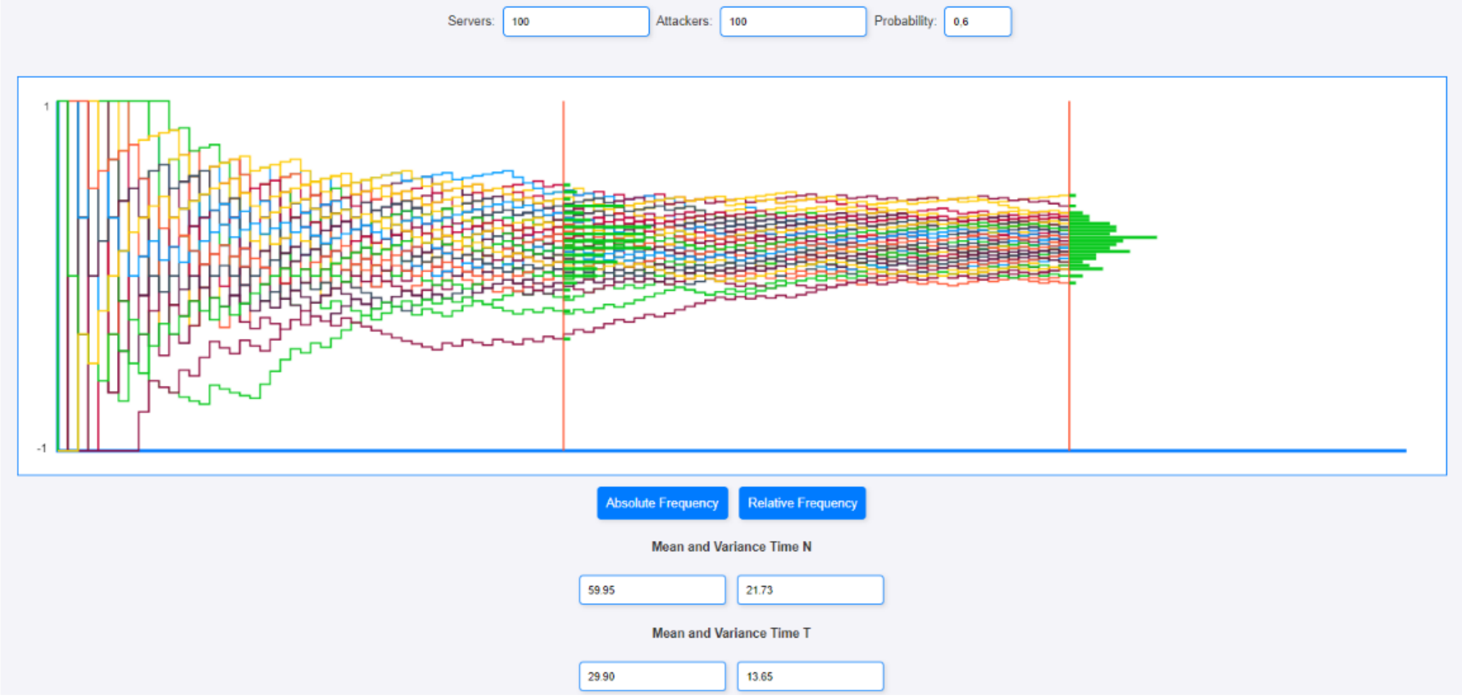
Distribution Shape:

In the graph, between $t = \frac{n}{2}$ and $t = n$, typical behavior of a normal or Gaussian distribution is observed, characterized by an increase in the dispersion of the data. At this stage, the distribution gradually widens and moves away from the mean, assuming a flatter shape. In the context of the normal distribution, this phenomenon reflects the property of **increasing dispersion**: initially, the observations are more concentrated around the mean, but as one approaches $t = n$, one notices that the observations gradually move away from the mean, creating an increasingly broader distribution with less steep tails.

Info

Curve Width: In the normal distribution, the width of the curve is determined by the standard deviation σ. A larger value of σ results in a wider curve, as an increase in variance leads to greater dispersion of data around the mean. Between t = n/2 and t = n, the random system (which depends on the variables n: number of servers, m: number of attackers, and r: probability of breach) produces increasing variability among the values, which in turn raises the variance σ² and consequently the width of the normal curve.

Relative



Servers: 100 Attackers: 100 Probability: 0,6

Mean and Variance Time N

59.95 21.73

Mean and Variance Time T

29.90 13.65

As we can see, with the same input data the mean and variance values change if we use another type of distribution called: "frequency distribution". It is composed by -1 and +1 jump.

Mean Comparison:

- **Time T: Mean = 29.90**: This indicates that, on average, each attacker successfully compromised about 30 systems. Given the number of systems (100) and the success probability (0.6), this means attackers are performing reasonably well but not exceptionally.
- **Time N: Mean = 59.95**: The significant increase in mean performance at the final time indicates that attackers have improved, on average, to nearly 60 compromised systems. This suggests either learning effects or more effective strategies over time.

Variance Comparison:

- **Time T: Variance = 13.65**: A lower variance indicates that the attackers' performances are relatively consistent. Most attackers are achieving similar levels of success.
- **Time N: Variance = 21.73**: The increase in variance signifies a growing disparity in performance among attackers. This could mean that while some attackers have become very effective, others may have lagged behind, leading to a wider range of outcomes.

Distribution Shape:

When we increase the number of trials, the distribution of relative frequencies tends to stabilize. This means that the proportions of successes become more constant and less subject to random variation.

⚠ Important

The variance is a measure of how spread out the data is. Even though the bell is narrower at time N than at time N-2, the data points are more spread out from the mean.
the variance at time N is calculated over the entire dataset, while the variance at time N-2 is calculated over a smaller subset of the data. This means that the variance at time N will be influenced by all the data points, even those that are far from the mean.

> Finally, we note how the histograms representing the Gaussian curve are reversed between the first graph and the second graph. This is due to two different behaviors of the distributions identified by the variance calculation.

Welford Recursion

The Wellford recursion is a technique used to efficiently compute the sample variance of a dataset in a numerically stable manner. It provides a robust solution for calculating sample variance, particularly useful in statistical programming and data analysis where performance and precision are critical.

We start by stating:

$$\sigma^2(x) = \frac{\sum(x_i - \overline{x})^2}{n}$$

We know that:

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

to derive the recursive formula of variance, we find the recursive formula of standard deviation.

1. We write the standard deviation so that it depends on the $(n-1)$-th element:

$$\sigma_n(x) = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \Rightarrow \sigma_n(x) = \sum_{i=1}^{n-1}[(x_i - \overline{x}_{n-1})^2] + (x_n - \overline{x}_n)^2$$

2. Using the recursive average formula, we can rewrite:

$$\overline{x}_n = \overline{x}_{n-1} + \frac{1}{n}(x_n - \overline{x}_{n-1}) \Rightarrow (x_i - \overline{x}_n) = (x_i - (\overline{x}_{n-1} + \frac{1}{n}(x_n - \overline{x}_{n-1})))$$

$$(x_i - \overline{x}_n)^2 = (x_i - \overline{x}_{n-1})^2 - \frac{2}{n}(x_i - \overline{x}_{n-1})(x_n - \overline{x}_{n-1}) + \frac{1}{n^2}(x_n - \overline{x}_{n-1})^2$$

$$\sigma_n(x) = \sum_{i=1}^{n-1}[(x_i - \overline{x}_{n-1})^2 - \frac{2}{n}(x_i - \overline{x}_{n-1})(x_n - \overline{x}_{n-1}) + \frac{1}{n^2}(x_n - \overline{x}_{n-1})^2] + (x_n - \overline{x}_n)^2$$

3. Decomposing the summation we obtain:

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

which turns out to be $\sigma_{n-1}(x)$.

$$-\sum_{i=1}^{n-1}\frac{2}{n}(x_i - \overline{x}_{n-1})(x_n - \overline{x}_{n-1}) = \sum_{i=1}^{n-1}(x_i - \overline{x}_{n-1}) = 0 \qquad \text{All this parts is cancelled}$$

$$\sum_{i=1}^{n-1}\frac{1}{n^2}(x_n - \overline{x}_{n-1})^2$$

does not depend on i, consequently we can rewrite it as:

$$(n-1)\frac{1}{n^2}(x_n - \overline{x}_{n-1})^2$$

6. Rewriting the entire formula we thus obtain:

$$\sigma_n(x) = \sigma_{n-1}(x) + (n-1)\frac{1}{n^2}(x_n - \overline{x}_{n-1})^2 + (x_n - \overline{x}_n)^2$$

$$(x_n - \overline{x}_n)^2 = (x_n - (\frac{n-1}{n})(\overline{x}_{n-1} + \frac{1}{n}x_n))^2$$

1. Collecting:

$$(x_n(1-\frac{1}{n})-\frac{n-1}{n}(\overline{x}_{n-1}))^2=(\frac{n-1}{n}x_n-\frac{n-1}{n}\overline{x}_{n-1})^2$$

2. We further collect by $\frac{n-1}{n}$ thus obtaining:

$$(\frac{n-1}{n}(x_n-\overline{x}_{n-1}))^2=(\frac{n-1}{n})^2(x_n-\overline{x}_{n-1})^2=\frac{(n-1)^2}{n^2}(x_n-\overline{x}_{n-1})^2$$

$$\sigma_n(x)=\sigma_{n-1}(x)+\frac{n-1}{n^2}(x_n-\overline{x}_{n-1})^2+\frac{(n-1)^2}{n^2}(x_n-\overline{x}_{n-1})^2$$

9. Collecting by $(x_n-\overline{x}_{n-1})$ we obtain:

$$\sigma_n(x)=\sigma_{n-1}(x)+(x_n-\overline{x}_{n-1})(\frac{n-1}{n^2}+\frac{(n-1)^2}{n^2})$$

1. Collecting the last term for $\frac{n-1}{n^2}$ we obtain:

$$\frac{(n-1)(n-1+1)}{n^2}=\frac{(n-1)n}{n^2}=\frac{n-1}{n}$$

10. Let's rewrite:

$$\sigma_n(x)=\sigma_{n-1}(x)+\frac{n-1}{n}(x_n-\overline{x}_{n-1})^2=\sigma_{n-1}(x)+\frac{n-1}{n}(x_n-\overline{x}_{n-1})(x_n-\overline{x}_{n-1})$$

1. Again from the recursive averaging formula, we know that:

$$\overline{x}_n=\frac{n-1}{n}(\overline{x}_{n-1})+\frac{1}{n}x_n$$

$$(x_n-\overline{x}_n)=x_n-(\frac{n-1}{n}(\overline{x}_{n-1})+\frac{1}{n}x_n)$$

3. Collecting:

$$x_n(1-\frac{1}{n})-\frac{n-1}{n}\overline{x}_{n-1}=\frac{n-1}{n}x_n-\frac{n-1}{n}\overline{x}_{n-1}$$

4. Collecting further for $\frac{n-1}{n}$:

$$\frac{n-1}{n}(x_n-\overline{x}_{n-1})=(x_n-\overline{x}_n)$$

11. Substituting then we obtain:

$$\sigma_n(x)=\sigma_{n-1}(x)+\frac{n-1}{n}(x_n-\overline{x}_{n-1})(x_n-\overline{x}_{n-1})=\sigma_{n-1}(x)+(x_n-\overline{x}_n)(x_n-\overline{x}_{n-1})$$

Final Formula:

- **Standard Deviation**:

$$\sigma_n(x)=\sigma_{n-1}(x)+(x_n-\overline{x}_{n-1})(x_n-\overline{x}_n)$$

- **Variance**:

$$\sigma_n^2(x) = \frac{\sigma_{n-1}(x) + (x_n - \overline{x}_{n-1})(x_n - \overline{x}_n)}{n}$$