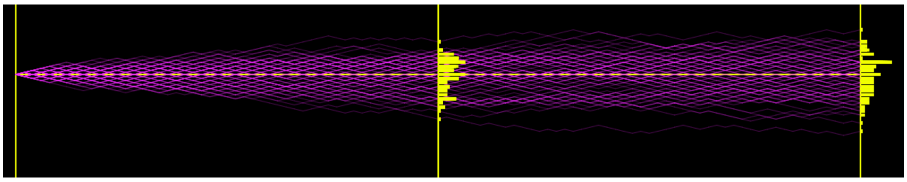
# Homework 2

## Relative Frequency and Absolute Frequency

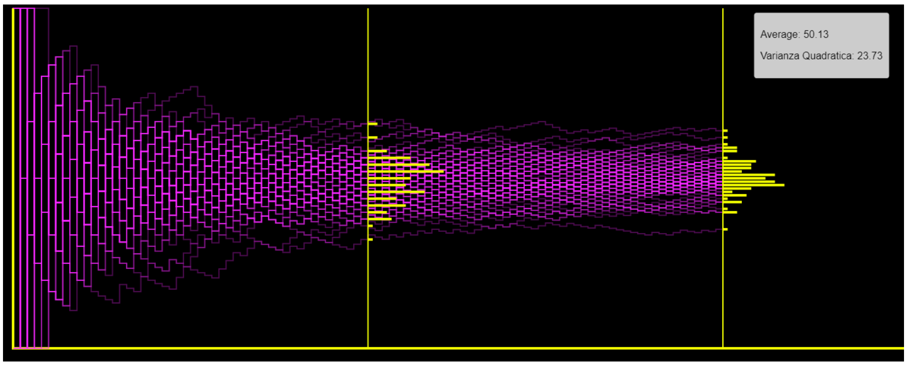
- **Absolute Frequency**: is the number of times a given event or result occurs within a data set. It is an integer value greater than $0$:

$$f_a = \text{counting \ events}$$



- **Relative Frequency**: is the absolute frequency of an event divided by the total number of observations. It is expressed as a fraction or percentage of the total, thus assuming values in the range $[0,1]$:

$$f_r = \frac{\text{total number of observations}}{f_a}$$



## Deviation

The **deviation** in statistics is a measure that expresses how much a value deviates from a middle value like average, median or fashion of a data set. In other words, it indicates the difference between an observed data and a reference value.

There are different types of deviation, depending on the context and needs:

1. **Simple Deviation**: this is the difference between a single value and the average of the data set:

$$d_i = x_i - \mu$$

   where $x_i$ is the observed value and $\mu$ is the sample or population average.
2. **Absolute Deviation**: is the absolute difference between an observed value and the average. It serves to prevent positive and negative deviations from cancelling each other out:

$$|d_i| = |x_i - \mu|$$

3. **Standard Deviation**: is a more complex measure that represents the square root of the variance and indicates how much the data is dispersed around the average. The more the data are concentrated close to the average, the lower the standard deviation:

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

where $\sigma$ is the standard deviation, $N$ is the total number of data, and $x_i$ represents the individual values.

4. **Average Deviation**: is the average of the absolute deviations of each value from the average of the data set. It is used to quantify the dispersion of the data from the average, without the use of squares:

$$\text{Average Deviation} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu$$

### Interpretation:

- **Null deviation**: if the deviation of a value is zero, it means that that value is exactly equal to the average of the data set.
- **Positive or negative deviation**: if the deviation is positive, it means that the value is above the average; if it is negative, it means that it is below the average.
- **High standard deviation**: indicates that the values are very dispersed around the average, so there is a lot of variability in the dataset.
- **Low standard deviation**: indicates that the values are concentrated close to the average, so there is little variability.

---

## Variance

Variance, represented by $\sigma^2$, is a measure of the dispersion of data around the average. Variance is a fundamental statistic that describes how much the values of a set of data differ on average from the average. More formally, variance is the average of the squares of the deviations of each value from the average.

### Variance Formula:

For a population, the variance is calculated as:

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

Where:

- $\sigma^2$: variance.
- $N$: total number of elements in the sample or population.
- $x_i$: single value.
- $\mu$: average of the sample or population.

If you calculate the variance for a sample (instead of for the whole population), you use a slightly different formula, called **sample variance**, which corrects the estimate by dividing by $N-1$ instead of by $N$:

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

Where:

- $s^2$: sample variance
- $\overline{x}$: sample average.

### Relation to Standard Deviation:

The variance is simply the square of the **standard deviation** ($\sigma$). Therefore, to obtain the standard deviation from the variance, we take the square root:

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

This implies that the variance and the standard deviation are closely related, but the variance is expressed in terms of squared units with respect to the original data.

### Interpretation of Variance:

- **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.

## Demonstration of the Recursive Formula of Variance

We start by stating:

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

We know that:

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

Therefore, to derive the recursive formula of the variance, we must first find the recursive formula of the standard deviation. Below is the procedure with an explanation of the various steps:

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:

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

3. Elevating to the square we obtain:

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

4. Let us rewrite the summation:

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

5. Decomposing the summation we obtain:

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

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

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

      but we know that :

      $$\sum_{i=1}^{n-1}(x_i - \overline{x}_{n-1}) = 0$$

      so the whole section is cancelled.

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

      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:

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

7. We again use the recursive average formula to rewrite the last term:

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

   1. Collecting by $x_n$ we obtain:

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

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

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

8. By substituting in the general formula:

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

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:

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

    2. And so:

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

    3. Collecting by $x_n$ we obtain:

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

    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**:

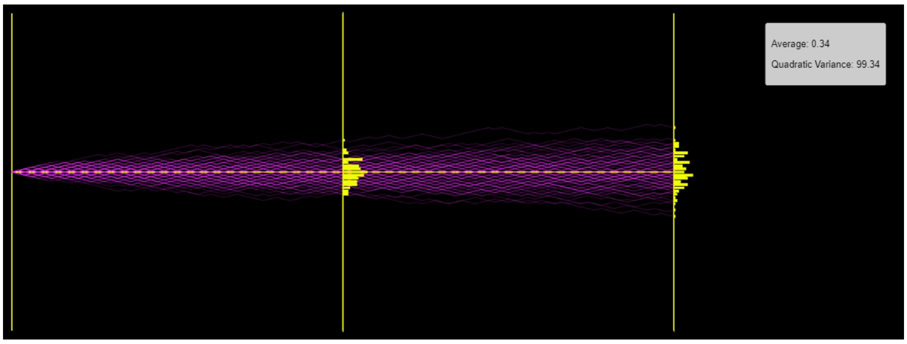
$$\sigma_n(x) = \sigma_{n-1}(x) + (x_n - \overline{x}_{n-1})(x_n - \overline{x}_n)$$

- **Variance**:

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

## Observations

### Chart 1



As time progresses, the randomness and variability of the variables involved tend to amplify the dispersion of values. This results in a progressive widening of the distribution between $t = n/2$ and $t = n$, where the distribution moves away from the average and assumes a broader shape. This phenomenon is consistent with the law of large numbers and the effect of the summation of random variables, which, with the passage of time, leads to a greater dispersion of the data around the average, until a flatter Gaussian curve is formed.

The behaviour of the graph has to do with the **normal distribution** and the formula that describes it. The spread observed between $t = n/2$ and $t = n$ is in fact related to the properties of the **normal distribution** (or Gaussian), which is described by the following formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

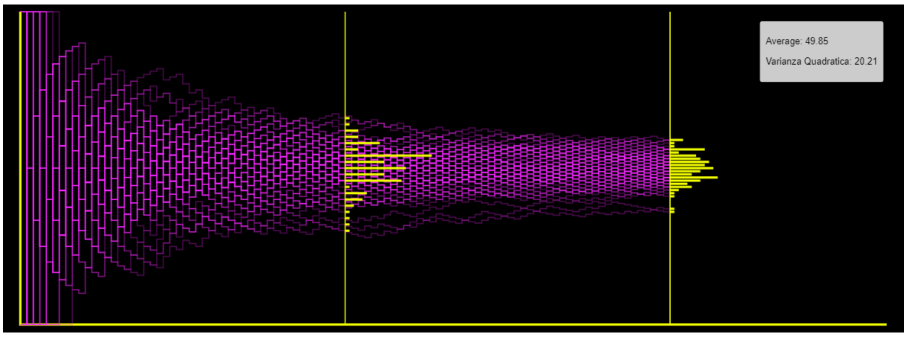
Where:

- $\mu$: **average** of the distribution.
- $\sigma$: **standard deviation**.
- $\sigma^2$: **variance**.
- $x$: random variable.

### How does this relate to the graph?

1. **Curve broadening**: in the normal distribution, the **width** of the curve depends on the **standard deviation** $\sigma$. The larger $\sigma$ is, the wider the curve is 'spanned' or broader, as the variance increases the dispersion of the data around the mean. Between $t = n/2$ and $t = n$, the random system (dependent on the variables $n$: number of servers, $m$: number of attackers, and $r$: probability of breaching) generates an increasing **variability** between the values, increasing the variance $\sigma^2$ and, consequently, the width of the normal curve.
2. **Sum of random variables effect**: the normal distribution emerges naturally when we sum many independent random variables (central limit theorem). If every fit, position or variable in your system follows a random pattern, over time individual deviations accumulate, widening the distribution and leading to the formation of a wider Gaussian curve (higher variance).
3. **Tendency towards a flatter curve**: When the variance $\sigma^2$ increases over time, the term $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ in the normal formula leads to a reduction in the height of the curve and a widening of its base. This reflects a greater scattering of the data and a 'spreading' phenomenon observed between $t = n/2$ and $t = n$.

### Chart 2



As we can see from the box on the top right, with the same input data the mean and variance values change if we use a frequency distribution.

- **Average**: the system simulates the movement of the attackers, the average value indicates that, as the random process evolves, most of the attackers tend to stabilise or converge towards a central position.
- **Variance**: the reduction in variance indicates that, over time, the data is stabilising around the new mean. The distribution is ‘tighter’, with less variation between the positions of different attackers or simulated units. The reason for the decrease in variance is due to a process of ‘normalisation’ or ‘clustering’ of attackers in positions closer together, which reduces the overall dispersion of the system.

Finally, we note how the histograms representing the Gaussian curve are inverted between the first graph and the second. This is again due to the fact that the variance in the first case tended to increase, while in the second it tended to decrease and, as explained above, the variance determines the ‘skewing’ of the curve.