**Lesson 4**

**Researches about theory (R)**

5\_R. Explain a possibly unified conceptual framework to obtain all most common measures of central tendency and of dispersion using the concept of distance (or "premetric", or similarity in general). Discuss why it is useful to discuss these concepts introducing the notion of distance. Finally, point out the difference between the mathematical definition of "distance" and the properties of the "premetrics" useful in statistics, pointing out trhe most important distances, indexes and similarity measures used in statistics, data analysis and machine learning (such as for instance; Mahalanobis distance, Euclidean distance, Minkowski distance, Manhattan distance, Hamming distance, Cosine distance, Chebishev distance, Jaccard index, Haversine distance, Sørensen-Dice index, etc.).

In Statistics, we have a lot of methods to extract data about a population or just a sample of that. When we have a set of items, let's take the sets of all (10 in this example :)) students in the Cybersecurity course (as always), we can apply all of these functions to take informations about them. The central tendency is the thing that tries to describe in a meaningful way the distribution we're analyzing. So, the goal of this measures is to extract few data/values that can be so much representative of the larger set of objects. Central tendency is often called Central location or Summary Statistics. The mean, the median and the mode are the most common ways to extract some meaningful data. Maybe the mean (also called average) is the most common and famous. We've seen in previous lessons some types to compute the arithmetic mean (but there are also other types of these, like the geometric mean), this value explains just the average value of the larger sets. The most common way to compute this is the naive computation (so summing all the values and then divide it by the number of the set), but we've seen this is a good way to get errors (from the compiler). Other formulas/methods to computer in a good way the mean/average is using the Kahan Sum or the Knuth algorithm. So, just to take a general visual example, let's take and measure the students.

| Student | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Heigth | 160 | 161 | 163 | 158 | 160 | 157 | 169 | 170 | 200 | 210 |

Let's visualize the distribution of the students' height.

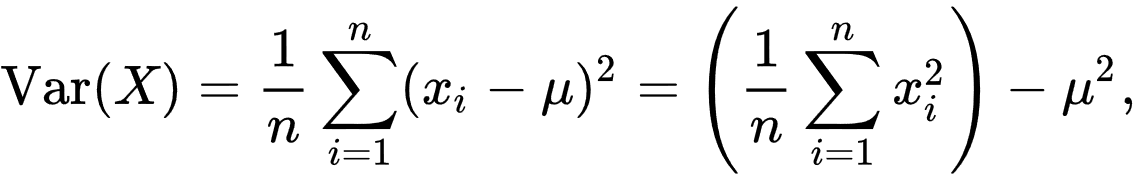
Chart, bar chart

Description automatically generated

We can immediately see that there are two outliers in this sample of data. There are 8 students with an arithmetic mean of the heigth of 162.25cm, and other two student with a mean of 205cm. The mean of the total sample is JUST 170.8cm. So, this isn't a good value for this distribution, we must apply other types of measures of the central tendency. One other method is the median. The median is the value in the middle of a sorted list of data. So, if we take our students and we sort the data, we have this:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 157 | 158 | 160 | 160 | 161 | 163 | 169 | 170 | 200 | 210 |

Now we can see there isn't a value just in the middle, we have two values in the middle because they have 4 elements on the left and 4 elements on the right. Or better, 4 values less or equal than them and 4 values greater or equal than them. So, we have two median values, but they are 161cm and 163 cm and these values don't describe in a good way our distribution. Another type of measurement is the mode. The mode is the most common and recurrent value in the distribution. As we can see in the previous chart (histogram) the mode is 160cm, because is the only value present two times.  
But we have other types of measures of central tendency that help us to understand better the data. I'm calling about Variance and Standard Deviation. As written on Wikipedia: *"In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers is spread out from their average value. Variance has a central role in statistics, where some ideas that use it include descriptive statistics, statistical inference, hypothesis testing, goodness of fit, and Monte Carlo sampling. Variance is an important tool in the sciences, where statistical analysis of data is common. The variance is the square of the standard deviation, the second central moment of a distribution, and the covariance of the random variable with itself, and it is often represented by sigma^2 or Var(X)."* So, the actual Variance is computed summing from i=1 to n all the (xi - mean)^2, all divided by n.



The variance of this distribution is Variance (σ^2): 313.76cm^2. And finally, we have the Standard Deviation. By Wikipedia:  
*In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values.[1] A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range. Standard deviation may be abbreviated SD and is most represented in mathematical texts and equations by the lower-case Greek letter sigma σ, for the population standard deviation, or the Latin letter s, for the sample standard deviation.*  
The standard deviation or sigma is obtained computing the square root of the sigma^2 or Variance, so in our case is 17.713271860388 cm.

Diagram, schematic

Description automatically generated

The measure of dispersion is important because it can show us the margin of error when we have a sample of a larger population and we want to make inferences, these measures play an important role in any dataset. These measures go along the measures of central tendency and show us the variability of our data. Measures of central tendency, as we have seen before, summarize the data with one or few values to explain better the population/sample. As we have seen before on the previous points, especially with the practical examples of the Cybersecurity class, these measures of central tendency (like mean, median, mode etc..) can't show us exactly or in a good and realistic way the data. So, these measures must be correlated with the measures of dispersion (Range, Average deviation, Variance, Standard Deviation...).  
The greater the dispersion in a sample, the more space you'll need to work within the margin. In other words, greater the dispersion, less representative your central tendency is (as we can see in the practical example).

We have different meanings of “distance”, from Wikipedia: *Distance is a numerical measurement of how far apart objects or points are. In Physics or everyday usage, distance may refer to a physical length or an estimation based on other criteria (e.g., "two counties over"). The distance from a point A to a point B is sometimes denoted as |AB|{\displaystyle |AB|}. In most cases, "distance from A to B" is interchangeable with "distance from B to A". In mathematics, a distance function or metric is a generalization of the concept of physical distance; it is a way of describing what it means for elements of some space to be "close to", or "far away from" each other. In psychology and social sciences, distance is a non-numerical measurement; Psychological distance is defined as "the different ways in which an object might be removed from" the self along dimensions such as "time, space, social distance, and hypotheticality.*

*In mathematics, the Euclidean distance between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, therefore occasionally being called the Pythagorean distance. The distance between two objects that are not points is usually defined to be the smallest distance among pairs of points from the two objects. Formulas are known for computing distances between different types of objects, such as the distance from a point to a line. In advanced mathematics, the concept of distance has been generalized to abstract metric spaces, and other distances than Euclidean have been studied. In some applications in statistics and optimization, the square of the Euclidean distance is used instead of the distance itself.*

**Applications / Practice (A)     [work on this at least 30' a day, all days]**

6\_A. (For this exercises use only 1 language chosen between C# or VB.NET, according to your preference)  
  
Prepare separately the following charts: 1) Scatterplot, 2) Histogram/Column chart [in the histogram, within each class interval, draw also a vertical colored line where lies the true mean of the observations falling in that class] and 3) Contingency table, using the graphics object and its methods (Drawstring(), MeasureString(), DrawLine(), etc).  
Use them to represent 2 numerical variables that you select from a CSV file. In particular, in the same picture box, you will make at least 2 separate charts: 1 dynamic rectangle will contain the contingency table, and 1 rectangle (chart) will contain the scatterplot, with the histograms/column charts and rug plots drawn respectively near the two axis (and oriented accordingly).

Video + code

**Researches about applications (RA)**

4\_RA. Do a personal research about the real world window to viewport transformation, and note separately the formulas and code which can be useful for your present and future applications.

Window to Viewport Transformation is the process of transforming a 2D world-coordinate objects to device coordinates. Objects inside the world or clipping window are mapped to the viewport which is the area on the screen where world coordinates are mapped to be displayed.

A picture containing text, whiteboard

Description automatically generated

General Terms:

* World coordinate – It is the Cartesian coordinate w.r.t which we define the diagram, like Xwmin, Xwmax, Ywmin, Ywmax.
* Device Coordinate – It is the screen coordinate where the objects is to be displayed, like Xvmin, Xvmax, Yvmin, Yvmax.
* Window – It is the area on world coordinate selected for display.
* ViewPort – It is the area on device coordinate where graphics is to be displayed.

Taken (Xw, Yw) as a point on Window, we must find the corresponding point on ViewPort.

A picture containing text, shoji, whiteboard, public

Description automatically generatedA picture containing text, shoji, wall, whiteboard

Description automatically generated

where, sx is scaling factor of x coordinate and sy is scaling factor of y coordinate

A screenshot of a computer

Description automatically generated with low confidenceA screenshot of a computer

Description automatically generated with low confidence

Also, on this resource I also found a good program to compute the window to viewport transformation

// C# program to implement

// Window to ViewPort Transformation

using System;

class GFG

{

// Function for window to viewport transformation

static void WindowtoViewport(int x\_w, int y\_w,

int x\_wmax, int y\_wmax,

int x\_wmin, int y\_wmin,

int x\_vmax, int y\_vmax,

int x\_vmin, int y\_vmin)

{

// point on viewport

int x\_v, y\_v;

// scaling factors for x coordinate

// and y coordinate

float sx, sy;

// calculatng Sx and Sy

sx = (float)(x\_vmax - x\_vmin) /

(x\_wmax - x\_wmin);

sy = (float)(y\_vmax - y\_vmin) /

(y\_wmax - y\_wmin);

// calculating the point on viewport

x\_v = (int) (x\_vmin +

(float)((x\_w - x\_wmin) \* sx));

y\_v = (int) (y\_vmin +

(float)((y\_w - y\_wmin) \* sy));

Console.Write("The point on viewport: " +

"({0}, {1} )\n ", x\_v, y\_v);

}

}

// This code is contributed by PrinciRaj1992

**References**

[Laerd Statistics - Measures of Central Tendency](https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php)  
[WikiPedia - Summary Statistics](https://en.wikipedia.org/wiki/Summary_statistics)  
[WikiPedia - Central Tendency](https://en.wikipedia.org/wiki/Central_tendency)  
[WikiPedia - Variance](https://en.wikipedia.org/wiki/Variance)  
[WikiPedia - Standard Deviation](https://en.wikipedia.org/wiki/Standard_deviation)

[WikiPedia - Distance](https://en.wikipedia.org/wiki/Distance)  
[GeeksForGeeks - Window to Viewport Transformation in Computer Graphics with Implementation](https://www.geeksforgeeks.org/window-to-viewport-transformation-in-computer-graphics-with-implementation)