**Lesson 3**

**Researches about theory (R)**

4\_R. Explain what are marginal, joint and conditional distributions and how we can explain the Bayes theorem using relative frequencies. Explain the concept of statistical independence and why, in case of independence, the relative joint frequencies are equal to the products of the corresponding marginal frequencies.

To help us with the goal of understanding these types of distributions (marginal, joint, and conditional) I'm gonna show the graphic view of **contingency table**. I’m gonna show these things referring to a bivariate distribution. As we've already seen in the first lesson, a univariate distribution it's a distribution considering only one characteristic of the analyzed population/sample (for instance, the age of the Statistics course students). The bivariate distribution shows two characteristics about the population/sample, in general we can say multivariate distribution (n characteristics, with n>=2). Let's take an example to explain these types of data. We take our Statistics class and make a bivariate distribution for it (fake data, just an example). We have 300 students (this is our population) and we want to show the distribution considering the age and the sex of the students.

| A\S | M | F |
| --- | --- | --- |
| 0-23 | 43 | 28 |
| 24-26 | 77 | 36 |
| 27-30 | 39 | 20 |
| 31-37 | 18 | 19 |
| 38+ | 8 | 12 |

So, in the first row we have some labels that are telling us: on the first column you have the whole population divided by sets of age values, instead on the first row you have the division by the sex. In the central values (so all the other cells) we have the i,j value; so the value involving the i-th row and the j-th row. For example, we can read that the 24-26 years old males are 77, and so on...

| A\S | M | F | d(A) |
| --- | --- | --- | --- |
| 0-23 | 43 | 28 | 71 |
| 24-26 | 77 | 36 | 113 |
| 27-30 | 39 | 20 | 59 |
| 31-37 | 18 | 19 | 37 |
| 38+ | 8 | 12 | 20 |
| **d(S)** | 185 | 115 | 300 |

Now instead I added a column and a row, specifically the last column and the last row. If you make some little calculations, you can see the cells of the last column contain the sum of the previous two values. Instead, the cells on the last row contain the sum of the previous integer numbers. So, we can see the last column and the last row contain respectively the univariate distribution of the age variable, instead the last row contains the univariate distribution of the sex. Obviously, the last cell of the last row (300) it's the total number of the items inside the population. This graphical view of a bivariate distribution is called contingency table.  
In this view the last row and the last column represent two marginal distributions. The distribution of age and sex, and they're called marginal because we can find them on the margins. Instead, the conditional distribution is composed by all the cells with a specified column. So, for example, I can consider only the column under the M symbol, so considering only the male students. This means my population now is considerely littler. So, we can refer at it as "conditional" because this distribution is conditioned by the fact, we consider only the male students. I can do the same also on the rows, isolating only for instance the row 24-26. Here my population is littler, and I consider only 113 students, which 77 are male and 36 females. The last definition is joint distribution. This refer to the value inside a cell having the condition that the row is the i-th and the column is the j-th. So, for example we know the 0-23 years old male students are 43. "Joint" because we join the data about two variables.  
Now, let's consider only the conditional distribution conditioned by the fact that we consider only the female students. We can write the relative frequencies dividing all the values in that column by 115, so the total of the out conditioned population. So, reading from the top to the bottom we will have: 28/115, 36/115, 20/115, 19/115 and 12/115. Now I see I can express 36/115 as 36/300 x 300/115. The first parameter is just the joint distribution, so the female students being 24-26 years old. I can rewrite that expression in this way: 36/115 = (36/300) / (115/300). The second parameter is just the relative frequency of the female univariate distribution. This is a very interesting relationship, and obviously I can do the same for the rows and I obtain the same relationship (with different values of course).  
Now, let's summarize and make some general assertions. Let n be the population cardinality and n(i,j) the joint distribution of the i-th row and the j-th column, let n(i) be the population of the i-th row and n(j) the population of the j-th column. We can say that n(i,j) / n(j) = ( n(i,j) / n ) / ( n(j) / n ). The first object of the second part is the relative frequency of the joint distribution n(i,j) and the second object of the second part is the relative frequency of n(j). So, the object of the first part is the relative frequency of the i-th row conditioned by the fact we consider the population of the j-th column.  
This relationship is called the **Bayes theorem**.

Independence is a fundamental notion in probability theory, as in statistics and the theory of stochastic processes. Two events are independent, statistically independent, or stochastically independent if the occurrence of one does not affect the probability of occurrence of the other (equivalently, does not affect the odds). Similarly, two random variables are independent if the realization of one does not affect the probability distribution of the other.  
We know that the conditioned probability of E | F is generally different from the P(E). So, knowing that F has occurred, changes the possibility that E has occurred. In the case we know P(E | F) is equal to P(E), we say these two events are indipendent. Because of P(E | F) = P(E + F) / P(F), we say that E and F are indipendent if P(E + F) = P(E)\*P(F).

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

4\_A. Create a program to read data from a CSV file and store it into a suitable collection of suitably designed objects, for further processing. Compute mean and standard deviation and frequency distribution for at least one of the variables, and for one pair of variables.

TODO  
  
5\_A. Compute a frequency distribution of the words from any text file and create a personal graphical representation of the corresponding "word cloud" (can use animation if you wish), keeping into account the frequencies of the words.

TODO

**Researches about applications (RA))**

2\_RA. Do a review about charts useful for statistics and data presentation (example of some: [StatCharts.txt](https://www.datatime.eu/public/cybersecurity/Statistics_2020_21/StatCharts.txt) ). What is the chart type that impressed you most and why?

I know that the "data visualization" studies the differents ways to present/show data, because there are meaningful ways to show different type of data. For example, with a percentage I would like to use a Pie chart, bacause it's the most understandable way to see a 100% total divided into some categories. Let's see some examples of these different charts.

* Line charts
  + When you want to make predictions based on a data history over time.
  + When you want to show trends. For example, how house prices have increased over time.
  + When comparing two or more different variables, situations, and information over a given period of time.

Chart

Description automatically generated

* Bar charts
  + When you want to display data that are grouped into nominal or ordinal categories (see lesson 2)
  + To compare data among different categories.
  + Bar charts are ideal for visualizing the distribution of data when we have more than three categories.

Chart, bar chart

Description automatically generated

* Pie charts
  + When you want to create and represent the composition of something.
  + To show percentage or proportional data.
  + When comparing areas of growth within a business such as profit.

Chart, pie chart

Description automatically generated

* Histograms
  + When the data is continuous.
  + When you want to represent the shape of the data’s distribution.
  + To summarize large data sets graphically.

Chart, histogram

Description automatically generated

* Area charts
  + When you want to show trends, rather than express specific values.
  + To show a simple comparison of the trend of data sets over the period.
  + To compare a small number of categories.

Chart, histogram

Description automatically generated

I can continue as much as I want, because on the website I took these informations there are listed more than 15 chart types. By the way I listed only the charts I saw more often. I think my favorite charts are the pie chart and the bar chart because in my opinion you see quickly the distribution of categories and you get a fast real idea about what you are seeing.

3\_RA. Do a comprehensive research about the GRAPHICS object and all its members (to get ready to create any statistical chart.)

In the .NET Framework library, six namespaces define managed GDI+: System.Drawing, System.Drawing.Design, System.Drawing.Drawing2D, System.Drawing.Imaging, system.Drawing.Printing, and System.Drawing.Text. Figure 1.3 shows these namespaces. To use any of the classes defined in these namespaces, you must include them in your application.

Diagram

Description automatically generated

**The System.Drawing.Design Namespace**  
As its name suggest, the System.Drawing.Design namespace provides additional functionality to develop design-time controls such as custom toolbox items, graphics editors, and type converters. The System.Drawing.Design namespace also define a few interfaces, delegates, and enumerations.  
**The System.Drawing.Drawing2D Namespace**  
The System.Drawing.Drawing2D Namespace defines functionality to develop advanced two-dimensional and vector graphics applications. This namespace provides classes for graphics containers, blending, advanced brushes, matrices, and transformation. The System.Drawing.Drawing2D Namespace provides dozens of enumerations.  
**The System.Drawing.Imaging Namespace**  
Basic imaging functionality is defined in the System.Drawing.Imaging namespace. The System.Drawing.Imaging namespace provides functionality for advanced imaging. Before an application uses classes from this namespace, it must reference the System.Drawing.Imaging namespace.  
**The System.Drawing.Printing Namespace**  
The System.Drawing.Printing namespace defines printing-related classes and types in GDI+. Before an application uses classes from this namespace, it must include the namespace.  
**The System.Drawing.Text Namespace**  
The System.Drawing.Text namespace contains only a few classes related to advanced GDI+ typography functionality. Before an application uses classes from this namespace, it must include the namespace.