LUISS Guido Carli

LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Department of *Political Science*

Bachelor in *Politics*, *Philosophy* and *Economics*

Chair of *Statistics*

What about Italy abroad?
A statistical analysis about *Il Bel Paese* among young international people

Supervisor

Prof. *Roberto Rocci*

Candidate

Isotta Mormile

076882

Academic Year *2016/2017*

What about Italy abroad?
A statistical analysis about *Il Bel Paese*
among young international people

CONTENTS

Introduction

Paradox: "*a situation, fact or statement which seems impossible and/or difficult to understand because it contains two opposite facts or characteristics*".
(Cambridge International Dictionary).

Italy is a paradox.
Our country, with its traditions, its history and its people has always been considered controversial: if on the one hand Italians are globally esteemed as smart, creative and brilliant, on the other they are simultaneously associated with the words corruption, illegality and crisis.
Indeed Italy is the birthplace of Michelangelo and Dante but at the same time it is the native land of political chicaneries and plots.
The image of the Pope together with that of the boss *mafioso*, the Church and the manifestation of illegal organizations, the beauty of our landscapes and the waste crisis.
How did this discrepancy come about? Why are Italians subjected to prejudices and preconceptions? Why are we loved and hated at the same time?
This thesis is aimed at analyzing in which way Italy is conceived abroad.
It thus consists of a quantitative analysis; we will firstly construct a questionnaire, we will distribute it to a sample composed only by international students and we will analyze it by putting in relation the variables of the questionnaire.
The driven idea for this dissertation comes in the aftermath of the experience as an Erasmus student at Uppsala University, in Sweden. There, I was given the possibility to come into strict contact with a multitude of different people, who came from all around the world. Being absorbed in an international and dynamic environment helped me to understand how differently my native land, Italy, is considered. Some people did not know anything about its geography or its history, others were confused by the most important socio-political Italian personalities, others nourished a remarkable passion for its beauty while others considered Italians as a brilliant and intellectually vibrant people.
"*Rome is esthetically ugly*" was the statement pronounced by a Maths student that particularly touched me. I started thinking, is Rome truly ugly? Has he ever been to Rome? Does he really know the City or maybe he pronounced these words *just to say*, influenced by some form of prejudice?
The dissertation is divided into two chapters; the first consists in an introduction of the statistical tools used to carry out the quantitative research and presents the statistical framework we referred to, the second is the analysis itself of the data and draws the conclusions obtained by the data examination.

1. Statistics: *"The Art and Science of Learning from Data"*

1.1 Statistics

In this thesis Statistics will represent our Dictionary.
The definition that the Oxford Dictionary offers for the word "dictionary" is the following: "*a book or electronic resource that lists the words of a language (typically in alphabetical order) and gives their meaning, or gives the equivalent words in a different language, often also providing information about pronunciation, origin, and usage*".
What Statistics does is essentially what a dictionary can do; the latter translates words into words of another language whereas Statistics converts data into knowledge.
Statistics is the dictionary we will use to carry out our analysis.
To the question *Why would you use Statistics?* the most adequate answer can be given by mentioning a sobering thought by the Chief Economist at Google, Hal Varian "*The sexy job in the next ten years will be statisticians. Because now we really do have essentially free and ubiquitous data. So the complimentary factor is the ability to understand that data and extract value from it*".
Therefore, a scientific research is not only about numbers and calculations, it describes a concept, an idea, a fact and we can *extract value from it*.
The data analysis process falls into two phases: *exploratory* and *confirmatory statistics*. The *exploratory* phase "*isolates patterns and features of the data and reveals these forcefully to the analyst*" (HOAGLIN, MOSTELLER and TURKEY, 2000), during this phase the analyst constructs a new theory.
On the other hand, *confirmatory analysis* "*quantifies the extent to which [deviations from a model] could be expected to occur by chance*" (GELMAN, 2004). During this phase, the analyst can accept or reject an existent theory.
Alan Agresti and Christine Franklin define Statistics as "*the art and science of designing studies and analyzing the data that those studies produce. Its ultimate goal is translating data into knowledge and understanding of the world around us. In short, statistics is the art and science of learning from data*" (AGRESTI and FRANKLIN, 2014). This overlapping of Art and Science offers us the right and proper basis for our thesis, using information found out from studying numbers and variables, analyzing them under a critical point of view and eventually drawing final conclusions is what this thesis in Statistics aims to carry out.

Statistical methods help us investigate questions in an objective and structured way. This process involves four principal steps:
(for further information, consult BLACK, 1999).

(1) **Formulation of a statistical question**, in this first phase of the research, the first aspect to investigate is the relation between the theory and the research that is "*structured in logically sequential phases, according to a substantially deductive approach (theory precedes observation), that strives to support the previously formulated theory with empirical data*". (CORBETTA, 2003). Within this framework, a systematic evaluation of the literature has a pivotal role as it provides the theoretical hypothesis on which the assumptions of the research are based.
A further aspect concerns the relationship between the researcher and the subjects observed. In quantitative research, observation derives from a position that is external to the subject analyzed, it is neutral, detached, merely and uniquely scientific.
The last concept related to the *research planning step* concerns the physical interaction between the researcher and the subjects; "*quantitative research does not envision any physical contact between the researcher and the subject*". (CORBETTA,2003). Efforts are made to reduce the interaction between the subjects studied and the researcher to a minimum.

(2) **Data collection**, in this second stage, the quantitative research is characterized by a structured and closed research design which precedes the research itself. Therefore, all subjects taking part in the research receive an identical treatment which means that the data-collection tool is the same for all cases.(for additional details, consult GOODE and KRUSKAL, 1954). This is realized because the information obtained will be used to create a data-matrix in which the same information is coded for all the cases.
The final point to mention is the nature of the data which are expected to be precise, unequivocal and *hard* (CORBETTA, 2003). Data need to be objectively understood which means that they should lend themselves neither to subjective understanding by the researcher nor to explicit subjectivity of the individual studied. Data should essentially be *standardized* (CORBETTA, 2003) in order to make comparisons among them possible and reasonable.

(3) **Data analysis**, in this third phase, we will refer to statistical tools, "*together with a whole array of tables, graphs, statistical tests, etc., as well as the full set of technological equipment (computers, files, data banks, software, etc.").* (CORBETTA, 2003).

Data analysis is implemented on variables and in an impersonal manner, what we refer to are means, percentages, correlation and association among variables. The aim of the research consists in explaining variation and association among different variables, that is the reason why we will be guided by statistical tools and techniques. (for additional details, consult MAXIM, 1999).

(4) **Interpretation of results**, the ultimate scope of a quantitative research is the generalizability of the results, the conclusion drawn from a quantitative study has as a final end the provision of higher-order synthesis, only in this way theory can be linked to the research.
By breaking down the subjects studied into variables, this form of research achieves a "*preliminary synthesis by correlating these variables (which can be synthesized into numerical indexes such as the correlation coefficient). It then achieves a higher level of conceptualization in the causal model and, in the most successful cases, in the formulation of synthetic expressions that come close to the 'laws' of the natural sciences*". (CORBETTA, 2003). In this regard, the data are presented and exposed in an economical, succinct and compact way. Tables, graphs, histograms, pie charts, frequency tables serve as the adequate synthetic representation of the data obtained.

As an example of the aforementioned four steps, let us describe our research project:

(1) What about Italy abroad? How is Italy conceived and perceived by young international people?
(2) We collect our data through a self-administered questionnaire shared on the social networks.
(3) We analyze our data by constructing frequency and contingency tables on *SAS University Edition* software.
(4) We interpret our data by observing the statistics we implement on *SAS University Edition.*

Statistical tools can be classified into three main categories:

(1) **Design**
(2) **Description**
(3) **Inference**

Design refers to planning how gather data in a way that is efficient and useful in relation to our question.
Description means summarizing and describing the data that we have obtained, and highlighting specific patterns in the data.

Inference refers to the idea of making decisions and predictions based on the data with the intention to answering to the initial statistical question.
Description and inference are complementary steps in the data investigation process, if on the one hand description provides useful summaries and specific patterns in the data, on the other, inference helps us make predictions and determine whether observed patterns are meaningful.

1.2 How to collect data*: Google Forms*

Google Forms (https://docs.google.com/forms), together with *Docs*, *Sheets* and *Slides* is an integrative part of Google's online apps suite and it represents the data collection method implemented for our experiment. It is generally considered an intuitive way to save data directly to a spreadsheet. Initially it was only a characteristic of *Google Sheets*; it was given to users the possibility to add a *form* to a spreadsheet and format it in another sheet. Only in 2016, did Google decided to add more features to *Forms* and eventually create its own stand-alone app. Today *Google Forms* consists in a full-featured *Forms* tool that can be used for free with any Google account.

- How to build a *Google Form*

The initial step in starting to constructing a *Form* is by entering *Google Forms* app by selecting the link: docs.google.com/forms.
Once the page is available there is the possibility to either choose a template or start a new *Form*.
There is also the possibility to select a link to *Google Forms* in *Docs*, *Sheets* and *Slides* by clicking *File* →*New*→*Form* to initiate a blank *Form* (https://gsuite.google.com/learning-center/products/forms/get-started).
Once the *Form* has been opened, it will fill the center of the screen, with a designed area for a title and a description followed by *Form* fields.
By clicking a *Form* field, we are able to edit it and add questions.
In order to choose the field type (short answer, multiple choice, checkboxes) we need to use the *dropdown box* close to the field.

- *Google Forms* Field Options

Google Forms has a multitude of settings options available, specifically it includes 12 field types; 9 different question typologies together with text, video and photo fields. In order to add a new question it is necessary to select the + icon in the right sidebar or click the *text*, *video*, *photo* icons

to attach media to the form. Moreover, each specific field has a *copy* button to duplicate the field with the aim of making easier adding similar questions to the *Form*.

- What does each field type offer?

Title and Description: these two options are added automatically to each *Form* and field, and we are given the chance to add an extra title block anywhere with the *Tt* button. It is indispensable to fill in the main *Form* title whereas it is possible to leave the title and description empty on questions.

Short answer: this field aims to ask for small portions of text as, for example, the name/email/age of the interviewee.

Paragraph: this field is designed for long-form text. We can use this field when we want to collect a detailed feedback from the interviewee.

Multiple choice: this type of question offers the interviewee a list of options, only one can be selected.

Checkboxes: refers to a list of answers and users are given the chance to select as many as they want.

Dropdown: this question works as a multiple choice field but the answers are listed in a menu, the aim of this option is to keep the form compact when there are various answer options.

Linear scale: this field let users pick a number in a range, the scale can be set from 0 or 1 to 2 or 10, with specific labels for the highest and lowest options.

Multiple choice grid: this field organizes questions as rows and the related answers as columns.

Date: if the researcher wants a specific date or time, this field can be used. A date, a month, a year and the time can be selected as answers.

Time: this field requests a length of time in hours, minutes and seconds.

Image: this option gives the interviewer the capacity to upload an image, select it from a link, *Google Drive* or webcam.

Video: this field supports *Youtube* videos.

- *Form* Sections and Logic

Sections break the *Form* up into blocks to answer one set of questions at a time. Indeed each section includes its own specific title and description together with an *arrow* button at the top to display or hide questions.
Full sections cannot be rearranged but questions can be switched between sections. Additionally, sections can be duplicated.
In order to ask interviewee follow-up questions based on their previous answer, there is the possibility to introduce sections with the optional questions.

- Design the *Form*

Google Forms offers the possibility to choose the color/image/theme/color shade of the *Form* by clicking the *color palette* icon in the top right. Colors at disposition are 15.

- Store *Form* of Responses in a Spreadsheet

Answers to the designed questions will be stored and saved by default by *Google Forms* which shows summary graphs and lists of answers with the aim to better analyze the data obtained. We can link the form to a *Google Sheets* spreadsheet, by clicking the green *Sheets* icon in the *Response tab* or selecting response destination in the menu to create a new spreadsheet to store the answers. Luckily, *Google Forms* has been set to keep a full copy of all the data *Form*, if accidentally something from the spreadsheet has been deleted, it can be easily recovered.

- *Form* Sharing Settings

When time to share the *Form* has arrived, in the response options, the researcher can let users submit another response, modify their answers or receive a summary of all responses. A progress bar related to the number of sections completed can be set to give the user a sense of the length of the *Form*.

- Share *Forms* online

The *send* button in the top right of the interface allows the *Form* to be shared via email, social networks or as a copy of the link to the *Form*. The researcher can choose whether to receive notifications via email whenever a *Form* has been completed or not.

1.3 How to analyze data: *descriptive* and *inferential statistics*

- Data

Variables are the subject characteristics observed in a particular study, the term *variable* emphasizes that data values/categories *vary*. In particular, the data values/categories that we observe carrying out a research are denominated observations. Each observation can be either a number or a category.

This thesis will consider uniquely *categorical variables* that are defined as such "*if each observation belongs to one of a set of categories*". (AGRESTI and FRANKLIN, 2014).

Graphs and numerical summaries describe the main patterns of a variable, for *categorical variables*, the key aspect to take in consideration is the relative number of observations in the various categories. The category with the highest frequency is the modal category while "*the proportion of the observations that fall in a certain category is the frequency (count) of observations in that category divided by the total number of observations. The percentage is the proportion multiplied by 100. Proportions and percentages are also called relative frequencies and serve as a way to summarize the measurements in categories of a categorical variable*". (AGRESTI and FRANKLIN, 2014).

What we use to graphically represent how the observations are distributed into several values/categories is a *frequency table* which consists in "*a listing of possible values/categories for a variable, together with the number of observations for each value*". (AGRESTI and FRANKLIN, 2014).

When we are willing to examine our data on two variables, the first step is to distinguish between the *response variable* and *explanatory variable*. The former is the outcome variable on which comparisons are made, the latter is the variable that explains changes in the *response variable*.

To put it differently, the data analyst investigates how the outcome on the *response variable* is dependent on, or is determined by, the value of the *explanatory variable*.

What a quantitative research aims to investigate is whether there is an association between the variables or not and define the nature and the degree of that association. An association exists between the variables whenever a particular value for one variable is more likely to occur with certain values of the other variable. The information about the degree and form of the association between two variables is contained in their joint distribution. A *contingency table* displays the joint distribution of two categorical variables; "*its rows list the categories of one variable and its columns list the categories of the other variable. Each entry in the table is the number of*

observations in the sample at a particular combination of categories of the two categorical variables". (AGRESTI and FRANKLIN, 2014).
Any time we distinguish between an *explanatory variable* and a *response variable*, it is a consequence to create *conditional proportions* which are based on the *explanatory variable*, for categories of the *response variable*. The *cell* represents the combination of each row and columns in the table, while the operation of taking a data file and looking for the frequencies for the cells of a *contingency table* is the *cross-tabulation* of the data.
At this point is necessary to make a distinction between two different methods which are respectively *description* and *inference* in statistical analysis to comprehend how we conducted the research.

- **Descriptive Statistics** indicates a method of summarizing the data obtained. The summaries can be represented by numbers, percentages, averages and graphs.

We will list four statistical indexes that are theoretically relevant in descriptive statistics. They measure the intensity of the association between two variables being 0 when it is absent.

I. *Chi-square Statistic*

"*Introduced by Karl Pearson in 1900, the chi-square statistic is symbolized by* χ^2". (GOODMAN and KRUSKAL, 1954).
It takes non-negative values, the bigger the value is, the strongest the association is among the variables.
If our sample has *v* observations and two variables A and B, we define ν_{ab} as the number of observations presenting category *a* of A and *b* of B. We also indicate with $\nu_a. = \sum_b \nu_{ab}$ the number of observations presenting category *a* of A. Analogously, with $\nu._b = \sum_a \nu_{ab}$ we indicate the number of observations presenting category *b* of B.
This is the formula:

$$\chi^2 = \sum_a \sum_b \frac{(\nu_{ab} - \nu_a.\nu._b/\nu)^2}{\nu_a.\nu._b/\nu} \qquad (1.3.1)$$

It takes values between 0 (independence) and *v* [max(α-1, β-1)] (complete dependence), where α and β are the number of categories of A and B, respectively.
Its main drawback is that its value depends on ν.

II. *Phi Coefficient*

This coefficient has been proposed by Karl Pearson and it is also referred to as the *mean square contingency coefficient* and is denoted by φ. The idea is to divide the *chi-square* by *v* in order to obtain an index that is independent of *v*.

It takes values in between 0 (independence) and $\sqrt{\max(\alpha-1,\beta-1)}$ (complete dependence).
This is the formula:

$$\varphi^2 = \frac{\chi^2}{\nu} \tag{1.3.2}$$

or

$$\varphi = \sqrt{\frac{\chi^2}{\nu}} \tag{1.3.3}$$

Where *v* is the total number of observations.

III. *Coefficient of Contingency*

Additionally, Pearson idealized *C*, which is simply a variation of φ. We use *C* in order to interpret the value of the *chi-square* or φ^2.
It takes values between 0 and 1, when is 0 determines independence among the variables.
This is the formula:

$$C = \sqrt{\frac{\chi^2/\nu}{1+\chi^2/\nu}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}} \tag{1.3.4}$$

The main drawback is that it is less than 1 even if between the two variables there is a relation of complete dependence.

IV. *Cramér's V*

This coefficient is based on Pearson's *chi-square statistic* and published by Harald Cramér. It is denoted by φ_c and can take values 0 and 1. Whether φ_c is determined by 0, there is no association among the variables, if it is 1, the association is complete.

This is the formula:

$$\varphi_c = \sqrt{[\chi^2/\nu]/\mathrm{Min}(\alpha - 1, \beta - 1)} \qquad (1.3.5)$$

- **Inferential Statistics** is uniquely used when data are a random sample. The scope is making decisions or predictions about the entire population, based on data collected from a sample of that population. The final purpose is to evaluate whether two categorical variables are independent or not in the population by using only the data in the sample. A *significant test* helps us to answer this question. A *significance test* is a method of using data to summarize the evidence about a hypothesis. Through the *test statistic* we can examine whether the data support some predictions or not. These predictions are hypothesis made on the population.

A test procedure to make decisions on our data, i.e. decide if the two variables are independent or not, is composed of 5 steps:

I. *Assumptions*

The variables are categorical.

The sample is randomly selected.

II. *Hypothesis*

The hypothesis are statement about the population. For our test they are:

(1) The null hypothesis (H_0): the two variables are independent.

(2) The alternative hypothesis (H_1): the two variables are associated and thus dependent, it specifies how the null can be false.

The idea is to reject H_0 if in the sample there is enough evidence against it.

What the *test* does is essentially to compare the cell counts in the frequency table with counts we would expect to observe if H_0 were true.

How do we do it? We create *expected* cell counts in case of independence to be compared to *observed* cell counts in the *test statistics*.

III. *Test Statistics*

The *Test Statistic* measures the amount of evidence against the null contained in the sample. This is done by measuring how close the *observed* cell counts fall to the *expected* cell counts. The formula has been given previously by the explanation of the χ^2 (see formula $^{(1.3.1)}$).

The *expected* cell counts are those values that satisfy the null hypothesis of independence. For a specific cell, the *expected* cell count is found by multiplying the row total and the column total and dividing it by the total sample size. In formulas, $\frac{v_{a\cdot}v_{\cdot b}}{v}$ is the *expected* count corresponding to the *observed* count v_{ab}.

Example, analysis of the number of *males* and *females* who come from *Italy* or *USA*.

I. Frequency table with *observed* cell counts

Gender	**Where are you from?**		
	Italy	USA	Total
MALE	1	2	3
FEMALE	3	4	7
Total	4	6	10

(2) Frequency table with *expected* cell counts

Gender	**Where are you from?**		
	Italy	USA	Total
MALE	1.2	1.8	3
FEMALE	2.8	4.2	7
Total	4	6	10

The numbers in the cells of the two tables are not equal, the first table indicates the number of observations that we have obtained while the second tables computes the *expected* cell counts. The values in the two tables are not the same, this means that the two variables *Gender* and *Country of origin* are not independent.

SAS University Edition in the PROC FREQ function calculates also the *Likelihood Ratio Chi-Square Test Statistics* that "*involves the ratios between the observed and expected frequencies*". (SAS INSTITUTE INC., 2016).

The *likelihood ratio chi-square* is computed as

$$G^2 = 2 \sum_a \sum_b v_{ab} e\ (v_{ab}/e_{ab}) \quad (1.3.6)$$

where v_{ab} is the *observed* frequency in the table cell (*a*,*b*) and e_{ab} is the corresponding *expected* frequency.

IV. *P-value*

It refers to the probability to observe values more extreme -greater- than the one *observed* when H_0 is true. Small values of this probability indicates a large amount of evidence against the null hypothesis.

V. *Conclusion*

After reporting the *p-value*, we have to interpret it in the context of the study. Based on the *p-value*, we can make a decision about H_0. Before examining the data, we establish how small the *p-value* would need to be to reject H_0. This cutoff point is the *significance level*. It is important to recall that the *significance level* corresponds to the probability of rejecting the null when it is true.
The *significance level* we set is 0.05. H_0 is rejected when *p-value* ≤ *significance level* (in our case 0.05).

1.4 How to implement a statistical analysis: *SAS University Edition*

This thesis will realize a quantitative analysis through the support of a specialized and advanced statistical software: the *SAS University Edition*.
This paragraph will explain what *SAS University Edition* is and how specifically worked in relation to our research.

- What *SAS University Edition* is

SAS, which stays for Statistical Analysis System, is a software developed by SAS Institute for statistical analysis, multivariate analysis, data analysis and management and predictive analytics. (https://www.sas.com/en_us/software/university-edition.html). The North Carolina State University is where *SAS* has been idealized.

What *SAS* is able to do is mining, managing and retrieving data from different sources and execute statistical analyses on it.
In order to use the Statistical Analysis System, the data that the researcher has obtained must be collected in a *SAS* format or spreadsheet table format, and *SAS* provides a graphical point-and-click user interface. *SAS University Edition* can be downloaded for free, directly from *SAS* and once it has been downloaded, it works locally on the PC using a virtualization software and the browser.
No internet access is required.

- *VirtualBox*: the Virtualization software

A virtualization software allows any PC to *host* virtual environments, what it does is emulating an operating system in order to permit a *guest* operating system to be run.
The *host* operating system "*is the operating system of the physical computer on which VirtualBox was installed*" (ORACLE CORPORATION, 2017). A Virtual machine (VM) "*is the special environment that VirtualBox creates for your guest operating system while it is running*" (ORACLE CORPORATION, 2017).
VirualBox is defined as a free and open-source hypervisor for x86 computers. (ORACLE CORPORATION, 2017). It has been implemented by Oracle Corporation and it can be installed on a number of *host* operating systems among which *Linux*, *Windows* and *macOS*.
VirtualBox can support the creation and management of *guest* virtual machines which run versions of *Linux*, *Windows* and other operating systems.

- *SAS University Edition* for Virtualization Software

In order for *SAS* to rightly and efficiently perform, a virtualization software is required.
We followed the following steps to make *SAS* work in our *Microsoft Windows 7* operating system (http://support.sas.com/software/products/university-edition/index.html):

I. Operating system requirements

-64-bit hardware with 1GB of *Random Access Memory* (minimum) and one or two processors

-Virtualization software: *Oracle VirtualBox 4.3* or later

-Web browser: *Google Chrome, Internet Explorer*

II. Virtualization Software Installation

-install *Oracle VirtualBox* by clicking this link: https://www.virtualbox.org/wiki/Downloads

-get *SAS University Edition vApp* and save this file in the *Downloads* folder

III. Add the *SAS University Edition vApp* to *VirtualBox*

-after opening the *VirtualBox*, select *File→Import Appliance*

-from the *downloads* folder, select the *OVA file* for *SAS University Edition vApp* and open it

-click *Next* and then *Import*

IV. Data and results folder creation

-create on the local PC, a folder "*SASUniversityEdition*" and a subfolder *myfolders* where we saved all our *SAS University Edition* files

-in the *VirtualBox*, select the *SAS University Edition vApp* and then choose *Machine→Settings*

-select *shared* folder in the navigation pane on the *Settings* dialog box and select the *adding* button

-select *myfolder* in the browser for folder window

-click *OK* to close the *settings* dialog box

V. *SAS University Edition vApp* starting

-open *VirtualBox* and select the *SAS University Edition vApp*

-select *Machine→Start*

VI. *SAS University Edition* opening

-type http://localhost:10080 in a web browser in the local PC and click *Start SAS Studio*

- The FREQ PROC: Measure of Associations

As declared in paragraph 1.3, the data we collected represent categorical type of data.
In order to count, display and analyze these data, PROC FREQ (SAS INSTITUTE INC., 2016) is an essential procedure within *BASE SAS*.

For two-way tables, PROC FREQ calculates measures and tests of associations while for n-way tables, it provides analysis by calculating statistics within and across strata. We will create one-way tables and tabulations and afterwards the PROC FREQ will provide *goodness-of-fit tests* for equal proportions or null proportions. For *contingency tables*, PROC FREQ can compute statistics to evaluate and investigate the relation between two categorical variables. We can determine the intensity of any association between two variables and by computing the *chi-square test* we can examine whether an association exists or not. While selecting measures of association to be implemented in the analysis of two-way tables, we firstly considered the study design which specifies whether the column and row variables are independent or dependent , the type of association that each measurement is planned to detect and any assumptions required for valid measures interpretation.

- Measures of Associations

This function calculates the measures and creates the tests that we described in the part regarding *descriptive Statistics* (paragraph 1.3).

2. A quantitative analysis about Italy

2.1 What about Italy abroad?

As mentioned in the Introduction, this dissertation is an experimental exercise aimed at analyzing how Italy is conceived abroad.
The idea came about while I was in Erasmus; I got in contact with people from all around the world and this gave me the possibility to get to know different mentalities, cultures and opinions.
I reckoned it would result interesting to conceive an analysis directed at finding out what opinion international young people hold regarding Italy. I thought it could be enriching and useful to carry out a research with the aim of clarifying to what extent the idea that international people have about Italy is linked to their actual knowledge of the country or rather if it is related to some kind of prejudice.(for further information, consult http://www.esteri.it/mae/it/sala_stampa/archivionotizie/approfondimenti/2010/03/20100325_immagine_italia_estero.html and VAAN ALDEREN, 2015) That is the reason why our questionnaire has been realized to collect personal information about our interviewees, to examine what level of knowledge they have on Italy and to investigate whether these two factors are related to their impressions and conception of the country or not.
We designed the questionnaire by referring to an inspiring research conducted by the *Intercultura Foundation for Intercultural Dialogue and International Youth Exchanges* in collaboration with *Ipsos* in December 2008 entitled "*The image of Italy abroad*" (http://www.fondazioneintercultura.org/it/Ricerche-pubblicate/L%27immagine-dell%27Italia-all%27estero). This research shared a project whose aim was to quantify and qualify how the principal international newspapers and magazines deal with the subject matter of *Italy*. The objective of the analysis was to break out of the stereotypes and truly face the phenomenon "*Italy*". Frequently we have heard about *Il Bel Paese* as an epithet referred to Italy as it is globally considered the native-land of the artistic beauty, of historical culture, of the tasty cuisine, of the beautiful landscapes. However too often these opinions overlap and confuse themselves with commonplace ideas.
The aim here is to statistically analyze what is the opinion held by young international people on Italy. We conducted the survey in such a way as to follow the *fil rouge* of the research conducted by *Intercultura* and *Ipsos* and we will be able to probe how and why, young international people judge *Il Bel Paese*.

2.2 Introduction to our sample

"*The sample is the set of n (sample size) sampling units (which we call cases) selected from among the N units that make up the population, and which represent that population hence the expression 'representative samples') for the purpose of our study*". (CORBETTA, 2003).
Sampling stands for the process of "*collecting only a subset of data regarding the population*" (MONTI, 2008). Clearly sampling discloses only an estimate, an approximate value of the examined population, however, a representative sample does represent the population accurately. Our sample is composed of 324 subjects and our aim was to collect information from the biggest number of people at two essential conditions:

1. No Italian nationality
2. Age between 14 and 25 years old

Our research is known as *sample survey* (CORBETTA, 2003) which is a type of non-experimental study. We selected a sample of subjects from a population and started to collect data from them. This typology of questionnaire, "*tool aimed at the data-collection*" (BORRA and DI CIACCIO, 2008) attempts to count some features about the people in the population.
The process was quite fast; we succeed in obtaining a sufficient number of responses in less than 3 weeks. For a sample survey to be representative, it is necessary that the sample obtained reflects the population well. Our sample can be considered informative as we received feedback from all over the world.
The platforms we used to spread the questionnaire were principally *Facebook* and *Messenger*. The questionnaire was sent privately to many international young people, many *Facebook groups* of Erasmus students revealed themselves very useful to our research. Each time we posted the questionnaire publicly on *Facebook* or we sent it privately, we accompanied the questionnaire with a brief text aimed to inform the potential interviewee that the questionnaire would have been anonymous.
Of course some *biases* took place, some of the respondents were did not specify clearly the country of origin, others misunderstood some questions. An additional *bias* we encountered is the so called *no response bias* which means that some sampled subjects could not be reached or refused to be interviewed. Moreover, even among those who answered the questionnaire, some did not respond to some of the questions turning out in *nonresponse bias* due to missing data.

Therefore before analyzing the data in *SAS University Edition*, we spent time organizing, de-codifying and eliminating data by working on an *Excel* file. For example, we codified each city or place of origin of the respondent with the country and continent of reference.
Similarly, to the question *Where have you been in Italy?* we constructed *a posteriori* a table listing the 20 Italian regions in order to have a clearer and easier list of places to examine. For instance, whenever the interviewee affirmed to have been to *Costiera Amalfitana*, *Rome* and *Todi*, we de-codified the answers by selecting *Campania*, *Lazio* and *Umbria* as variables to take into consideration.
All the modifications and de-codifications were done on the same *Excel* file which represented the table of data that we uploaded as our *SAS* data file.

2.3 What about Italy abroad: presentation of questions

This paragraph will present the questions we designed for our research.
The questions are 15 and the answers are defined statistically as *categorical variables*. We collected 324 answers. All the questions were compulsory in the questionnaire, a part from the one asking *Where have you been in Italy?* (of course the interviewee who had not visited Italy could not answer this question). The format we used for the creation of the questionnaire is the multiple choice question, and for some questions, the alternative *Other* has been offered.
Only one question reports images of the *UNESCO sites* that we took from the *Official United Nations website* (http://en.unesco.org) and only the question *Where are you from?* presents the formula *Short Answer*; the interviewee is required to add a small portion of texts.
The questionnaire is divided in three Sections the we nominated *About YOU*, *What do YOU know about Italy?* and *What do YOU think about Italy?*.
We will exhibit the three Sections separately and for each Section each question will be reported together with its table of frequencies, in a descending order for categories, and a textual description.

- Section 1: *About YOU*

The first Section aims to collect basic information about the interviewee. The questions are 4 and in a multiple choice format.

1. *Where are you from?*

The first question investigates the place of origin of the interviewee. Our purpose was to create a sample of international young people and we received feedback from people coming from 52 different countries; our variegated sample can be regarded as informative.
In order to render the analysis clearer, we de-codified *a posteriori* each *Country* with its *Continent* of origin. For example, if the interviewee answered *Cambodia*, we assigned it to the category *Asia*. Specifically approximately 68% of our sample comes from *Europe*, 17% from *America*, 7% from *Asia*, 5% from *Australia* and less than 1% from *Africa*. *Africans* are not representative for their *Continent*.
The majority of respondents from *Europe* are from *Sweden* which represent approximately 12% of the entire sample size. Immediately after appears *France* with 38 interviewed.
We did expect such a big affluence of responses from *Swedish* people as *Sweden* was the country where I had been during Erasmus; I got in contact with many *Swedish* young people and many *Facebook groups* where we posted the questionnaire were composed by *Swedish* people.
We report both the table which displays the variable *Continent* and the one showing the variable *Country*.

2.3.1	**Where are you from?**			
Continent	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Europe	223	68.21	221	68.21
America	57	17.59	278	85.80
Asia	25	7.72	303	93.52
Australia	16	4.94	319	98.46
Africa	3	0.93	322	100.00

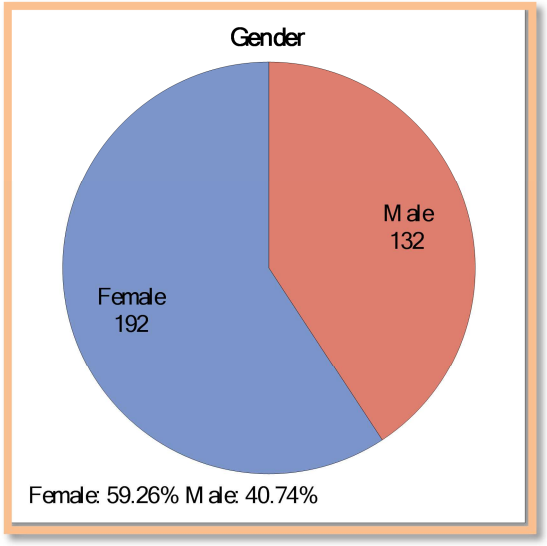
2.3.2	Where are you from?			
Country	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Sweden	40	12.35	40	12.35
France	38	11.73	78	24.07
Spain	26	8.02	104	32.10
USA	24	7.41	128	39.51
Greece	21	6.48	149	45.99
The Netherlands	19	5.86	168	51.85
UK	17	5.25	185	57.10
Australia	15	4.63	200	61.73
Germany	15	4.63	215	66.36
Canada	7	2.16	222	68.52
India	7	2.16	229	70.68
Argentina	6	1.85	235	72.53
Belgium	6	1.85	241	74.38
Finland	6	1.85	247	76.23
Austria	5	1.54	252	77.78
Colombia	5	1.54	257	79.32
Saudi Arabia	5	1.54	262	80.86
Portugal	4	1.23	266	82.10
Belarus	3	0.93	269	83.02
China	3	0.93	272	83.95
Iran	3	0.93	275	84.88
Mexico	3	0.93	278	85.80
Peru	3	0.93	281	86.73
Russia	3	0.93	284	87.65
Slovakia	3	0.93	287	88.58
Bolivia	2	0.62	289	89.20
Brazil	2	0.62	291	89.81
Croatia	2	0.62	293	90.43
Cyprus	2	0.62	295	91.05
Korea	2	0.62	297	91.67

Norway	2	0.62	299	92.28
Poland	2	0.62	301	92.90
Romania	2	0.62	303	93.52
Turkey	2	0.62	305	94.14
Ukraine	2	0.62	307	94.75
Bangladesh	1	0.31	308	95.06
Bhutan	1	0.31	309	95.37
Chile	1	0.31	310	95.68
Congo	1	0.31	311	95.99
Czech Republic	1	0.31	312	96.30
Denmark	1	0.31	313	96.60
Ecuador	1	0.31	314	96.91
Iceland	1	0.31	315	97.22
Jordan	1	0.31	316	97.53
Kenya	1	0.31	317	97.84
New Zealand	1	0.31	318	98.15
Pakistan	1	0.31	319	98.46
Serbia	1	0.31	320	98.77
South Africa	1	0.31	321	99.07
Switzerland	1	0.31	322	99.38
Venezuela	1	0.31	323	99.69
Vietnam	1	0.31	324	100.00

2. *What is your gender?*

Clearly this question wants to explore how many *Males* and how many *Females* compose the sample. Nearly 60% are *Females* and the remaining 40% are *Males*. The difference among the number of *Females* and *Males* can be considered a *bias*, while in Erasmus I meet more *Females* than *Males.* It is likely that the majority of respondents are people who had been come directly into contact with me.

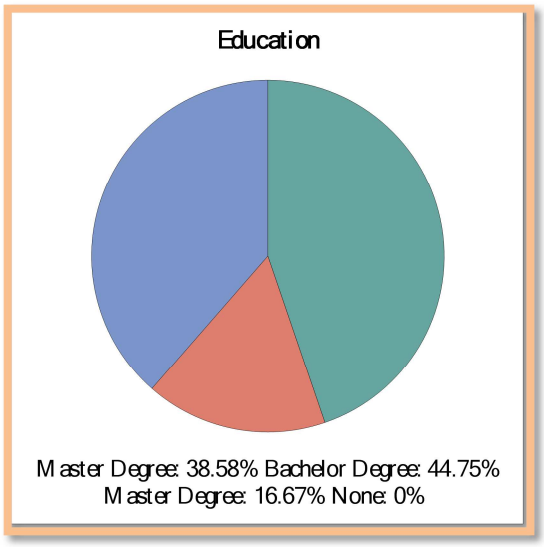
2.3.3	What is your Gender?			
Gender	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Female	192	59.26	192	59.26
Male	132	40.74	324	100.00



3. *Which level of education have you currently completed?*

Here, the interviewee is asked to state which *level of education* he or she has completed at the moment of filling in the questionnaire. The options offered in the questionnaire were *None*, *High School*, *Bachelor Degree*, *Master Degree*. The majority of respondents have obtained a *Bachelor Degree*, about 45% of our sample, while the lowest percentage is represented by students who obtained a *Master Degree*, 17%. The remaining part of the sample refers to students who completed *High School*. Nobody selected the option *None* which means that the sample is completely formed by people who completed at least *High School* education.

2.3.4	Which level of education have you currently completed?			
Level of Education	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Bachelor Degree	145	44.75	145	44.75
High School	125	38.58	270	83.33
Master Degree	54	16.67	324	100.00



4. *Which field is closer to your study plan?*

In the last part of the first Section, the purpose is to collect information on the *study plan* of the interviewee. We offered three specific options, *Social Science*, *Scientific*, *Fine Arts* or *Classics*, while the fourth makes reference to the emerging *Interdisciplinary* courses and it proposes even the option *Other* for whoever does not identify himself in any of the previous categories. More than half of our sample did *Social Science* studies, while the minority selected the option *Interdisciplinary or Other* (5.25%). We did expect such a relevant affluence from *Social Science* students as, while in Erasmus, I had the opportunity to encounter many people who were enrolled in the same courses as I was.

2.3.5	Which field is closer to your study plan?			
Study Plan	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Social Sciences (Political Science, Economics, Law, Sociology, Business…)	169	52.16	169	52.16
Scientific (Engineering, Mathematics, Chemistry, Medicine..)	104	32.10	273	84.26
Fine Arts or Classics (English, History, Literature…)	34	10.49	307	94.75
Interdisciplinary or Other	17	5.25	324	100.00

- Section 2: *What do YOU know about Italy?*

The second Section embodies the core of the research as it concerns the knowledge that the interviewees have about Italy. The goal is to determine to what extent our subjects know and have knowledge about the most important factors of the Country. Analyzing this section gives us the possibility to determine whether the conception and the opinion that our sample has about Italy is related and is a consequence of their understanding of the country or not.

1. *Have you ever been to Italy?*

This first question asks to the interviewee whether he or she has ever visited Italy. 249 out of 324 have been at least once to Italy. They make up the majority by constituting more than 76% of the entire sample.

2.3.6	**Have you ever been to Italy?**			
Yes/No	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Yes	249	76.85	249	76.85
No	75	23.15	324	100.00

2. *Where have you been?*

This question require the interviewee to insert a small portion of text. We organize each answer before being able to analyze our data; we de-codified the place, the city, the town that the interviewee wrote as individual answer by referring to the Italian region of provenience.
We reckoned that this method would facilitate our analysis by making it easier to interpret.

The highest proportion of the sample who answered *Yes* to the question *Have you ever been to Italy?* visited *Lazio*. More than 72%. This finding is not surprising as *Lazio* is the region where the Italian Capital City is located, where the *Vatican City* is and where the Pope lives. *Lazio* represents one of the most alluring touristic poles both for religious reasons and for cultural and artistic attractions.

A relevant portion of our sample stated to have been to *Veneto*, exactly 43.15%. We did imagine such a relevant portion as *Veneto* is where *Venice* is placed. Still the largest part of the sample has not been to *Veneto*.

A significant fraction of the sample stated to have been to *Toscana*, exactly 41.13%. We expected this percentage as *Toscana* is where *Florence* is and where many attractive, cultural sites are located.

Remarkably only 37% affirmed to have been to *Lombardia*. We did not expect such a low rating as *Lombardia* is where the city of *Milan* is located. *Milan* is one of the richest Italian cities and one of the most industrially developed. 37.50% thus determines a small part of the sample.

The percentage of the sample who has been in *Campania* is about 22% which represents a statistically small part of it. However, *Campania* represents the only Southern Italian region that counts a statistically relevant numbers of visits. The majority of the respondents who have been to *Campania*, declared to have been to *Naples, Amalfi Coast* and *Pompei*.

Around 90% of the respondents have not been to *Sicilia.* Therefore the portion of subjects who visited this region represents a very low percentage.

Less than 10% of the interviewees state that they have been to *Puglia*.

Nearly 9% of the interviewees have been to *Emilia-Romagna*. All of them have been to *Bologna*.

The big majority of our sample has never been to *Liguria*. Only 8% of the sample declared to have visited this region and in particular *Cinque Terre.*

The big majority of our sample have not been to *Piemonte*, exactly 92.74%. We did not imagine this scarce affluence in *Piemonte* as it is the region where *Turin*, one of the most developed Italian cities, is located.

Only 18 out of 324 affirmed to have been to *Trentino-Alto Adige*.

About 94% of the sample declared not to have been to *Sardegna*, it is a significantly high percentage.

Only 11 respondents stayed in *Friuli-Venezia Giulia*.

Only 4% of the respondents have visited *Umbria*, the majority of them inserted as an individual answer *Assisi* which represents the birthplace of the Franciscan religious order.

The proportion of visitors in *Marche* is very low, about 2% of the entire sample.

Only 4 people of our sample have been to *Abruzzo*. This represents a very low percentage.

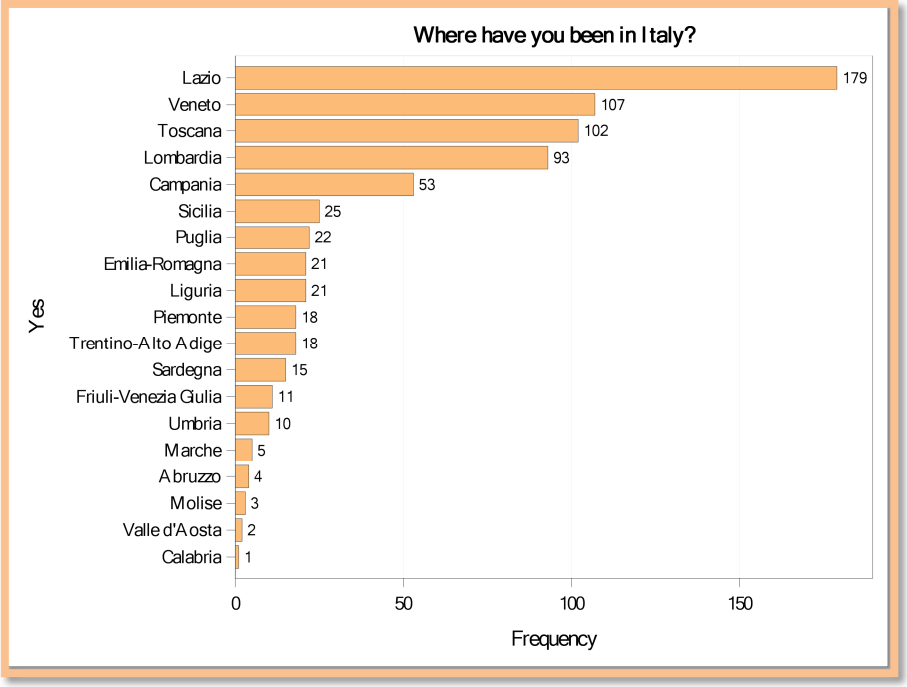
Only 3 people have been to *Molise*.

Only 2 subjects went to *Valle d'Aosta*. We are not surprised by this result as this region is the smallest in Italy.

Only 1 interviewee visited *Calabria*, this proportion is extremely low.

None visited *Basilicata*. This is quite a surprising finding as *Basilicata* is where one of the most important Italian *UNESCO* sites, *Sassi di Matera*, is located. Moreover the Capital City of *Basilicata*, *Matera*, has been nominated *European Capital of Culture 2019* (http://www.matera-basilicata2019.it/it).

The following graphs summarize for each Italian region the frequencies and percentages of *Yes.*

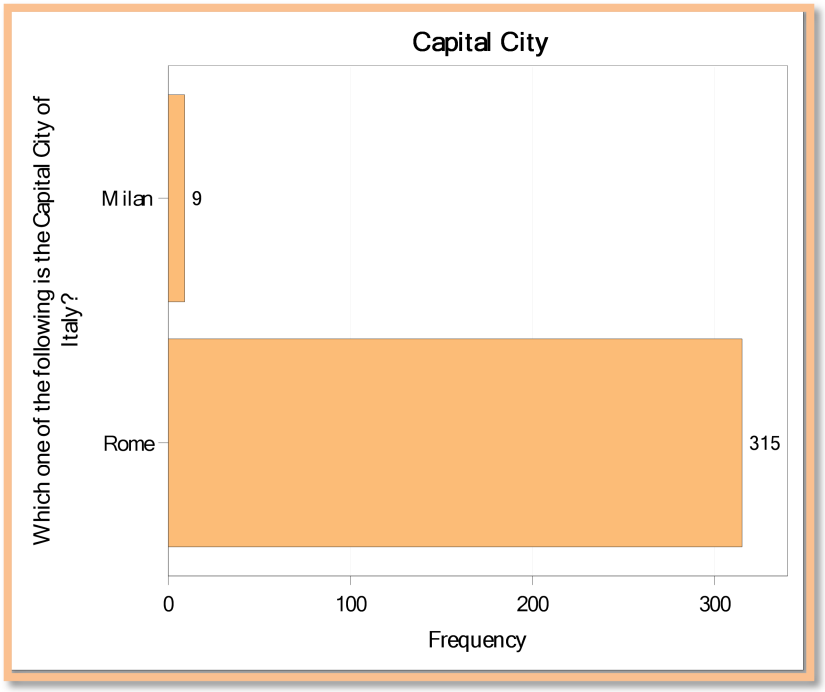


3. *Which one of the following is the Capital City of Italy?*

This is the first question that tests the interviewee's knowledge about Italy.

97% of the interviewees selected the right answer, *Rome*, whereas only less than 3% chose *Milan*. None selected the options *Porto, Palermo, Naples, Venice*. A vast majority of respondents, even those who have never been to Italy, answered correctly. This first finding regarding the knowledge that our sample has on Italy is positive.

2.3.8	Which one of the following is the Capital City of Italy?			
Capital City	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Rome	315	97.22	315	97.22
Milan	9	2.78	324	100.00



4. *Please, indicate in which part of Europe Italy is located*

This question tests the geographical knowledge of the interviewee. 87% of the entire sample answered correctly, *Southern Europe*. However the remaining 13% selected a wrong answer. The majority of the respondents who answered wrongly, selected *Western Europe* as individual answer.

2.3.9	Please, indicate in which part of Europe Italy is located			
Part of Europe	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Southern Europe	282	87.04	282	87.04
Western Europe	34	10.49	316	97.53
Eastern Europe	6	1.85	322	99.38
Northern Europe	2	0.62	324	100.00

5. *How many inhabitants are there approximately in Italy?*

204 interviewees out of 324 picked the right answer, in fact Italy counts approximately *60 million* people, but a relevant 20% opted for the option *25 million*.

2.3.10	How many inhabitants are there approximately in Italy? (in million)			
Numbers of inhabitants	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
60	204	63.16	204	63.16
25	70	21.67	274	84.83
12	32	9.91	306	94.74
100	10	3.10	316	97.83
150	4	1.24	320	99.07
6	3	0.93	323	100.00
Frequency Missing = 1				

6. *Which one of the following pictures represents an Italian UNESCO site?*

The following question is the only one that presents the interviewees a picture taken from the *Official United Nations site* (http://en.unesco.org). Indeed we created a list of six pictures each illustrating a *UNESCO site* and only one picture portrays an Italian *UNESCO site*. Almost 48% answered correctly, 20% selected a Portuguese *UNESCO site*, 13% selected a Spanish *UNESCO site*, 12% selected a French *UNESCO site*, 3% of the sample opted for the Cambodian *UNESCO site* of Angkor and the remaining 2% for a Japanese *UNESCO site*.

The Italian *UNESCO site* that we chose is located in *Basilicata* which, in the question 1, represented the only region where none of the interviewees have ever been. The fact that the majority of respondents answered correctly, even though none had been to Italy, is positive. Nonetheless still 53% of the whole sample opted for one of the wrong options, this percentage is high. In addition 1.54% answered *Don't know* to the question. To note is the fact that the biggest share of wrong answers lies in the options showing a European *UNESCO site*, such as the *Monastery of Alcobaça* in Portugal and the *Alhambra* in Spain.

Sassi di Matera

(Italy)

Monastery of Alcobaça

(Portugal)

Alhambra

(Spain)

Palace of Versailles

(France)

Angkor

(Cambodia)

Hiroshima Peace Memorial

(Japan)

2.3.11	Which one of the following pictures represents an Italian UNESCO site?			
UNESCO site	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Sassi di Matera (Italy)	155	47.84	155	47.84
Monastery of Alcobaça (Portugal)	66	20.37	221	68.21
Alhambra (Spain)	43	13.27	264	81.48
Palace of Versailles (France)	40	12.35	304	93.83
Angkor (Cambodia)	8	2.47	312	96.30
Hiroshima Peace Memorial (Japan)	7	2.16	319	98.46
Don't know	5	1.54	324	100.00

- *Section 3: What do YOU think about Italy?*

In this Section, the aim of the research is to obtain information regarding the opinion that the interviewee holds about Italy. We used as guiding line the research mentioned in paragraph 2.1.

1. *Which word would you associate with "Italy"?*

Intercultura and *Ipsos*'s aim was to analyze the reputation that Italy has abroad through the analysis of 6 characteristics:

1. *Culture*
2. *Economy*
3. *Politics*
4. *Illegality*
5. *Sport*
6. *(Daily) Chronicle*

The research found out that the subject matter that received more attention is *Culture*, identified by factors such as *Made in Italy*, artistic culture, fashion and gastronomic culture.

The second characteristic is *Economy*, understood principally as *Economic Crisis*.

The third factor is *Politics*, meant as the Italian political life, its reforms, its laws and *Silvio Berlusconi* as central personality.

The fourth aspect *Illegality*, intended as criminal facts, important scandals and action carried out by the *Mafia*.

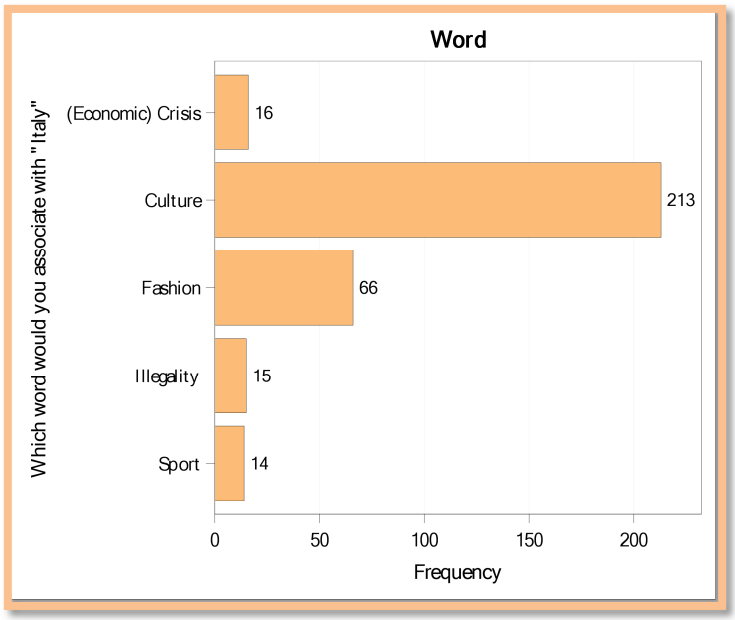
The fifth element is represented by the Italian *Chronicle* and principally by the figure of the *Pope*.

The last factor that received attention is *Sport*, whose determinant factors are the game of *Football* and the figure of *Valentino Rossi*.

In our research we offered 6 options as possible answers: *Culture*, *Fashion*, (*Economic) Crisis*, *Illegality*, *Sport* and *(Daily) Chronicle*.

We found out that much more than half of our sample indicated *Culture* as *word* associate with Italy. This result is statistically very high and it does confirm the one obtained by the research carried out by *Intercultura* and *Ipsos*. *Economic Crisis*, *Illegality* and *Sport* received less than 5% in total. None selected the alternative *Daily Chronicle*, probably this alternative was not clear enough.

2.3.12	**Which word would you associate with "*Italy*"?**			
Word	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Culture	213	65.74	213	65.74
Fashion	66	20.37	279	86.11
(Economic) Crisis	16	4.94	295	91.05
Illegality	15	4.63	310	95.68
Sport	14	4.32	324	100.00



2. *Which idea, concept would you relate with "Italianness"?*

In this question we still take as referring source the research conducted by *Intercultura* and *Ipsos*; we selected some *concepts* which are globally recognized as Italian and which, according to us, could represent some of the characteristics analyzed by *Intercultura* and *Ipsos* (listed in the previous question). These are *Coffee*, *Vespa*, *Mafia*, *Colosseum*, *Football* and *Pope.*

Coffee, *Vespa* and *Colosseum* are related to *Culture*, *Mafia* is related to *Illegality*, *Football* is linked to *Sport* and *Pope* is related to *(Daily) Chronicle* (as mentioned in the research conducted by *Intercultura* and *Ipsos*).

We asked the interviewee to select one *idea* that he or she would relate to the *concept* of "*Italianness*". Even if two Italian products, *Coffee* and *Vespa* together overstepped half of the responses in the sample, the alternative *Mafia* determines nearly 20% of the entire sample and this proportion is statistically significant. This means that, even if Italian products are representative characteristics of the *Made in Italy*, many people still recognize the *Mafia* as an inherent Italian *concept*.

2.3.13	**Which idea, concept would you relate with "*Italianness*"?**			
Idea	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Coffee	107	33.02	107	33.02
Vespa	71	21.91	178	54.94
Mafia	64	19.75	242	74.69
Colosseum	36	11.11	278	85.80
Football	23	7.10	301	92.90
Pope	23	7.10	324	100.00

3. *Which Italian brand do you know better?*

As mentioned above, the *Made in Italy* is known all around the world. Through this question we test to what extent our international sample has knowledge about it. We offered a list of possible answers: *Armani*, *Lavazza*, *Ducati*, *Bialetti*, *Jacuzzi* and *Other* along with the possibility to specify a brand. *Armani* received nearly 60% of the total of responses. *Lavazza* comes after with 20% of the total of the responses. *Ducati* received only 7% responses, *Bialetti* and *Jacuzzi* did not reach individually 3% of the responses. The remaining part of the sample opted for the alternative *Other*. Among those who selected *Other*, some answered *None*, others answered *Barilla*, *Ferrero*, *Ferrari*,

Gucci or *Prada*. Initially we offered the alternative *Other* not expecting such a big relevance of responses, but the significant affluence of responses indicating *Other* makes us conclude that the *Made in Italy* is actually known even among young people. The variety of answers for this question has made this question not useful for the analysis discolsed in paragraph 2.4.

2.3.14	Which Italian brand do you know better?			
Italian Brand	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Armani	188	58.02	188	58.02
Lavazza	77	23.77	265	81.79
Ducati	23	7.10	288	88.89
Bialetti	8	2.47	296	91.36
Jacuzzi	7	2.16	303	93.52
Other	6	1.85	309	95.37
Barilla	3	0.93	312	96.30
Ferrero	3	0.93	315	97.22
None	3	0.93	318	98.15
Ferrari	2	0.62	320	98.77
Gucci	2	0.62	322	99.38
Prada	2	0.62	324	100.00

4. *Please, select the Italian personality that you know better*

In the questionnaire, we presented the interviewees a list of *Italian personalities* among whom *Silvio Berlusconi, Valentino Rossi, Chiara Ferragni, Laura Pausini, Renzo Piano, Sergio Marchionne, Salvatore Riina, Lapo Elkann,* and *Other* as an alternative option.
Surprisingly the alternative *Other* has received more than 10% of the total of the responses.
The option *Silvio Berlusconi* has been selected by more than a half of the sample, *Valentino Rossi* by 20%, *Chiara Ferragni* and *Laura Pausini* received about 7% of the responses, *Renzo Piano* less than 3%, *Sergio Marchionne* has been chosen by 1.54% of our sample and *Lapo Elkann* by no more than 1% and none selected *Salvatore Riina*.
It is clear than the majority of the sample knows better the *Italian personality* which is related to the world of the Italian politics, he has been at the center of the Italian political system for almost 20 years. However, *Silvio Berlusconi* is also strongly linked with the *Sport* sphere and is involved in

the entrepreneurial world as well. Immediately after appears *Valentino Rossi* which is linked with the *Sport* sphere. *Chiara Ferragni* is the leading representative of the Italian *Fashion* worldwide, while *Laura Pausini* and *Renzo Piano* are referred to the world of the Italian *Culture*. *Sergio Marchionne* is connected with the *Economy* framework and *Lapo Elkann* is one of the protagonists of the Italian *Chronicle*.

2.3.15	Please, select the Italian personality that you know better			
Italian Personality	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Silvio Berlusconi	173	53.40	173	53.40
Valentino Rossi	56	17.28	229	70.68
Chiara Ferragni	23	7.10	252	77.78
Laura Pausini	21	6.48	273	84.26
Other	16	4.94	289	89.20
None	14	4.32	303	93.52
Renzo Piano	9	2.78	312	96.30
Sergio Marchionne	5	1.54	317	97.84
Lapo Elkann	3	0.93	320	98.77
Andrea Pirlo	2	0.62	322	99.38
Leonardo da Vinci	2	0.62	324	100.00

5. *What adjective best describes Italian people?*

In the last question of our questionnaire, our purpose is to find out what international people think of the Italian people. Our scale of analysis consists in three degrees of evaluation: positive, neutral, negative. We associated two different *adjectives* to each evaluation. *Friendly* and *Creative* stand for a positive opinion. *Talkative* and *Gourmet* refer to a neutral idea regarding Italian people. *Rude* and *Corrupt* highlight a negative conception of the Italian people. 52% of our sample selected *Talkative*, nearly 30% selected *Friendly*. Important to note that the two *adjectives* related to a negative opinion, *Rude* and *Corrupt* obtained a total of less than 5%.

2.3.16	What adjective best describes Italian people?			
Adjective	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Talkative	169	52.16	169	52.16
Friendly	94	29.01	263	81.17
Gourmet	33	10.19	296	91.36
Creative	13	4.01	309	95.37
Rude	9	2.78	318	98.15
Corrupt	6	1.85	324	100.00

To sum up, after presenting our sample through the questions description, we can conclude by delineating the *interviewee-type* by taking into consideration the *modes* for each variables. The *mode* is the value that occurs most frequently in the data distribution. (AGRESTI and FRANKLIN, 2014).

Our *interviewee-type* is a *European Female* who obtained a *Bachelor Degree* in *Social Sciences* and who has been at least once to Italy, specifically to *Lazio*. Our *interviewee-type* knows that *Rome* is the Italian *Capital City*, that Italy is located in *Southern Europe*, there are *60 million inhabitants* and *Sassi di Matera* is one of the most famous Italian *UNESCO site*. Our *interviewee-type* would associate the *word Culture* with "*Italy*", the *idea* of *Coffee* to the *concept* of "*Italianness*". She recognizes *Armani* as leading *Italian brand*, *Silvio Berlusconi* as principal *Italian personality* and she would describe *Italian people* as *Talkative*.

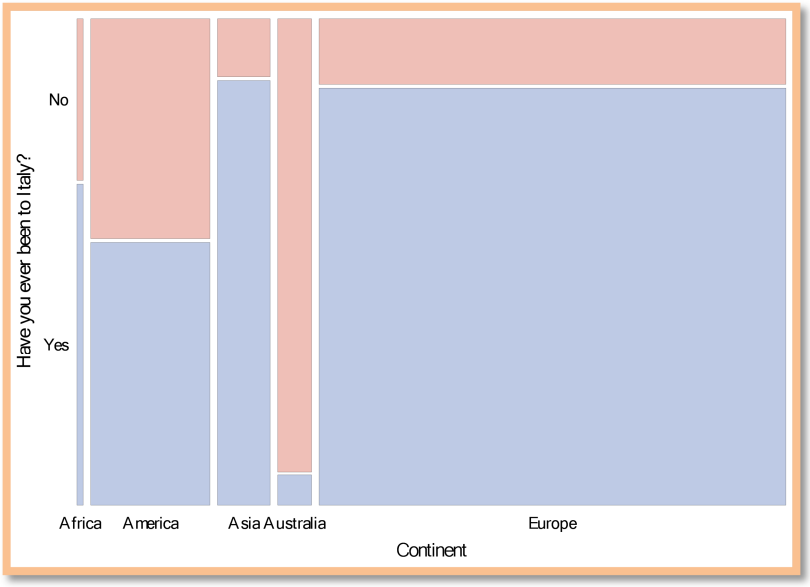
Evidently, our *interviewee-type* has some characteristics similar to mine. This can be definitely considered a *bias*.

2.4 What about Italy abroad: association analysis

This paragraph embodies the core section of the dissertation as it will proceed with the data analysis. We will be able to offer an idea on what a sample composed uniquely by international young people think about Italy. We will conduct the analysis in the following way; we will put in relation the variables of the questionnaire, we will construct for each pair of variables a *contingency table* that we will report in the analysis together with a specific comment. As mentioned in paragraph 1.4, we will use the *SAS* FREQ PROC (SAS INSTITUTE INC., 2016) to carry out the computations.

- The first aspect we want to analyze is whether the fact that our interviewees have been to Italy is explained by the *Continent* of origin or not. This means that the explanatory variable is the variable *Continent* while the response variable is *Have you ever been to Italy?*.

2.4.1	**Have you ever been to Italy?**					
Yes/No	**Continent**					
Frequency Col Pct	**Africa**	**America**	**Asia**	**Australia**	**Europe**	**Total**
No	1 33.33	26 45.61	3 12.00	15 93.75	30 13.45	75
Yes	2 66.67	31 54.39	22 88.00	1 6.25	193 86.55	249
Total	3	57	25	16	223	324



Statistics Table

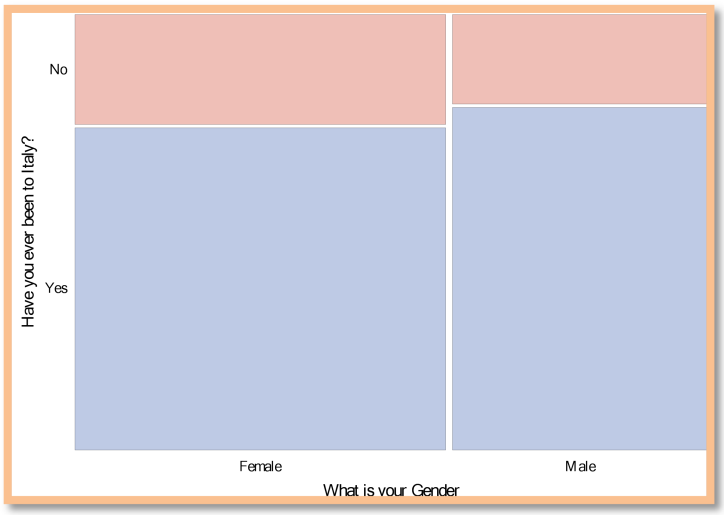
Statistic	DF	Value	Prob
Chi-Square	4	74.7072	<.0001
Likelihood Ratio Chi-Square	4	66.2524	<.0001
Phi Coefficient		0.4802	
Contingency Coefficient		0.4329	
Cramèr's V		0.4802	
WARNING: 30% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that the *Continent* of origin is determinant for the fact of ever being to Italy. This statement can be made by seeing the *p-value* which assumes a highly significant value (<.0001).
- The values that the *Phi-coefficient*, the *Contingency Coefficient* and the *Cramèr's V* assume highlight a strong dependence among the two variables, considering that their values range between 0 and 1.
- In particular, the fact that the interviewee comes from *Asia* or *Europe* is positively related to the fact of ever been to Italy, whereas who comes from *America* and *Australia* is more likely not to have been to Italy.
- In the Statistics table, the WARNING Section, advises the analyst that for some cells, the observations are less than 5 and this means that the *chi-square* may not be a valid test. In this particular case, the observations regarding the categories *Asia/No, Australia/Yes, Africa/No, Africa/Yes* are less than 5. This can be considered to be biased due to the fact that we had many more difficulties in reaching people coming from these *Continents*.

- The second aspect we want to consider in our analysis is whether the fact that our interviewees have been to Italy is explained by the *gender* or not. This means that the explanatory variable is the variable *gender* while the response variable is *Have you ever been to Italy.*

2.4.2	Have you ever been to Italy?		
Yes/No	**Gender**		
Frequency **Col Pct**	**Female**	**Male**	**Total**
No	48 25.00	27 20.45	75
Yes	144 75.00	105 79.55	249
Total	192	132	324



Statistics Table

Statistic	DF	Value	Prob
Chi-Square	1	0.9085	0.3405
Likelihood Ratio Chi-Square	1	0.9175	0.3381
Phi Coefficient		0.0530	
Contingency Coefficient		0.0529	
Cramér's V		0.0530	

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We accept, or to be more precise, we do not reject H_0, which indicates that the two variables are independent. This means that the *gender* is not determinant for the fact of ever being to Italy. This statement can be made by seeing the *p-value* which assumes a no significant value (.3405).
- The dependence among the two variables, detected by a non-zero *Cramér's V* , is due only to random sampling. Being the two variables independent, there is no link between their categories.

In the previous analysis's section (paragraph 1.3), we studied what is the perception that our sample has about Italy. Now, on the basis of the info collected, from table 2.4, we will try to understand *who thinks what* regarding Italy and the reasons *why*. From table 2.4.3, we will investigate what is the perception that our sample has acquired about Italy. We want to find out whether this perception is given by the interviewees' knowledge of Italy, if it is related to their personal information (for example the *Continent* of origin, the *gender*, the *level of education*) or rather is not related to any specific fact. In order to do that we will put in relation each variable from the *Section 3* with variables from *Section 1* and *Section 2*.
We will construct *contingency tables* on *SAS* with the variables listed in the table below as response and explanatory variables. We decided to calculate the majority of intersections with the variable *Which word would you associate with "Italy"?* as we found it more useful to our analysis's purpose. It may be that the interviewees understood this question better than the others. On the other hand, as mentioned in paragraph 2.3, we will not use the variable *Italian brand* and the variable *adjective describing Italian people,* in the construction of the bivariate analysis as we did not find useful to our analysis the outcomes that we could have obtained through the potential associations.

Response Variable	**Explanatory Variables**						
Section 3	**Section 1**				**Section 2**		
	Where are you from?	**What is your Gender?**	**Which level of education have you currently completed ?**	**Which field is closer to your study plan?**	**Have you ever been to Italy?**	**How many inhabitants are there approximately in Italy?**	**Which one of the following options represents an Italian UNESCO site?**
Which word would you associate with *"Italy"* ?	Independence (Table 2.4.3)	Dependence (Table 2.4.5)	----	Dependence (Table 2.4.6)	Independence (Table 2.4.4)	Independence (Table 2.4.7)	Independence (Table 2.4.8)
Which idea, would you relate with *"Italianness"* ?	----	Dependence (Table 2.4.9)	Dependence (Table 2.4.10)	----	----	----	----
Which is the Italian personality that you know better?	Dependence (Table 2.4.11)	----	----	Dependence (Table 2.4.12)	----	----	----

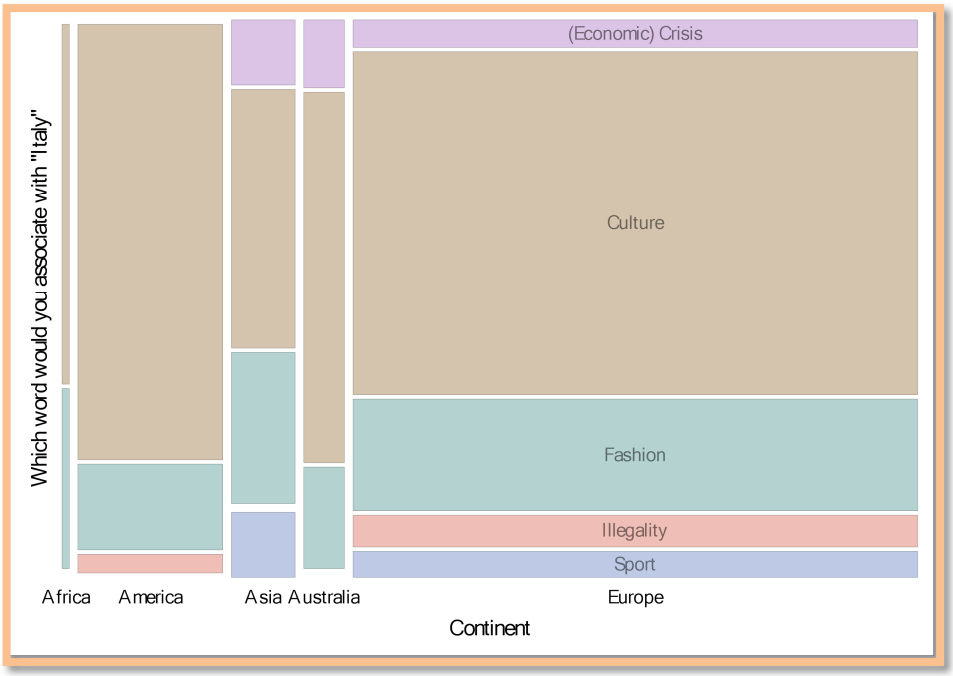
---- →we did not report the *contingency tables* created on *SAS* for those associations as they revealed not to be significant (independent) and we retained them not interesting to the purpose of the analysis.

Independence → there is no association among the variables.

Dependence →there exists association among the variables.

- Table 2.4.3 shows the association between the *word* that the interviewee would associate with "*Italy*" and the *Continent* of Origin. The explanatory variable is defined by the question *Where are you from?* and the response variable by the variable *word*.

Table 2.4.3	Which word would you associate with "*Italy*"?					
Word	**Where are you from?**					
Frequency **Col Pct**	**Africa**	**America**	**Asia**	**Australia**	**Europe**	**Total**
(Economic) Crisis	0 0.00	0 0.00	3 12.00	2 12.50	11 4.93	16
Culture	2 66.67	46 80.70	12 48.00	11 68.75	142 63.68	213
Fashion	1 33.33	9 15.79	7 28.00	3 18.75	46 20.63	66
Illegality	0 0.00	2 3.51	0 0.00	0 0.00	13 5.83	15
Sport	0 0.00	0 0.00	3 12.00	0 0.00	11 4.93	14
Total	3	57	25	16	223	324



Statistics Table

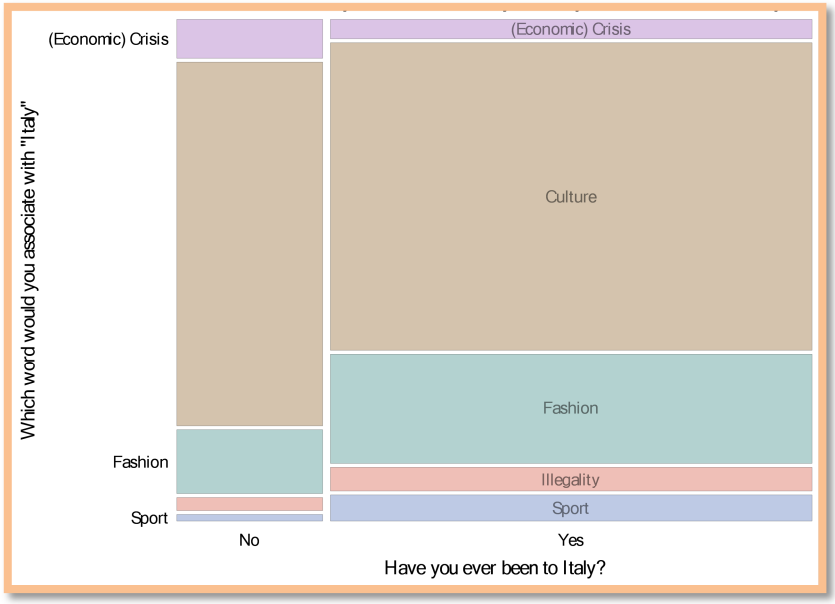
Statistic	DF	Value	Prob
Chi-Square	16	21.9957	0.1433
Likelihood Ratio Chi-Square	16	27.7522	0.0338
Phi Coefficient		0.2606	
Contingency Coefficient		0.2521	
Cramèr's V		0.1303	
WARNING: 60% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- H_0 is not rejected, which indicates that the two variables are independent. This means that the *Continent* is not determinant for the selection of one of the *words*. This statement can be made by seeing the *p-value* which assumes a non-significant value (.1433).
- The dependence among the two variables, detected by a non-zero *Cramèr's V*, is due only to random sampling. Being the two variables independent, there is no link between the categories *Continent/word*. The fact that the interviewee has chosen a *word* which displays a positive aspect of Italy (for example *Culture*) or not is not a consequence of his or her place of origin.
- In the Statistics table, the WARNING Section, advises the analyst that for some cells, the observations are less than 5 and this means that the *chi-square* may not be a valid test. In this particular case, the observations regarding the categories *America/(Economic) Crisis*, *America/Illegality*, *America/Sport*, *Asia/ (Economic) Crisis*, *Asia/Illegality*, *Asia/Sport*, *Australia/ (Economic) Crisis*, *Australia/Fashion*, *Australia/Illegality*, *Australia/Sport* and all the categories regarding *Africa* count less than 5 observations. This means two facts: *Africans* are not representative for their *Continent* and the variables *Culture* and *Fashion* have almost always received the majority of responses (they count less than 5 observations only for the categories *Australia/Fashion* and when associated with the category *Africa*).

- Table 2.4.4 puts in relation the *word* as response variable to the fact of ever being to Italy as explanatory variable. We studied how the distribution of the variable *word* varies when the variable *Have you ever been to Italy?* varies.

Table 2.4.4	**Which word would you associate with "*Italy*"?**		
Word	**Have you ever been to Italy?**		
Frequency **Col Pct**	**No**	**Yes**	**Total**
(Economic) Crisis	6 8.00	10 4.02	16
Culture	56 74.67	157 63.05	213
Fashion	10 13.33	56 22.49	66
Illegality	2 2.67	13 5.22	15
Sport	1 1.33	13 5.22	14
Total	75	249	324



Statistics Table

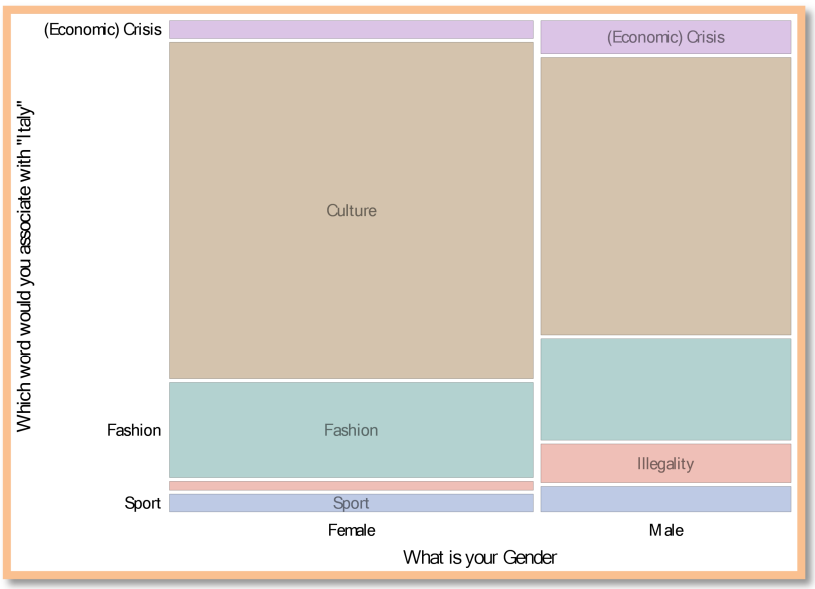
Statistic	DF	Value	Prob
Chi-Square	4	8.2359	0.0833
Likelihood Ratio Chi-Square	4	8.8986	0.0637
Phi Coefficient		0.1594	
Contingency Coefficient		0.1574	
Cramèr's V		0.1594	
WARNING: 30% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We do not reject $H_{0,}$ the two variables are independent. This means that the fact of ever having been to Italy does not influence the *word* that our respondents have chosen as associated with Italy. This statement can be made by seeing the *p-value* which assumes a non-significant value (.0833).
- The dependence among the two variables, detected by a non-zero *Cramèr's V*, is due only to random sampling. Being the two variables independent, there is no link between the categories *Have you ever been to Italy?/word*. The fact that the interviewee has chosen a *word* which displays a positive aspect of Italy (for example *Culture*) or not is not a consequence of having visited the country. However, even if there is not dependence among the categories, we can notice to what extent the variables *Culture* and *Fashion* overcome all the others, together they represent 86% of the whole sample.
- In the Statistics Table, the WARNING Section advises the analyst that 30% of the cells count less than 5 observations and consequently the *chi-square* may not be a valid test. Again in this case, the categories that count less than 5 observations are not those that mention *Culture* or *Fashion.*

- Table 2.4.5 shows how the distribution of the response variable *word* changes when the *gender* of the interviewee, which is the explanatory variable, varies.

Table 2.4.5	Which word would you associate with "*Italy*"?		
Word	**What is your Gender?**		
Frequency **Col Pct**	**Female**	**Male**	**Total**
(Economic) Crisis	7 3.65	9 6.82	16
Culture	136 70.83	77 58.33	213
Fashion	38 19.79	28 21.21	66
Illegality	4 2.08	11 8.33	15
Sport	7 3.65	7 5.30	14
Total	192	132	324



Statistics Table

Statistic	DF	Value	Prob
Chi-Square	4	10.6279	0.0311
Likelihood Ratio Chi-Square	4	10.5515	0.0321
Phi Coefficient		0.1811	
Contingency Coefficient		0.1782	
Cramèr's V		0.1811	

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that the *gender* is determinant for the selection of the *word* to associate with "*Italy*". This statement can be made by seeing the *p-value* which assumes a value highly significant (.0311).
- The values that the *Phi-coefficient*, the *Contingency Coefficient* and the *Cramèr's V* assume highlight a weak dependence among the two variables, considering that their values range between 0 and 1.
- Particularly, the fact that the interviewee is a *Male* is positively associated with the selection of the *words (Economic) Crisis*, *Fashion*, *Illegality* and *Sport*. *Females* are more likely to recognize as representing word for Italy only *Culture*.

- Table 2.4.6 presents the relation existing between the response variable *word* and the explanatory variable *study plan*.

Table 2.4.6	**Which word would you associate with "*Italy*"?**				
Word	**Which field is closer to your study plan?**				
Frequency **Col Pct**	**Fine Arts/ Classics**	**Interdisciplinary/ Other**	**Scientific**	**Social Sciences**	**Total**
(Economic) Crisis	1 2.94	3 17.65	3 2.88	9 5.33	16
Culture	26 76.47	11 64.71	58 55.77	118 69.82	213
Fashion	7 20.59	2 11.76	29 27.88	28 16.57	66
Illegality	0 0.00	0 0.00	6 5.77	9 5.33	15
Sport	0 0.00	1 5.88	8 7.69	5 2.96	14
Total	34	17	104	169	324

Statistics Table

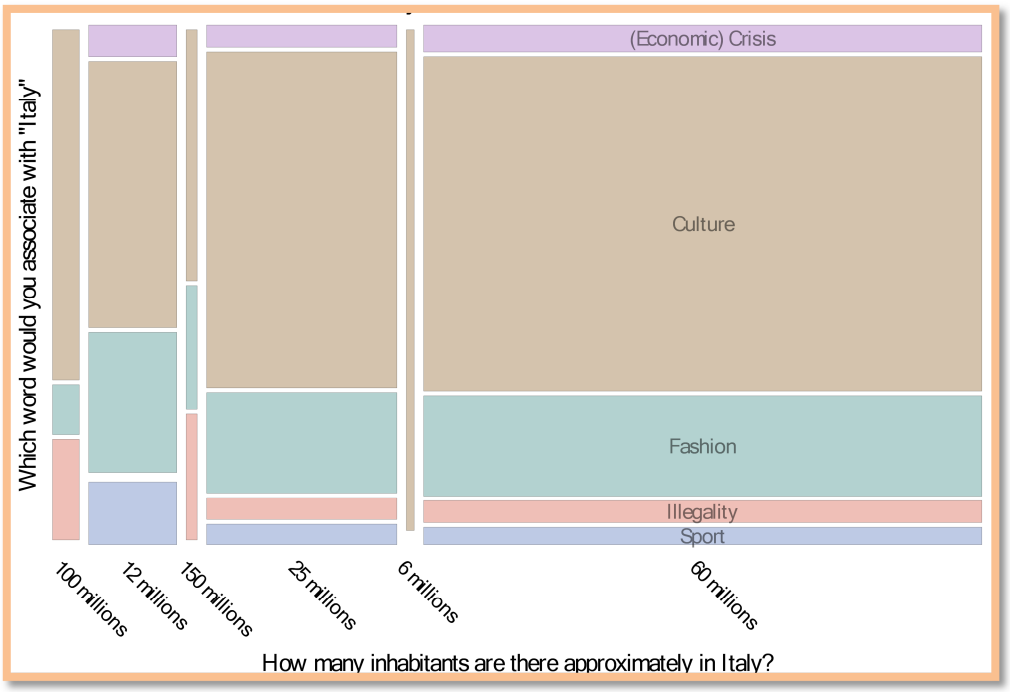
Statistic	DF	Value	Prob
Chi-Square	12	21.9309	0.0383
Likelihood Ratio Chi-Square	12	23.1977	0.0261
Phi Coefficient		0.2602	
Contingency Coefficient		0.2518	
Cramèr's V		0.1502	
WARNING: 45% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that the *field of study plan* is determinant for the selection of the *word* to associate with "*Italy*". This statement can be made by seeing the *p-value* which assumes a value highly significant (.0383).
- The values that the *Phi-coefficient*, *the Contingency Coefficient* and *the Cramèr's V* assume highlight a quite low dependence among the two variables, considering that their values range between 0 and 1.
- Particularly, the fact that the subjects in the sample are studying *Fine Arts or Classics* is positively related with the words *Culture* and *Fashion*. Who is enrolled in an *Interdisciplinary* course is more likely to recognize *(Economic) Crisis* and *Sport* as defining categories for Italy while who carries out *Scientific* studies is more likely to select the word *Sport*, *Fashion* and *Illegality*. Eventually who studies *Social Sciences*, is positively associated with the selection of the words *(Economic) Crisis*, *Culture* and *Illegality*.
- In the Statistics Table, the WARNING Section reveals to the analyst that 45% of the cells count less than 5 observation and thus the *chi-square* may not represent a valid test. Once again *Culture* is the only category that always counts more than 5 observations

- Table 2.4.7 embodies the relation between the response variable *word* and the explanatory variable *How many inhabitants are there approximately in Italy.*

Table 2.4.7	Which word would you associate with "*Italy*"?						
Word	**How many inhabitants are there approximately in Italy? (in million)**						
Frequency Col Pct	**100**	**12**	**150**	**25**	**6**	**60**	**Total**
(Economic) Crisis	0 0.00	2 6.25	0 0.00	3 4.29	0 0.00	11 5.37	16
Culture	7 70.00	17 53.13	2 50.00	47 67.14	3 100.00	137 66.83	213
Fashion	1 10.00	9 28.13	1 25.00	14 20.00	0 0.00	41 20.00	66
Illegality	2 20.00	0 0.00	1 25.00	3 4.29	0 0.00	9 4.39	15
Sport	0 0.00	4 12.50	0 0.00	3 4.29	0 0.00	7 3.41	14
Total	10	32	4	70	3	205	324



Statistics Table

Statistic	DF	Value	Prob
Chi-Square	20	21.2086	0.3850
Likelihood Ratio Chi-Square	20	19.1932	0.5093
Phi Coefficient		0.2558	
Contingency Coefficient		0.2479	
Cramèr's V		0.1279	
WARNING: 67% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We do not reject H_0, the two variables are independent. This means that the fact of knowing how many *inhabitants* there are in Italy does not influence the selection of a more positive or negative *word* to associate with Italy. We can make this statement by seeing the *p-value* which assumes a non-significant value (.3850).
- The dependence among the two variable, detected by a non-zero *Cramèr's V*, is due only to random sampling. Being the two variables independent, there is no link between their categories.
- In the Table Analysis, the WARNING Section advises the analyst that 67% of the cells count less than 5 observations and thus the *chi-square* may not be not a valid test and for the first time with count less than 5 observations even for the category *Culture*.

- Table 2.4.8 puts in relation the *word* to associate with Italy as response variable to the knowledge of an *UNESCO site* as explanatory variable.

Table 2.4.8	Which word would you associate with "*Italy*"?							
Word	**Which one of the following options represents an Italian UNESCO site?**							
Frequency Col Pct	**Don't know**	**Angkor (Cambodia)**	**Sassi di Matera (Italy)**	**Palace of Versailles (France)**	**Monastery of Alcobaça (Portugal)**	**Alhambra (Spain)**	**Hiroshima Peace Memorial (Japan)**	**Total**
(Economic) Crisis	0 0.00	0 0.00	8 5.16	3 7.50	3 4.55	2 4.65	0 0.00	16
Culture	4 80.00	5 62.50	105 67.74	28 70.00	38 57.58	27 62.79	6 85.71	213
Fashion	1 20.00	3 37.50	29 18.71	6 15.00	18 27.27	9 20.39	0 0.00	66
Illegality	0 0.00	0 0.00	6 3.87	2 5.00	4 6.06	2 4.65	1 14.29	15
Sport	0 0.00	0 0.00	7 4.52	1 2.50	3 4.55	3 6.98	0 0.00	14
Total	5	8	155	40	66	47	7	324

Statistics Table

Statistic	DF	Value	Prob
Chi-Square	24	12.3935	0.9751
Likelihood Ratio Chi-Square	24	15.2828	0.9122
Phi Coefficient		0.1956	
Contingency Coefficient		0.1919	
Cramèr's V		0.0978	
WARNING: 66% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

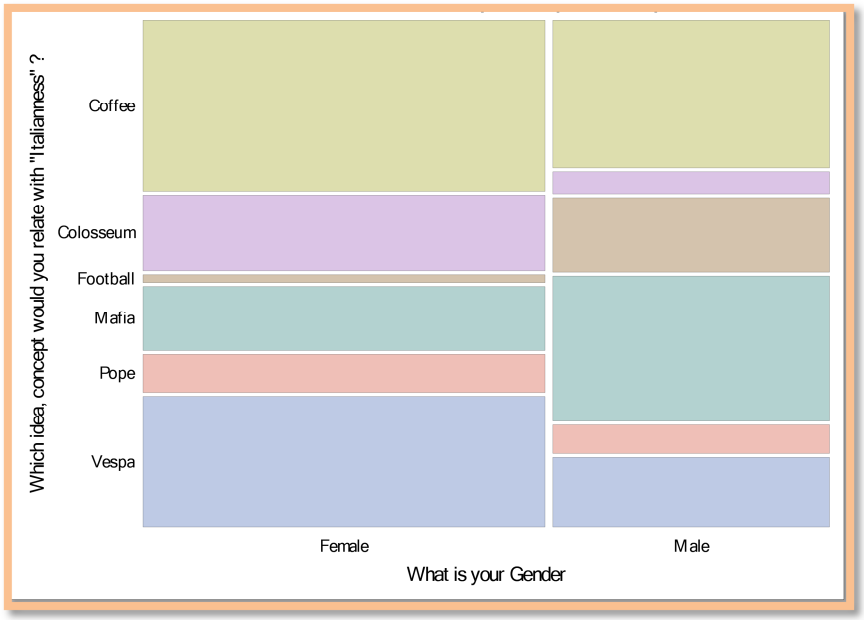
- H_0 is not rejected, the two variables are independent. This means that recognizing the Italian *UNESCO site* is not influential for the selection of the *word* to associate with Italy. This statement can be made by seeing the *p-value* which assumes a non-significant value (.9751).
- The dependence among the two variables, detected by a non-zero *Cramèr's V*, is due only to random sampling. Being the two variables independent, there is no link between their categories.

- Even if the variables are not dependent, we have noticed that almost 68% of who answered *Sassi di Matera* to the *UNESCO site* question has selected *Culture.* It is a statistically high portion of the sample.
- In the Statistics Table, the WARNING Section advises the analyst that 66% of the cells count less than 5 observations and thus the *chi-square* may not be a valid test. In this specific case, again the categories *(Economic) Crisis*, *Illegality* and *Sport* are the ones that count the lowest number of observations.

From table 2.4.9 to table 2.4.10 we will investigate how Italy is conceived abroad by offering to our sample six different possible answers. We found adequate to our thesis's object to take into consideration the *concept* of *Italianness* which can be defined as the quality, the style or the characteristics of being Italian. As mentioned in paragraph 2.3, *Coffee*, *Colosseum*, *Mafia*, *Vespa*, *Pope* and *Football* are the *concepts* we inserted in the questionnaire as representative of *Italianness.*

- Table 2.4.9 presents the relation existing between the response variable *Which idea, concept would you relate with "Italianness"?* and the explanatory variable defined by the *gender* of the respondent. Therefore we will analyze to what extent the distribution of the variable *idea, concept related to Italianness* changes on the difference of the *gender* of our sample.

Table 2.4.9	**Which idea, concept would you relate with "*Italianness*"?**		
Idea/Concept	**Gender**		
Frequency **Col Pct**	**Female**	**Male**	**Total**
Coffee	67 34.90	40 30.30	107
Colosseum	30 15.63	6 4.55	36
Football	3 1.56	20 15.15	23
Mafia	25 13.02	39 29.55	64
Pope	15 7.81	8 6.06	23
Vespa	52 27.08	19 14.39	71
Total	192	132	324



Statistics Table

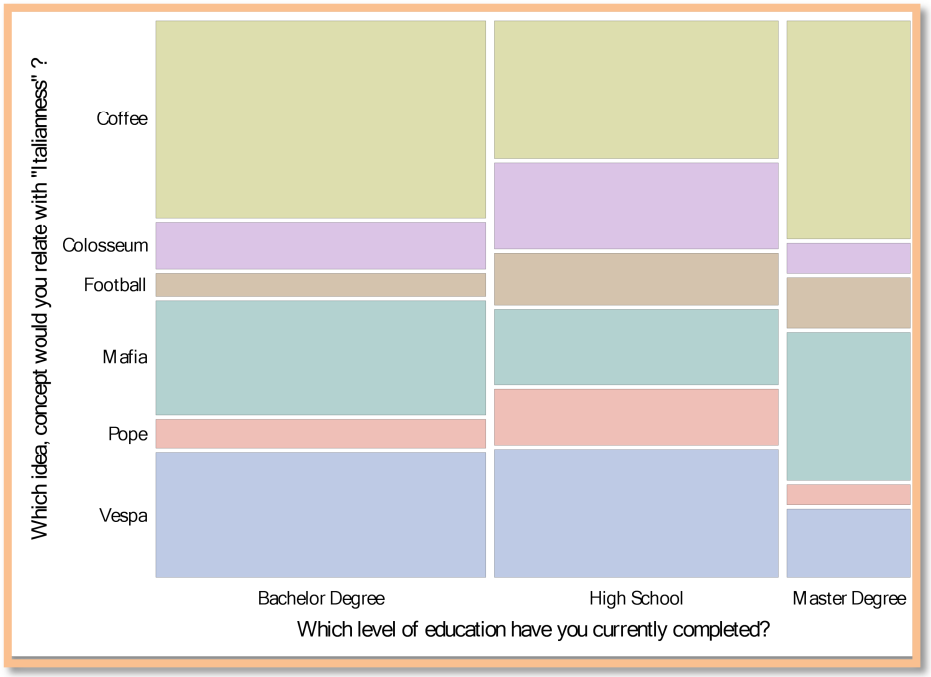
Statistic	DF	Value	Prob
Chi-Square	5	46.3890	<.0001
Likelihood Ratio Chi-Square	5	48.4474	<.0001
Phi Coefficient		0.3784	
Contingency Coefficient		0.3539	
Cramèr's V		0.3784	

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that the *gender* is determinant for the selection of the *idea* or *concept* to associate with *Italianness*. This statement can be made by seeing the *p-value* which assumes a value highly significant (.0001).
- The values that the *Phi-coefficient*, the *Contingency Coefficient* and the *Cramèr's V* assume highlight a quite high dependence among the two variables, considering that their values range between 0 and 1.
- Particularly, the fact that the interviewee is a *Male* is positively associated with the choice of the categories *Football* and *Mafia*, while *Females* are more likely to associate *Colosseum* and *Vespa* to the *concept* of *Italianness*. *Coffee* remains the answer mainly selected.

- Table 2.4.10 represents the association that we idealized between the variables *idea/concept* as response variable and *level of education* of the interviewee as explanatory variable.

Table 2.4.10	**Which idea, concept would you relate with "*Italianness*"?**			
Idea/Concept	**Which level of education have you currently completed?**			
Frequency Col Pct	**Bachelor Degree**	**High School**	**Master Degree**	**Total**
Coffee	53 36.55	32 25.60	22 40.74	107
Colosseum	13 8.97	20 16.00	3 5.56	36
Football	6 4.14	12 9.60	5 9.26	23
Mafia	31 21.38	18 14.40	15 27.78	64
Pope	8 5.52	13 10.40	2 3.70	23
Vespa	34 23.45	30 24.00	7 12.96	71
Total	145	125	54	324



Statistics Table

Statistic	DF	Value	Prob
Chi-Square	10	21.0963	0.0204
Likelihood Ratio Chi-Square	10	21.7269	0.0166
Phi Coefficient		0.2552	
Contingency Coefficient		0.2472	
Cramèr's V		0.1804	

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that *level of education* affects the selection of the *idea* or *concept* to associate with *Italianness*. This statement can be made by observing the *p-value* which assumes a value highly significant (.0204).
- The values that the *Phi-coefficient*, the *Contingency Coefficient* and the *Cramèr's V* assume highlight a quite low dependence among the two variables, considering that their values range between 0 and 1.
- Especially, we note that the higher the *level of education* of the respondents, the higher the likelihood that he or she opts for *Mafia* as a defining word for *Italianness*: 14% of the *High School* respondents selected *Mafia*, this percentage increases by 7% for *Bachelor Degree* respondents and eventually arrives at 30% with the respondents who have a *Master Degree*. It may be that the more the respondents are educated, the more knowledge they have of some negative social aspects of our country, such as the *Mafia* phenomenon.
- The same happens for the category *Coffee*, the more the respondents are educated, the more they tend to select this idea to the one of *Italianness*. Therefore we can affirm that *Coffee* still remains a remarkable product that describes the Italian style. However, this is not a result of the positive trend with the category *level of education*, rather we can interpret it as linked with the age of the interviewees.
- On the other hand, we observe the opposite result for the option *Pope*; the higher the *level of education*, the lower is the inclination for choosing *Pope* as a defining concept for *Italianness*.

From Table 2.4.11 to Table 2.4.12 the aim is to study what variable is responsible for the selection of the *Italian personality* that our respondents know better. We offered a list which includes *Silvio Berlusconi, Valentino Rossi, Chiara Ferragni, Laura Pausini, Renzo Piano, Sergio Marchionne, Salvatore Riina, Lapo Elkann, and Other* as alternative option. The results that we will comment here will always bring us to draw the same conclusion: *Silvio Berlusconi,* which represents 53% of the answers in the sample, always prevails over the other categories. Considering that we also offered the option *Other*, this proportion is incredibly high.

- Table 2.4.11 specifically shows what kind of relation exists between the response variable *the Italian personality that you know better* and the explanatory variable *Where are you from?*. The purpose is to determine whether the selection of the *Italian personality* that our respondents know better depends on the place of origin of the interviewee or not.

Table 2.4.11	**The Italian personality that you know better**					
Italian personality	**Continent**					
Frequency **Col Pct**	**Africa**	**America**	**Asia**	**Australia**	**Europe**	**Total**
Andrea Pirlo	0 0.00	0 0.00	1 4.00	0 0.00	1 0.45	2
Chiara Ferragni	0 0.00	3 5.26	2 8.00	0 0.00	18 8.07	23
Lapo Elkann	0 0.00	1 1.75	1 4.00	0 0.00	1 0.45	3
Laura Pausini	0 0.00	10 17.54	2 8.00	1 6.25	8 3.59	21
Leonardo da Vinci	0 0.00	0 0.00	0 0.00	0 0.00	2 0.90	2
None	0 0.00	8 14.04	2 8.00	0 0.00	4 1.79	14
Other	1 33.33	3 5.26	1 4.00	0 0.00	11 4.93	16
Renzo Piano	0 0.00	1 1.75	1 4.00	0 0.00	7 3.14	9
Sergio Marchionne	0 0.00	1 1.75	2 8.00	0 0.00	2 0.90	5
Silvio Berlusconi	1 33.33	16 28.07	5 20.00	2 12.50	149 66.82	173
Valentino Rossi	1 33.33	14 24.56	8 32.00	13 81.25	20 8.97	56
Total	3	57	25	16	223	324

Statistics Table

Statistic	DF	Value	Prob
Chi-Square	40	134.3802	<.0001
Likelihood Ratio Chi-Square	40	110.7797	<.0001
Phi Coefficient		0.6440	
Contingency Coefficient		0.5414	
Cramèr's V		0.3220	
WARNING: 80% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that *Continent* of origin affects the selection of the *Italian personality.* This statement can be made by seeing the *p-value* which assumes a value highly significant (.0001).
- The values that the *Phi-coefficient*, the *Contingency Coefficient* and the *Cramèr's V* assume highlight a quite strong dependence among the two variables, considering that their values range between 0 and 1.
- We can affirm that the option *Silvio Berlusconi* has been selected by the majority of *Europeans* interviewees, so the categories *Europe/Silvio Berlusconi* are positively correlated. This can be explained by the fact that probably more *Europeans* (67%) are interested in the Italian socio-political life than, for example, *Australians* are (13%).
- We observe the exact opposite for the category *Valentino Rossi* which received the smaller number of responses from *Europeans* interviewees. We can conclude that *Valentino Rossi* is recognized as an Italian symbol globally whereas *Silvio Berlusconi*, who undoubtedly received the majority of responses from the sample, is mainly known in *Europe*.
- In the Statistics Table, the WARNING Section advises the analyst that 80% of the cells count less than 5 observations and as a consequence the *chi-square* may be not a valid test. In this specific case, *Lapo Elkann*, *Sergio Marchionne*, *Andrea Pirlo* and *Leonardo da Vinci* count always less than 5 observations for each category.

- Table 2.4.12 shows the relation between the response variable *Italian personality that you know better* and the explanatory variable *study plan*. Our question is: Does the *study plan* that our interviewees carried out influence the choice of the *Italian personality* they know better?

2.4.12	The Italian personality that you know better				
Italian personality	**Which field is closer to your study plan?**				
Frequency **Col Pct**	**Fine Arts/ Classics**	**Interdisciplinary/ Other**	**Scientific**	**Social Sciences**	**Total**
Andrea Pirlo	0 0.00	0 0.00	1 0.96	1 0.59	2
Chiara Ferragni	3 8.82	0 0.00	5 4.81	15 8.88	23
Lapo Elkann	0 0.00	0 0.00	2 1.92	1 0.59	3
Laura Pausini	1 2.94	1 5.88	3 2.88	16 9.47	21
Leonardo da Vinci	0 0.00	0 0.00	1 0.96	1 0.59	2
None	1 2.94	0 0.00	9 8.65	4 2.37	14
Other	3 8.82	0 0.00	2 1.92	11 6.51	16
Renzo Piano	3 8.82	2 11.76	3 2.88	1 0.59	9
Sergio Marchionne	0 0.00	0 0.00	3 2.88	2 1.18	5
Silvio Berlusconi	18 52.94	9 52.94	49 47.12	97 57.40	173
Valentino Rossi	5 14.71	5 29.41	26 25.00	20 11.83	56
Total	34	17	104	169	324

Statistics Table

Statistic	DF	Value	Prob
Chi-Square	30	46.1075	0.0303
Likelihood Ratio Chi-Square	30	47.5103	0.0222
Phi Coefficient		0.3772	
Contingency Coefficient		0.3530	
Cramèr's V		0.2178	
WARNING: 68% of the cells have expected counts less than 5. *Chi-Square* may not be a valid test.			

By observing the Statistics Table that we constructed on *SAS*, we can conclude that:

- We reject H_0, the two variables are dependent, which means that the *study plan* does affect the selection of the *Italian personality*. We can make this statement by seeing the *p-value* which assumes a value highly significant (.0303).
- The values that the *Phi-coefficient*, the *Contingency Coefficient* and the *Cramèr's V* assume highlight a weak dependence among the two variables, considering that their values range between 0 and 1.
- Observing the table, we can state that if a positive association exists between the categories *Silvio Berlusconi/Social Sciences*, this means that who carries on studies such as Political Science, Economics, Law, and Sociology is more likely to recognize *Silvio Berlusconi*, character in the political Italian limelight for almost twenty years, as a leading *Italian personality*.
- The same positive relation exists between the categories *Renzo Piano/Fine Arts* or *Classics* and *Renzo Piano/Interdisciplinary* or *Other*; who is involved in a study plan that envisages historical, cultural and artistic courses is more likely to associate an Italian artist (*Renzo Piano*) as a *personality* who defines Italy.
- In the Statistics Table, the WARNING Section advises the analyst that 68% of the cells count less than 5 observations and as a consequence the *chi-square* may be not a valid test. In this specific case only the categories *Silvio Berlusconi* and *Valentino Rossi* have received for each intersection with the variable *study plan* always 5 or more observations.

Conclusions

What about Italy abroad?

After conducting the quantitative research, we can give an answer to the question we formulated at the beginning of this dissertation. We can provide an overview on what foreign young people think about Italy. Obviously we are talking about a generalization regarding the population but the results that we obtained from the sample can be considered quite uniform and homogeneous (among the sample).

The first evidence to be highlighted is the fact that foreign young people effectively have some knowledge about Italy. This conclusion is implied by the fact that a very significant percentage of our sample always answered correctly to the section of the questionnaire listing questions regarding some fundamental Italian indexes, such as the *Capital City*, the *number of inhabitants*, the localization in Europe, the *UNESCO site*. Italy is acknowledged even by who has never visited it.

A second important aspect regards the interest that foreign young people nourish for Italy; 77% of our sample has visited Italy at least once. Italy is universally regarded as a *must-see place*. To be precise we can affirm that Central and Northern Italy enhance much more interest than the Southern of Italy does. An exception to this however, is the region of *Campania* which has been visited a significant number of times by our interviewees. Additionally, there is an association between the interviewees' place of origin and the fact of ever having been to Italy; *Europeans* and *Asians* are more likely to have visited Italy than *Americans* and *Australians* are. On the other hand, not surprisingly, there is no association between the *gender* of the interviewee and the fact that he or she has ever been to Italy.

But what do they think about Italy?

To answer this question we set out a number of questions that include the selection of a *word*, an *idea*, a *person* or an *adjective* that our interviewees recognize as defining for Italy. By analyzing the answers, the bivariate associations and the number of observations for each variable we can draw the following conclusions.

Italian people are positively considered. The two *adjectives* with a negative connotation (*Corrupt* and *Rude*) correspond to less than 5% of the total of the responses to the question *Which adjective best describes Italian people?*. Instead we are considered a *Talkative* and *Friendly* people. This positive finding was quite unpredictable as we knew how heavy the theme of the Italian corruption is abroad and to what extent it attracts the attention of the international politics.

A second aspect to highlight is the fact that even foreign young people recognize Italian *Fashion* as an intrinsic characteristic that defines Italy. We did not expect that to the question *Which Italian*

brand do you know better? our sample would answer by proposing such a relevant number of other *brands*, among which *Gucci*, *Barilla*, *Ferrari*, *Ferrero*, *Luxottica*, *Prada*, *Fiat*, *Lamborghini*, *Dolce&Gabbana*. Therefore *Made in Italy* is a reality known even among the youngest; we are able to confirm that Italy is famous for its style, its design, its fashion industry and its influencing trends and we can add that Italy is recognized as such even by young foreign people.

A further underlying factor is represented by the fact that there is a relationship between the *gender* of the interviewees and the *idea* that they relate to *Italianness*, *Males* tend to think about *Football* while *Females* associate *Vespa* to the Italian style. The game of *Football* represents the most popular Italian sport activity, recognized as such in the collective imagination, whereas *Vespa* has always embodied the characteristics and the qualities of the Italian style. It may be a compelling explanation that our feminine sample recalled the image of the romantic scene in *Vespa* from the memorable movie filmed in Rome, "*Roman Holidays*".

Moreover, investigating the better known *Italian personalities* by our sample we found out an interesting factor which lies in the percentage of foreign young people who know *Silvio Berlusconi*. They make up more than 50% of the whole sample. This is a very high statistic. In this case, contrary to the one that we will present afterwards regarding the *Mafia* phenomenon, we are not able to affirm the reasons why our sample selected *Silvio Berlusconi* as the better known *Italian personality*, we cannot tell whether the inclination for this answer has been formed by an actual knowledge of the *personality* or rather of the socio-political dynamics that characterize our country. However our research disclosed two positive associations.

The first association exists between the selection of *Silvio Berlusconi* as a defining *Italian personality* and the fact that our interviewees are *Europeans* or *Asians*. It may be possible that *Europeans* are more engaged in the Italian political life than *Americans* or *Australians*. The latter revealed to be much more interested in the person of *Valentino Rossi*, who is one of the symbols of the Italian *Sport* worldwide.

The second positive association appears between choosing *Silvio Berlusconi* as a defining *Italian personality* and the *Social Science* study *field*; such a relation was expected, it is logical that who studies subjects such as Political Science, Economy, Sociology is more absorbed in socio-political environments and dynamics of which Berlusconi was at the center for almost 20 years. Likewise, we noticed the same positive relationship between the selection of *Renzo Piano* as an *Italian personality* and the *Fine Arts* and *Classics* study plan. Equally here, it is natural that those who carry out these kinds of studies know who *Renzo Piano* is and choose him as the *Italian personality* that he or she knows the best.

Again, evaluating the *idea* that foreign young people associate with Italy, we found a strong association between the *idea* itself and the *level of education* completed by the interviewee. The higher the *level of education*, the more likely the interviewee is to associate *Mafia* to the *concept* of *Italianness*. Around 30% of the *Master Degree* interviewees selected *Mafia* as a defining *concept* of *Italianness*, whereas who has obtained a *Bachelor Degree* selected this *word* only for the 21% of the total of the responses and the percentage drops to 14% with *High School* interviewees.
How do we interpret the high percentage of *Master Degree* interviewees selecting *Mafia*? We can state that the more our sample is educated, the more knowledge they have of some negative social aspects of Italy. Clearly *Mafia* is an existent reality that symbolizes a negative aspect of our country, but our question at this point is why anybody in the sample selected *Salvatore Riina* as a *personality* defining Italy? Riina, the "*boss of bosses*", has personified for almost 20 years the leading head of *Cosa Nostra*, he is globally known to be responsible for hundreds of murders and for having been in hiding for 25 years.
Does *Mafia* represent a well-known and studied phenomenon that foreign young people associate with Italy because they truly have knowledge about it, or does it rather represent a criminal organization not really known, that foreign young tend to associate with Italy as a prejudice? Strong is the media role played by the cinematographic industry in the delineation of the *Mafia* phenomenon and as a consequence *Mafia* remains a vivid and allegoric icon associated with Italy. We would claim that it is more likely that foreign young people mention *Mafia* as an Italian stereotype, they have demonstrated not to know a fundamental aspect of the *Mafia* phenomenon such as its most important man.
The final result which we consider the more relevant, is represented by the fact that foreign young people mainly associate the *word Culture* to Italy. 67% of the sample opted for this *word*, Italy is conceived as a place of culture and thus of history, of art, of beauty. The remarkable aspect is that it does not exist any dependence between the *word Culture* chosen as the *word* to associate with Italy and the fact of ever having been to Italy. We only noticed that who studied *Fine Arts* and *Classics* is more likely to recognize the cultural side of Italy than students of *Scientific* courses. However, the strong relevance of the statistic suggests that Italy is synonymous of culture. 213 out of 324 young people indicated *Culture* as *word* that illustrates and describes Italy.
To sum up we can conclude this dissertation feeling pleased by the fact that the perception foreign young people have on Italy is generally positive; they know the main indexes that characterize Italy, they recognize the popularity of Italian *Fashion*, they are more likely to have visited the country, especially the Northern and the Central territorial areas, and they have a very positive opinion about the people. Additionally, it may be possible that the *Mafia* phenomenon is a commonplace term

used to define Italy as a country, as our sample does not know it very well and, independently of the fact of ever having visited the country or not, Italy is 213 times synonym of *Culture*.
After all, already 600 years ago, Dante coined the idiomatic expression *Bel Paese* as alternative noun for Italy!
(for deeper knowledge, watch https://www.youtube.com/watch?v=14E0hlJCwzo).

Bibliography

AGRESTI A., and FRANKLIN M., (2014), *Statistics: The Art and Science of Learning from Data*, Harlow, Pearson New International Edition.

BLACK, T.R., (1999), *Doing Quantitative Research in The Social Sciences: An Integrated Approach to Research Design, Measurement and Statistics*, London, Sage.

BORRA S., and DI CIACCIO A., (2008), *Statistica: Metodologia per le Scienze Economiche e Sociali*, Milan, McGraw-Hill Education.

CORBETTA P., (2003), *Social Research: Theory, Methods and Techniques*, London [etc.], Sage.

GELMAN A., (2004), "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13 (4): 755-779.

GOODE, W.J. and HATT, P.K., (1952), *Methods in Social Research*, New York, McGraw-Hill Education.

GOODMAN L.A. and KRUSKAL W.H.(1954), "Measures of Association for Cross Classifications." *Journal of the American Statistical Association* 49 (268): 732-764.

HOAGLIN D.C., MOSTELLER F., and TUKEY J.W, (2000), *Understanding of Robust and Exploratory Data Analysis*, New York, Wiley Classics Library edition.

MAXIM, P.S. (1999), *Quantitative Research Methods in the Social Sciences*, Cambridge, Cambridge University Press.

MONTI, A.C., (2008), *Introduzione alla Statistica*, Naples, Edizioni Scientifiche Italiane.

SAS INSTITUTE INC., (2016), *SAS/STAT® 14.2 User's Guide: The FREQ Procedure*, SAS Institute Inc.

ORACLE CORPORATION, (2017), *Oracle VM. VirtualBox®: User Manual*, Oracle Corporation.

VAN AALDEREN M., (2015), *Il bello dell'Italia: Il Bel Paese visto dai corrispondenti della stampa estera*, Rome, Albeggi Edizioni.

Sitography

Facebook in
https://www.facebook.com (Accessed 04-04-2017)

Google Forms in
https://docs.google.com/forms (Accessed 31-03-2017).

Google Suite Learning Center in
https://gsuite.google.com/learning-center/products/forms/get-started (Accessed 3-04-2017).

La Cultura Italiana secondo Roberto Benigni in
https://www.youtube.com/watch?v=14E0hlJCwzo (Accessed 2-06-2017).

La Farnesina e l'Immagine dell'Italia all'Estero in
http://www.esteri.it/mae/it/sala_stampa/archivionotizie/approfondimenti/2010/03/20100325_immagine_italia_estero.html (Accessed 18-04-2017)

L'immagine dell'Italia all'Estero in
http://www.fondazioneintercultura.org/it/Ricerche-pubblicate/L%27immagine-dell%27Italia-all%27estero (Accessed 6-04-2017).

Matera 2019. Capitale Europea della Cultura in
http://www.matera-basilicata2019.it/it (Accessed 28-05-2017)-

SAS Resources in
http://support.sas.com/software/products/university-edition/index.html (Accessed 29-03-2017).

SAS University Edition in
https://www.sas.com/en_us/software/university-edition.html (Accessed 29-03-2017).

UNESCO official site in
http://en.unesco.org (Accessed 8-04-2017).

VirtualBox in
https://www.virtualbox.org (Accessed 29-03-2017).

ABSTRACT

INTRODUZIONE

La Statistica come metodo: le quattro fasi di un'analisi statistica

Le riflessioni che vogliamo condurre nel corso della tesi mirano a comprendere quale sia la reputazione di cui l'Italia gode all'estero. L'idea alla base di questo elaborato è stata concepita durante il periodo di studio in Erasmus, dove ho avuto modo di verificare quanto forte sia l'interesse che il nostro Paese susciti agli occhi dei giovani studenti stranieri.
A tal fine, si è deciso di condurre un'indagine scientifica fondata su un processo *esploratorio* di analisi dei dati, il che significa che non si è fatto riferimento ad una teoria già esistente con l'intento di confermarla o rifiutarla bensì si è reputato più interessante considerare, *ex novo*, la tematica che prende in esame l'immagine dell'Italia all'estero.
I motivi alla base della scelta di preferire il processo statistico *esploratorio*, rispetto a quello *confirmatorio*, muovono dal riscontro di una certa difficoltà nel raccogliere una bibliografia corposa che trattasse in maniera approfondita ed analitica l'argomento e in seguito abbiamo creduto personalmente più stimolante e costruttivo eseguire una ricerca di tipo sperimentale con lo scopo di esaminare cosa pensasse una specifica popolazione composta da *giovani-stranieri.*
Si è ritenuto adeguato rapportarsi alla Statistica come macro materia in quanto questa funge da dizionario che traduce i valori numerici ottenuti durante la raccolta dati in comprensione e cognizione di un concetto qualitativo, come l'immagine dell'Italia all'estero.
Riportare le quattro fasi secondo cui si suddivide un'analisi statistica, delucida il modo in cui si è sviluppata la ricerca:

I. **Formulazione della domanda statistica** → in questa prima fase, l'obiettivo è chiarire quale sia l'argomento che si vuole approfondire e , nel nostro caso, come è reputata l'Italia all'estero?

II. **Raccolta dati** → il secondo *step* di un'analisi quantitativa prevede la pianificazione di un metodo di ricerca strutturato ed organizzato: tutti coloro che hanno partecipano alla ricerca sono stati analizzati e valutati secondo le stesse modalità, ricevendo nello stesso ordine le medesime domande. In questa ricerca, abbiamo personalmente ideato e costruito un questionario anonimo su un'applicazione online di Google che prende il nome *Google Forms*.

III. **Analisi dati** → questa fase, che si identifica come la parte più corposa dell'indagine, si riferisce all'utilizzo dei metodi statistici che vengono adoperati per analizzare e decifrare i dati. Lo scopo è quello di verificare l'esistenza di una associazione tra variabili e, nel caso si trovasse, descriverne il grado e le caratteristiche. Questa tesi vede come metodo di analisi dei dati la costruzione di tabelle di frequenza e di contingenza mediante il software statistico chiamato *SAS University Edition*.

IV. **Interpretazione dei risultati** → generalizzare i risultati è lo scopo ultimo della ricerca quantitativa. Abbiamo portato a termine questa ultima fase dell'esperimento interpretando le statistiche implementate su *SAS University Edition* e commentandone i risultati. A questo fine ci siamo serviti di grafici e tabelle dimodoché il lettore possa farsi un'idea immediata dei risultati ai quali siamo giunti.

RACCOLTA DATI

Il nostro campione

Le *variabili* costituiscono le caratteristiche osservate in uno specifico studio, il termine *variabile* sottolinea il fatto che i valori *variano*. I valori che studiamo conducendo una ricerca statistica sono tecnicamente denominati *osservazioni* che possono essere rappresentate sia da numeri che appartenere a categorie. Nel nostro studio, le variabili si manifesteranno unicamente come *osservazioni categoriche*. Quando vogliamo analizzare i valori ottenuti rispetto alla relazione esistente tra due variabili, è necessario precisare la differenza tra *variabile risposta* e *variabile esplicativa*, la prima è la variabile su cui vengono effettuate le comparazioni mentre la seconda è quella che spiega e genera le variazioni nella *risposta*.
La regola è studiare come cambia la distribuzione della *variabile risposta* al variare della *variabile esplicativa*.
Come riusciamo ad osservare se la distribuzione di una variabile cambia in base a cambiamenti dell'altra? Come raccogliamo informazioni riguardo al grado e alla tipologia dell'associazione tra le due variabili? Analizziamo la loro distribuzione congiunta rappresentata in tabelle di contingenza implementate su *SAS*. Questa tipologia rappresenta la combinazione, per riga e per colonna, di ogni categoria della *variabile risposta* con ogni categoria della *variabile esplicativa*.
Il campione che abbiamo ottenuto attraverso la condivisione del questionario è composto da 324 soggetti, il nostro intento era quello di ricevere un riscontro dal maggior numero di persone che rispettassero due condizioni:

1. Non essere italiane
2. Avere un'età compresa tra i 14 e i 25 anni.

Si sono ovviamente verificate alcune distorsioni, *bias*: alcuni intervistati non hanno reso noto in maniera chiara il paese di provenienza, altri non hanno compreso la domanda, etc. Abbiamo dunque proceduto, prima di caricare i dati sulla piattaforma *SAS*, all'organizzazione e alla decodificazione dei dati in *Excel*.

Il nostro questionario

Il questionario creato su *Google Forms* è composto da 15 domande a risposta multipla (solamente una prevede l'inserimento di una porzione di testo).

È suddiviso in tre Sezioni:

I. La prima Sezione è focalizzata sulla raccolta dei dati personali dell'intervistato:
   - Da dove vieni?
   - Specifica il tuo Sesso
   - Che livello di istruzione hai completato?
   - Quale tra i seguenti campi è il più simile al piano di studi che stai conducendo? [Scienze Sociali/Scientifico/Classico/Interdisciplinare o Altro]

II. La seconda Sezione esplora quanto profonda sia la conoscenza che i nostri intervistati hanno dell'Italia:
   - Sei mai stato in Italia?
   - Dove sei stato in Italia?
   - Quale è la capitale Italiana?
   - Sapresti indicare dove è localizzata l'Italia?
   - Sapresti indicare approssimativamente quanti abitanti ci sono in Italia?
   - Sapresti indicare quale tra le seguenti immagini rappresenta un sito UNESCO Italiano?

III. La terza e ultima Sezione è stata pianificata per cercare di studiare quale sia la percezione che il nostro campione ha dell'Italia:
   - Quale parola assoceresti alla parola "*Italia*"? [Cultura/Fashion/Crisi Economica/Illegalità/Sport/Cronaca]

   - Quale idea relazioneresti al concetto di "*Italianità*"? [Caffè/Vespa/Mafia/Colosseo/Calcio/Papa]

- Quale tra i seguenti brand Italiani conosci meglio?
  [Armani/Lavazza/Ducati/Bialetti/Jacuzzi/Altro]

- Seleziona il personaggio Italiano che conosci meglio
  [Silvio Berlusconi/Chiara Ferragni/Laura Pausini/Renzo Piano/Sergio Marchionne/Salvatore Riina/Lapo Elkann/Altro]

- Quale tra i seguenti aggettivi descrive meglio una persona Italiana?
  [Loquace/Amichevole/Buongustaio/Creativo/Maleducato/Corrotto]

ANALISI DEI DATI

SAS University Edition

L'indagine che abbiamo condotto è stata implementata su *SAS University Edition*, acronimo di *Statistical Analysis System*, un software sviluppato dal SAS Institute per effettuare diverse tipologie di analisi statistiche. I dati raccolti attraverso il questionario creato su *Google Forms*, una volta riordinati e decodificati su un foglio *Excel*, sono stati caricati in *SAS*.
È necessario specificare che, per installare e poter usufruire (gratuitamente) di questo software, è stato necessario scaricare un software che funge da Macchina Virtuale che permette a ogni tipologia di computer ti ricreare un ambiente virtuale capace di emulare il sistema operativo desiderato. La nostra Macchina Virtuale è la *Oracle VM VirtualBox*.

Strumenti Statistici

A questo punto della ricerca, una volta chiarita la domanda da approfondire, somministrato il questionario creato su *Google Forms* e caricati i dati in formato *SAS* su *SAS University Edition*, dobbiamo menzionare gli strumenti statistici che abbiamo utilizzato:

La Statistica Descrittiva → è il ramo della statistica che analizza i criteri di rilevazione, classificazione, sintesi e raffigurazione dei dati ottenuti durante lo studio di una popolazione o di un campione. Nella nostra tesi, attraverso la statistica descrittiva presentiamo e commentiamo i dati una volta portato a termine il processo di *raccolta*. In questo caso specifico, facciamo riferimento a quattro indici statistici che misurano l'intensità della associazione tra due variabili ed assumono il valore 0 nel caso di indipendenza (assenza di associazione):

I. *Chi-square Statistc* → $\chi^2 = \sum_a \sum_b \frac{(\nu_{ab} - \nu_a . \nu_{.b}/\nu)^2}{\nu_a . \nu_{.b}/\nu}$

II. *Phi Coefficient* → $\varphi^2 = \frac{\chi^2}{\nu}$

III. *Coefficient of Contingency* → $C = \sqrt{\frac{\chi^2/\nu}{1+\chi^2/\nu}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$

IV. *Cramèr's V* → $\varphi_c = \sqrt{[\chi^2/\nu]/\mathrm{Min}(\alpha - 1, \beta - 1)}$

La Statistica Inferenziale → viene utilizzata con il fine di fare predizioni riguardo l'intera popolazione avendo dei dati che si identificano in un campione casuale. Come esplicitato in precedenza, l'obiettivo di questa ricerca è valutare se e come due variabili categoriche siano indipendenti o dipendano l'una dall'altra. Il *test di significatività* è un metodo che prevede l'utilizzo dei dati per riassumere evidenza contro un'ipotesi (una previsione riguardo la popolazione).
La *Verifica di ipotesi*, che si usa per verificare la bontà di un'ipotesi, segue 5 fasi:

I. Determinare le *Assunzioni*, nel nostro caso le variabili sono categoriche e il campione è *random*.

II. Formulare le *Ipotesi*, quest'ultime sono due affermazioni sulla popolazione:
   1. H_0 (Ipotesi Nulla)→le due variabili sono indipendenti.
   2. H_1 (Ipotesi Alternativa)→le due variabili sono associate e dipendenti e dunque H_0 è falsa.

   L'idea è rifiutare H_0 se nel campione esiste sufficiente evidenza contro questa.

III. Procedere con il calcolo della *statistica test*, che misura quanta evidenza esiste contro H_0, usiamo la formula del *chi quadro* (χ^2) sopra ri
IV. portata.

V. Calcolare il *p-value*, la probabilità di osservare valori più grandi di quelli osservati sotto H_0. Piccoli valori del *p-value* indicano grande evidenza contro H_0.

VI. Trarre le *Conclusioni*, questo ultimo passaggio prevede l'interpretazione del valore del *p-value* e la conseguente decisione riguardo H_0.

Conclusioni

Attraverso l'interpretazione delle tabelle di frequenza e di contingenza realizzate su *SAS*, siamo in grado di trarre delle conclusioni che possono essere generalizzate alla popolazione *giovani-stranieri*.

Il nostro campione, rappresentativo della popolazione *giovani-stranieri*, conosce i fondamentali aspetti territoriali dell'Italia, come la Capitale, dove è localizzata in Europa, il numero di abitanti e riconosce uno tra i più importanti siti *UNESCO* Italiani.

Abbiamo inoltre riscontrato un forte interesse per il nostro Paese osservando il numero di persone che hanno visitato l'Italia: il 77% del campione ha dichiarato di essere stato in Italia almeno una volta. In aggiunta, abbiamo notato che il Nord e il Centro Italia sono le zone territoriali che attraggono il maggior numero di visitatori. Il Sud non rappresenta un polo attrattivo per i *giovani-stranieri*, a esclusione della *Campania* che è l'unica regione meridionale con un numero di visite statisticamente rilevante.

È stato osservato attraverso le associazioni bi-variate che gli *Europei* e gli *Asiatici* sono più inclini a visitare l'Italia rispetto ad *Americani* e *Australiani*. (i 3 *Africani* intervistati non sono rappresentativi del loro *Continente*).

Per quanto concerne l'opinione che i *giovani-stranieri* hanno dell'Italia, abbiamo individuato che il popolo degli Italiani è giudicato positivamente, meno del 5% del *sample* ha selezionato gli *aggettivi Corrotto* o *Maleducato* come tipici del popolo italiano. Di contro, siamo identificati come *Loquaci* e *Amichevoli* da più dell'80% del campione.

Un altro aspetto che abbiamo individuato interpretando i nostri dati riguarda la moda italiana. Moltissimi intervistati hanno proposto una consistente varietà di *brand* non menzionati dalle opzioni risposta, selezionando l'opzione *Altro* alla domanda *Quale brand italiano conosci meglio?*. Non ci aspettavamo un riscontro così positivo e statisticamente tanto rilevante. Anche i *giovani-stranieri* riconoscono il *Made in Italy* come protagonista nel panorama internazionale del *Fashion*.

L'*idea* che maggiormente si associa al *concetto* di *Italianità* è il *Caffè*.

Tuttavia, il risultato interessante riguarda le associazioni che abbiamo trovato tra l'*idea* da relazionare all'*Italianità* e il *genere* dell'intervistato; i *Maschi* riconoscono nel gioco del *Calcio* una caratteristica tipica italiana mentre le *Femmine* identificano nella *Vespa* il simbolo dello stile italiano. Forse quest'ultima relazione si può spiegare facendo riferimento all'icona vivida nell'immaginario collettivo del film "*Vacanze Romane*" dove la *Vespa* gioca un ruolo simbolico e caratterizzante dell'Italia.

Analizzando le *personalità italiane meglio conosciute* dai *giovani-stranieri*, *Silvio Berlusconi* supera di gran lunga tutte le altre *personalità* proposte nella domanda. Non siamo in grado di

spiegare se l'attenzione per questo personaggio sia dovuta ad una reale conoscenza della vita politica dell'Italia o sia dettata dalla risonanza mediatica delle vicende legate a questo personaggio. L'unica delucidazione che può essere portata in superficie risiede nel fatto che, tra i nostri intervistati, chi ha completato un *piano di studi* affine alle *Scienze Sociali* e dunque alle Scienze Politiche, alla Sociologia o all'Economia, è più probabile che conosca *Silvio Berlusconi*. Riteniamo questo dato conforme alle nostre aspettative. Ugualmente abbiamo notato un'associazione positiva tra la selezione *Silvio Berlusconi* e candidati *Europei* o *Asiatici*, al contrario *Americani* e *Australiani* riconoscono in *Valentino Rossi* il personaggio simbolo dell'Italia più di quanto lo riconoscano *Europei* e *Asiatici.*

Un dato singolare si riscontra nell'associazione positiva tra la scelta del *concetto* di *Mafia* e il *livello di istruzione* raggiunto dall'intervistato: più alto è il *livello di istruzione*, più il nostro intervistato propende a selezionare *Mafia* come risposta alla domanda *Quale idea relazioneresti al concetto di "Italianità"?.* Sembrerebbe che chi è più istruito, conosca meglio le dinamiche sociali di questa organizzazione criminale. Tuttavia, ci siamo domandati come mai nessuno abbia selezionato *Salvatore Riina* come persona simbolo dell'Italia, come mai tra quel 28% di studenti che hanno ottenuto un *Master* e che hanno indicato *Mafia* come *concetto* simbolo dell'Italia, nessuno abbia selezionato la *personalità*-simbolo della *Mafia.* Possiamo affermare che la *Mafia*, più che essere un fenomeno realmente conosciuto e studiato dai *giovani-stranieri*, rappresenti una forma di stereotipo da attribuire all'Italia. Difatti, è forte e influente il ruolo mediatico giocato dall'industria cinematografica mondiale nella rappresentazione del fenomeno mafioso.

L'ultimo risultato che ci permette di affermare che i *giovani-stranieri* giudicano positivamente l'Italia, consiste nel fatto che il 67% del campione abbia selezionato *Cultura* alla domanda *Che parola associeresti all'Italia?*. L'aspetto che riteniamo interessante è che non ci sia alcuna associazione tra *Cultura* come *parola* selezionata e il fatto di aver visitato il nostro Paese. Indipendentemente dal *genere* dell'intervistato, dalla sua nazionalità e dalla sua conoscenza del paese, l'Italia per 213 volte (su 324 intervistati) è apparsa sinonimo di *Cultura*.