POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Gestionale

Tesi di Laurea Magistrale

Star and Prolific Inventors: Empirical Analysis of Differences in Education and Career

Relatore:

Federico CAVIGGIOLI

Candidata:

Sara FERRAGATTI

Aprile 2019

To those who believe in me and fight with me every single day of my life

To those who have left this world but will never leave my heart

Acknowledgments

I would like to thank all those who have allowed the accomplishment of this path, starting from the first year at university to these six months employed to work on the thesis. I thank all those who have allowed the success of this work.

I would like to thank prof. Federico Caviggioli, supervisor of this Thesis for the infinite availability and all the time and the energies invested, for transmitting the passion and enthusiasm for the thesis to take shape day after day and to have guided me in the management of time and in the most appropriate choices.

I also would like to thank my shoulder Davide, for the moral support and for giving me the tenacity and the love I needed in order to face this work to the fullest.

Finally, a special thanks to my parents, Flavio and Daniela, my granny Anna and my beloved aunt Dada for their support and their continuous and encouragement thanks to which I reached the finish line. Thank you for sharing with me the joys, the sacrifices, the successes but also the moments of despair and the failures related to this path. The affection and the empathy you have shown make this success even more extraordinary.

Contents

List of Figures

List of Tables

List of Equation

Chapter 1

1 Introduction

The study of innovation has spread over the last fifty years in an ascending climax of importance. The process of technological innovation has been studied in depth by numerous academic fellows in order to understand how it is composed and on which factors it relies on. From Schumpeter [28], the innovation pioneer, to Rogers [29] : academics question themselves about the phenomenon that, etymologically, *introduces new systems, new orders, and new methods*. However, the question of who conceived the invention of the century regarding his origins, formation and professional career is a much more recent issue.

This dissertation aims to investigate the correlation between becoming a Star Scientists and the professional career and education.

The work is divided into two chronologically different phases: the first phase of information gathering, and database creation and the second analysis of data collected through statistical tools such as multivariate regression.

This thesis proposes to collect two samples of inventors: the first sample consisting of successful inventors who participated as finalists or won the EPO award (*Star Inventors*) and the control sample consisting of the *Prolific Inventors*, i.e., those who have patented a lot during their career none of their inventions has achieved scientific but commercial success.

The peculiarity of this thesis consists primarily in the construction of the database: data about the inventors' candidates has been collected for the EPO award from 2010 to 2014 and information about the Prolific Inventors with a deep research through social networks, academic articles, search engines and online encyclopedias. This first phase of data collection turned out to be smoother for the EPO sample and harder for the sample containing the prolific inventors: the information on the Star inventors is widely available

because of their fame due to their inventions; on the other hand, prolific inventors are mainly traced through the information they share on social networks. Therefore, the timing dedicated to the data collection of the prolific inventors proved to be long and with a success rate of just over 40%: in order to obtain 114 complete subjects, more than 251 subjects were analyzed.

Once the first phase of data collection and database creation has been completed, the statistical analysis and regression phase of the information retrieved begins. Firstly, a statistical analysis is carried out using Excel on the basis of the samples analyzed and then a logistic model on multivariate regression analysis is chosen to confirm this analysis. The regression on Stata is carried out on the basis of the EPO **dependent variable**, which represents a characteristic of being a Star Scientists if it is set equal to 1, the **variables of interest** which represent characteristics studied starting from the database regarding the main factors of education and career and finally the **control variables** that is the biographical characteristics of the inventors.

The ultimate purpose of this thesis is, therefore, to understand the relationship between being a successful inventor (*Star Scientist*) and the previous personal path in terms of education and career, considering however the biographical factors.

1.1 Structure of Thesis

This dissertation consists of 6 chapters, structured as follows:

- **Chapter 2** presents an overview about the state of the art of innovation and the ways in which it is rewarded
- **Chapter 3** introduces the methodology used to create the database and the difficulties encountered in this regard.
- **Chapter 4** explains the statistical analysis in the two samples according to the two main factors, i.e., education and career
- **Chapter 5** presents several models of multivariate regression analysis performed using the Stata software

- **Chapter 6** exposes the conclusions and possible future developments of this dissertation

Chapter 2

2 Background

This chapter deals with the main theoretical concepts introduced in this study: it describes the main points of view of the literature drafted up to now on topics that will be dealt with following. Starting from the most general topics such as innovation and invention from different points of view: the difference between the two concepts and the link that binds them; and then introduce the patents, the awards and their alternation and / or substitutability.

After illustrating and discussing the different main concepts concerning the invention - innovation binomial, with a specification regarding breakthrough innovation, it will follow an accurate description of the literature on current awards which pays homage to innovation Finally, a summary of the main concepts concerning patents, and patenting offices.

Moreover, this chapter provides details on the European Inventor Award and introduces the two categories that will be presented in the next sections: the "Star Inventors" and the "Prolific inventors", the distinction between the two groups of scientists based on the success of their inventions.

2.1 Description of the underlying concept

2.1.1 State of the art of innovation

In this section it will be introduced the general framework of the project: what inventiveness is and how it has been defined in numerous models, including the linear one and the Schumpeterian one, to then define how it leads to innovation and how relevant it is to the work. Innovation is a fundamental factor in economic progress that benefits

consumers, businesses and the economy as a whole. In economic terms, innovation indicates the development and application of ideas and technologies that improve goods and services or make production more efficient and also achieve social improvement.

Nowadays **Invention** and **Innovation** are key words in business contexts: firms and companies operate in increasingly complex environments. The field is super-competitive, with many challenges that force all the actors to work to make the differences. Innovation and invention have a fundamental role in this mechanism.

It is necessary to distinguish these two concepts: according to the Organization Economic Co-operation and Development (OECD) Innovation is the "*implementation of a product, or of a process, new or considerably improved, of a new marketing method, rather of a new organizational method.*"[1] Instead, we can define the Invention as *"something such as a device or process that has been created or made up, or the process of creating or making up something or figuring out a way to do something."* [2] Thus, from both definitions it is clear the difference: Invention becomes innovation if goes through the production phase, the marketing phase and, finally, it is widespread on the market. In any case, an innovation is always linked to an invention. It is therefore possible to state that an invention is a necessary condition for an innovation, which is nothing more than an invention marketed successfully.

In the literature the innovations can be subdivided according to various categorizations, in this thesis the one based on the novelty of the results will be adopted. Three types of innovation can then be distinguished:

- *Incremental innovation*: modification, consolidation and improvement of products, processes, services and activities production and distribution already existing. Most innovations belong to this category;
- *Radical innovation*: the introduction of new products or services that are due to considerable changes within an entire enterprise, and engine for creation of new values.
- *"Breakthrough" innovation* (or revolutionary innovation) point on the surprise they generate in people.

These types of innovation they represent rare events, deriving from scientific or engineering intuitions, and this is why they are considered revolutionary, they realize what many people do not they thought possible. Revolutionary innovations imply the creation of something new, or they satisfy a need that had never been discovered before.

The latter are considered the most rare and valuable and at the base of technological changes.

Then it will be explained how inventions become innovations through a simplified model; to try to explain the innovation process the first model used to understanding the relationship between technology, science to economy is the linear model of Innovation as shown in Figure 2-1.

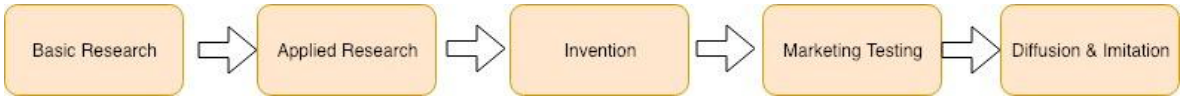

Figure 2-1 – Linear Model of Innovation

In short, the model explains how inventions become innovations in three steps. The process starts with the first step: the research, articulated in "basic research" and "applied research". Then the second step is the experimental development. Finally, the third step is articulated in production and market phase: i.e., whole series of prototypes are produced, then the results can be patented and finally they will go to the market.

From the philosophical point of view Joseph Aloïs Schumpeter (1883-1950) a pioneer in the definition of innovation, in 1939 outlines the innovative process as "Creative Destruction" to underline the attempt of man to simplify his life through the cancellation of previous technologies. Schumpeter articulates the innovation process in three concatenated concepts:

1. the invention, that is an expression of knowledge that can be articulated or not in application;
2. the actual innovation;
3. the diffusion of innovation, that is the imitation of the latter.

On the other hand, if Schumpeter focuses on the phenomenon of innovation, other scientists such as Daniel W. Surry and Donald P. Ely and then Everett Rogers in his seminar [21] focus on the diffusion of this phenomenon. The latter in fact outlines the spread of the innovation as the development of a disease, almost a Gaussian: on the y-axis the number of enterprises Adopters and on the x-axis the Time. Based on this graph Everett Rogers (1962) identifies 5 categories, like in Figure 2-2 of companies: Innovators, Early Adopters, Early Majority, Late Majority and Laggards.

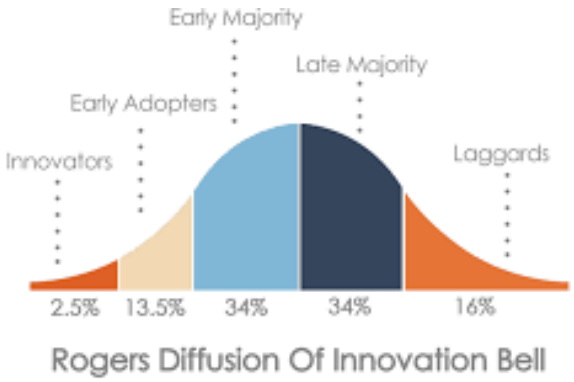

Figure 2-2 - Rogers Diffusion of Innovation Bell

In the 70s then the study of this new subject leads Abernathy & Utterback in 1971 to argue that innovation is the natural continuation of an invention. They see innovations as interdependent evolutions through distinct phases that correspond to structures of different sectors and, consequently, to different types of competitive advantage. Finally, the literature comes to define the curve 's', that is to say that the life cycle of the technology the graph in Figure 2-3 where we can see the percentage of technology presumably achieved as a function of time or cumulative R&D. From the embryonic state of a technology in the state of maturity that will soon be surpassed by another technology or the evolution of the previous one.

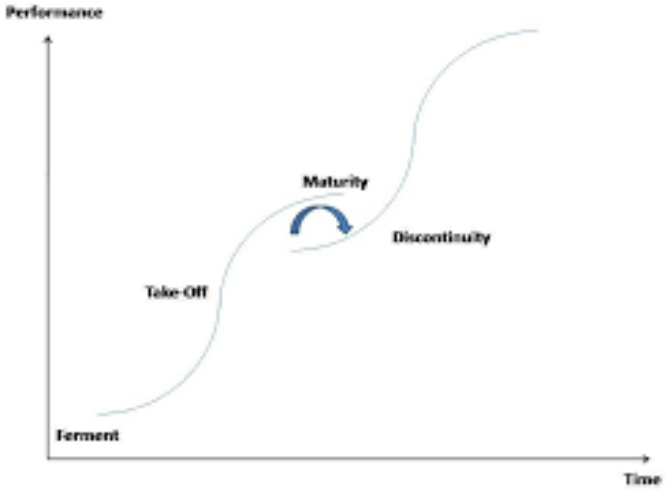

Figure 2-3 - Innovation Curve

In conclusion Innovation is therefore a very important dynamic phenomenon that must be studied from all points of view. In this regard, inventors can be considered the engines of innovation that is etymologically those who "put the phenomenon in motion".

2.1.2 Star Scientist versus Prolific Inventor

Those who we have recently called "innovation engines" in a metaphor are the broadest category that we will split into "Prolific Inventor" and "Star Scientist" for the purposes of our study. The aforementioned chiasm is not recognized by most of the literature as a clear division, that is, on the one hand those whose inventions are revolutionary from the commercial and scientific point of view, and on the other those whose inventions have not achieved great success as regards the marketability and in the scientific field.

In this paragraph we will discuss the dichotomy “Star Scientist” / “Prolific Inventor” in the current state of art according to some experts, who have conducted several researches in order to document and prove the dichotomy itself from different points of view.

There are many inventors and many inventions, but few of these can change the world with their masterpieces: they are called call they Giants, in this article of Jan Hohberger:

“The phrase “standing on the shoulders of giants” is often used by economists, sociologists and historians to describe progress in science and technology. At the core of this statement lies the notion of a process in which inventors and scientists develop ideas, based upon the discoveries of other inventors or scientists, and where

new ideas add to an existing stock of knowledge. However, there may be a second 'truth' within the statement. Research in science and innovation has demonstrated that a small group of individuals, often called 'star scientists' or 'star inventors', is associated with generating a disproportionately large amount of scientific and technological ideas (Lotka, 1926; Narin and Breitzman, 1995; Zucker et al., 2002). Therefore, 'stars' might be the 'giants' in knowledge development on whose 'shoulders' we stand.". [4]

According to this paper Star Scientist are a small group of people who detained a lot slice of knowledge. Usually these "Giants" are awarded prizes for their remarkable inventions and thanks to this factor we can distinguish them from the prolific inventors, i.e. those who, despite their numerous patents, have never left a footprint with their inventions. The data that will confirm our theory, presented later in the third chapter, are based on an illustrious award given to the star inventors: the EPO award. According to the essays examined [11], [12], [13],[14],[15] it is possible to conclude that the literature does not identify a true chiasm between the prolific inventors and the star inventors, although studies on this differentiation have been carried out and the two categories have been examined; never arrived at a clear definition of two groups. Despite this, the literature proposes a clear improvement from the assumption of a star scientist in an R&D unit as evidenced by the study [11] where the performances were examined before and after the arrival of a Star Inventor in a company in the field of biology. The same argument also supports another essay [12] concerning the biotechnology sector where a regression analysis on the Star Inventors with a control sample of prolific inventors shows a positive correlation between the presence of a star inventor in a team and the team performance itself. But can a "Giant" also influence a young student's career? A study on this topic conducted at major Chinese and Korean universities [13] demonstrates how a collaboration with a star inventor could lead to a successful career in most cases; collaboration in terms of tutoring, PhD superadvisoring, relatoring. In the same study the principal methodologies that are still used today to measure the success of a scientist are analyzed in an accurate way and all this is important in view of the Thesis to which we want to reach this volume. First of all, the number of publications and citations of the subject under analysis is analyzed, taking care to specify if the number of the last ones contains also self-citations and finally a recent tool frequently used in these statistical studies is the Hirsch Index. The H index is introduced in 2005 in order to replace the bibliometric indicators and takes into

account both the publication number and the average number of citations and the sum of all citations and is defined as "scientist has index h if h of his or her Np papers have at least h citations each and the other (Np - h) papers have ≤h citations each" [14]. The h-index is relatively quicker to calculate with the right information, but it is not suitable for this job since it is indicated to compare inventors of the same age, in a similar field of study.

In conclusion in the literature we can find many studies on what are called "giants" or "Star inventor" regarding the positive influence on the performances of the projects in which they participate, on their previous studies, on their participation in research centers or in universities but there is a shortage of studies concerning those we have called dubbed "prolific inventors" to reinforce the thesis that a clear difference between the two groups is not documented. As anticipated, the lackluster literature on this topic argues that the prolific inventors have produced a plethora of patented secondary inventions but with little economic impact; although the fact that they are patented guarantees the requirement of novelty and 'marketability'. The study of Levine [15] by means of a bibliometric analysis proves the aforementioned thesis. The number of citations of the articles verifies at least partially the impact of an invention and through this it has been possible to conclude that the prolific inventors are ignored with respect to the Star Scientists although they are generally gifted with great technical creativity.

In light of all that has been said about star scientists and prolific inventors, our work proposes a career and education analysis of two samples: one of scientists participating in the EPO award and opposed to this the control sample containing inventors prolific selected by the European patent office. The analysis aims to understand by means of a regression analysis whether there is a substantial difference between the two classes regarding the university they have attended, the state of origin, the accomplishment or not of PhD, Master and many other variables. For this work analyzing inventor's carriers is basic, so with an accurate research we tried to discover experiences abroad (in the EPO sample before and after prize), training or carrier inside a University or Research Center.

2.2 The relationship between the binomial relationship / education and innovation

In the previous paragraph it has not been studied in depth how the state-of-the-art education / career binomial has been examined in light of inventiveness and innovation.

In general, it seems that in history it is documented and ascertained that the role of innovators has led to both economic and social growth, but we do not have much information on the education of innovators. The question to be answered is: can the academic and work path influence the inventiveness? There seems to be little evidence to support this theory. In this regard, therefore, the essay [21] by William J. Baumol, Melissa A. Schilling, and Edward N. Wolff proposes to collect evidence from a bibliographic database on the potential role of education in inventive and innovative processes. The paper premised an initial hypothesis before the data collection phase: "*Standard educational approaches, especially those that are more rigorous and technical, rather than helping, tend to impede innovative entrepreneurship by constraining heterodox thinking and exercise of the imagination.*" [21] As this hypothesis is heavy it would require an equally heavy demonstration; in fact, the essay does not arrive at this claim that would be revolutionary, but by specializing on innovative innovators, it comes to the conclusion that, in order to become such, a type of preparation is necessary. However, it is not a specific university that makes the difference but only the rigor of the structure itself and how it is dealt with by the candidate. The conclusion of this article is that in all countries we can see an increase in entrepreneurs / innovators who have attended high school, college, masters, and PhD. degrees therefore a climax of degrees of instruction over time. Moreover, a percentage of educated inventors higher than the level of innovative entrepreneurs: to support this thesis it is possible to cite as an example all the Nobel prizes to science are endowed with at least a PhD., while entrepreneurs who have changed the history of technology often they do not even have a degree; is the example of Steve Jobs, William Gates and Lawrence Ellison. An unresolved issue in the scientific literature remains open: does in-depth education and long-term experience foster innovation in the aforementioned field?

Even on this the literature appears dichotomous: on the one hand there are those who say that before making a revolutionary discovery in a field it is necessary to train in that field through a scientific baggage relevant to the discipline (Simonton, 1999a, 1999b) sometimes also quantifying the necessary experience (Simon and Chase (1973)); while the opposing faction asserts that an individual's studies and experiential baggage may even stop or inhibit the creative problem-solving mechanism inherent in him. (Wertheimer, 1945/1959). Therefore, automating the formula for solving a problem does not spur the individual to find a better one. It can be concluded that there is evidence to support both theories and what our thesis proposes instead to demonstrate is to reveal what makes an inventor, a star inventor? We will try to understand if we are dealing with factors related to education and career.

Generally, one can identify a star scientist with respect to a prolific scientist on the basis of the social and economic impact of an invention. An invention of a certain importance is usually rewarded through scientific acknowledgments; and it is on the basis of this parameter that we have established to distinguish between star scientist and prolific inventors in our project. Precisely because the literature does not identify a real chiasm between the two categories identified, it was necessary to arbitrarily establish a distinction to perform a statistical analysis: The star inventors in this work are those who have received recognition (or have been finalists) of the award given by the European Patent Office (EPO) and will be our main champion; opposed to the second category examined the Prolific Inventor, i.e. the control sample; those who have often obtained numerous patents but none of an economic and social importance. In the following paragraphs therefore the above EPO award is analyzed, which distinguishes the star scientists who have achieved it and for completeness all the awards and prizes awarded in the scientific field to star scientists.

2.3 The most important innovation Award

Often the prizes and awards not only accompany patents but are the alternative to them; and the very fact of receiving a recognition is a recognition.

"*Individual have an innate desire to distinguish themselves from other individuals.*" [6], receiving a prize is not only a monetary reward, but it is intrinsic in human nature to desire a social ascension, in order to distinguish itself: being recognized in a research field has always symbolized elevation of role and prestige in society. But apart from the pleasant feeling of being rewarded, for the research the award achieved serves to distinguish the categories identified: "Star scientists" and "prolific inventors". For completeness it will be presented the major institutions and their prize level. The first but also the most famous is the **Nobel Foundation**, which releases the award of the Nobel Prize, which honors living people who have brought the most important benefits to humanity in economics, chemistry, physics, medicine, literature and finally for the peace. Another illustrious recognition is the so-called **Academy Award**, the most ancient film award in the world, followed by the American **Pulitzer** prize, which honors journalism. Speaking about arts the **Grammy** and **Emmy** awards for music and for television programs are the most known. Entering in the field of science a well-known prize is the **Field Medal**, the most coveted recognition for mathematics. Above all in the economic field, innovation prizes and awards have aroused since the eighteenth century in order to encourage technological creativity and research activity of inventors. In 1714 the English parliament decreed a law that provided for economic recognition to those who had found the way to measure the longitude of the Earth (Longitude Act), this symbolically became the first innovation prize going back to the modern era. This has become the emblem of many awards and rewards that celebrate the spirit of innovation from the eighteenth century to the contemporary era. Some current examples of these competitions are found in the United States in particular: Innovation Celebration is an organization that allocates funds for innovation awards aimed at students, start-ups and businesses in the most distinguished American colleges [7], another example eloquent is the National Medal of Technology and Innovation (NMTI), which is conferred by the President of the United States to the major American innovators who have contributed to the development of very relevant technologies, represents the highest US honor for technological progress. In the United States as well as in many European countries the awards that celebrate innovation have increased over time.

In recent decades there has been a tendency to substitute the patent with alternative mechanisms such as awards and research contracts as explained and demonstrated by a

mathematical model the scientist D. Wright in his essay "*The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*" [8]: this paper analyzes the three forms of recognition and shows the best benefits, how much and how it is preferable to one another.

Many other examples of authority competitions and recognitions unknown to the public celebrate innovation in many different nations, why? There has always been some sort of debate between patent and alternative mechanisms such as awards and research contracts. As explained and demonstrated by a mathematical model the scientist *Dorell Lawrence Wright (1985)* in his essay "*The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*" [8] are all forms of protection and encouragement for innovation with different nuances. This paper analyzes the three forms of recognition (patents, awards and research contracts) and discloses the best benefits, how much and how it is preferable to each other. Research grants and contracts seem appropriate if the inventors are not financially motivated, otherwise due to the strong informational asymmetry between them and the research bodies could derive an illicit advantage. As far as patents are concerned, the main concern is that for which they are the main source of monopolies in the economy, particularly in the pharmaceutical sector, a recent development in this field is that of buy-out patents, which limits this series of questions.

2.4 The European Inventor Award

Among all the forms of recognition analyzed, the EPO award was chosen, as it refers to patents, which as explained in the previous paragraphs are a measure of innovation.

The EPO Award is "*one of the most prestigious competitions of its kind, the European Inventor Award pays tribute to the creativity of inventors the world over, who use their technical, scientific and intellectual skills to make a real contribution to technological progress and economic growth and so improve people's daily lives*". This is the description of European Inventor Award taken directly from the official web page of the European Patent Office [5].

This competition started in 2006 and since then it has been proposed to encourage innovation and patenting. In the course of 2010 in 2010 the public is to define the winner from the finalists of a peculiar category in itself called "Popular Prize". The voting mechanism takes place through social media (in particular on the official web page or on the EPO website) and grants the right to participate in a lottery.

In EPO prize there are five categories:

- **Industry**: this category includes inventors who have made an invention typically with a huge economic impact, and on behalf of big companies, described by two parameters, that is to say a minimum of 250 employees and a minimum of annual turnover of €50 Million;
- **Small and medium-sized enterprises (SMEs)**: this award honors the inventions behind small and medium-sized businesses; i.e. companies with an annual turnover of less than 50 million euros and a number of employees less than 250 at the time of granting the patent. The invention in question has ideally had a significant commercial economic impact that enabled it to expand its market
- **Research**: this award is presented to inventors working at research centers and universities. Historically, the winning inventions and finalists of this category have led to considerable technological advances and have helped to increase the reputation of the inventor's research institute.
- **Non-European Countries**: This category is open to all inventors outside the 38 EPO states regardless of whether they come from a research institute or company and the size and turnover of the latter. However, the commercial success achieved by the invention in Europe is relevant.
- **Lifetime achievement:** This award honors the contribution of a long-term European inventor whose efforts, sometimes witnessed by numerous patents, have had a major impact on technology and on society in general.

The selection mechanism is simple: candidates send their application through the EPO website, among all the candidates some are selected, and few will become "Nominated" at the European Invention Award. After this process, a carefully discerned jury examines the applications of the nominated candidates and choose a narrow group of finalists (which can vary between two and four) and, finally, nominates the winner for each category. The EPO expert and the independent international Jury evaluate the applications and their related innovation, taking into consideration not only the technological originality, but also the social-economic impact.

The Jury is usually composed on an average of twelve distinguished experts in science, politics, intellectual property, business, media and research, among them one is the chair, usually the one with the brightest Curriculum Vitae.

For the last thirteen years the ceremony takes place in a different European city, respectively since 2006 Brussels, Munich, Ljubljana, Prague, Madrid, Budapest, Copenhagen, Amsterdam, Berlin, Paris, Lisbon, Venice, Paris and finally in 2019, Vienna.

Precisely because the EPO award is based on patents as a measure of innovation, it is necessary to find out about the basic functions and regulations related to patents, also in terms of knowing how to read and interpret the specific codes related to patents and the main world organizations in the database to protect these.

2.5 Awards Versus Intellectual property

As already mentioned, the role of prizes and patents in innovation is ambiguous: sometimes exclusive, sometimes complementary. The traditional view is the second, as patents are a complement to the prizes, which increase its value and increase its technical developments. However, in the literature there are opposing examples such as awards as an alternative to IP. Intellectual property contributes first of all to giving a high level of protection as regards the right to exclusivity and also provides a way to ensure a higher price for the product in question, considering that at the marginal cost of the product an extra cost must be added

due to exclusivity given by the patent. As for the prizes, instead, they confer less rigid protection and a flat-rate recognition.

2.6 Patents

Intellectual (or industrial) property (*IP*) can be defined as "*a set of legal rights aimed at protecting inventions and in general the fruit of creation and human intellect*". [9] When a company launches a line of success, competitors will probably try to launch similar lines that, sometimes, becomes identical after a certain period of time. However, the innovating companies invests in R&D a large amount of money that is of course reflected in the selling price, while the competitors, who have not invested or made any research effort, can keep a lower price and, sometimes, make agreements with the distributors. Intellectual property helps protecting the companies' drawings, brands, etc. The IP system provides property at work and exclusive rights for production control, import and counterfeiting.

The intellectual property can be divided into two categories: the **copyrights** that mainly protect music, cinematographic, literary works and architectural structures, and proper **intellectual property** that covers patents, utility models and trademarks. This work focuses on patents as unit of measure of innovation activities.

The patent is a kind of contract between the applicant and the society: the applicant is interested to benefit from his invention and the related pecuniary performance, while the entity that issues the patents and the scientific community have an interest in supporting the technological innovation, provide the protection they need the companies investing in R&D activities in order to keep the economy competitive disclosing the details of new inventions in order to promote improvements of the latter and, ultimately, facilitate technology transfer. In conclusion, companies are seeking protection for innovation in exchange for disclosure of the same, this social contract takes shape in the patent law. In fact, in order for a patent application to be accepted, the invention should possess a set of requirements that varies according to the country. In any case, irrespective of the nationality of the inventor, all inventions should normally be new, with an innovative step compared to the state of the art and must be applied to the industrial application. To distinguish what is patentable from what is not, it is necessary to consult *Article 52 and 53* of the *European*

Patent Convention (ECB). In order to be able to read and interpret the database and in particular the patent number of the inventions is been necessary studying patent's classification, International Patent Classification (IPC) and Cooperative Patent Classification (CPC). Word Intellectual Property Organization (WIPO) manual defines Classification as a "*specific system which subdivides technology into distinct units*" [16]; the classification symbol is like an ID for the patent, therefore it is printed on the first page of the patent document and recorded in a specific database. The International Patent Classification (IPC) was the first attempt to unify the national classifications made until 1968, and today is applied to patent publications of almost all jurisdictions worldwide. The IPC is regularly revised to include new technologies, or to separate existing classification units into several sub-units with a more narrowly defined scope. Classification symbols are therefore usually accompanied by version indicators. Despite this sorting United States and European patents have different code, so in 2010 the USPTO (United States Patents and Trademarks Office) and EPO (European Patent Office), respectively the patent offices of the United States and in Europe, decided to unify in the Cooperative Patent Classification (CPC).

2.7 Alternatives to patent

The patent is a mechanism of protection of industrial property with a very high appropriation regime and a high initial and maintenance cost, the main issue of the patent is the fact that the invention is revealed. An alternative, based on appropriateness schemes, is *industrial secrecy* and it is advisable if innovation is not visible from the finished product and is unlike the ideally eternal patent.

An opposite mechanism of protection is the *public defensive*: if someone does not want to patent an invention, but at the same time she wants to prevent someone else doing it, she can divulge certain information, for example, on an exchange platform or in a specialized article. In this way the invention will belong to the state of the art, i.e., it will be in the public domain and can no longer be patented. The inventor can continue using it, but she will not be able to prevent his competitors from doing the same.

Chapter 3

3 The Methodology

The peculiarity of this thesis is the database, as it is created from scratch through the search for each individual inventor. A meticulous and precise research for all 424 inventors analyzed, of which only 251 are found in full.

In this chapter it will be analyzed the structure of Database: its composition detailed from EPO Versus Prolific Inventor, the macro-section realized, the methods of realization, the issues that occurred during the research and the main sources of information. Therefore, the dynamic development of the tables will be discussed: why firstly they were separated and why then they had been united and in which way. The link between the database and the *leitmotif* of the present work will also be argued: the main education and carrier of both categories of inventors.

Finally, in the next chapter the evolution of the database as a function of statistical analysis and regression will also be discussed.

3.1 Database Creation

3.1.1 How to build a database?

First of all, it is defined the modern concept of database, according with Umesh Maheshwari [17] as "*in the most general sense, is an organized collection of data. More specifically, a database is an electronic system that allows data to be easily accessed, manipulated and updated. In other words, a database is used by an organization as a method of storing, managing and retrieving information.*"

During the Brainstorming phase, it is necessary do some research about "how it is made a Database" and "How it is built". Then the data are organized into macro parts, that are explained below, which, in turn, are divided in more detailed sections with the aim of being able to easily search for data once the work is over. The tables are created with the support of the Excel program[1], with the support of which it is then possible to perform a statistical analysis smoothly.

3.1.2 Taxonomy of the database

Firstly, two split databases are created because they contained partially different information: *EPO Database* and *Prolific Inventor Database*. EPO Database contains all the information of the inventors participating in the EPO Award from 2010 to 2014 for a total of 140 inventors analyzed. On the other hand, the second database contains those we have called "prolific inventors", namely inventors who obtains a variable, but significant number of patents taken from a database containing the 'TOP 5000' according to an EPO study between 2006 and 2018. From this last one a sample of 287 inventors is taken and examined with a random selection from the original 5000 collection. The two cases of the two databases (the main EPO awarded inventors, and the second containing the star inventors) are similar: they coincide with the macro categories Education and Career, but differ from the moment it does not appear in the first the address of the participants and the link to

[1] *Microsoft Excel for Mac, 16.20 version*

google patent used to identify them, but in the second one completely lacks the macro category that shows all the info of the EPO award.

The Prolific inventors are those who didn't became star scientists and, as explained in Chapter 2, often they have the tendency to patent inventions of secondary commercial and economical importance, however, having the characteristics of patentability.

This feature makes them generally less known than the star inventors; therefore, it was quite difficult to find data concerning the prolific inventors. At the point that on a sample of 287 subjects analyzed, only 114 can be considered complete. Since only 40% of the detected samples are found, it is considered appropriate for the statistical and regression analysis to use a "clean" version of the database where the incomplete data are eliminated. The criterion used for the number of inventors in the data collection is a progressive advancement of the researches on the prolific inventor until a comparable number (therefore equal or with little discard) is identified with respect to the sample of EPO inventors (5 years analyzed and 140 inventors).

3.2 General considerations on the structure and dynamic evolution of the database

The corpus initially conceived for the work undergoes many changes during the course of building the database, because of the problems of availability of information, and the operations for which the database is designed. It is therefore necessary to build multiple versions of the same database, of which we try to trace a line of general evolution and the main reasons.

3.2.1 The issue of age estimation

The first problem, in chronological order, revealed the lack of information on the age and/or date of birth of the inventors, preponderant as regards the participants of the EPO prize, total as regards the prolific inventors. Firstly, it is thought to use the Azure Face

APIs[2] [19] to estimate the age of the inventors but given the difficulty in dating the photos it is not a plausible strategy. Therefore, an assumed year of birth is added to the *Date of birth* column, in which a formula is implemented. By means of the latter formula, if the authentic data is found, that is reported, otherwise the bachelor's degree is assumed to be taken at 22 years old, assuming the average age for a three-year degree. If even the bachelor's degree does not reveal to be available then we go to consider the year of master's degree, which is assumed to be taken at 24. Finally, if any of this information is available, it is marked as missing information. It is necessary to specify how information is arbitrarily decided in a coherent way in order to make clear and for a correct use of the database. With regard to the specific degree durations, such as Law-school, Medicine, Psychology and Veterinary, the degree wasn't presented as a Bachelor, but as master's degree. However, the hypothetical student in question has not only attended the two years of a master's degree but a longer training path, which varies from five to seven years. The latter is treated according to the law in force in Europe, but there are exceptions concerning degrees that are certainly not 3 years long. For example, there is the Bachelor of Medicine, that it is not an actual degree in Medicine at the end: in each country there are different versions of the same degree, for the sake of this work we have made the assumption that all the degrees are treated following the European laws. Therefore, it was decided to complete both the bachelor column and the 'Master' column with a Boolean value completing with the same date of graduation and using the master as an indicator of age.

3.2.2 The Homonymy and the Triangulation mechanism

An easily predictable obstacle is the homonyms' analysis, made even harder by the lack of an accurate date of birth information. This issue has been solved by means of a Triangulation algorithm: in order to verify the reliability of the information, the research is carried out by consulting multiple sources. However, sometimes, subjects' information is completely unavailable, it is therefore essential to look for information in the original

2 https://azure.microsoft.com/en-us/services/cognitive-services/face/

language, using the support of translators. Comparing the multiple sources, it is possible to gather the information of many inventors; where it is not possible, and the data is hypothetical it is indicated as missing.

During the research of the inventors the difficulties of dealing with homonyms has caused the presence of a small percentage of missing data, especially regarding South American and Iberian inventors, where the presence of multiple names makes the research harder.

3.2.3 Degree Specialties' Categorization

The database columns increase when, in view of a future statistical study, university degrees, intended as fields, and universities as a physical structure, are a plethora and very different from each other. Therefore, it is thought for universities to group them by country of origin in order to obtain a sort of ranking, while for fields to group them into 5 generic macro-categories:

- *STEM:* acronym for *Science, Technology, Engineering and Mathematics*, includes scientific-technological courses of study (e.g., engineering, mathematics, physics, ...)
- *Life Science (or biological): it* includes the branches of science that involve the study of organisms and life (e.g. biology, pharmacy, medicine, genetics, etc.)
- *ICT:* acronym for *Information and Communications Technology,* it includes the fields of study of transmission, reception and communication of data. It can be divided into two branches: information technology and telecommunications, some examples computer engineering, electronics, optics.
- *Others:* fields that are not relevant to innovation and technology, mainly humanistic faculties such as law and philosophy.

3.2.4 Modifications of the template database according to the availability of data

At the beginning of the project, we set ourselves the target of collecting information about the career and the academic path of the EPO inventors by splitting the "Before Prize" information from the "After Prize" ones. However, the lack of data retrieved from the sources force us to change the database: international experiences of study, work, and experience as entrepreneurs, or start-upper become a Boolean variable independent from information regarding the EPO award. Moreover, due to lack of information the columns which reported the "years of work to win / participation in the prize", "and "last work Before Prize" are eliminated. The elimination of these columns is due to the difficulty in finding such information, and to the doubtful reliability of the data obtained; secondly to the fact that in order to compare with the control sample they do not seem to mean. Sometimes it is chosen to add ad-hoc columns, for example to distinguish the degree of medicine, that is not usually a bachelor, but rather it is a single-cycle degree with a duration of six/seven years, and the specialist degrees that cannot be properly defined master's degree.

Additional columns are added during the construction of the database also to include all degrees of education, even those less frequent in the context of innovation: MBA and PhD students mainly. However, considering the numbers of the data collection, those who hold a PhD or MBA were very few, therefore it was considered appropriate for consistency not to eliminate these columns, as they were not ambiguous or incomplete, but given the overwhelming minority it was not considered significant to consider them in the statistics.

3.2.5 Standardized Sector for each field

In order to standardize this database for a possible statistical analysis, the sector in which the inventor operates is chosen from a predefined list divided into *Sectors,* each sector contains in turn the *Fields* as explained in Table 3-1.

Sector	Field	Sector	Field
Chemistry	Basic materials chemistry	Instruments	Analysis of biological materials
Chemistry	Biotechnology	Instruments	Control
Chemistry	Chemical engineering	Instruments	Measurement
Chemistry	Environmental technology	Instruments	Medical technology
Chemistry	Food chemistry	Instruments	Optics
Chemistry	Macromolecular chemistry, polymers	Mechanical engineering	Engines, pumps, turbines
Chemistry	Materials, metallurgy	Mechanical engineering	Handling
Chemistry	Micro-structural and nano-technology	Mechanical engineering	Machine tools
Chemistry	Organic fine chemistry	Mechanical engineering	Mechanical elements

Chemistry	Pharmaceuticals	Mechanical engineering	Other special machines
Chemistry	Surface technology, coating	Mechanical engineering	Textile and paper machines
Electrical engineering	Audio-visual technology	Mechanical engineering	Thermal processes and apparatus
Electrical engineering	Basic communication processes	Mechanical engineering	Transport
Electrical engineering	Computer technology	Other fields	Civil engineering
Electrical engineering	Digital communication	Other fields	Furniture, games
Electrical engineering	Electrical machinery, apparatus, energy	Other fields	Other consumer goods
Electrical engineering	IT methods for management	Other fields	Materials
Electrical engineering	Telecommunications	Electrical engineering	Semiconductors

Table 3-1 - Sectors and related fields used in databases

3.3 Main sources employed in the accomplishment of the database

To find information about education and carrier, several search engines and social networks were used.

The name of the inventor was searched on Google for a first selection of information and for the appearance of scientific sites; in case of candidates of EPO award, it has been discovered with the information in the EPO award website, in case of prolific inventors limited to the address and the patent number. The data have been found exclusively from secure and trusted sitography, which is limited to ordinary or scientific search engines such as Google Scholar, Scopus, Google Patent, and major social networks such as LinkedIn and secondly also Facebook and Google+.

After researching numerous inventors mainly belonging to the prolific inventor category, the peculiarities linked to the subject's nationality have come to light: firstly, the totality of subjects of South-Eastern Asia origin, who have not moved to America or Europe or have in any way immediately indirectly influences of globalization, do not have western social media, but rather a local one, in local language. Therefore, unless their discovery is of primary importance it is impossible to extrapolate the data. In the second place, it can also be noticed that, considering the age of the subjects, those of the American culture the date of birth never appears in the *resume*, sometimes they also hide the graduation year so that it is not possible to track back the actual age. The American *resume,* in fact, does not have any personal reference such as photos, hobbies or age in order to prevent discrimination based on those data.

As it will be seen in Chapter IV, which is broadly based on statistics, an overwhelming majority of German authors emerges in the control sample. In the research of the German inventors there was a certain difficulty due, on the one hand the presence of complex composed names and the occurrence of numerous homonyms, on the other to the intrinsic conservativity and the love for the mother tongue that often leads them to express themselves on social media exclusively in German. For these reasons it was difficult to extrapolate the data useful for the project, therefore by doing a more in-depth research it was necessary to subscribe to XING [20], a social network aimed at the business (a kind of

German LinkedIn) which in recent years is taking foot in Europe with a network of 7 million users. Thanks to this tool it was possible to extrapolate a data collection of greater importance because often the data shared on this social network were not made public elsewhere. Therefore, the most used site to find information is undoubtedly *LinkedIn*, using a triangulation mechanism through *Google Patent,* comparing the information derived from the patent to be sure do not have to face misunderstandings due to the homonymy. However, it is also possible to mention other frequently recurring websites in the data collection: *Wikipedia* for example is one of the most famous online encyclopedias where, however, information can only be found about the most famous inventors, who have left a trace in the science of technology. A reason of which *Wikipedia* is recurrent as a source of data in the database of star scientists and absent in the database of the control sample.

As already mentioned in the previous paragraph, the social network for business *XING* has also been useful for innovators of the German people. It was also possible to enrich the database through information drawn from scientific sitography, scientific articles, papers dealing with the EPO award or, in general, innovation and from scientific blogs such as the most popular *ResearchGate[24]*, very useful for our research. As for the innovators who are also University Professors, or in any case those who work in research centers, often the sites of the universities provide public information, or, sometimes, even the entire Curriculum Vitae; they are therefore unlikely to have the incomplete data collection. As for those who have their own company or start-up (especially for the SMEs category of the EPO award), often the small-medium business sit-even counts the biography of the founders or the most important collaborators.

3.4 The two databases in detail: similarities and differences

In this paragraph a careful analysis of analogies and differences of the two databases follows, divided into three sub-paragraphs: the first analyzes the common parts, in the second the peculiarities of the EPO Database and in the third the Database of the control sample.

After having illustrated the methodology adopted for the construction of the database, it is necessary to concentrate on the structure of the two databases in the details. The following paragraphs are treated analytically because the database is not provided as a document in the appendix due to its prominent extension, so it follows a detailed description of both, with some extracted statements.[3]

3.4.1 Common features of the two databases

The common characteristics include the two macro categories: Education and Biographical Data.

As regards the *Biographical data* macro-section, it includes four sub-categories:

- Gender
- Year of birth
- Assumed birth year
- Nationality

Regarding the category *"Assumed year of birth"* for those whose year of birth is not documented, it is estimated using the algorithm explained in paragraph 3.2.1. Regarding the "*Nationality*", it differs from "*Country*" related to the macro-category "*About the prize*" because the first refers to the country of origin of the inventor, while the second refers to the country in which it is made the invention, then properly where the inventor works (often the two categories express the same information).

The education sub-category contains all the degrees of education achievable starting from the first level degree (called bachelor's degree), the second level master's degree, and finally the PhD (acronym for Doctor of Philosophy). For each of these degrees of education the field, the year of graduation and the name of the university attended are indicated. In order

3 It should in any case be specified that the database will be available as an external attachment to the thesis

to standardize the *Fields* are broken down into four main categories, as explained in detail in paragraph 3.2.3, in order to use a language as standard as possible and to facilitate the normalization. Given that once the statistics on the universities attended by 251 inventors selected (of which 114 prolific inventors and 137 star inventors) are completed, more than 80 different universities are surveyed, it is decided to group universities by Country and to recalculate the statistics. In addition to these two main categories in both databases there is the link of the main sources used in the research of the inventors. In paragraph 3.3 the topic of the sources was dealt with details.

3.4.2 Peculiarities of EPO database

This paragraph describes the characteristics of the EPO database, that is to say the main sample of this work, it is presented the structure that we would then use to incorporate and record the main information about inventors which differentiate this database from that of the prolific inventor.

The analysis focuses on the *Winners* and *Finalists* of the EPO award since 2010, this year is chosen because before there are different rules and conditions in the Award, so to ensure uniformity and consistency in the data concerning the prize. After which the candidates of the years 2011-2014 are also analyzed in the same way.

Name (first Last)
URL webpage (source1: finalists)
URL webpage (source2: all candidates)
Year
Award category
Rank
Team size
Country
Company/Research Centre
Patent Number
Technical field / Sector
The invention in a nutshell
Societal benefit
Economic benefit
Gender
Date of birth
Year of birth (assumed)
Nationality
Bachelor (BA)
BA_year
BA_field
BA_Macrocategories
BA_University
BA_University_Nation
Master (MA)
MA_year
MA_field
MA_Macrocategories
MA_University
MA_University_Nation
Specialistic Degree
Year of speciaistic degree
Specialistic Degree University
Specialistic Degree field
PhD
PhD_year
PhD_field
PhD_Macrocategories
PhD_university
MBA
MBA_year
MBA_univ
Second Degree
Other Degrees
International work experience
International study experience
First experience as Entrepreneur / startupper / founder of companies
Number of years working AT PRIZE DATE
Industry experience (main sector of work)
Worked in a company
Worked in UNIVERSITY/Research centre
Last Job name
Last Company
Link_info

Figure 3-1 - EPO database columns

The Database consists of four Macro-Categories of information:

- About the prize
- Biographical Data
- Education
- Carrier

Biographical data and *Education* are explained in the paragraph above as identical for both databases; as far as the category *About the prize* is concerned, it refers only to the EPO database, whereas the *Career* category presents some subtle differences but present in both, as can be seen by comparing the two templates in Figure 3-1 and Figure 3-2.

About the prize is a macro-section that contains data that were easily found, mostly on the EPO website. [5] Nevertheless there were problems in obtaining information belonging to this section due to the dynamic evolution of the website and the presence (or absence) of particulars depending on the reference year of the award. Specifically, for 2010, 2011 no invention had the patent number on the EPO site, so it was necessary for all the invention finalists and winners of these years and part of the invention of 2012-2014 to go back to the patent number. As explained in the Chapter II, section 2.6, a patent may correspond to several IPC or PCP codes that refer to multiple versions: it was complicated to guarantee that the code found was the right one. The query has been implemented by Google Patents [18], a search engine by Google, which helps to find patent numbers searching with the inventor's extended name and a brief description of the invention by means of keywords. However, there may be many patent numbers that correspond to several versions or even different patents, from here a Triangulation algorithm has been applied. The patents provided by the EPO (and therefore corresponding to the IPC code dating back to the exact version and extrapolated directly from the EPO site [5]) are easily distinguishable because they have the prefix **EP-**, instead the others found through the mechanism described above have other prefixes, for example **US-**, if they are US patents, or **W-** if they are conferred by the World Intellectual Property Organization.

We can note that in this section are present a lot of descriptive information about the invention in order to understand the economic and societal impact of this and explain how it works. This typology of Data is not significative to the Regression Analysis or Statistical studies however it helps to understand which profile of the inventor to look for in order to avoid the homonyms. It becomes part of the triangulation mechanism explained in detail in the course of this chapter. On the other hand, in this section there are information used in Statistics and Regression Analysis like "Country", "Technical Field", "Award Category", "Team size".

Below is a list of all the items in this macro category in detail:

- URL of the webpage of all candidates of the category and of the finalists,
- award category,
- rank, (Winner or Finalist)
- team size,
- country,
- company/research center,
- patent number,
- technical field,
- the invention in a nutshell.
- societal benefit,
- economic benefit.

The *'Career'* category in the EPO database includes:

- International work experience (Works in more than 1 country);
- International study experience (Studies in more than 1 country);
- First experience as Entrepreneur / startupper / founder of companies;
- Number of years working AT PRIZE DATE;
- Industry experience (main sector of work);

- Worked in a company;
- Worked in UNIVERSITY/Research center;
- Last Job name;
- Last Company.

This category is aimed at portraying the professional profile of each inventor, starting from the experiences that have formed the inventor, until the last job (which if she is not retired coincides with the current work). Therefore, for each inventor the professional path and the career achievements are available.

3.5 Prolific Inventor Database

Unlike the EPO database, a series of Top 5000 inventors selected by the European patent office are used to create the prolific inventor database. On the basis of a heuristic method database is created: through the full name, the Google Patent link and sometimes the address by means a triangulation mechanism the subject is searched through a thorough search on search engines, social media, scientific articles.

The Prolific inventors, by definition of the literature they do not enjoy fame, so their traceability is discouraging. Out of 424 subjects analyzed we have complete information for only 114. As explained in the previous paragraphs the research of information about them creates more issues and is more lasting.

ID_PERSON
Name
Address
Country
Random_code
Birth Date
Year of Birth
Gendre
Nationality
Patent Number
Granted EP patents since 2006
Link Google Patents
Bachelor (BA)
BA_year
BA_field
BA_ Macrocategories
BA_University
BA_University_Nation
Specialistic Degree
Year of speciaistic degree
Specialistic Degree University
Specialist Degree_field
Master (MA)
MA_year
MA_field
MA_Macrocategories
MA_University
MA_University_Nation
PhD
PhD_year
PhD_field
PhD_Macrocategories
PhD_university
MBA
MBA_year
MBA_univ
Second Degree
...
International work experience
International study experience (Previous studies in more than 1 country)
Experience as Entrepreneur / startupper / founder of companies
Industry experience (main sector of work)
worked in a company
worked in UNIVERSITY/Research centre
Last Job name
company of last job name
Link_info

Figure 3-2 - Star Inventors database columns

As far as the control sample database is concerned, obviously the information in the "*About the prize*" category does not appear, but instead of this additional information appears

regarding the patent corresponding to the invention. Therefore, the macro-category scheme appears as follows:

- Biographical Data
- About the patent
- Education
- Carrier

As for the *Biographical data* and *Education* categories, they appear identical in both databases and are analyzed in detail in paragraph 3.4.1. There is only one peculiarity regarding the biographical data, for the prolific inventor database the address of the latter is also indicated. The Prolific Inventors' file provides some personal data including the full name and address and some information on patents such as the patent code, the number of patents obtained and the Google Patent link. As for the career section, it looks very similar to the EPO database one, without the information regarding the prize.

As already mentioned above, with regard to the prolific inventors, creating the database is more difficult, as it is not possible to rely on the information shared by them in the first person on social networks, coming from universities and companies in rare cases. Therefore, it is for this reason that to get to the complete profile of 114 inventors, much more are analyzed, to come to a ratio of inventors found / inventors analyzed equal to 27.12%. As argued by most of the literature summarized in Chapter II, prolific inventors are unknown to the public precisely because their inventions while being patentable and although often numerous, are not considered commercially nor scientifically attractive to become part of the history of the invention and be rewarded by any kind of recognition. Therefore, with the following two Chapters of the thesis based on the statistical analysis of the results obtained from the database, we aim to understand if there are substantial differences concerning the academic preparation, life experiences and career that distinguish these two groups of inventors.

Chapter 4

4 Statistical Analysis

This chapter reports the first results obtained by analyzing in detail the database with a statistical analysis with spreadsheet. In this chapter the main sample and the control sample are inspected in order to distinguish the Star Inventors from the control sample of the prolific inventors according to their education and career.

The statistics are carried out using mainly Microsoft "Excel" in a first analysis and different nuances of the two main themes "Education" and "career" are examined. This chapter is divided into three sections, which are the main categories of statistics: Biographical data, Education and Career.

In order to avoid ambiguity, it is necessary to specify that in this chapter abbreviations are used to indicate the two samples examined: the winning / finalists' inventors of the EPO award are referred to as *EPO* sample while the prolific inventors as sample *NO EPO*. This abbreviation is necessary to mark concisely tables and graphs. Therefore, also abbreviations such as *MA* for master's degree and *BA* for bachelor's degree are frequently used. It is also necessary to clarify that the missing data count is 2% for prolific inventors and 3% for EPO inventors in the latest version of the database. For Prolific inventors the percentage rate of missing data is so lower, since there was a selection because the percentage of missing data was very high as explained in Chapter 3.

4.1 EPO and Prolific Inventor Statistics

The first data to be specified here is the samples analyzed and those found: with regards to the data of the EPO prize, the finalists and winners of the 5 categories of the award were analyzed in the years 2010-2014 with a total of 137 inventors identified as can be seen in Table 4-1.

EPO Winners/Finalist	Prolific Inventors
137	114

Table 4-1 - Total sample of analyzed Inventors

As for the prolific inventors, that is to say the control sample, 251 subjects are examined and the comparison for subjects with a complete dataset amounts to 114. Therefore, for the control sample of 251 inventors examined for only about 45%, the complete data set is found.

As expected, the research of the EPO inventors is easier since the winning / participation in the EPO prize has increased the reputation of the aforementioned inventors, by interviews, biographical and scientific articles about the invention on them. As for the Prolific Inventors, which tend not to have much scientific relevance, the information available basically reflects what they have on social profiles (in the first line LinkedIn, the first social network for work in Europe), therefore they exclude some categories: inventors prolific older people who are not familiar with these recent means or those whose culture has particular social network mainly the oriental cultures of China and Japan.

4.2 Biographical statistics

Regarding the biographical data the gender, the nationality and the year of birth are searched. The data that creates more problems in terms of availability is the date of birth which is, in most case, estimated starting from the bachelor or master's degree, as explained in Chapter III. In some cultures, age, hobbies and other biographical specifications are considered extremely personal and, therefore, should not be disclosed to the public as a possible reason for discrimination, for example when exposed in a Curriculum Vitae or in a Social Network.

4.2.1 Consideration on the gender and nationality of the two samples

In a very predictable way, it is possible to find an overwhelming majority of the male gender that manifests itself both in the sample of the Inventors taking part in the EPO award and in the control sample, as shown in Table 4-2. In the EPO sample, however, there is a slight superiority regarding the female presence compared to the other sample.

Gender	EPO	NO EPO

Female	8,03%	7,08%
Male	91,97%	92,92%

Table 4-2 - Gender distribution in the two control samples

Analyzing the statistics on nationality, some very low percentage rates are noted; therefore, it is established to report in the following graphs only the ten most common nationalities for both samples.

TOP 10 Nationality	
Nationality	*NO EPO*
German	**45,45%**
Swedish	9,09%
US	8,18%
French	5,45%
Swiss	4,55%
Chinese	3,64%
Italian	3,64%

Turkish	3,64%
Austrian	2,73%
British	2,73%

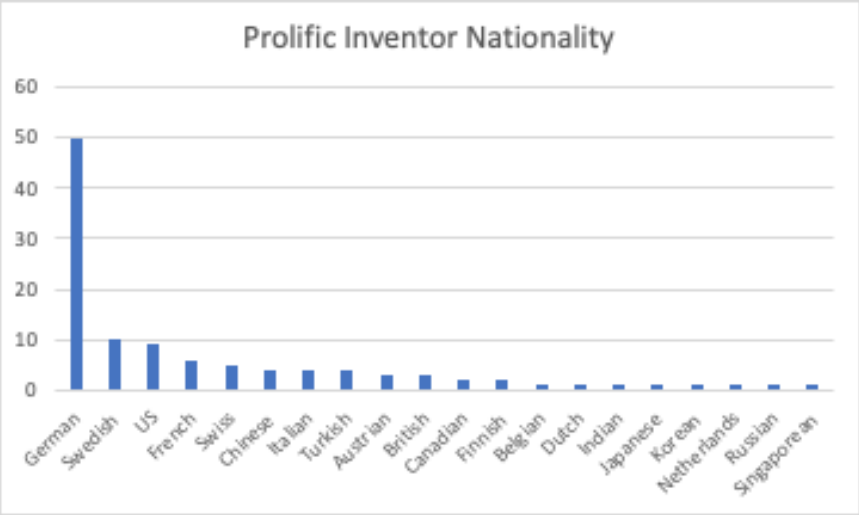

Table 4-3 - Top 10 Nationalities of Prolific Inventor

Among the nationalities of the prolific inventors stands out the German people with 45.45% and it is natural to wonder why this overwhelming majority. To justify it, it is first of all possible to make a consideration of the European population: out of about 500 million of the European population, 80 million are of German origin. The German people is the largest in Europe, so this is certainly a factor to be taken into consideration which, however, does not explain such a preponderant percentage. In addition, it is necessary to take into account the investment of Germany in Research and Development (R&D): very important compared to other European countries. In fact, in 2016, Italy invested € 21.6 billion in R&D, while Germany € 92 billion, about 4 times as much. [21] Finally, it must be considered, in agreement with Gambardella's article [22], that German inventors have a clear idea of the economic value of their patents and are, on average, the best-rewarded inventors for their right to invention. Also, in accordance with *The value of European patents*

by Alfonso Gambardella [22], the sample average of VALUEM[4] for German patents is € 5.6 million, an extremely high value. For all these reasons, inventors in Germany are more likely to patent than other European countries and, therefore, this country holds the highest number of European patents produced every year. Finally, it is possible to notice a strong percentage of northern and central Europe (Italy, Sweden, Switzerland) and the United States.

TOP 10 Nationality	
Nationality	*EPO*
French	**15,27%**
German	13,74%
US	9,92%
Belgian	6,11%
Danish	6,11%
Swedish	6,11%
British	5,34%

[4] the midpoint of the value intervals

Italian	5,34%
Spanish	5,34%
Australian	4,58%

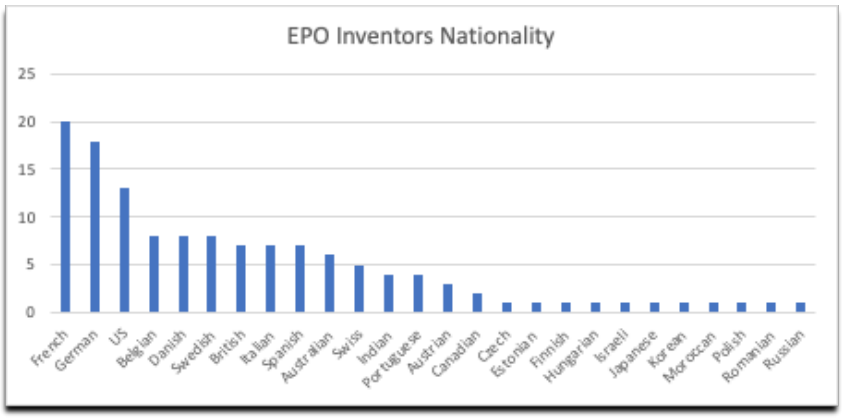

Table 4-4 - Top 10 Nationalities of EPO Inventors

As regards the nationality of the EPO participants, no major percentage is observed, as in the case of the control sample patents, even if French and German nationalities dominate with a negligible difference of 2 percentage points. From the comparison of the Table 4-3 and Table 4-4 it can be deduced that, although the German produces many patents, many of these do not have a commercial success or a particular scientific value. Therefore, in conclusion, Germany is the second most important in the EPO participants, however, with a much lower percentage compared to the control sample.

In order to understand whether the statistics concerning the nationality of the inventors belonging to the two samples generally reflect the nationalities of the inventors on a much larger sample (that is extrapolated from the original EPO statistics on the applications of the annual patents), the data collection of this thesis with the 2015 annual report of the European patent office.[26]

Origin	2014 Patents	2013 Patents	% change 2014 vs. 2013	Share in total applications 2014
EPO	75.180	73.575	2,2%	49%
Germany	25621	26510	-3,4%	17%
France	10557	9835	7,3%	7%
Netherlands	6844	5852	17,0%	4%
Switzerland	6833	6742	1,3%	4%
United Kingdom	4687	4587	2,2%	3%
Sweden	3837	3674	4,4%	3%
Italy	3613	3706	-2,5%	2%
Other EPO member states	13.188	12.669	4,1%	9%
United States	36.491	34.011	7,3%	24%
Japan	22.018	22.405	-1,7%	15%

Table 4-5 - 2014 Annual Report of the European Patent Office about European patent applications per country of origin

European patent applications per country of origin

Nearly half (49%) of all European patent applications in 2014 came from the EPO member states, with European firms filing 2.2% more European patent applications than in the previous year. The leading countries of origin overall were the US, Germany, Japan, France and the Netherlands.

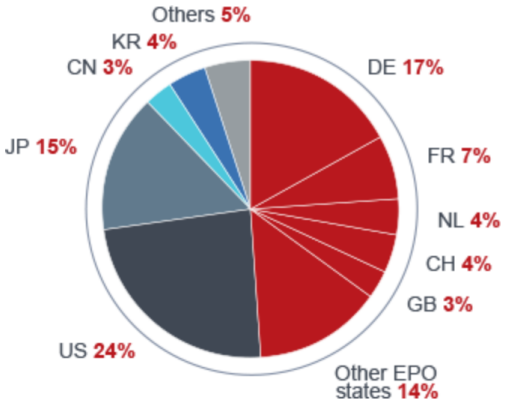

Figure 4-1 - Annual report 2014, European patent applications per country of origin

Figure 4-1 and Table 4-5 show that the highest percentage of patenting for European countries are in Germany and this can be matched by the EPO sample and the Prolific Inventor sample. The second nationality which registers more patents is the France: it matches the EPO sample. In the Top 10 nationality analyzed in this thesis and in the top 10 nationality drawn up by the EPO also appear the same country list: Italy, Switzerland, UK and United States. Therefore, this shows a consistency in the research carried out, despite the diversity of the sample size. However, there is a serious discrepancy: in the count of the EPO patents a significant presence of Japan appears that does not appear in any of the two samples and it is possible that these in the samples of the thesis are part of the missing data, discarded in the sample of the Prolific inventors and stored in the EPO sample.

4.2.2 *Considerations about the age estimate of the two data set analyzed*

As said before, date of birth, and consequently the age, is one of the most difficult data to be found. Out of 214 subjects, 47 have been found by research on social networks and

scientific articles and are referred to as "Real", while the remaining subjects' age have been estimated by means of the graduation year, using the Bachelor's year and, if not available, the Master's year; they are labeled as "From BA" if the estimate come from the Bachelor's and "From MA" if the estimate come from Master's degree. Finally, for only one subject, the PhD is used as an indicator of presumed age, with the wording in the database "From PhD".

Age	EPO	NO EPO
Under 30	0,00%	2,20%
31-40	3,88%	18,68%
41-50	17,48%	**49,45%**
51-65	**39,81%**	25,27%
Over 65	38,83%	4,40%

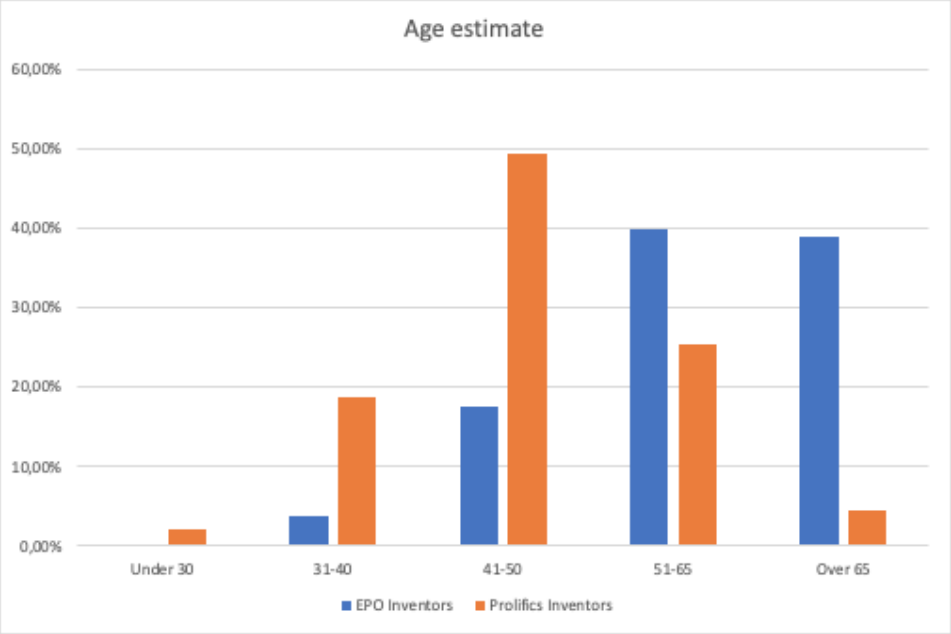

Table 4-6 - Estimating the age of the inventors participating in the EPO prize compared to the prolific inventors

In this way information about age, which is an important control factor in a statistic based on education and career, contains a low percentage of missing, both in the main and the control sample, taking into consideration the difficulties in finding the above-mentioned data.

As it is possible to see in Table 4-6 five age groups of about 10 years each are designed. Therefore, the estimate of the age starting from the graduation year is in itself an inaccurate data as it depends on contingent factors; however, it is acceptable inserted in the purpose for which it is used. In an unexpected way, it has been found out from Table 4-6 that the inventors who win or participate in the EPO award are older than the prolific inventors. Indeed, while 49.45% of prolific inventors fall within the age group between 40 and 50 years, only 17.48% of the participating inventors or winners of the EPO award fall into the same category. As well as in the category between 31 and 40 years the prolific inventors amount to 18.68% while the EPO to 3.88%. In the two oldest categories ranging from 51 to "over 65" the total percentage of participants in the EPO prize is 78.64% while for prolific inventors it amounts to 29.67%. This comparison therefore reveals a very significant difference to which we have tried to provide a plausible explanation that follows. Firstly, it is possible to make a consideration on the origin of the data: the chosen sample

of the analyzed Star Scientists was extrapolated from the participants of the EPO prize between 2010 and 2014, while the Prolific Inventors come from a list drawn up by the European patent office in 2016. Therefore, a possible reason for the difference in age can already be found in the data source. Moreover, as regards the Star Scientists, the category of the EPO "Lifetime Achievement" "*[...] (it) honors the long-term contribution of an individual European inventor whose dedication and tireless efforts*". The caption just mentioned describes the category of the EPO award and it makes clear how recognition is given to subjects who have dedicated their whole existence to a certain discovery, which has revolutionized a sector. Therefore, analyzing the aforementioned database category only one inventor was born in 1957, one subject belongs to the age group between 51 and 65 years while all the other winners / finalists of this category today are over 65 years old, some of them are 80 years, others passed away since years. This is because the category rewards discoveries that have revolutionized science years later as it takes time to clearly determine the scientific or technological impact of an innovation that it is a drug for a specific disease that is cured or LCD technology. Finally, we can consider the thesis that many prolific, yet young inventors can in the near future become Star Scientists through an invention that changes the fates of science.

4.3 Statistics concerning the level of Education

The purpose of this work is to determine how a person's education and career can affect whether or not it becomes a Star Scientist. Therefore, while biographical statistics are factors of control[5], the person's education and career (subject of this paragraph and the following) are the variables of interest[6] of this study.

[5] An effective control variable is one that, if included in the regression, makes the error condition unrelated to the variable of interest. Keeping the control variable (s) constant, the variable of interest is randomly assigned "as is".

[6] Also called variable dependent or explained

The first level of education that is examined is a first level academic qualification which, depending on the country, may last three or four years and is conventionally called bachelor's degree. This title indicates both the BA, abbreviation for Bachelor of Arts, and the BSc., abbreviation for Bachelor of Science. As it was easy to expect a very high percentage of inventors participating in the EPO award and does not hold the Bachelor's Degree, as it is possible to observe in Table 4-7 and Table 4-8

	EPO	NO EPO
Has BA	87,59%	94,69%
Has not BA	12,41%	5,31%

Table 4-7 - Percentage of Inventors who hold bachelor's degree

It is necessary to specify that in the calculation of those who do not hold the bachelor's degree there is also included that small percentage of missing data, as explained at the beginning of this chapter. Therefore, taking into account the missing data, in the case of the Prolific inventors it is almost all of the percentage of those who hold the bachelor, while in the case of participants of the EPO award not exactly. There are, in fact, sporadic cases in which innovation evidently does not depend on the level of education because although they do not hold any academic qualification they have come to win / or as finalists of the EPO award. In any case, they are usually very old subjects.

BA Macro Categories	EPO	NO EPO
ICT	**34,48%**	30,69%

Life Science	30,17%	**31,68%**
STEM	33,62%	27,72%
OTHERS	1,72%	9,90%

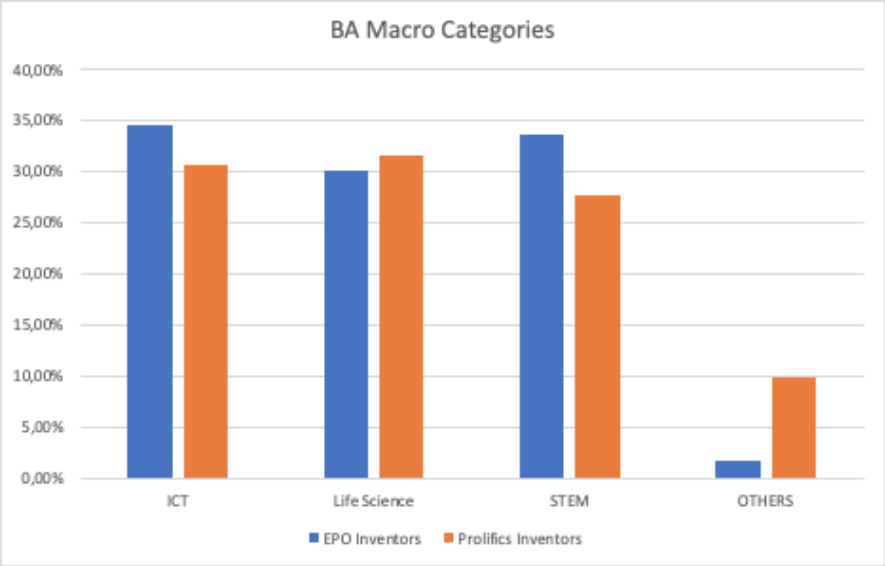

Table 4-8 - BA Macro Categories of EPO inventors

How it is possible to observe in the Table 4-8, there is not a predominant field in which the winners of the EPO award get a Bachelor: there is a slight prevalence in the ICT sector, followed by a single subject of difference from the STEM sector. Also, regarding the control sample no net prevalence of the sector is established, even if the most recurrent field in which the prolific inventors obtain the Bachelor is the *Life Science* category.

The second level of education taken into consideration is the master's degree; it normally requires previous studies at the bachelor's level, either as a separate title or as part of an integrated degree course. It should be noted that the single-cycle degrees (mostly present in Europe) such as Law, Medicine, Veterinary and Pharmacy, have been treated as a

master's degrees, because, according to the country, their duration is equal or greater than five years. There are many types of MA, depending on the constitution of the state and the branch undertaken; in this dissertation we do not deal with dividing the types in detail. Compared to the first level degree, the MA has lower achievement rates in both the examined samples: they remain significantly higher, especially for the prolific inventors, as is noted in Table 4-9. In the percentage rate in which the subjects do not hold the master's degree the missing data is counted. The study therefore shows that the prolific inventors are statistically more likely to obtain the MA. A hypothetical explanation could be that a certain type of didactic training is needed to patent a lot; however, to patent a brilliant, marketable and scientifically valid invention, a high level of education is not a necessary condition, but several contingent factors such as creativity, genius and timing act in this process.

	EPO	NO EPO
Has MA	56,93%	70,80%
Has not MA	43,07%	29,20%

Table 4-9 - Percentage of Inventors who hold master's degree

Moreover, analyzing the macro categories of the MA, a predominance of the ICT field is found in both samples, followed by the STEM field. Percentage rates appear quite similar in the two groups as it is possible to note in Table 4-10.

MA Macro Categories	EPO	NO EPO

ICT	**38,66%**	**39,48%**
Life Science	22,67%	25,00%
OTHERS	4,00%	7,89%
STEM	34,67%	27,63%

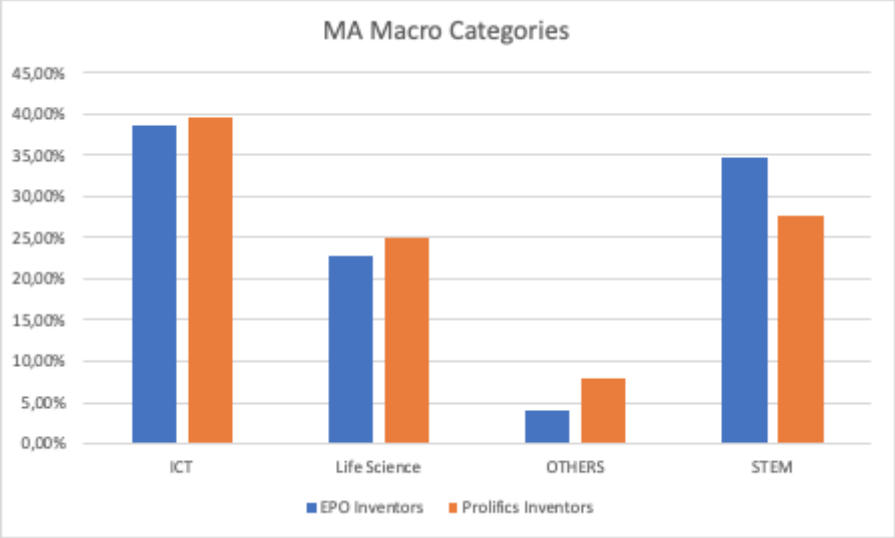

Table 4-10- MA Macro Categories of EPO inventors

The last degree of education examined is the PhD, acronym for Doctor of Philosophy and represents the highest degree of university education obtainable. Therefore, to carry out a survey on two different groups of subjects in which one patents successful inventions and the other is not the same as asking how education affects the success of an inventor.

	EPO	NO EPO
Has PhD	51,09%	48,67%

No	48,91%	51,33%

Table 4-11 - Percentage of Inventors who hold master's degree

Table 4-11 shows that more than half of the sample containing the winning or EPO prize-winning inventors holds a PhD, while less than half of the sample containing the prolific inventors holds it. Therefore, there is a difference of about 2.42 percentage points between the two groups. During the next chapter, it will be checked whether the difference is statistically significant or not through a multivariate analysis.

PhD Macro Categories	EPO	NO EPO
ICT	25,71%	25,93%
Life Science	**34,29%**	**44,44%**
OTHERS	5,71%	5,56%
STEM	**34,29%**	24,07%

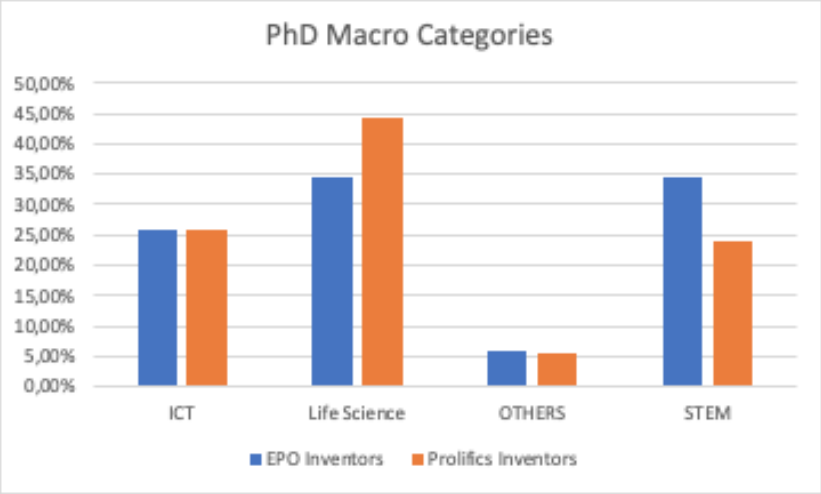

Table 4-12 - PhD Macro Categories of EPO inventors

Table 4-12 shows that for EPO sample there are two equally predominant Macro Categories: *Life Science* and *STEM,* while for the other sample there is a clear prevalence of the *Life Science* category. From the comparison between the two samples it is possible to notice an identical percentage rate for *ICT*, a percentage rate of *OTHERS* almost 6% slightly higher for the Prolific Inventors and a different percentage distribution for the remaining Life Science categories and *STEM*.

The statistical analysis of the Education samples did not reveal any significant discrepancies: as regards the second level degree there is a higher diffusion for the sample of the Prolific Inventor, while for the diffusion of the PhD it is more marked in the sample EPO. From the analysis of the Macro Categories of degree does not emerge any peculiarity or even from the distribution of the various degrees of education in the three main fields (*STEM, ICT, Life Science*).

4.4 Statistics about Career

The career of an inventor is the professional path of the individual: the career includes the experiences of education and work abroad, if the individual has or not obtained an increase in salary and level of company classification, if he has initiated a start-up or a company. We

will analyze the experiences abroad of the inventors, the entrepreneurial activity and the professional placement. The latter includes information, such as whether the subject worked in a company or in a research center and / or in a university, and the field of study / work. On the other hand, the EPO sample doesn't not analyze the job position before and after the award; salary increase and other information, interesting but objectively difficult to find in curriculum and social media.

4.4.1 Statistics about work placement

Firstly, the two samples are analyzed in the light of the work placement: whether they worked in a company, in research centers, whether they are entrepreneurs or both.

Worked in a company	EPO	NO EPO
Yes	78,10%	98,23%
No	21,90%	1,77%

Table 4-13 - Percentage of Inventors who work/worked in a Company

Worked in a University	EPO	NO EPO
Yes	58,39%	41,59%
No	41,61%	58,41%

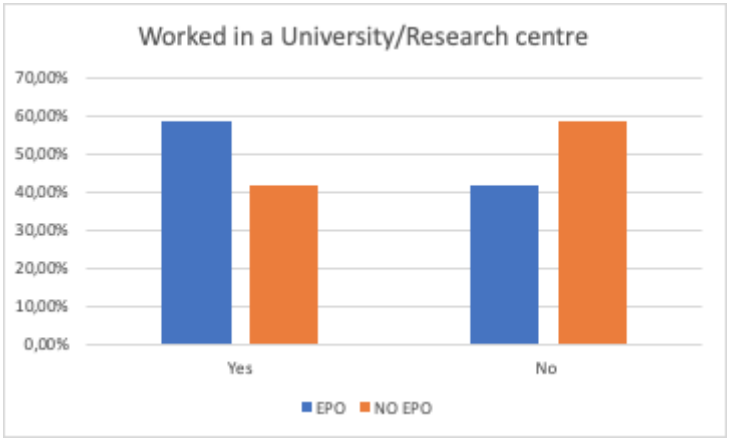

Table 4-14 - Percentage of Inventors who work/worked in Research center or University

Company & University	EPO	NO EPO
Worked in both	44,53%	39,82%
Just one/No one	55,47%	60,18%

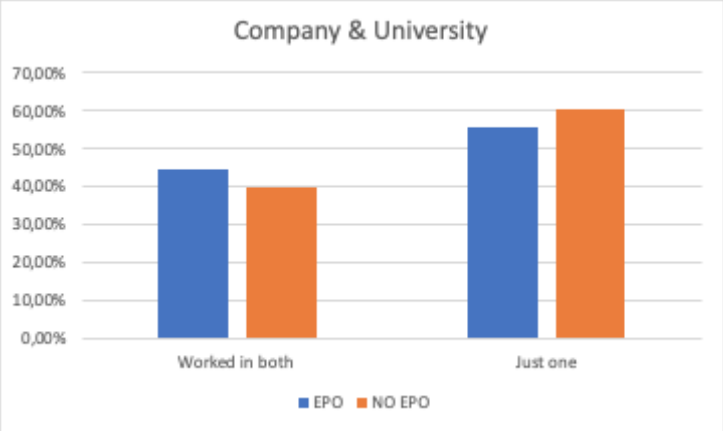

Table 4-15 - Percentage of Inventors who work/worked in both

Although the percentage of inventors working in universities is very high in both samples, it is possible to find a significant difference between the two: almost all the prolific inventors (98,23%) and against 78.10% of the inventors belonging to the EPO sample, as it is possible to observe in Table 4-13. While inventors working at a university or research center are more numerous in the EPO sample (58.39%) than the other sample (41.59%) as it can be noticed from the comparison of Table 4-13 and Table 4-14. The fact that the winners / finalists of the EPO award work more in research centers / universities than the control sample could suggest that because research centers and universities have the means and resources aimed at creating scientific knowledge. R&D is their core competence, while for companies it is a department counted among others and hardly constitutes the core business. In addition, companies focus research and development on a range of flagship products, while to university researchers is usually given more freedom in the field of research. However, there is a certain flexibility in changing from one company to a research center in both samples, as a high percentage has worked in both places: respectively 44.53% in the EPO sample and 39.82% in the sample control.

First experience as Entrepreneur	EPO	NO EPO

Yes	27,74%	6,19%
No	72,26%	93,81%

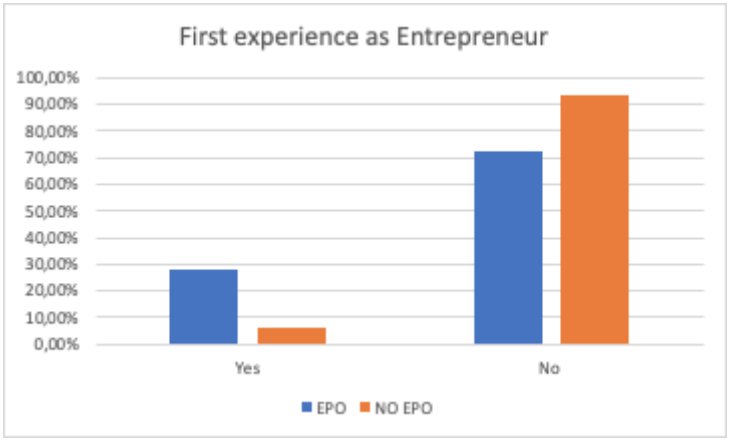

Table 4-16 - Percentage of inventors in the analyzed samples with experience of entrepreneurs and startuppers

Furthermore, in the EPO sample there is a significantly higher percentage of inventors who have experience as entrepreneurs or startuppers during their career. Interpreting the results this could be more a consequence: those who patent a commercially successful invention statistically are more likely to find a company or to find a start-up because the company's core business could be based on that patented invention. For example, an informatic realizes an app for smartphones to be able to pay via mobile, decides to patent it, the app is successful, and the founder decides to found a startup to manage the new core business and update it.

4.4.2 Statistics about jobs' field

Information about the field in which the inventors work is collected for both samples; in order to standardize and make changes, standard fields are used as explained in Table 3-1. For the sake of this work the ten technical fields in which more patents are deposited are considered. From the comparison between the two groups it is possible to establish in which sectors many patents are registered without scientific validity and in which many patents are statically important for the history of the science and innovation.

Top 10 Main field of Job	EPO
Digital Communications	**18,85%**
Medical technology	13,93%
Computer technology	11,48%
Environmental technology	9,02%
Pharmaceuticals	7,38%
Biotechnology	5,74%
Chemical engineering	4,10%
Telecommunication	4,10%
Civil Engineering	3,28%
Engines pumps, turbines	3,28%

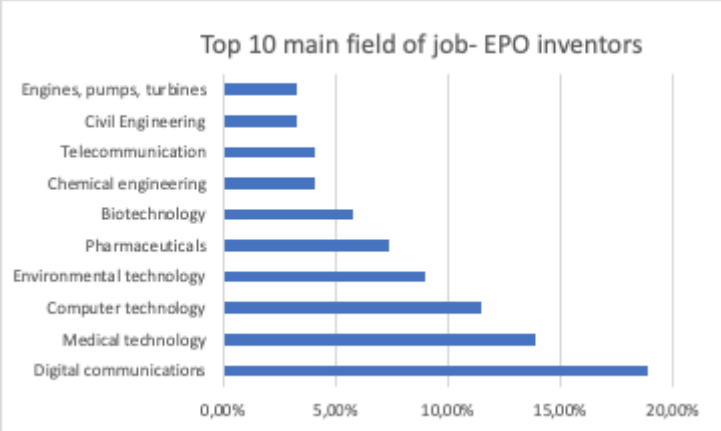

Table 4-17 - Top 10 Main Technical field of EPO inventors

Top 10 Main field of Job	NO EPO
Chemical engineering	**13,27%**
Pharmaceuticals	10,62%
Telecommunications	9,73%
Computer technology	8,85%
Medical technology	7,96%
Audio-visual technology	6,19%
Digital communication	6,19%
Mechanical elements	5,31%

Biotechnology	4,42%
Basic communication processes	3,54%

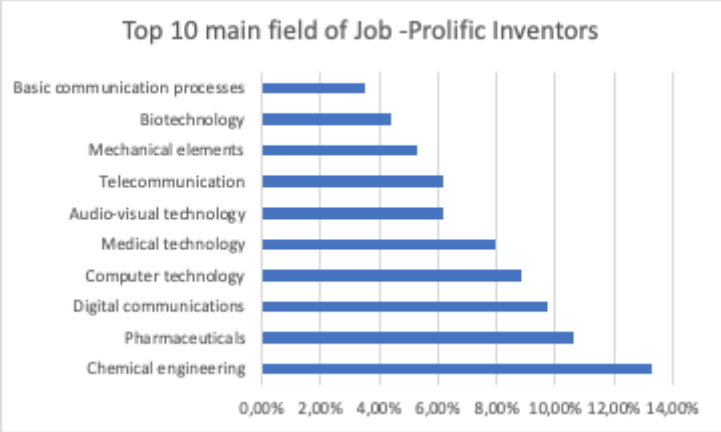

Table 4-18 - Top 10 technical field of Prolific Inventors

From Table 4-17 and Table 4-18 it is noted that there are in the first positions some recurring technical fields: the "winning" patents are numerous in the ICT field and in the biomedical field, while it is clear that the patents of chemical engineering are very numerous but not commercially or scientifically meaningful. However, it is necessary to remember that this data is extrapolated from a search on 251 inventors of which 137 have won or participated in the EPO prize while the others have patented inventions that have not been successful until now. Therefore, this study about technical fields, based on the inventions of the EPO award, undoubtedly reflects the fields where the most brilliant inventions take place; however, it is interesting to know whether they also reflect the overall annual percentage of patents issued by the EPO. In order to answer this question, the reports of the 2014-2015 European Patent Office were analyzed and the reports for 2017 appear in an imminent future. In order to verify the information found in the collection of data, it is appropriate to consider Table 4-18 taken directly from the official report of the European patent office [26]

Technology	2011 Patents	% change 2011 vs. 2010	2012 Patents	% change 2012 vs. 2011	2013 Patents	% change 2013 vs. 2012	2014 Patents	% change 2014 vs. 2013
Medical technology	10.628	-4,6%	10.502	-1,2%	10.782	2,7%	11.124	3,2%
Electrical machinery, apparatus, energy	8.693	1,9%	9.746	12,1%	10.138	4,0%	10.944	8,0%
Digital communication	8.261	-1,8%	9.809	18,7%	9.398	-4,2%	10.018	6,6%
Computer technology	8.194	-5,3%	8.540	4,2%	9.158	7,2%	9.869	7,8%
Transport	6.448	1,3%	7.002	8,6%	7.443	6,3%	7.533	1,2%
Measurement	6.448	-4,0%	6.633	2,9%	6.779	2,2%	7.228	6,6%
Organic fine chemistry	6.935	-9,6%	6.588	-5,0%	6.215	-5,7%	6.132	-1,3%
Biotechnology	5.870	-24,0%	5.539	-5,6%	5.269	-4,9%	5.905	12,1%

Engines, pumps, turbines	4.802	5,4%	5.874	22,3%	5.494	-6,5%	5.318	-3,2%
Pharmaceuticals	6.081	-12,0%	6.309	3,7%	5.568	-11,7%	5.270	-5,4%

Table 4-19 – Annual Report 2015 from EPO regarding the number of patents from 2011 to 2014

As can be observed from the comparison of the Table 4-17 and Table 4-19 the most recurrent technical fields according to the analysis of this thesis are in descending order **Digital communication**, **Medical Technology**, **Computer Technology**, which in the annual report which considers from 2010 to 2014 are respectively in third, first, and fourth position. Therefore, we can state that, considering the diversity of the sample size, (this thesis has a total sample of 251 inventors while the annual report averages about 100k inventors) the results analyzed are consistent. Also from the comparison with the sample of the prolific inventors (Table 4-19) some analogies emerge, although they are less evident: according to the analysis of this thesis the technical field where the most frequent patenting is chemical engineering and then follow Telecommunications, Computer technology, Medical technology respectively in the third, fourth and fifth position.

In order to monitor the evolution of the technical fields in which more patents are registered, the 2017 statistics were also analyzed, where the ranking is almost unchanged: first in medical technology, second in digital communications and third in computer technology. Furthermore, in 2017 there is a general increase in annual patents.[28]

4.4.3 Statistics about entrepreneurial venture

In order to provide a complete picture of the professional career of the analyzed subject it was considered appropriate to investigate also the dependent or independent condition of the subject. The entry of the "First experience as Entrepreneur / startupper / founder of

companies" template answers if the subject has ever founded an entrepreneurial activity. At first there were two split conditions for the EPO sample Before and After prize; however, the identified data collection was not considered reliable.

First experience as Entrepreneur / startupper / founder of companies	EPO Inventors	Prolific Inventors
Yes	27,74%	6,14%
No	72,26%	93,86%

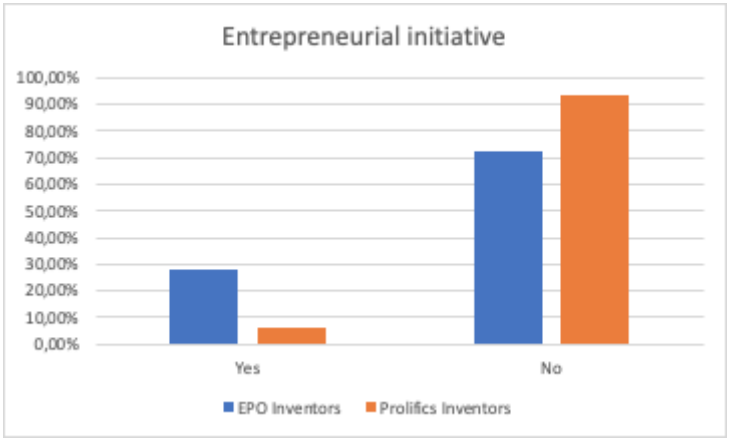

Table 4-20 - Representation of Entrepreneur initiative in sample of Star Scientists Vs Prolific Inventors

Table 4-20 shows that the analyzed Star scientists have a marked inclination to find their own activity compared to the prolific inventors. A plausible explanation could be that once a commercially successful invention has been patented, the exclusive right[7] is assured,

[7] From the art. 66 of the Industrial Property Code: "The patent rights for industrial invention consist in the exclusive right to implement the invention and profit from it within the territory of the State, within the limits and under the conditions set forth in this code."

which can be fully exploited by coming from a company. As an alternative to the exploitation of the patent there are the sale of the patent itself, the granting of the patent and the licensing of the same. Surely facing such alternatives, the main source of income comes out from the first choice. It can be assumed that the creation of the company is subordinated to the patented star invention, however, this relationship could also be reversed: an entrepreneur has incentives to patent its core products, in order to avoid problems of plagiarism and imitation.

In the following chapter the statistics will be analyzed in depth through the multivariate regression analysis in order to confirm or deny the results found and in order to find a connection between being a Star Scientist and education and career.

Chapter 5

5 Logistic Regression

In order to deepen the statistical analysis about the Star Scientists based on Education and Career factors a multivariate regression analysis has been developed. This chapter will, therefore, deal with the statistical analysis carried out starting from a theoretical summary of the implemented models, reaching then the actual implementation and, finally, explaining the results found.

It is necessary to point out that with *outcome variable* and *response variable* we always refer to dependent variable, on the other hand, with *variable of interest* we refer to independent variable.

5.1 Theoretical explanation of Logistic Regression

5.1.1 Why Logistic Regression?

Often, in order to better understand the results gathered from a statistical survey, multivariate analysis is used. In particular, in this work it has been decided to use the branch of multivariate statistics called logistic regression. It is a particularly suitable method for

this type of thesis as it allows to analyze a dichotomous dependent variable in functions of many variables of interest and controls themselves in a dichotomous form.

As David W. Hosmer explains in "Applied Logistic Regression" regression methods "have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variable" [27] . The *response variable* mentioned above is the dependent variable, in other words, the variable which is a participant of a subordination relationship with other variables, called independent. In this thesis the *outcome variable* is EPO, which is coded with a value of zero if the subject is a Prolific inventor, or 1 to indicate that the subject is a Star Scientists who participated as finalist or won the EPO award. The independent variables are the key factors of education path and career, such as a qualification achieved, or being a founder of Company. It represents the relationship between being a Star scientist (EPO = 1) and a series of variables of interest concerning education and career under the control of some factors such as gender, age group and Nationality.

For this type of analysis, the logit model has been chosen, also called logistic regression, as a non-linear relationship between dichotomous variables is studied. The objective of the model used here is therefore to establish the probability with which the main characteristic, expressed by the dependent variable, manifests itself (i.e. the fact of becoming a Star Scientist) as a function of the variables of interest.

The choice falls on the non-linear regression also because the epsilon residue of a binary dependent variable cannot have normal distribution. Therefore, the models of logistic regression, despite the greater complexity compared to the linear models, represent better the case analyzed in this thesis, as they allow to overcome some limits present in the linear regression, concerning for example observation errors, heteroskedasticity disorders and related perturbations.

5.1.2 The mathematical model of logistic regression

The logit model belongs to the class of generalized linear models, as well as the log-linear model and the probit model from which it is distinguished by function (formula 5.1).

Since the variable outcome is dichotomous it can only assume the values 0 and 1, thus it is necessary to limit the domain of this function:

$$1.\quad F(x_i'\beta) : R \rightarrow [0,1]$$
$$2.\quad \begin{cases} \lim_{x_i'\beta \rightarrow +\infty} F(x_i'\beta) = 1 \\ \lim_{x_i'\beta \rightarrow -\infty} F(x_i'\beta) = 0 \end{cases}.$$

Equation 5-1 - Domain of F and definition of Logit model

The Formula, reported in Equation 5-1, is therefore necessary in order to guarantee that the dependent variable can only assume the values 0 and 1.

In the logit model, F () is represented by a cumulative probability function (or distribution function) of a logistic type. In all statistical studies there is always a residual term of error indicated by the letter **ε**, in particular in the logit model it is assumed that the distribution of the error term is a random logistics variable.

It is also possible to define the probability of becoming a Star Scientist as illustrated in Formula reported in Equation 5-2 in which Λ represents the logistic distribution function.

$$P(Y_i = 1 | X_{1i}, \dots, X_{ji}) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} = \frac{1}{1 - e^{x_i'\beta}} = \boldsymbol{\Lambda(x_i'\beta)},$$

Equation 5-2 - Probability of becoming a Star Scientist

5.2 Implementation of the logit model using Stata

In order to accomplish regression analysis *Stata* by StataCorp has been used, a general-purpose statistical software package.

In order to perform the multivariate analysis, the *logit* command on Stata is used as it adapts to a logistic model for a binary response with maximum likelihood and models the probability of a positive result given a set of regressors. Furthermore for the logistic regression analysis also the use of the Probit command would be equally correct.

5.2.1 The modification of database and selection of variables in the model

For ease the databases of the two samples of inventors are kept separate during construction, however in order to implement the analysis it is necessary to combine them into a single data collection where the un-useful data are eliminated for analysis: as regards the EPO database the fields in which the invention is described, the link to the page of the patent office, while for the sample of the prolific inventor the address and the reference link to Google Patents. Instead an identification code is added for the two samples in order to distinguish them in an easier way.

As explained in the introductory paragraph of this Chapter, the logit model requires dependent and independent variables of the dichotomous type: they may be 0 or 1. Since the survey of this dissertation includes also string variables defined by ASCII characters it is necessary to transform them into dichotomous variables with the same meaning.

In this regard, the following algorithm is applied for each variable to be transformed: a dummy variable is generated and placed equal to zero, a variable in itself dichotomous, the value of the dummy equals to 1 is then changed according to the case.

By means of this algorism, therefore, all the dichotomous variables necessary to proceed with the analysis have been created.

It is possible to distinguish three type of variables:

- Dependent Variable or outcome variable
- Independent Variables or variable of interests
- Control Variables

The *dependent variable* is "EPO" and represents the characteristic of belonging to Star Scientists or not, therefore EPO = 1 the subject is a Star Scientist, EPO = 0 the subject is a Prolific Inventor. The *variables of interests* are several therefore divided into two groups according to the two main notions of the analysis: Education and Career.

As far as education is concerned, qualitative variables are used to express the level of education of the subject, the field of study, the year of obtaining the degree, the university attended and the study abroad experiences. Of all these variables are used in the logit model and therefore transformed into binary:

- Education level: BA, MA, PhD;
- Field of study expressed by Macro Categories for each of the educational levels: STEM, ICT, Life Science, Other;
- Study experiences abroad.

As for work the qualitative variables used to express the subject's career are the type of work (in the company or at a research institution), the experiences abroad, the experiences in-house, the sector of experience, the last company where he has worked and with what role. From this list, not all variables are counted among those used in multivariate analysis, followed by a list of those used in the regression:

- Work international experience;
- Personal work path, includes the types of work of the subject: whether he works at a company, or at research institution, or both during his career;
- if he is an entrepreneur or a startupper or not.

Finally, the *control variables* represent the background, i.e. information that does not regard either the professional life or the education of an individual, but they influence the dependent variable. The background information common to the two samples are all used in the regression and are:

- Gender: male or female;
- Nationality: For statistical significance dividing the provenance into European, US or other;
- Age range: the subjects have been divided into 4 main age groups: under 30, 30-40 years, 50-60 years, over 60;
- Sector of employment divided into macro categories: ICT, STEM, Life Science, Other.

5.2.2 . Iteration of regression analysis

In order to deepen the link between becoming a Star Inventor and a personal path in terms of education and career, the logit model is used. In order to understand if there is a significant link between two variables (called "Robust" in statistics) it is necessary to test the same condition several times and see if it appears recurrent in the models.

Therefore, 4 models are chosen to test the supposed links present:

- Model containing only the **control factors**
- Model containing the **control factors** and **education** variable of interests
- Model containing the **control factors** and **career** variable of interests
- Model containing the **control factors** and **education** and **career** variable of interests.

The first model is elaborated in order to test if there are control variables with recurring values in the first or in the second sample, even if it is not the purpose of the analysis of this thesis. Instead, in the other three models the combinations of the variables of interest have been tested always in relation to the control factors, in order to heuristically search for strong links between the variables. Within the three models the regression is iterated several times depending on whether there are correlations or not, if the factors are present in a percentage so high as to be omitted, or simply to test the robustness of a statistic.

In the following paragraph the most significant models will be shown, and the results will be interpreted according to the single value and also in a global perspective.

5.3 Stata Output and interpretation of the results

Before introducing the actual results, it is necessary to make some clarifications regarding the variables used typical of logistic regression. Moreover, it is necessary to specify that in Equation 5-3 and Equation 5-4 the values of the regression coefficients are shown and in the round brackets the values of the statistics z with an asterisk or two depending on the

significance of the relationship between the variables. Therefore, the comment is used to use the *z statistic* and not the *p-value*.

5.3.1 Pseudo R^2 and the goodness of the model

The measure of the goodness of the model is expressed by the variable Pseudo R^2, that is a kind of R^2 as the name suggests. In models *OLS* R^2 can be interpreted as explained variable, or improvement from the null model to the estimated one or finally as a correlation square.

Although some alternatives have been proposed, pseudo R^2 remains the best to measure the goodness of the model for binary choice models. R^2 is explained in Equation 5-3 using the function free log-likelihood $\ell(\beta^)$ and of the constrained log-likelihood given by $\ell(\beta)$. The cases of this index indicate that, when in the free model the estimated coefficients are zero, free and constrained log-likelihood coincide, therefore pseudo-R^2 is zero. The pseudo-domain $R2 \in [0, 1)$ can also be delimited.

Finally, it can be said that the goodness of the model is proportional to the pseudo R^2: R^2 would be zero if all the coefficients of the regression were equal to zero, R^2 would be equal to 1 in the ideal condition. In the following Equation 5-3 and Equation 5-4 it is observed that R^2 is always decimal and varies from a minimum of 0,1369 to a maximum of 0,3389. This measure of goodness of the adaptation appears minimal in the poorer model of variables where they appear in addition to the dependent variable only those of control and appears instead maximum in those models where there are also in addition the independent variables which explain the education and the career. These results are in fact consistent with the meaning of the pseudo R^2, since it represents the measure of how much the regressors correctly preach the dependent variable.

$$R^2 = 1 - \frac{1}{1 + 2N[\ell(\hat{\beta}) - \ell(\tilde{\beta})]}$$

Equation 5-3 - Pseudo R² Formula

5.3.2 Likelihood and Log-Likelihood in Logit Model

It is possible to define *likelihood* as measure of the goodness with which a set of data "supports" a particular value of a parameter. In fact, David W. Hosmer defines the likelihood of a specific parameter value (θ) "*the probability of obtaining the observed data if the parameter (X) was equal to that specific value (x).*" as shown in Equation 5-4 [25]

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x)$$

Equation 5-4 - Likelihood function

By itself this value is therefore not very significant, it becomes when it is compared with others. Therefore, since likelihood is a probability, a probabilistic model is born that seeks its maximum value. Moving to the substance of the logistic regression, sometimes logarithm natural of the likelihood function, called log-likelihood is more suitable. The logarithm is a strictly growing function therefore the logarithm of a function and the function itself present the maximums in the same points; for this reason, log-likelihood is used in the estimation of maximum likelihood. This transformation appears to be a simplification since finding the maximum of a function means performing the derivative and often the derivative of the probability of the logarithm is simpler than the derivative of the authentic probability as it transforms the products into sums and reduces the parameters to the exponents in the most distributions.

Log likelihoods are always negative because probabilities are decimals including between 0 and 1, the logarithm function is negative in the region. Finally, it is possible to conclude

that in general higher likelihood means that the model has a better relative chance of producing the data obtained.

5.3.3 The relationship between becoming a Star Scientist and Education

Table 5.1 shows the results of the regression analysis of several models implemented on Stata. The relationships between the dependent variable and one or more variables of interest may be Non-significant, Significant or Very Significant; the significativity is expressed by the regression coefficient and by the z statistic. If you report the same significant or very significant relationship in different models, it is considered *Robust*.

- Model 0: regression between the outcome variable with the control variables, that are the dummies indicating the origin, age, gender and field of work of the inventors.
- Model 1: regression among the dependent variable, some variables of interest (e.g., master's degree, PhD and study abroad experience) and the control variables, that are always present in the subsequent reiterations of the analysis.
- Model 2: regression among the outcome variable, some independent variables (e.g., education and work experience in research centers or universities) and control variables.
- Model 2A: regression that, with respect to the one present in Model 2, adds the variables of interest with regard to the field of study of the obtained bachelor's degree.
- Model 2B: regression that, with respect to the one present in Model 2, adds the variables of interest with regard to the field of study of the obtained master's degree.

These five models are the outcome of a selection resulting from numerous iterations of the logistics model on Stata, based on the most significant results. For example, the variable indicating the obtaining of the Bachelor is removed because it appears that almost the all

the samples have obtained it: only 2 inventors have not obtained it, 228 have obtained it while 21 are missing data in this field as shown in Figure 5-1.

```
. tab Bachelor
```

Bachelor (BA)	Freq.	Percent	Cum.
0	2	0.87	0.87
1	228	99.13	100.00
Total	230	100.00	

Figure 5-1 - Description of Bachelor variable

Moreover, in Table 5-1 it is possible to notice that the variables representing the acquisition of PhD and work in a research Institution / University are never used together. This is because the two variables are strongly correlated, as shown in Figure 5-2, so in order not to alter the regression they are never used together. Anyway, this correlation is lawful because often who obtains the PhD title aims to an academic career.

```
. pwcorr PhD WorkedinUNIVERSITYResearchce
```

	PhD	Worked~e
PhD	1.0000	
WorkedinUN~e	0.4674	1.0000

Figure 5-2 - Correlation between holding a PhD and working in a Research center / University

Variables	Model 0	Model 1	Model 2	Model 2A	Model 2B
Master		0,242	0,242	-0,042	0,201
		(0,5)	(-0,51)	(-0,08)	(-0,28)
PhD		0,992*			
		(2,56)			
BA ICT				2,074*	
				(2,08)	
BA Life Science				1,488	
				(-1,51)	
BA STEM				1,14	
				(-1,19)	
Worked in			1,212**	1,026**	1,185**
University or Research center			(3,21)	(2,75)	(3,15)
MA STEM					-0,121
					(-0,18)
MA ICT					0,502
					(-0,77)
MA LS					-0,232
					(-0,3)
International study experience		0,393			
		-0,98			
Gender	0,487	1,04	0,388	0,29	0,501
	(0,85)	(1,53)	(-0,53)	(0,41)	(0,7)
Nationality (from USA)	1,144	1,845*	1,855*	1,826*	1,698*
	(-1,91)	(2,41)	(2,37)	(2,47)	(2,32)
Nationality (from Europe)	0,877*	1,356*	1,468*	1,289*	1,269*
	(1,98)	(2,16)	(2,35)	(2,30)	(2,24)
Age range	-2,088**	-2,572**	-2,698**	-2,527**	-2,624**
(30-45 years old)	(5,52)	(4,92)	(5,11)	(5,08)	(5,20)
Age range	-0,258	-0,741	-0,854	-0,557	-0,673
(46-60 years old)	(-0,76)	(1,64)	(-1,8)	(-1,26)	(-1,53)
ICT	-1,932	9.42	-0,458	0,625	-0,936
	(-1,74)	(+1,35)	(+0,62)	(+1,46)	(+1,24)
STEM	-1,747	0,495	-0,459	0,643	-0,497
	(-1,56)	(+1,06)	(+0,59)	(+1,39)	(+0,66)
Life Science	-2,57*	-1,91	-1,331	-1,331	-1,41
	-2,33	(2,14)	(1,81)	(1,81)	(0,209)
Constant	2,155	-1,366	-0,579	-1,599	-0,209
	(1,68)	(1,67)	(0,57)	(1,94)	(0,22)
N of observation	251	186	189	197	197
Log likelihood	-146,73	-149,25	-103,6894	-108,58	-107,74
Pseudo R^2	0,1663	0,1369	0,2066	0,2041	0,2103

Table 5-1 - Analysis of the relationship between the dependent variable and Variable of interest representing Education

Legend:

*** p<0.01
** p<0.05
* p<0.10

The relationships that are significant in Module 0 column show the database composition with respect to the biographical data of the inventors. As shown by the value of the z-statistic, a prevalence in the sample of the winning Star scientists emerges in the Module 0 of the EPO prize of subjects of European nationality and averagely younger subjects aged between 30 and 45 years, while in the control sample emerges a majority of scientists working in the life science sector. In particular, the relationship between the dependent variable and the dummy *'Age range (30-45 years old)'* interest variable is robust in all models and it is also found in the statistical analysis of Chapter 3: prolific inventors, therefore, are averagely younger than the EPO one.

With regard to the variables of interest related to education, the relationship between *being a Star Scientist* and *working at a Research institution / University* is Very Significant and robust in all models in Table 5-1. The relationship between having a PhD and being a Star Scientist is also Significant and Robust, as shown Appendix A, Table 5-1 (Modul 1) and Table 5-2 (Modul 3+1). Table 5-1 shows some models that, as anticipated, are the result of a selection of many iterations of the Logit on Stata regression model. In particular models 2, 2A, 2B contain the variable of interest "*Worked in University or Research Center*" and not "*PhD"* because the two variables are correlated, the one with the strongest correlation with the dependent variable is chosen. Furthermore, in the model presented in Table 5-2, only significant and very significant relationships are considered in the one presented in Table 5-2.

Therefore, with regard to the relationship between being a Star Scientist and education it can be inferred that those who hold a PhD and / or work in a Research entity statistically have more chances to become a Star Scientist. In conclusion, to the question "*Ranking and type of university and degree of education influence on becoming a successful inventor?"*

Based on this statistical study the answer is: **Partially Yes**.

5.3.4 The relationship between becoming a Star Scientist, Education and Career

Table 5-2 shows five models which express the relationship between being a Star Scientist, education and career path. The models are the result of a selection among numerous models implemented in Stata, only the most significant are shown.

- Model 3: regression between the outcome variable appears with the variables of interest about career, the dummies indicating the working condition as an employee in a company or Research center or as an entrepreneur and work experience abroad and the factor variables.
- Model 3+1: regression among the outcome variable with the variables of interest about career, some variable of interest regarding Education such as Master, PhD and international study experience and the factor variables.
- Model 3+2: regression among the outcome variable appears with the variables of interest about career, some variable of interest regarding Education such as Master, Worked in University/Research Center and the factor variables.
- Model 3+2A: regression among the outcome variable appears with the variables of interest about career, some variable of interest regarding Education such as Master, Worked in University/Research Center, field of bachelor's degree, and the factor variables;
- Model 3+2B: Regression among the outcome variable appears with the variables of interest about career, some variable of interest regarding Education such as Master, Worked in University/Research Center, field of master's degree, and the factor variables;

Variables	Model 3	Model 3+1	Model 3+2	Model 3+2A	Model 3+2B
Master		-0,07	0,162	-0,042	0,013
		(-0,13)	(-0,31)	(-0,07)	-0,02
PhD		0,974*			
		(2,30)			
BA ICT				2,047	
				(-1,77)	
BA LS				1,679	
				(-1,43)	
BA STEM				1,258	
				(-1,1)	
Worked in			0,912*	0,633	0,872*
University			(2,25)	(-1,47)	(2,02)
MA STEM					-0,242
					(-0,31)
MA ICT					0,537
					(-0,7)
MA LS					0,174
					(-0,19)
International work	0,564		0,533	0,703	
experience	(-1,47)		(-1,2)	(-1,45)	
International study		0,355		0,233	0,581
experience		(-0,78)		(-0,45)	(-1,29)
First experience	1,952**	2,503**	2,423**	2,234**	2,374**
as Entrepreneur	(3,62)	(3,73)	(3,55)	(3,22)	(3,38)
Worked in a Company	-2,375**	-2,775**	-2,337**	-2,14*	-2,388**
	(2,68)	(2,95)	(2,58)	(2,39)	(2,64)
Gender	0,42	0,371	0,455	-0,036	0,394
	(-0,54)	(-0,45)	(-0,55)	(-0,04)	(-0,47)
Nationality	1,281	1,797*	1,779*	2,095*	1,889*
(from USA)	(-1,75)	(2,11)	(2,07)	(2,36)	(2,18)
Nationality	1,174*	1,637*	1,831**	2,136**	2,005**
(from USA)	(2,02)	(2,32)	(2,58)	(2,82)	(2,71)
age range	-2,385**	-2,39**	-2,517**	-2,384**	-2,619**
(30-40 years old)	(4,79)	(4,07)	(4,36)	(4,00)	(4,41)
age range	-0,312	-0,405	-0,611	-0,357	-0,605
(50-60 years old)	(-0,72)	(-0,78)	(-1,19)	(-0,66)	(-1,18)
ICT	1,88*	1,491	1,28	0,74	0,983
	(2,15)	(1,51)	(1,33)	(0,72)	(0,97)
STEM	1,33	1,013	1,049	1,041	0,928
	(1,50)	(1,00)	(1,06)	(1,04)	(0,93)
LS	0,48	0,248	0,095	-0,15	-0,008
	(0,57)	(0,27)	(0,10)	(0,15)	(0,01)
Constant	0,106	0,024	-0,377	-2,051	-0,345
	(0,09)	(0,02)	(0,26)	(1,13)	(0,24)
N. of observations	218	186	189	188	189
Log likelihood	-104,4	-86,86	-98,71	-86,98	-88,08
Pseudo R^2	0,3084	0,3226	0,3211	0,3389	0,326

Table 5-2 - Analysis of the relationship between the dependent variable and Variable of interest representing Education and Career

Legend:

*** p<0.01
** p<0.05
* p<0.10

Before commenting the results of this study, it is necessary to make some clarifications regarding some variables used or not in the models under analysis. During the implementation of several models on Stata it has been observed that some variables have a strong correlation, and, for this reason, they are never used together. This is the case of "*international study experience*" and "*international work experience*", which are strongly correlated as shown in Figure 5-3; this correlation also seems plausible.

```
. pwcorr Internationalworkexperience Internationalstudyexperience

             | I~work~e I~stud~e
-------------+------------------
I~workexpe~e |   1.0000
I~studyexp~e |   0.4094   1.0000
```

Figure 5-3 - Correlation between the two variables which represent the experience abroad in terms of study and work

The variable *"international work experience"* it is also correlated with the variable that represents having a PhD, as proved in Figure 5-4 even if in a marked way, but in order to avoid alterations of the results they are never used together.

```
. pwcorr PhD Internationalworkexperience

             |      PhD I~work~e
-------------+------------------
         PhD |   1.0000
I~workexpe~e |   0.2082   1.0000
```

Figure 5-4 - Correlation between the two variables which represent the acquisition of PhD title and work experience abroad

First of all, the five models presented have been chosen because the most representative, are in fact similar to each other, for the main variables of interest, but differ for some other variables. The purpose of this second study is to define if there are links between becoming a Star scientist and having a certain kind of career and education. In Model 3, this study is carried out through variables of interest related to the career while in the following models independent variables related to the education are used.

Considering the results of the multivariate analysis shown in Table 5-2 there are no significant relationships between having won / participated in the EPO award and therefore being a Star Scientist and the experiences abroad of study or work.

Taking into account both the analysis tables and evaluating the relationship between the dependent variable and "*Worked in UNIVERSITY Research Center*", it is observed that only the part Education as variables of interest is examined, it is very significant; instead when only the career or both are considered is just significant. Only in the Model 3 + 2A it is not significant, probably because there are two related variables: "*international work experience"* and "*international study experience"*, overall the result can be considered *Robust*.

Furthermore, it seems that the relationship between the dependent variable and "*First experience as Entrepreneur*" is very significant, so the two variables are positively correlated. According to this statistical analysis, therefore, those who started a company on the basis of an invention are more likely to become Star Scientists. The latter assertion also makes sense in common thought: an invention is considered brilliant when it has not only scientific but commercial success.

In conclusion, the relationship between the dependent variable and "*Worked in a Company*" is analyzed, which has a negative regression coefficient, therefore the Prolific Inventor group is positively correlated to the variable of interest under discussion.

5.3.5 Global Interpretation of Stata analysis

After iterated the regression in all the possible models, robust results are achieved, which in part distinguish the two groups of inventors analyzed.

The statistical study analysis can be concluded by distinguishing the two samples based on the results found. Furthermore, it is more probable for a subject to become a Star Scientist if he holds the PhD title and if he has work experience at a University or Research institution; in addition, there is a strong correlation between being a Star Scientist and being an entrepreneur or a Startupper.

Finally a longer university career ease the path to become a star scientist; it is also partially confirmed as a research entity and a patent company with completely different impact objectives: company patents are used to protect commercial interests and to manipulate the balances of competition with *incremental inventions;* on the other hand the inventions carried out at Research institutions or Universities appear *radical invention*.

Chapter 6

6 Conclusion

This final chapter outlines the conclusions on the work, on the main revelations obtained through statistical analysis and draws the possible future developments related to this dissertation.

6.1 Final Considerations

From the beginning the main objective of this thesis is defined: to understand the relational balances that bind the Star Scientists and their personal path in terms of education and career. In order to find out which factors related to education and career ease the way to become a Star Scientist, a database has been created containing 137 Star Scientists and 114 Prolific Inventor through a personal search using search engines, scientific articles, periodicals and social networks. After completing this first research phase, the statistical analysis phase begins, starting with a simple results analysis study using Excel, and proceeding with a multivariate regression analysis with a logit model implemented with *Stata* software.

Therefore, with the statistical analysis of the database using *Excel* the proportions were defined considering the two different inventor samples regarding the biographical data, the degrees obtained, the sectors in which they studied / work, the experiences outside and the rate of entrepreneurs. While through the multivariate analysis carried out with Stata studying a single database, the peculiarities of both samples emerge: the composition in biographical terms, and the sought-after factors of education and career. It is from this last analysis that the most prominent conclusions can be drawn. To the question "*is there a*

relationship between the success of a Star inventor and his personal path in terms of education and career?"
There is finally an answer.

It is possible to grasp the differences between the Star Inventor sample and the one relating to the prolific inventors along the dimension of whether or not holding a PhD and along different aspects of the professional path.

Certainly, as far as the education factor is concerned, a more lasting academic path that culminates with a PhD promotes becoming a Star Inventor. Furthermore, it can be said that those who work in universities or research centers are more likely to conceive a successful invention than those who work in a company. This last finding is legitimate because in a company the patents are thought to direct the balances with the competitors and preventing that the core products are not imitated: the inventions tend to be incremental. In a university or research organization is published with objectives of greater impact on society and, in most cases, the means available to a research entity are different with those in a company. In the end, statistically, those who have conceived an invention worth of European recognition are more likely to become entrepreneurs. Even if the cause-effect relationship in detail is not possible to define here as it is not understandable from this research, probably as a result of a scientifically patented successful invention, it is a natural process to construct what concerns the commercial part in order to exploit the patent through a start-up or a new company.

6.2 Future Work

The research supported by this thesis could be carried out by extending the two samples in order to expand the theories reported.

It could also be interesting to analyze the intellectual impact of the inventions that won the EPO award and the most important invention of Prolific inventors, in terms of citations and bibliometric analysis. Indeed, the quantitative analysis of bibliographic citations and scientific articles documents the intellectual connections between documents and reveals the intensity of such intellectual connections. In other words: the number of citations

related to a specific document provides an indicator of its cognitive impact. Finally, a Hirsch-index-based analysis can also be performed, in order to establish a ranking of importance of the inventors and their own publications.

7 Bibliography & Sitography

[1] http://www.oecd.org/innovation/

[2] https://en.oxforddictionaries.com/definition/invention

[3] Godin, Benoît. "The linear model of innovation: The historical construction of an analytical framework." *Science, Technology, & Human Values* 31.6 (2006): 639-667.

[4] Hohberger, Jan. "Does it pay to stand on the shoulders of giants? An analysis of the inventions of star inventors in the biotechnology sector." *Research Policy* 45.3 (2016): 682-698.

[5] https://www.epo.org/learning-events/european-inventor.html

[6] Frey, Bruno S. "Knight fever-Towards an economics of awards." (2005).

[7] http://innovationcelebration.com/

[8] Wright, Brian D. "The economics of invention incentives: Patents, prizes, and research contracts." *The American Economic Review* 73.4 (1983): 691-707.

[9] May, Christopher. The global political economy of intellectual property rights: The new enclosures. Routledge, 2015.

[10] https://euipo.europa.eu/ohimportal/it

[11] Agrawal, Ajay, John McHale, and Alexander Oettl. "How stars matter: Recruiting and peer effects in evolutionary biology." *Research Policy* 46.4 (2017): 853-867.

[12] Hohberger, Jan. "Does it pay to stand on the shoulders of giants? An analysis of the inventions of star inventors in the biotechnology sector." *Research Policy* 45.3 (2016): 682-698.

[13] Amjad, Tehmina, et al. "Standing on the shoulders of giants." *Journal of Informetrics* 11.1 (2017): 307-323.

[14] Bornmann, Lutz, and Hans-Dieter Daniel. "The state of h index research: is the h index the ideal way to measure research performance?." *EMBO reports* 10.1 (2009): 2-6.

[15] Levine, L. O. "Prolific inventors—A bibliometric analysis." *Scientometrics* 10.1-2 (1986): 35-42.

[16] https://wipo-analytics.github.io

[17] Maheshwari, Umesh, Radek Vingralek, and William Shapiro. "How to build a trusted database system on untrusted storage." *Proceedings of the 4th conference on Symposium on Operating System Design & Implementation-Volume 4*. USENIX Association, 2000.

[18] https://patents.google.com/

[19] https://azure.microsoft.com/en-us/services/cognitive-services/face/

[20] https://www.xing.com/app/startpage

[21] https://www.ilsole24ore.com/art/mondo/2017-12-01/spesa-rd-l-italia-investe-l-13percento-pil-quarto-germania-134159.shtml?uuid=AEq0BdLD&refresh_ce=1

[22] Gambardella, Alfonso, Dietmar Harhoff, and Bart Verspagen. "The value of European patents." *European Management Review* 5.2 (2008): 69-84.

[23] https://www.epo.org/learning-events/european-inventor/about/categories

[24] https://www.researchgate.net/blog/1

[25] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[26] Annual Report 2014, European patent applications

[27] Annual Report 2017, European patent applications

[28] Schumpeter, Joseph. "Creative destruction." Capitalism, socialism and democracy 825 (1942): 82-85.

[29] Rogers, Everett M. Diffusion of innovations. Simon and Schuster, 2010.