Foundations of Probability and Statistics: Project

A.A. 2019-2020

Lucia Ravazzi (matr. 852646) & Silvia Tamburini (matr. 813117)

14 settembre 2020

Contents

1 Introduction

The aim of this project is to analyze the status of ICT prevalence and usage in Italian companies; in particular, being Data Science students at University of Milano-Bicocca, we are particularly interested in studying the situation in the labour market for Computer Science and Data Science students. Our work will be focused in four different parts:

- General conditions of the labour market;
- Skills in the ICT fields: which are present and which are needed;
- Specific topics for Data Science students: big data and cloud computing usage;
- Future investments that could improve the situation.

We used a dataset publicly available at ISTAT website, and we will also provide a general description of the dataset and of the sample being analyzed.

2 Dataset description

We used an ISTAT dataset referring to a 2018 study made in conjunction with Eurostat and the statistical institutes of EU countries. This is a yearly analysis started in 2001 but it's mandatory since 2004. The dataset is named *Rilevazione sulle tecnologie dell'informazione e della comunicazione nelle imprese (ICT): Microdati a uso pubblico* and it is publicly available at [1], although the responsibility of the analysis is upon us and should not be attributable to ISTAT.

The provided dataset refers to Italian companies with 10 employees or more, which were active in 2018. The statistical population has nearly 200.000 individuals; ISTAT extracted a statistical sample with 33.000 units, although subsequent data cleaning processes reduced the number of elements to 22.000 observations (66.8% of the sample, 11% of the total population).

The method by which the statistical sample has been extracted is described in [2]; basically, for companies that have below 250 employees, ISTAT defined many sub-populations (one for every work field, number of employees and Italian region) and then extracted a random sample from every one of them, with the number of total extractions being determined by a statistical procedure; while all of the companies with 250 employees and more were present in the original extraction (apart from data cleaning processes). For this reason, we decided not to extract another sample and to work with all the observations that we have.

Data have been collected by a survey, which consists of almost 40 questions, with multiple answers; in the dataset, we have a variable for every answer, so that is why the dataset is characterized by a huge amount of fields, which are described by our source in an attached document. Our analysis will regard only a fraction of all the available attributes, and we will describe them later in the process. Here, we'll rather focus on the main topics of the dataset, which are the following:

- Skills of employees in the ICT field;
- Online activities (website, social media, public administration);
- Internet connection;
- E-commerce;
- Electronic invoicing;
- Specific applications, like 3D-printer, cloud computing, robotics and Big Data;
- Future ICT investments

It's important to underline that data are anonymized in two ways: by using a code which identifies each company, thus eliminating sensible information; and by discretizing specific attributes (such as earnings and number of employees) that could lead to the identification of the company.

Most attributes are binary (here, they appear as integers); plus, there are 4 double attributes (mostly percentages) and 5 character attributes, which are codes to identify information about the company, i.e. its location, its number of employees, etc. Variable types are summed up in table 1.

Table 1 – Variable types.

Types	Frequency
character	5
double	4
integer	215

Finally, the dataset has a lot of missing data values, but most of them are reasonable: for example, a lot of questions in the survey have to be answered only if a "yes" answer to another question has already been given. This can be understood further with the help of the following code: the below *miss* vector contains the number of missing values for every column. We will print here only the first 50 elements, and it can be seen for example that the 26th, 27th, 28th, 29th, 30th and the 31st attribute have the same number of missing values. This is a consequence of the explanation given above.

```
miss = c()

#Number of missing data for each column
for (i in 1:dim(data)[2]){
  miss = c(miss, data[,i] %>% is.na() %>% sum())
}

miss[1:50]
```

```
##  [1]     0     6     0    92    92    92    92 18993    92    92    92    92
## [13]    92    92    92    92   195   195  1039   195  5639  5639  5639  5633
## [25]   195  4698  4698  4698  4698  4698  4698   195 16493 16493 16493 16493
## [37]   177   177   177   177   177   177   177   177   177   177   177   177
## [49]   177   177
```

3 General information about the sample

In the following section, we are reporting information about the companies in the sample, with specific regard to how many employees they have, how much they earn, their area of expertise and where they are.

3.1 Number of employees

First of all, we can distinguish between small, medium and big enterprises on the basis of the number of employees. The absolute number isn't provided: it has been discretized into three intervals: $[10, 49]$, $[50, 249]$, $[250, +\infty]$. In table 2, the first interval refers to *small enterprises*,

the second refers to *medium enterprises* and the third refers to *big enterprises*. We note that small enterprises are the most frequent ones.

Table 2 – Company dimension percentage frequencies.

Company dimension	Frequency (%)
Small	63.5
Medium	23.0
Big	13.6

3.2 Income

Secondly, we studied the distribution of their income. Being a sensible information that could help recognizing the company, the income is also discretized into invervals that can be seen in the figure 1. As it can be seen from that graph, $[2000000, 4000000)$ is the most likely, i.e. the *mode*; followed by $[1.000.000, 2.000.000)$ and by $[5.000.000, 10.000.000)$. It has to be noted though that the intervals have not the same width.

Figure 1 – Distribution of income of enterprises.

3.3 Connection between number of employees and revenue

It can also be asked if there is a connection between the number of employees studied previously, and the revenue of the company. We'll use the Chi-Square test to answer to this question, especially we are testing the $H_0 : \chi^2 = 0$.

```
## Warning in chisq.test(cont): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont
## X-squared = 16439, df = 24, p-value < 2.2e-16
```

Assuming that the value of $\alpha = 0.05$, we observe that the p-value is lower than α and as a consequence, the H_0 hypothesis must be rejected: as might be expected, the big enterprises have the highest revenue, although there are also some small ones which have a high income.

Note that the function that we used to compute the Chi-Squared test throws a warning (*Chi-squared approximation may be incorrect*) whenever one of the expected counts is lower than five: as it can be seen from the histogram in figure 1, there are only a few enterprises with an income lower than 200.000 euros.

3.4 Location

Moreover, we studied the territorial distribution of the companies that we have in the dataset. As it can be seen from table 3, more than half of the enterprises that we are studying are in the Northern regions of Italy, while only 4% of them are located in the islands.

Table 3 – Company location percentage frequencies.

Location	Frequency (%)
North-West	35.7
North-East	26.6
Center	24.3
South	9.8
Islands	3.7

3.5 Areas of expertise

Finally, we are showing here the expertise areas percentage frequency distribution of the companies of interest. As it can be seen from figure 2, retail and mechanic companies are the most prevalent, followed by manifacturing and building ones: in total, they are more than half of the companies that we have. Plus, for some companies we don't have their expertise area, and they're represented as *Missing*.

Figure 2 – Expertise area percentage frequencies distribution.

4 General situation

In this part, we'll describe the general situation of ICT usage in Italian business. We'll focus on specific details later on in the analysis. We want to focus here on four specific points:

- Number of employees who use a device to work;
- Available internet connection;
- Number of employees who use a connected device to work;
- General information about other areas of interest in the dataset (the ones we described in section 1)

4.1 Percentage of employees who use an ICT device to work

We assume here that an important factor to take into account if we want to focus on ICT usage in enterprises is the percentage of employees who need at least one ICT device to work. This can track the importance of ICT in the workfield. As it can be seen from graph 3, in our sample, almost 25% of enterprises are such that all of their employees need at least one ICT device to work. It is also remarkable to note that the the percentages of enterprises within other bins occur with very rare frequency (less than 5%).

Figure 3 – Distribution of the percentages of employees who need at least one ICT device to work.

However, visualization 3 doesn't take into account the sector in which the enterprises work. Therefore, *conditional boxplots* in figure 4 show the distribution of the fraction of employees who use at least one ICT device to work against their expertise area; as it can be seen, there are three categories in which the median is almost 100%: professional, scientific and technical activities, ICT services and computer management. This is coherent with the required activities of the sector. Moreover, also the electrical energy, gas, steam and air conditioning category has a value

major of 90%, while the rental activities, travel agencies and business support activities have the worst median. Finally, it's noteworthy to note that for the top fourth boxplots there are many outliers. Moreover, almost all of boxplot show that the distributions are asymmetric.

Figure 4 – Boxplots show the distribution of percentages of employees who need at least one device to work splitted by their area of interest.

4.2 Internet connection

Secondly, we assume here that the presence and quality of internet connection is also an important factor to track ICT usage in this country.

4.2.1 Availability

In this first part, we are showing here the answer to the following question: "Do you have an internet connection in your business?". As it can be seen from table 4, it is very rare now not to have an internet connection, although some enterprises in our sample are in this situation.

Table 4 – Internet connection percentage frequencies.

Do you have an internet connection?	Frequency (%)
No	0.5
Yes	99.5

It may be asked if there is a connection between the presence (or absence) of an internet connection and the location of the enterprise. This hypothesis will be tested below.

```
## Warning in chisq.test(table(data$C1, data$Locations)): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data$C1, data$Locations)
## X-squared = 7.8235, df = 4, p-value = 0.09826
```

Due to the p-value, at the $\alpha = 0.05$, the H_0 mustn't be rejected and as a consequence, there may be independence between the two attributes. Fortunately, the sample may shows that there isn't a gap between different areas of our country with respect to the presence of internet connection.

4.2.2 Quality

Now, we want to study the quality of the available internet connection. Here, we are showing the answer to the following question: "If you have an internet connection in your business, do you have a broadband connection?". As it can be seen from table 5, most of the businesses that have an internet connection answer "yes" to this question.

We want to check now the velocity of this connection. It can be seen from table 6 that most of Italian enterprises do have access to a pretty fast connection: less than 3% of enterprises have a download velocity less than 2 Mbit/s, and more than half of enterprises have a download velocity between 10 and 100 Mbit/s.

Table 5 – Broadband connection percentage frequencies.

Do you have a broadband connection?	Frequency (%)
No	3.9
Yes	96.1

Table 6 – Download velocity percentage frequencies.

Download velocity	Frequency (%)
$<$ 2 Mbit/s	2.8
(2, 10) Mbit/s	24.0
(10, 30) Mbit/s	30.4
(30, 100) Mbit/s	25.9
$>$ 100 Mbit/s	16.9

Table 7 represents the contingency table of download velocities with respect to the dimension of the company. As it can be seen, big companies usually need faster connections, while small and medium companies might be able to work well even if slower connections are available.

Table 7 – Contingency table of Internet connection velocity and company dimension. Values are the percentage conditional distribution by column.

Download_velocity	Small (%)	Medium (%)	Big (%)
$<$ 2 Mbit/s	3.6	1.9	0.8
(2, 10) Mbit/s	29.0	18.4	10.7
(10, 30) Mbit/s	33.5	28.7	19.4
(30, 100) Mbit/s	23.0	30.0	32.5
$>$ 100 Mbit/s	10.9	21.0	36.6

The following computation of Chi quadro test shows the rejection of the null hypothesis of independence between variables of table 7 with $\alpha = 0.05$.

```
## Warning in chisq.test(fast_vel_cont[, c(2, 3, 4)]): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fast_vel_cont[, c(2, 3, 4)]
## X-squared = 30.778, df = 8, p-value = 0.0001539
```

4.3 Percentage of employees who use at least one connected ICT device to work

After studying the presence and quality of internet connection in Italian enterprises, we want to explore the percentage of employees who use at least one ICT device to work, versus the percentage of employees who use at least one *connected* ICT device to work. From graph 5, two important patterns emerge:

- A lot of enterprises have all of their employees who need at least one connected device to work, but only some of them also need an internet connection to work (this is the line at 100 percentage on the x axis);
- A lot of enterprises have all of their ICT devices connected to the internet, regardless of the percentage of employees who need them (this is the bisect). Specifically, as it can be seen from graph 6, these enterprises are almost 80% of the total ones.

Figure 5 – Percentage of employees with at least one device vs. percentage of employees with at least one connected device. Each point is an enterprise.

Figure 6 – Distribution of the ratio between the percentage of employees with at least one device and the percentage of employees with at least one connected device.

Finally, it would be interesting to show how the mean percentages of employees who have at least one connected ICT device changes with the macro-areas of our country – specifically, North-West, North-East, Center, South and Islands parts. To increase the interpretability of the results, a map is used as a tool to represent the *punctual estimation* for the mean of the proportion in each area. The map is shown in figure 7.

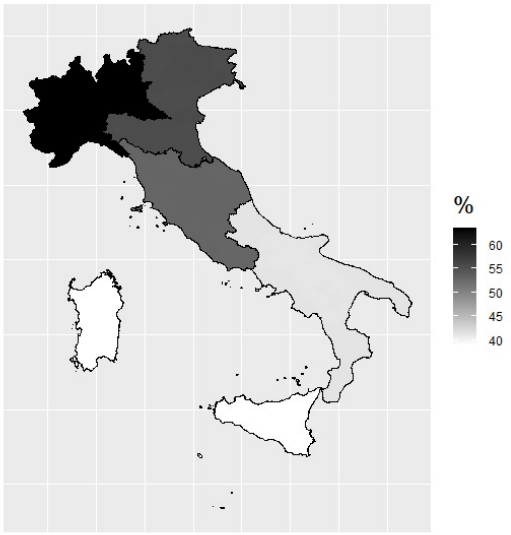

Figure 7 – Mean percentages of employees who use at least one connected device in several Italian macro-areas

As it can be seen from the map in figure 7, it seems that the percentage of employees with at least one connected device is higher in the North-east area, while south and island are characterized by a smaller rate.

4.4 General situation of other variables

We now want to analyze the general situation with ICT usage, especially with regards to specific services:

- Presence of employees specifically trained to use ICT devices;
- Update of ICT knowledge for employees;
- Web pages;
- Web advertisement;
- Cloud computing;
- 3D-printer;
- Robotics;
- Big Data Analytics;
- Electronic invoicing;
- E-commerce.

These information are collected in the dataset as answers to a question like the following: "Have you used *service* in *year*?" The *service* could have been one of the previous ones, while the *year* could have been 2018 (year of the survey) or 2017 (previous year); answers could have been *yes* or *no*. In figure 8, it is shown the percentage of enterprises that have answered "yes" to each of the questions; some fields are labeled with the year 2017 because the question referred to that period, otherwise the 2018 year is implicit.

Since we are managing a sample, it's fundamental to underline that the computation of the percentage relies on the *estimator* of the proportion which is correct, consistent and efficient:

$$\hat{P} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

where N is the number of statistical units and X_i the number of successes. Since the dataset has a huge number of data, we assume the validity of the *Central Limit Theorem* which allow us to state that the distribution of the estimator is *gaussian*. As a consequence, we can compute the *confidence interval* for each proportion and show it in the graph. This same procedure will be applied further, every time we will compute confidence intervals. We set $\alpha = 5\%$ for all of them.

Furthermore, we want to understand if two proportions p_1 and p_2 are statistically different: to do that, we should test the hypothesis $H_0 : p_1 - p_2 = 0$ with the following test:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(p(1-p))[\frac{1}{n_1} + \frac{1}{n_2}]}} \sim N(0,1),$$

where p_1 and p_2 are the proportions of the population, n_1 and n_2 are the sizes of samples and finally, $p = \frac{X_1 + X_2}{n_1 + n_2}$ with X_1 and X_2 as the number of *yes*. But this test should be performed only on the assumption that samples are independent. This is not the case in this project, because there are for example some companies that adopt more services at once; if connections between the samples could have been considered negligible, we would have performed this test anyway, with a certain degree of error. However, we verified that this isn't true in our case: therefore, we won't perform this test in this analysis and we will only give qualitative comments with regards to comparing our confidence intervals. Future developments of this project could consider performing a comparison test taking dependence of the sample into account.

Although most of enterprises have a web page and adopt electronic invoicing, only a few of them use 3D-printer, robot and big data analytics: this is reasonable because most of the time these technologies are technical and cost a lot of money, and so they are used in specific fields.

Figure 8 – Percentage of enteprises that use these services with confidence intervals.

5 Skills in the ICT fields

In this part, we want to analyze if ICT specialists are needed and if this need is satisfied.

5.1 Needs of an ICT specialists

Here, we are showing the answer from the following question: "In 2017, have you employed or tried to employ ICT specialists?". As it can be seen from table 8, most companies answered "No", and this could be discouraging for Computer Science students.

Table 8 – ICT specialists employment percentage frequencies.

Have you employed ICT specialists?	Frequency (%)
No	85.96
Yes	14.04

However, studying the same table with respect to the dimension of the company might be useful in this case.

Looking at the percentage conditional distribution by the answer to the question (table 9), it can be seen that most "No" answers come from small enterprises, while "yes" answers are equally distributed among the dimensions, with a small preference to big enterprises.

Table 9 – Contingency table of ICT specialists employment vs. the company dimension. Values are the percentage conditional distribution by answers (column).

Company Dimension	No (%)	Yes (%)
small	68.6	30.9
medium	21.3	33.7
big	10.1	35.4

Looking instead at the percentage conditional distribution by the dimension of the company, it can be seen from table 10 that the "No" answer is more frequent in all cases, but the "Yes" answer is more prevalent in big companies than in medium and small companies; therefore, Computer science students might consider keeping an eye on big companies.

Table 10 – Contingency table of ICT specialists employment vs. the company dimension. Values are the percentage conditional distribution by company dimensions (row).

Company Dimension	No (%)	Yes (%)
small	93.1	6.9
medium	79.5	20.5
big	63.6	36.4

Indeed, if a Chi-squared test of independence is used to test if the dimension of the company

and the needs of an ICT specialist are independent or not, due to the p-value, at $\alpha = 0.05$ the hypothesis of independence must be rejected.

```
##
##  Pearson's Chi-squared test
##
## data:  tables
## X-squared = 2020.8, df = 2, p-value < 2.2e-16
```

5.2 Problems in finding ICT specialists

Looking specifically at the "Yes" answers to the previous question, we are analyzing in table 11 the answer to the following question: "Have you encountered some problems finding the ICT specialists that you needed?". It's interesting to know that the frequencies to both answers is almost the same: finding an ICT specialist was a difficult task for almost half of the companies.

Table 11 – ICT specialists problems percentage frequencies.

Have you encountered problems finding ICT specialists?	Frequency (%)
No	53.21
Yes	46.79

Studying the double contingency table with respect to the dimension of the company, and in particular the percentage conditional distribution by row (from table 12), it can be seen that it was a difficult task for companies regardless of the dimension, although it was slightly easier for big companies. Therefore, Computer Science students are highly encouraged to apply for a job if the company is asking, because they could have a high chance of being hired. It has to be noted though that this analysis doesn't take into account the age of potential new ICT specialists, and this is notably an important factor for hiring a new employee – specifically, if previous experiences are taken into account.

Table 12 – Contingency table of ICT specialists employment problems vs. company dimension. Values are the percentage conditional distribution by company dimension (row).

Company Dimension	No (%)	Yes (%)
small	48.6	51.4
medium	55.0	45.0
big	55.5	44.5

5.3 ICT activities

Finally, we want to focus on ICT activities: if, in a particular business, they are carried out by internal staff, external staff or not carried out at all; as it can be seen from table 13, most activities are carried out by external staff. It is also remarkable that more than 30% of enterprises don't carry out *Web maintenance activities*, even if, as we have previously noted, almost 80% of Italian enterprises have a website.

It has to be noted though that the distribution showed in table 13 strongly depends on the sector. As an example, we are exploring here the activities regarding *device maintenance*; as it can be

Table 13 – Who carries out ICT activities in italian enterprises

Activities	Internal (%)	External (%)	No one (%)
Web development	15.5	53.6	30.9
Web maintenance	17.3	48.2	34.5
Software development	17.6	52.1	30.3
Software maintenance	24.0	53.0	23.0
Security	28.9	53.9	17.2
Devices maintenance	29.8	63.1	7.1
Office software support	43.3	43.7	13.0

seen from table 14, while building and food and accommodation services prefer to have to be carried out this activity by external staff by a large margin, computer management enterprises carry it out with internal staff, as it has to be expected.

Table 14 – Who carries out device maintenance in italian enterprises, by work areas

Work Areas	Internal (%)	External (%)	No one (%)
Building	15.7	70.2	14.1
Food and accomodation services	16.9	75.6	7.5
Water, sewerage and trash	22.1	73.3	4.6
Retail and mechanic	23.7	69.5	6.9
Development company	26.8	66.5	6.6
Transportation and storage	27.9	61.8	10.3
Rental agencies, travel agencies and business support	28.5	63.3	8.2
Manifacturing	38.3	58.2	3.5
Electrical energy, gas, steam and air conditioning	42.7	54.4	2.9
Professional, scientific and technical activities	43.3	52.6	4.1
Missing	50.0	46.7	3.3
Information and communication services	66.1	28.0	5.9
Computer management	80.5	14.6	4.9

6 Big data and Cloud computing

Being data science students, we now want to focus on main technologies regarding use and analysis of *big data*; specifically, big data sources and analysis, and cloud computing services, which could be really helpful if internal servers aren't enough, both for testing and development projects.

6.1 Big data

We will start focusing on Big Data Analytics. From figure 8, we know that this is still a rare technology in our enterprises, specifically, as it can be seen from table 15, it has been used in 2017 only by 13% of our sample.

With regards to the enterprise's features, 31% of the big enterprises have used Big Data Analytics

Table 15 – Big Data Analytics percentage frequencies.

Have you performed Big Data Analytics?	Frequency (%)
No	86.6
Yes	13.4

possibilities, while less than 8% of small companies have done the same, as it can be seen from table 16, showing the contingency table between the use of Big Data Analytics and the dimension of the company, and in particular the conditional distribution by the second variable (company dimension).

Table 16 – Contingency table: Big Data Analytics use and company dimension; percentage conditional distribution by company dimension.

Have you used Big Data Analytics?	Small (%)	Medium (%)	Big (%)
No	92.5	80.9	69
Yes	7.5	19.1	31

Furthermore, as it can be seen from table 17, a bigger fraction of electrical energy, gas, steam and air conditioning enterprises have used Big Data Analytics in 2017, followed by Information and communication services and professional, scientific and technical activities. Big Data students might consider keeping an eye on these companies.

Table 17 – Big Data Analytics usage splitted by services.

Work Area	No (%)	Yes (%)
Electrical energy, gas, steam and air conditioning	69.6	30.4
Missing	70.0	30.0
Information and communication services	72.5	27.5
Professional, scientific and technical activities	79.2	20.8
Transportation and storage	81.5	18.5
Water, sewerage and trash	83.1	16.9
Manifacturing	84.4	15.6
Rental agencies, travel agencies and business support	85.5	14.5
Food and accomodation services	89.9	10.1
Retail and mechanic	89.9	10.1
Computer management	90.2	9.8
Building	93.1	6.9
Development company	94.6	5.4

As for the sources of Big Data, it can be seen from figure 9 that data mostly come from sensors and intelligent devices (as information and communication services and scientific activities use them for various purposes), while geolocalization and Social media data are less used. Moreover, it's worthwhile to mention that it is more common that internal employees analyze big data: this could be because most companies that planned to use big data also have specific big data architectures at their disposal. Figure 9 was built on the basis of the percentage of *yes* or *no* to

the associated questions and the associated confidence intervals are built as above. Moreover, it's important to underline that different enterprises can choose different services (samples aren't independent).

Figure 9 – Percentages of enterprises that let big data analytics be performed from various sources (left) and by internal or external staff (right), with confidence intervals.

6.2 Cloud Computing

From figure 8, we know that this is still a fairly rare technology in our enterprises; specifically, as it can be seen from table 18, it has been used in 2017 only by 33% of enterprises in our sample.

Table 18 – Cloud Computing services percentage frequencies.

Have you bought Cloud Computing services?	Frequency (%)
No	66.8
Yes	33.2

As for why they've used it, it can be seen from graph 10 (on the left) that most enterprises that choose cloud computing services use them for email services; indeed, this solution is often used in order to retain the most important messages and email that may be lost if they are stored in a local personal computer. Moreover, cloud computing is also used for file storage, database hosting and software, although this happens less often. Figure 10, both the left and right side, was built on the basis of the percentage of *yes* or *no* to the associated questions and the confidence intervals are built as above.

Given the contingency tables between the dimension of the enterprises and the type of services (public and separately the private one) with the conditional percentage distribution with respect to the dimension of the company, the figure 10 (on the right) show the percentages of *yes* of that tables. As it might be seen, there isn't an important difference between the percentage of enterprises that choose public cloud computing (which means that the service is shared with others) and private cloud computing (reserved ICT resources - as CPU and memory). However, one characteristic is that the big companies use a public service more than the private one.

Figure 10 – On the left, figure shows the percentages of enterprises that use cloud computing for different services with confidence intervals, while on the right, graph shows the private and public cloud computing services, splitted by company dimension.

Furthermore, there are several companies which adopt both the public and the private cloud computing services.The confidence intervals for this percentage proportion are shown below.

```
## [1] 0.1818527 0.1999255
```

6.3 Both

Finally, it could be interesting to study the relationship between the usage Big Data Analytics and Cloud Computing. First of all, as it can be seen from the code below, the chi-squared test tells that the hypothesis of independence (at 95 of confidence) between the two technologies must be rejected: they are indeed connected.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(bigdatacloud$G1, bigdatacloud$`data$D1`)
## X-squared = 1089, df = 1, p-value < 2.2e-16
```

To investigate this connection, contingency tables will show percentages of companies that have answered positively to one of the following questions: "Have you performed Big Data Analytics?"

and "Have you chosen Cloud Computing services?". These percentages will be computed as percentage conditional distributions, with respect to the first or the second variable.

As it can be seen from table 19, showing the percentage conditional distribution with respect to big data, almost 60% of companies that adopted Big Data Analytics have also chosen Cloud Computing Services; while only 24% of companies that adopted Cloud Computing Services have also performed Big Data Analytics, as it can be seen from table 20. This could suggest an association rule like the following: "Big Data Analytics → Cloud Computing": if big data are being analyzed in the enterprise, cloud computing services will probably be adopted. However, this association rule has to be further tested, and this will not be performed here.

Table 19 – Big data and Cloud computing contingency table: percentage conditional distribution with regards to big data.

	Cloud Computing	
Big data (%)	No	Yes
No	70.9	29.1
Yes	40.1	59.9

Table 20 – Big Data and Cloud Computing contingency table: percentage conditional distribution with regards to cloud computing.

	Cloud Computing (%)	
Big data	No	Yes
No	91.9	75.8
Yes	8.1	24.2

7 Future ICT investments

Finally, we want to understand time evolution of ICT development in ICT enterprises. In the survey that we are analyzing, there are some fields that collect data about past, present and future investments in ICT areas. The questions were of the following type: "Have you purchased goods or devices that refers to *area* in *year*?". *Year* could be from 2016 to 2019, and *area* is one of the following:

- IoT (Internet of Things)
- 3D printer;
- Robotics;
- Programmable machines (or with sensors);
- Cloud Computing;
- Web applications;
- E-commerce;
- Big Data Analytics;
- Augmented reality and virtual reality (VR)
- Cyber-security

Answers could have been "yes" or "no". Note that investments with regards to 2018 and 2019 are planned ones.

In graph 11, percentages of enterprises that have invested in these services are shown, with their confidence intervals, for each year. There is a line that connects these points, to better show time evolution; line is dotted because we have a year granularity (the line shouldn't be interpreted as a trend line) and because companies are not the same for each year.

From graph 11, a trend emerges: although some confidence intervals overlap, investments go up between 2016 and 2017, go down between 2017 to 2018 and go up again in 2019. It has to be noted that cyber-security subverts this last part, although it is the ICT area where Italian enterprises have invested more.

Figure 11 – Percentages of enterprises that have invested (2016-2017) or planned to invest (2018-2019) in various ICT areas, with confidence intervals (0.95 of confidence level).

We also thought it could be interesting to analyze possible reasons to invest more. Table 21 answers the following question: "Do you think that X is a compelling reason to invest more in ICT technologies?", where X is a possible reason like tax incentives, help from Public Administration, and so on. As it can be seen, most of enterprises thought in 2017 that development and improvement of their technological infrastructures in $2018 - 2019$ could be carried out with the support of tax breaks, financing or tax incentives, the introduction of ultra-broadband internet connection and the reinforcement and improvement of technological skills of employees.

Table 21 – Possible reasons to invest more: percentage of enterprises and confidence intervals.

	Proportion (%)	Lower (%)	Upper (%)
Tax incentives	51.21	50.55	51.87
Ultra-broadband	37.13	36.50	37.77
Training courses for internal employee	26.67	26.08	27.25
Improve the digitalization	21.71	21.17	22.26
Don't know	19.18	18.66	19.70
New ICT employees	13.57	13.12	14.03
Help from PA	13.38	12.93	13.82
Collaboration with other enterprises	7.93	7.58	8.29
Digital doesn't matter	7.34	6.99	7.68
Other reasons	3.65	3.40	3.89

Unfortunately for computer science students, it has to be noted that companies usually prefer to train their existing employees instead of hiring new ones. However, looking at the company's dimension might be useful in this case. So, we have investigated answers to the following questions: "Do you think that employing new people is a reason to invest more in ICT technologies? What about training existing employees?". In figure 12, *yes* answers at the two questions, with their confidence intervals, can be seen, conditioned to the dimension of the company.

It can be seen that percentages of enterprises that will like to hire new ICT professionists are lower than the percentages of enterprises that will like to invest in training courses for their existing employees, in all companies' dimensions. However, the first percentages are bigger for big companies than for small and medium ones: this confirms once again the importance of big companies for computer science students.

Figure 12 – Percentages of enterprises that are going to hire new ICT professionists, vs. ones that are going to invest in training courses for their existing employees, splitted by size of the company.

8 Conclusions

In conclusion, ICT development and usage in Italian enterprises is still pretty much on going: while most of them have access to common-knowledge technologies, like internet connections and websites, other technologies, like cloud computing, 3D-printer, robotics and big data analytics, are still obscure and not so much around: this is probably because they are technical, specific and cost a lot of money. That is also the reason why computer science students and data science students might consider keeping an eye on big companies, because the drive towards innovation will probably come from them: they usually have more resources, more income and more interest towards new technologies. Data Science students are also encouraged to look specifically at energy companies, information companies and professional and scientific companies, and to apply for a job if they see a possibility.

We want to reiterate that this analysis doesn't take into account the age of company's employees, the rate of employment of students, or other variables that could be useful to predict and analyze hiring possibilities for students: this is just meant to be a sight at the labor market, developments and possibilities for new people, but further analysis might integrate these data. Furthermore, financial activities are excluded from this sample: they are a notorious example of Big Data Analytics development and usage, and a more complete analysis would also take them into account.

References

1. ISTAT. Rilevazione sulle tecnologie dell'informazione e della comunicazione delle imprese (ict): Microdati a uso pubblico. https://www.istat.it/it/archivio/177221. Accessed 1 Aug 2020.

2. ISTAT. Nota metodologica. https://www.istat.it/microdata/download.php?id=/import/fs/pub/wwwarmida/183/2018/01/Nota.pdf. Accessed 1 Aug 2020.