Corso di Laurea in Business Informatics
Anno 2017/2018

Statistical Methods for Data Science

Size and growth of firms: a statistical approach

Student Names
Armillotta Alessandro
Onesto Giuseppe
Shtjefni Mario

Contents

1 Dataset Description

The dataset we have analyzed has been extracted from the **AIDA Database**, a database with financial statements, company and commodities data of more than 900,000 Italian firms, except for banks, insurance companies and public entities.
On AIDA it is possible to perform queries by business sector, geographical area, financial statements etc.

The initial dataset contained 8,397,955 rows with 14 attributes:

- **ATECO**: ISTAT code for activity classification. The distinct codes are 1660;
- **Company Name**: the company name.
- **Legal Form**: kind of company (S.p.a.,S.R.L....). The distinct forms are 34;
- **Province**: the province where the firm is based. The distinct provinces are 110, corresponding to the total number of Italian provinces;
- **Region**: the region where the firm is based. The regions are 20, the same as the Italian regions;
- **Status**: indicates the firm status. The distinct statuses are 10;
- **TaxID**: an identification code. There are 1,266,379 different TaxIDs;
- **TradingProvince**: the province where the firm operates;
- **TradingRegion**: the region where the firm operates;
- **Year**: year associated with the row data. There is a time window of length 27 from 1990 to 2016;
- **Employee (E)**: number of Employees ;
- **EBITDA (B)**: a measure to evaluate a company's performance without having to factor in financing decisions, accounting decisions or tax environments.;
- **Profit (P)**: the firm profit;
- **Revenue (R)**: the firm revenue;

1.1 Dataset Analysis

In this section we provide statistics of the distribution of our data, starting to focus on the attributes that we've selected for deeper analysis that are described in the next chapters. As mentioned in the description part, we've started with a dataset containing 8,397,955 rows, each representing individual firm data for a single year. By analyzing the distinct row values of Region/TradingRegion and Province/TradingProvince, we've seen that they're exactly the same for each row; due to this, we've then removed the attributes TradingProvince and TradingRegion from our analysis.

Talking about the distribution by Region, the largest Regions in terms of firms are Lombardia(>22% of firms), followed by Lazio($\approx$ 13.5%), Veneto, Emilia Romagna and Campania; the smallest ones are Valle D'Aosta($\approx$ 0.18%) and Molise($\approx$ 0.36%). Since the number of Regions is too big for individual analysis, we've decided to map each firm into its own Geographic Area(North, Center and South). We've included Sardegna and Sicilia regions in the South GeoArea, then obtaining the distribution shown in the following picture.

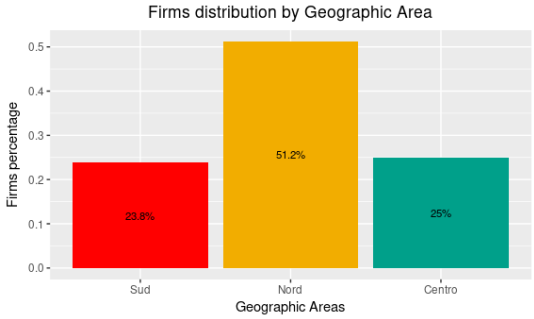
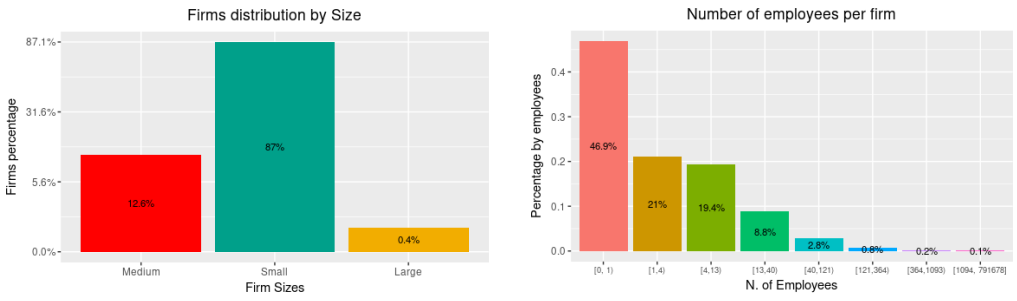


Figure 1: Firms distribution by Geographic Area

As you can see in Figure 1, the majority of firms are up North, while the remaining firms($\approx$ 49%) are equally divided between Center and South.
To better analyze the behaviour of the distinct sizes of firms (as we better explain next), we have splitted them into three categories: small, medium and large based on their number of employees[1]. In the next picture we're showing their distribution based on their Size and the number of Employees itself.



(a) Firms distribution by Size (b) Employees binning as Power Law

Figure 2: Firms distribution

The vast majority of firms have very few Employees(more than 46% have 0

Employees), and just a very small percentage($\approx 0.4\%$) has more than 250. Please be aware that Figure 2a is scaled as exp(0.4) to reduce the linear difference between the three intervals, that is much higher. Figure 2b, instead, shows the distribution of Employees by binning firm workers with growing width in exponential(base=3) way, as also shown in [2]. That is, by increasing the number N of Employees, the number of Firms having N Employees decreases in an exponential manner (more than Exp(3)) . This is consistent with what happens in literature[2] [3].

We have then analyzed the distribution of Firm Sizes by their Geographic Area (Figure 3) , reaching an interesting result: the North firms are more likely to be Large, with respect to the South and Center ones, as shown in the next barplot.

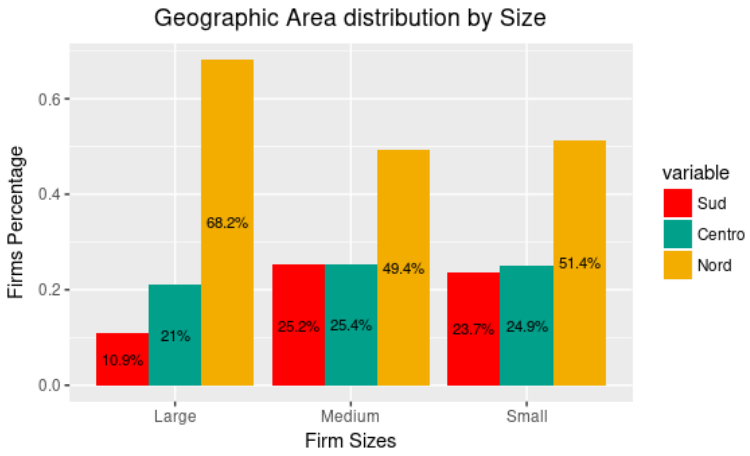


Figure 3: Geographic Area distribution grouping by Size

Finally, we've analyzed the distribution of records in the whole period(1990-2016), and as you can see from the histogram in Figure 4, the records of firms till 2006 are very limited. Hence, we've decided to analyze the records referring to years 2007-2015 (that're still more than 96% of the original data).

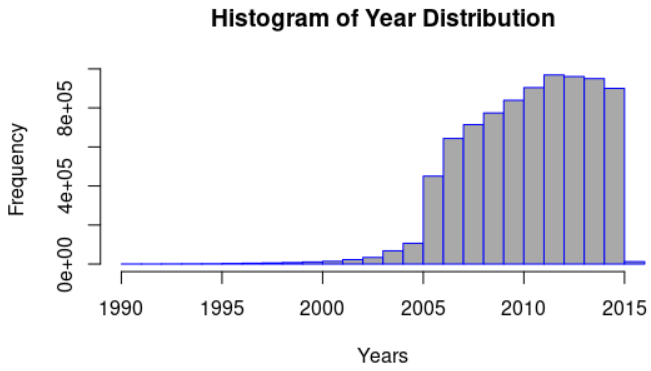


Figure 4: Histogram of Year Frequencies

1.2 Data cleaning

The original data are generally well formatted, and there don't appear to be many errors. We've found some strange rows(0,3%) containing negative Revenue values, that we've decided to remove from our data. It has also been crucial to normalize monetary columns(EBITDA, Profit, Revenue) by taking inflation rates into account: we've considered the inflation rates from ISTAT for the years 2007-2015, where for each year the inflation rate with respect to the previous year is calculated, so we have fixed it at 1 for 2007 and then applied it for all the following years, until 2015. Then, in our analysis of Employee variable distribution, we observed that there were quite a few($\approx 9.3\%$) missing values for Employee; we considered using some kind of regression to handle them, but the majority of these missing values are associated to firms having no-valid rows or just one-year-rows with a certain value for Employee and missing values for the other attributes. Because of this, we decided to remove these rows too. Finally, the cleaned dataset we have worked on contains 7,050,620 rows.

2 Correlation and Linear Regression

Correlation explores the relationship between two quantitative variables. It determines if one variable varies systematically as another variable changes. It does not specify that one variable is the dependent variable and the other is the independent variable.

For our analysis we test the correlation for AIDA and three different sectors (Manufacturing, Restaurant and Media). For this analysis we take in consideration only the Employee and Revenue attributes. We do not use EBITDA and Profit because these attributes can be influenced by some other external variables (i.e external costs, Employee Costs or Total Production Cost).

We have used a sample with 35.000 records for computational problems related to confidence interval testing with the bootstrap technique and also for a better data visualization. To make sure that we do not introduce bias , we perform bootstrapping. From Aida we draw a thousand times a sample of length 35.000 and for each sample we compute the correlation. At the end of the bootstrap we plot the 95% Confidence Interval and the empirical correlation.

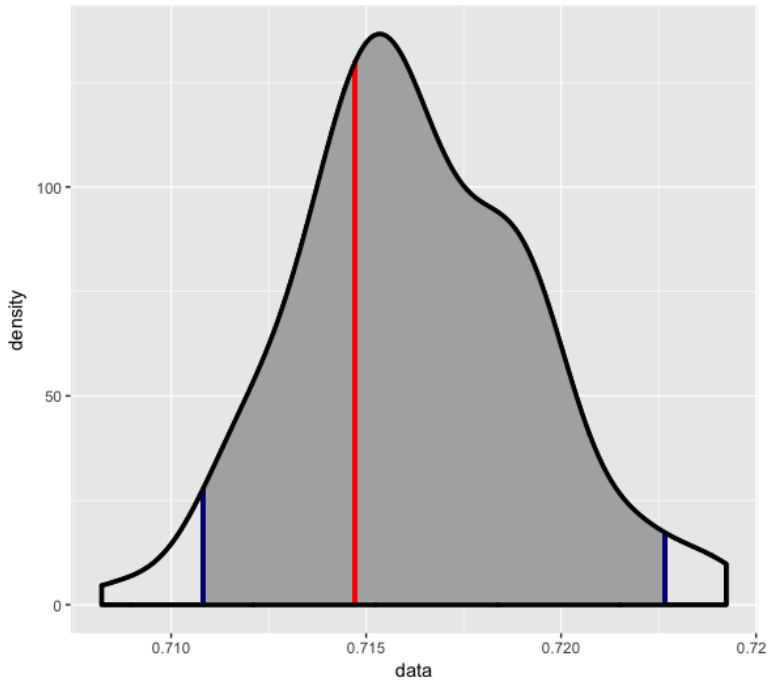


Figure 5: Confidence Interval

The *95% CI is [0.7098784,0.7219468]* and the empirical value is *0.7147149*. The empirical correlation is in the CI. For this reason we can accept the hypothesis that we do not introduce bias and our sample is good for the analysis.

2.1 Correlation Analysis AIDA

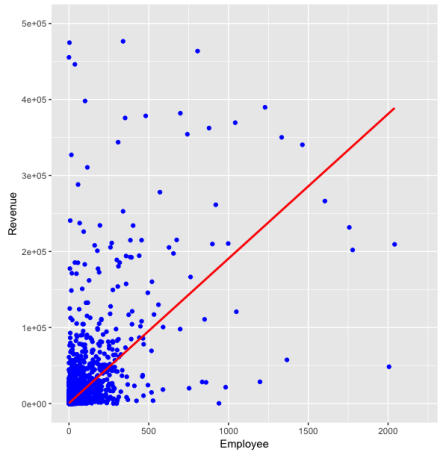


Figure 6: Employee - Revenue scatterplot

From Figure 6, we can see that the relationship on average is linear. The scatter plot shows a linear regression line, hence we are dealing with linear association between the two variables.

With the *ad.test* we have tested the **Normality Distribution** for Employee and Revenue.
The Hypotheses for the Anderson-Darling test are:

- *Null hypothesis*: the data are normally distributed;
- *Alternative hypothesis*: the data are not normally distributed;

In the Figure 7 we have performed **Anderson-Darling normality test** for Employee and Revenue. Both p-values obtained are smaller than Significance Level = 0.05, for this reason we reject the null hypothesis and we accept the Alternative Hypothesis.

```
        Anderson-Darling normality test

data:  x.aida$E
A = 12200, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  x.aida$R
A = 12313, p-value < 2.2e-16
```

Figure 7: Anderson-Darling normality test for Employee and Revenue

Based on the assumption of the Alternative Hypothesis, we won't use Pearson correlation coefficient, thus we have computed *Spearman correlation*:

- Spearman correlation is considered a non-parametric analysis. It is a rank-based test that does not require assumptions about the distribution of the data.

Figure 8: Spearman Correlation Employee - Revenue

In the Figure 8 we can see the Spearman *rho* . The correlation measure is greater than 0 and we can assume a certain positive correlation.
For the Spearman's correlation Test we have used the function *cor.test* in the stats package in R and we have assumed 2 hypothesis:

- *Null hypothesis*: rho value is equal to 0;
- *Alternative hypothesis*: rho value is not equal to 0;

```
        Spearman's rank correlation rho

data:  x.aida$E and x.aida$R
S = 2.0386e+12, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7147149
```

Figure 9: Spearman Correlation Test

In the result in Figure 9 p-value is less the Level of Significance(0.05) and we can reject the null hypothesis. This means that the correlation is different from 0 and therefore there is correlation. This means that Employee increases with Revenue.

For a better analysis we have plotted the Confidence Interval of Correlation with bootstrap technique. With boot function we have generated 10000 bootstrapped correlation coefficients and then we have drawn up the 95% CI.

The 95% CI [0.7085223, 0.7208407] and the empirical value is 0.7147149 and it's in the interval.

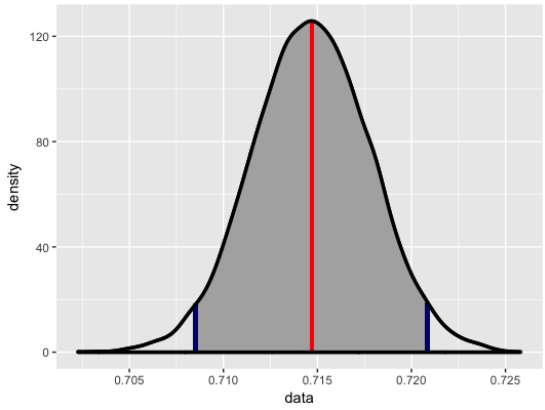


Figure 10: Correlation Confidence Interval

Another correlation computation is based on years for the same sample. For each year we have extracted the *Spearman Correlation* for Employee and Revenue.

For each year, the correlation coefficients is $0.56 \leqslant rho \leqslant 0.8$ and all coefficient test reject the null hypothesis. This means that the correlation is different from 0 and therefore there is linear correlation between Employee and Revenue in each year.

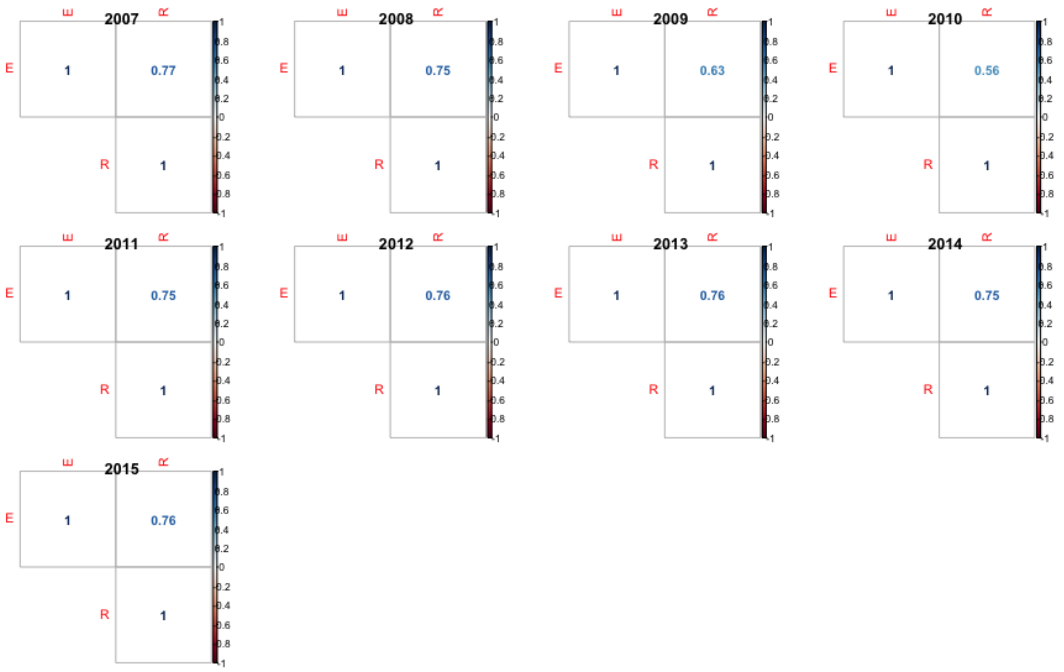
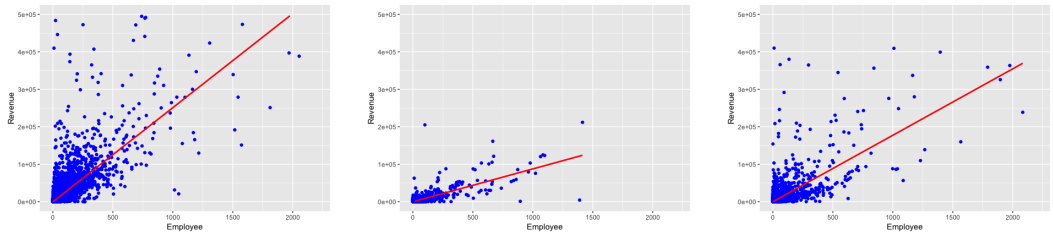


Figure 11: Correlation by years

In the results we can assume that the correlation is always positive.

2.1.1 Correlation Analysis by Sectors

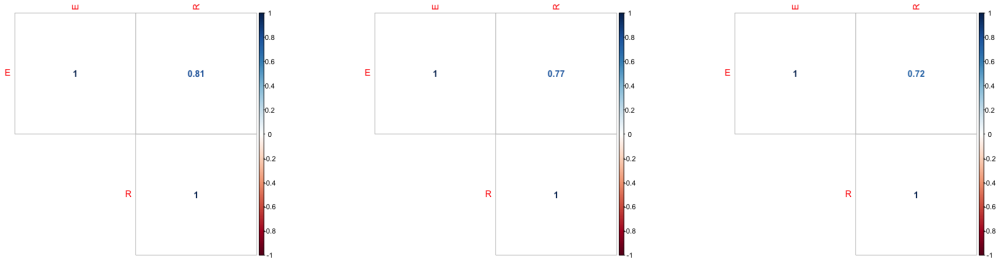
Another analysis correlation is made by sector. For this analysis we have used 35000 records as for AIDA sample.



(a) Manufacturing (b) Restaurant (c) Media

Figure 12: Employee - Revenue scatterplots by sectors

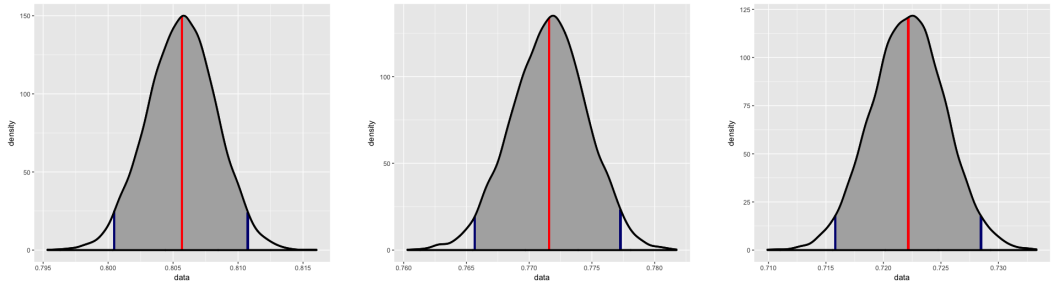
Based on Figure 12 we can assume a certain linear relationship between Employee and Revenue, maybe very similar to AIDA sample. We have tested the Normality Distribution with Shapiro-Wilk Test, but also here the distribution aren't Normal. All p-values are smaller than Significance Level = 0.05 and we reject the null Hypothesis.



(a) Manufacturing (b) Restaurant (c) Media

Figure 13: Correlation by sectors

In the Figure 13 we can see that a positive Spearman Correlation exists between E and R. The obtained correlation coefficients are : Manufacturing *rho*=0.81, Restaurant *rho*=0.77 and Media *rho*=0.72. For the Spearman Correlation Test we have used the *cor.test* function. All p-values are less than the Level of Significance 0.05 and for this reasons we reject the null hypothesis, hence all sectors have a correlation different from 0.



(a) Manufacturing (b) Restaurant (c) Media

Figure 14: Correlation Confidence Interval of distinct sectors

We have plotted the Confidence Interval for the Correlation coefficient (Figure 14) with the bootstrap technique. With *boot* function we have created 10000 bootstrapped correlation coefficients and then we have created the 95 % CI.

In Figure 14a, Manufacturing 95% CI is [0.8004605 - 0.8107432] and the empirical value is 0.8056803.
In Figure 14b, Restaurant 95% CI is [0.7656675 - 0.7772744] and the empirical value is 0.7715832.
In Figure 14c, Media 95% CI is [0.7158272 - 0.7284739] and the empirical value is 0.7221645.

In conclusion, after all these tests we can say that the correlation coefficients are different from 0 but each sector has a different coefficient. Manufacturing has the largest coefficient, with respect to Media and Restaurant. We can say that the three sectors have positive correlations but these are different from each other.

2.2 Linear Regression AIDA

Linear regression is a very common approach to model the relationship between two variables. The method assumes that there is a linear relationship between the dependent variable and the independent variable, and finds a best fit model for this relationship. Linear regression can then be used as a predictive model, whereby the model can be used to predict a y value for any given x, in our case the model can be used to predict Employee by Revenue.
We assume that there is linear relationship between Employee and Revenue.

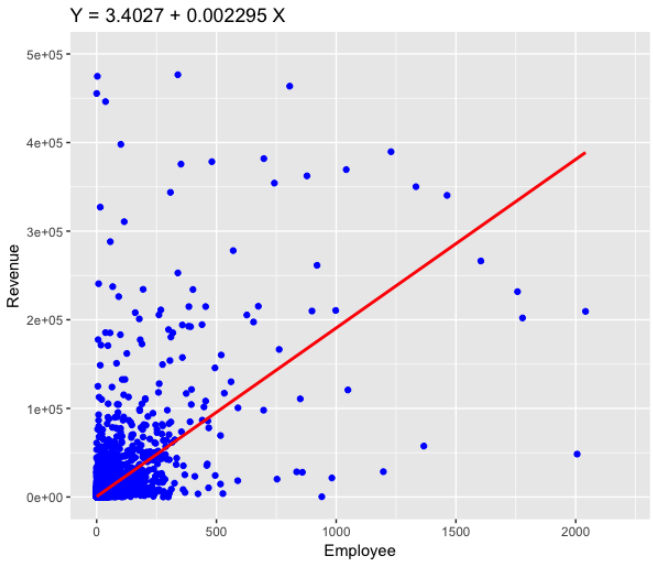


Figure 15: Linear Regression

In Figure 15 the estimated linear model has an **Intercept** equal to 3.402714 and a **Slope** equal to 0.002295.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.403e+00  9.308e-01   3.656 0.000257 ***
R           2.295e-03  1.735e-05 132.247  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 173.9 on 34998 degrees of freedom
Multiple R-squared:  0.3332,    Adjusted R-squared:  0.3332
F-statistic: 1.749e+04 on 1 and 34998 DF,  p-value: < 2.2e-16
```

Figure 16: Results for Linear Regression

In the summary statistics in Figure 16 we can see the p-value for statistical significance. In Linear Regression, the Null Hypothesis is that the coefficients associated with the variables is equal to zero. The Pr(>|t|) are less then Level of Significance 0.05 and this means that the coefficients are significantly different from zero. The Null Hypothesis is rejected.

The *R-squared* implies that the regression equation explains 33.33% variation of observed values around mean. If we add other variable in the model, R-squared increases.

After estimating the model it is necessary to verify the assumptions for the linear regression model. First of all we have to test that the mean errors is not significantly different from zero with the t-Student (Figure 17).

```
        One Sample t-test

data:  residui
t = 1.6962e-14, df = 34999, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.821675  1.821675
sample estimates:
   mean of x
1.576481e-14
```

Figure 17: T-Student test

Then we have to test the Normality distribution for errors with the Anderson-Darling test:

```
        Anderson-Darling normality test

data:  residui
A = 12190, p-value < 2.2e-16
```

Figure 18: Anderson-Darling test

We continue testing Homoscedasticity of residuals with Breusch-Pagan and the absence of Serial Correlation between residual with Durbin-Watson test:

```
        studentized Breusch-Pagan test

data:  modello
BP = 23931, df = 3, p-value < 2.2e-16
```

(a) Breusch-Pagan test

```
        Durbin-Watson test

data:  modello
DW = 2.002, p-value = 0.5741
```

(b) Durbin-Watson test

Figure 19: Homoscedasticity and Serial Correlation tests

Some test gave positive results (T-Student and Durbin-Watson) but others gave negative results (Anderson-Darling and Breush-Pagan). This means that our model is not good and we should use another one or other coefficients to predict the Employee number.

3 Firms Distribution

The aim of this chapter is to search, if any, for the (family of) distribution that best fits our data. We've decided to use a standard approach, by testing the fits of the data with 7 different distribution: Normal, Log-Normal, Gamma, Weibull, Log-Logistic, Exponential and Pareto.
We have used MLE to estimate parameters for each distribution and then we have done tests of Significance for it. MLE is a method of estimating the parameters of a statistical model, given observations. MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a maximum likelihood estimate.
Once we obtained the parameters for a distribution, we have done the Kolmogorov-Smirov Test between Empirical Distribution and Theoretical Distribution. Our test is based on 2 hypothesis:

- H0: the Empirical data are consistent with being drawn from that distribution;
- H1: the Empirical data are not consistent with being drawn from that distribution.

3.1 Firm Size Distribution: AIDA

These steps have been performed at different sample sizes. Through the test, we've found out that with decreasing the sample size, P-Values usually increase and AIC values decrease. It's been hard to find a large sample with a valid P-Value and good AIC. For this reason we have done these steps starting from all dataset to a sample of length 200. In the Tables 1, 2 e 3 results for AIC , P-Value and D are summarized.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	4324122	1852628	2179490	2035016	2366053	2003884	**1417141**
	R	8757437	5132371	**4964636**	4975135	22210677	5456659	5004718
10000	E	106119.50	53081.23	61169.42	58022.75	65342.82	57304.06	**40470.71**
	R	223300.7	145405.1	**141933.5**	142161.8	505809.2	155934.1	143008.7
1000	E	8986.035	5122.780	5757.543	5568.290	5929.781	5599.275	**3948.155**
	R	20652.44	14362.97	**14126.07**	14153.60	36129.38	15638.22	14368.28
500	E	4918.524	2516.332	2928.886	2777.796	3109.057	2684.933	**1812.464**
	R	11231.452	7161.005	**6913.384**	6957.492	29162.472	7572.172	6922.925
300	E	3753.497	1586.121	1970.129	1783.270	2284.622	1694.785	**1193.444**
	R	6871.286	4370.065	**4210.365**	4236.522	21857.045	4604.422	4214.115
200	E	2171.5261	1019.0436	1225.8569	1139.0324	1359.8834	1069.1187	**715.9961**
	R	4414.178	2895.272	**2807.983**	2822.809	11812.777	3070.882	2810.430

Table 1: Aida AIC values

For the AIC values (Table 1) , we can see that **Employee** Distribution can be explained by **Pareto** Distribution at different granularity levels. Instead the **Revenue** Distribution can be well explained by **Gamma** Distribution. For Revenue, AIC values are very similar (e.g. for *1000* records, the best AIC value is for Gamma (14126.07), but Weibull Distribution has a AIC value little bigger (14153.60) than Gamma, followed by LogNormal Distribution). With AIC values being very similar,we cannot determine with certainty which Distribution is the best.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
1000	E	0	0	0	0	0	0	0
	R	0	2.220446e-16	3.025358e-13	**1.494792e-09**	0	0	0
500	E	0	0	0	0	0	0	0
	R	0	6.228035e-09	1.602607e-12	**1.110453e-05**	0	0	0
300	E	0	**1.110223e-16**	0	0	0	0	0
	R	0	6.343246e-07	3.025946e-11	**0.001055125**	0	0	0
200	E	0	9.992007e-16	**1.409983e-14**	0	0	0	0
	R	0	0.001103599	1.626096e-06	**0.01906872**	0	0	2.618910e-09

Table 2: Aida P-values

For the P-Value in (2), we obtain a significant test for Revenue with a sample length 300 records. With a Significance Level 0.01 and this tiny sample size, we may assume that Revenue has a Weibull Distribution(with parameters *scale*=386.28 - 95%CI [273.6465, 519.689] and *shape*=0.365 - 95%CI [0.3364, 0.4], while for Employee we don't get any significant distribution.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	0.4664771	0.2691004	**0.2557069**	0.2947443	0.4036486	0.5000000	0.4763232
	R	0.48352161	0.13681492	0.17346340	**0.09600427**	0.35812671	0.50717213	0.23886476
10000	E	0.4295857	0.2697283	**0.2497772**	0.2886437	0.3899272	0.5000000	0.4789000
	R	0.45183950	0.13767793	0.14658112	**0.09889393**	0.35853462	0.50716786	0.23733258
1000	E	0.3882696	0.2639319	**0.2346117**	0.2727701	0.3320875	0.5000000	0.4630000
	R	0.4276174	0.1357360	0.1214903	**0.1025047**	0.3690231	0.5071779	0.2631165
500	E	0.4132818	0.2959946	**0.2619111**	0.3037062	0.4101220	0.5000000	0.5220000
	R	0.4276174	0.1357360	0.1214903	**0.1025047**	0.3690231	0.5071779	0.2631165
300	E	0.4506986	0.2757033	**0.2519397**	0.3213587	0.5115818	0.5000000	0.4666667
	R	0.4440810	0.1579233	0.2037742	**0.1121550**	0.3441738	0.5072740	0.2578178
200	E	0.4275803	0.2969875	**0.2854264**	0.3229250	0.4643375	0.5000000	0.5350000
	R	0.4334398	0.1369519	0.1872330	**0.1078524**	0.3424419	0.5072985	0.2261286

Table 3: Aida D values

The D values in Table 3 follow a Gamma Distribution for Employee and a Weibull Distribution for Revenue. Weibull distribution has a D value that is smaller with respect to other distribution, instead Gamma Distribution has a D value very similar to other distributions.

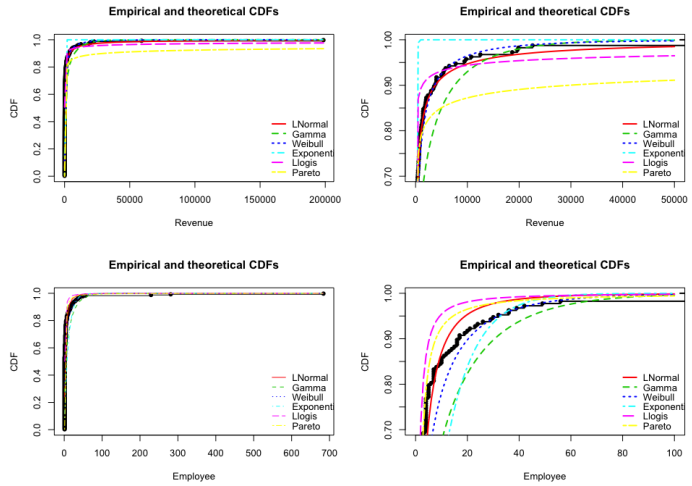


Figure 20: CDF for Revenue and Employee

One graphical technique is the CDF plot of the empirical cumulative distribution function for the data. The empirical cdf F(x) is defined as the proportion of X values less than or equal to x. This plot is useful for examining the distribution of a sample of data.We can overlay a theoretical cdf on the same plot to compare the empirical distribution of the sample to the theoretical distribution.

In the Revenue CDF in Figure 20 we can see that Empirical Distribution follows Weibull Distribution better but LogNormal distribution isn't distant from the Empirical one either. Looking at AIC values in Table 1 we can see that LogNormal, Weibull and Gamma Distributions have similar values, and this could be the reason why this distribution fit the Empirical data well.

For Employee Distribution,as we have already said, it isn't possible identify a significant fit because all P-Values are less α=0.01. Looking at CDF in Figure 20, we can see that Empirical Distribution is fitted for a Weibull/Log-Normal distribution.

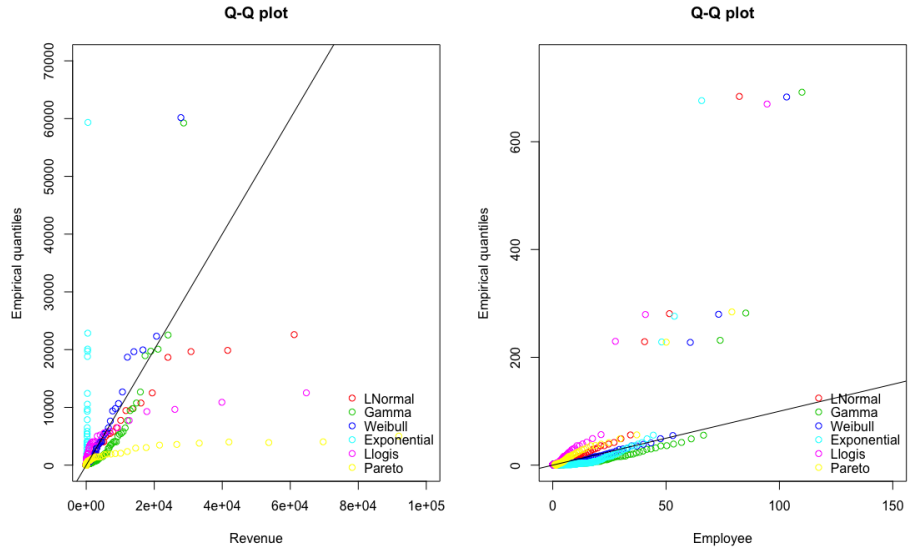


Figure 21: Q-Qplot for Revenue and Employee

Another graphical technique for determining if two data sets come from populations with a common distribution is QQplot. They can also be used to compare the distributions of one set of values with some theoretical distribution. If the values being plotted resemble a sample from a specific distribution, they will lie on a straight line.

In Figure 21, we can see that the Empirical Quantile doesn't fit Theoretical Quantile. Maybe a similar but not exact distribution is Weibull for Revenue and Employee.

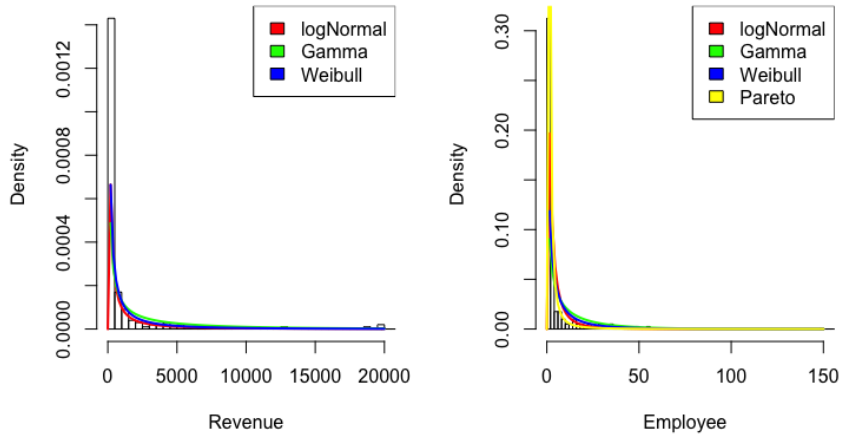
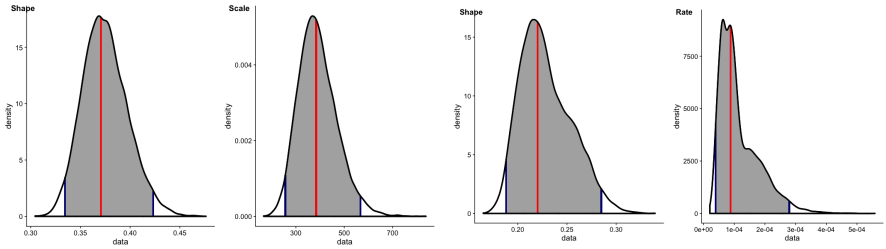


Figure 22: Histograms and Density for Revenue and Employee

A third graphical technique is to use histogram and density. In Figure 22 we can see that Employee is well fitted by Pareto Distribution. Instead for Revenue, Weibull and Gamma are good fits. Maybe Gamma Distribution is the best fit also looking at the AIC value.
In Figure 22 we show the closest distributions to our data, here on Revenue we've used Weibull with parameters: shape=0.371[0.3344109 0.4230618], scale=384.016[256.1458 567.1988]) (Figure 23a), and Gamma with parameters: shape=0.22[0.1882148 0.2848943], rate=8.827971e-05[0.0000390436 0.0002793919]). (Figure 57b)



(a) Weibull (b) Gamma

Figure 23: Confidence Interval for parameters of fits over Revenue

In summary we can say that Pareto yields the best fit on Employees, but with no Statistical significance. For Revenue we can say that Weibull Distribution seems the best family, but Gamma and Log-Normal also provide a very similar fit.

3.2 Firm Size Distribution By Sector

In this section we report the results we've seen over distinct sectors of our dataset. We've decided to use three different sectors to be compared, with the idea of using sectors that should be uncorrelated from each other, to see if there emerges some evident difference in their distributions. For this purpose, we've decided to use Manufacturing, Media and Restaurant(Ho-Re-Ca) sectors. For each sector we've proceeded the same way as shown on AIDA, trying to test which family of distribution eventually fits our data, and which yields the best results(in terms of P-Values, AIC scores and KS-D statistic).

3.2.1 Manufacturing sector

Manufacturing sector originally contains 1,127,003 data from 170,721 distinct firms.

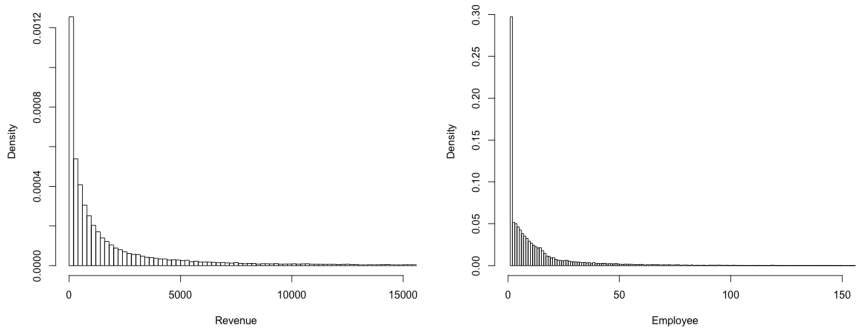


Figure 24: Histograms for Revenue and Employee

Its values of Revenue range between 0 and €20,837,757 and its Employees number range between 0 and 33636. 24

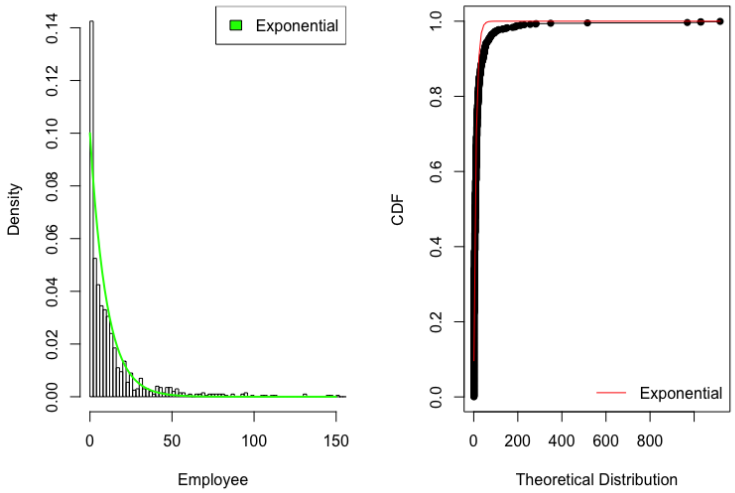


Figure 25: Histograms and CDF for Employee

For what concerns kurtosis, its value is 25869.22 over Revenue and 14287.03 over Employees, indicating the presence of very heavy tails; while skewness values are respectively 130.6277 and 85.18966.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	4392965	2544044	2730771	2643185	3178247	2906486	**2432501**
	R	8956722	6039703	6036602	**5943239**	434026877	6636684	6442627
10000	E	127918.81	73017.32	78833.96	76002.48	93984.70	83404.67	**69891.11**
	R	238850.7	172355.9	171478.7	**169531.4**	10657152.3	189357.3	183813.0
1000	E	11432.589	7287.197	7759.734	7552.420	8790.170	8367.784	**7032.369**
	R	24584.26	17200.47	17222.18	**16935.87**	1330074.37	18868.38	18311.74
500	E	5195.355	3555.042	3736.561	3666.924	4017.785	4079.903	**3404.473**
	R	12576.232	8655.368	8702.613	**8543.531**	887856.675	9467.743	9184.981
300	E	3204.231	2180.282	2301.930	2253.031	2579.751	2478.185	**2068.139**
	R	6852.294	5248.811	5234.197	**5176.447**	339263.604	5796.394	5636.637
200	E	1908.218	1446.208	1481.488	1470.778	1530.434	1698.583	**1443.998**
	R	4862.765	3473.749	3477.281	**3420.894**	251378.128	3826.546	3719.662

Table 4: Manufacturing AIC values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
1000	E	0	9.992007e-16	0	0	**1.276756e-14**	0	0
	R	0	0	0	**2.002571e-08**	0	0	0
500	E	0	5.240689e-10	5.**405735e-10**	1.425271e-11	4.789477e-10	0	0
	R	0.	1.260985e-10	3.552714e-15	**7.119650e-04**	0	0	0
300	E	0	**2.889664e-06**	6.026482e-07	7.618183e-08	6.052376e-07	0	0.
	R	0	1.059766e-04	9.440008e-06	**1.024693e-01**	0	0	0
200	E	1.110223e-16	2.384545e-03	1.852429e-02	7.356452e-03	**3.711021e-02**	0	1.070502e-09
	R	0.	0.0005704606	0.0004399813	**0.2057351094**	0	0	0

Table 5: Manufacturing P-values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	0.4339874	**0.1390639**	0.1643181	0.1740539	0.1413203	0.5000000	0.2364829
	R	0.47177415	0.15246529	0.14633356	**0.08299381**	0.85337668	0.57285472	0.38518182
10000	E	0.4370818	**0.1379583**	0.1685695	0.1769359	0.1395374	0.5000000	0.2347000
	R	0.44325756	0.15400359	0.12853075	**0.08237626**	0.85395899	0.57370379	0.38591905
1000	E	0.3930126	0.1326277	0.1542209	0.1648994	**0.1278374**	0.5000000	0.2230000
	R	0.44990448	0.16613302	0.15743044	**0.09596717**	0.84658686	0.56634769	0.37971541
500	E	0.3553268	0.1485347	**0.1484302**	0.1602099	0.1488374	0.5000000	0.2440000
	R	0.44951742	0.15325503	0.18425484	**0.08911021**	0.85161092	0.57173795	0.38500413
300	E	0.3528301	**0.1497081**	0.1581933	0.1687369	0.1581708	0.5000000	0.2533333
	R	0.39860144	0.12809787	0.14296678	**0.07037051**	0.86502784	0.58330865	0.39072561
200	E	0.30788066	0.12972947	0.10818756	0.11837783	**0.09983742**	0.50000000	0.23102102
	R	0.44541025	0.14284794	0.14510277	**0.07538601**	0.86179262	0.58354123	0.39307070

Table 6: Manufacturing D values

As seen on AIDA, the p-values tend to increase by decreasing sample size used. Here, Weibull distribution always yields the best results for Revenue both in terms of AICs and p-values; by the way, AICs of Gamma and lognorm distributions are very close to the ones yielded by Weibull. With regard to D values, they seem interesting as they are relatively low for Weibull even with larger sample sizes, while they tend to rise for Gamma and lognormal distributions. Referring to Employees, Pareto always shows the best results for AICs(lognormal and, then, Weibull have close AICs), while in terms of p-values Exponential, Lognormal and Gamma distributions seem to be the best, even if their results are very poor. (See table 4)
In Figure 25 we show the results of the Exponential(exp=0.1, 95% CI [0.1, 0.116]) distribution over E, with sample size=1000.

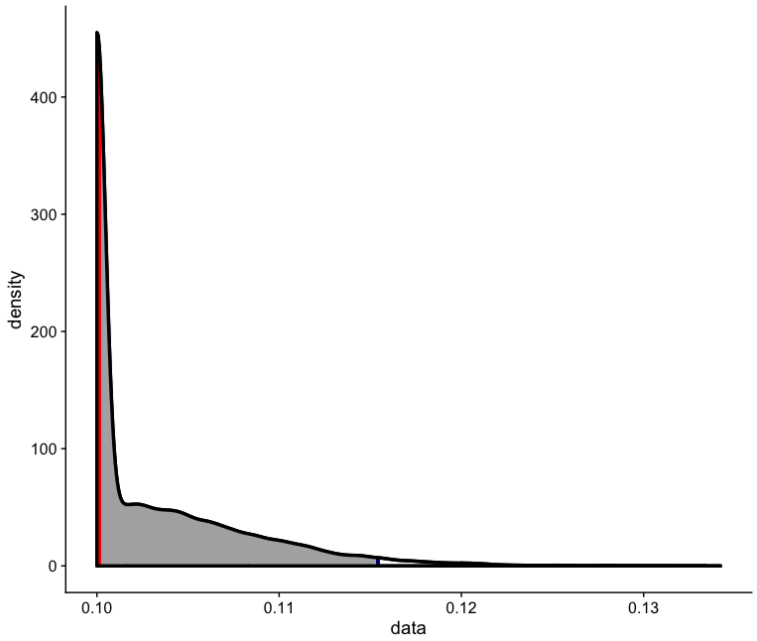


Figure 26: Confidence Interval for the Exponential parameter of 25

3.2.2 Media sector

Media sector originally contains 312,344 data from 56,125 distinct firms.

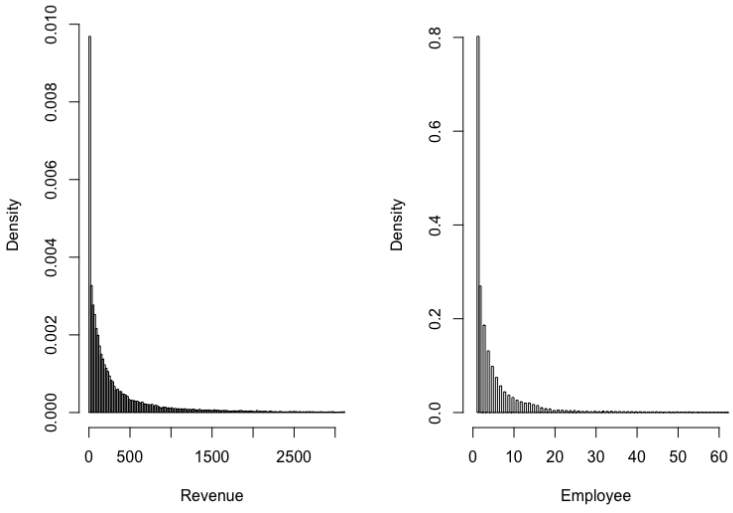


Figure 27: Histograms for Revenue and Employee

For what concerns kurtosis, its value is 22265.77 over Revenue and 25430.52 over Employees, communicating a huge significance of tails, even higher than the other sectors; while skewness values are respectively 135.425 and 147.046.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	2075535.3	783356.9	947980.3	870831.5	1033218.2	862673.4	**625740.7**
	R	3833075	2073115	2143190	**2054891**	8295872	2274594	2204987
10000	E	116854.25	52082.61	62118.80	57867.94	66734.00	57292.69	**41481.80**
	R	227306.9	137979.9	141081.5	**136678.1**	413383.4	151420.5	146797.6
1000	E	8421.369	4993.674	5592.048	5442.418	5688.805	5516.431	**3911.537**
	R	19462.56	13594.89	13590.90	**13417.44**	25686.02	14884.87	14421.15
500	E	5563.200	2572.487	3105.286	2884.393	3356.038	2817.753	**2028.589**
	R	10805.250	6858.090	7029.180	**6800.476**	20054.434	7538.207	7314.560
300	E	2383.775	1435.307	1606.946	1570.329	1621.933	1583.435	**1094.171**
	R	5589.829	3976.465	3950.595	**3915.270**	7199.750	4304.418	4150.494
200	E	2360.0995	1117.8246	1339.7728	1240.0763	1522.8590	1197.8635	**876.8221**
	R	4420.557	2859.404	2931.629	**2843.527**	11567.337	3133.908	3040.407

Table 7: Media AIC values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
1000	E	0	0	0	0	0	0	0
	R	0	0	1.134204e-12	**9.877149e-06**	0	0	0
500	E	0	0	0	0	0	0	0
	R	0	6.341832e-10	0	**1.105215e-04**	0	0	0
300	E	0	0	**2.464695e-14**	0	0	0	0
	R	0	4.852168e-06	2.**076767e-03**	1.074433e-03	0	0	0
200	E	0	6.**798186e-08**	1.909584e-14	2.220446e-16	0	0	0
	R	0	3.215937e-05	1.983141e-09	**4.270668e-03**	0	0	0

Table 8: Media P-values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	0.4828456	**0.2160859**	0.2538080	0.2935548	0.4020074	0.5000000	0.4004067
	R	0.48926591	0.13914077	0.22658549	**0.09016046**	0.29706436	0.51977372	0.35039180
10000	E	0.4553837	**0.2152657**	0.2491370	0.2911236	0.3805473	0.5000000	0.4007000
	R	0.46932933	0.13930311	0.19340924	**0.08692204**	0.29566869	0.52215430	0.35265954
1000	E	0.3719394	0.2359307	**0.2262489**	0.2656703	0.2946543	0.5000000	0.4240000
	R	0.41973621	0.14529912	0.11873977	**0.07816148**	0.29216772	0.50438163	0.33539822
500	E	0.4397224	**0.2132556**	0.2579852	0.2994338	0.4099030	0.5000000	0.4020000
	R	0.44826025	0.14789128	0.20152426	**0.09901236**	0.27738750	0.53477305	0.36398227
300	E	0.3630902	0.2653904	**0.2310444**	0.2706819	0.2930147	0.5000000	0.4600000
	R	0.3910738	0.1467948	**0.1070054**	0.1120202	0.2920201	0.4869634	0.3325420
200	E	0.4297650	**0.2073474**	0.2840897	0.3021988	0.4994924	0.5000000	0.4100000
	R	0.4377964	0.1661171	0.2276606	**0.1239872**	0.3304284	0.5563692	0.3831279

Table 9: Media D values

Over R, the results are almost the same as the Manufacturing ones, in fact the Weibull distribution seems to fit best both in terms of AIC and p-values; Gamma and lognorm also seem to fit pretty well with respect to other distributions. With regard to E, the results are very poor in terms of p-values, while with AICs Pareto distribution always seem to be the fit best. (Table 7).
We show the results on Revenue, in terms of histograms and ECDF for the Weibull distribution used(scale=393.389 - 95% CI [273.015, 563.244]), shape=0.399 - 95% CI [0.352, 0.479]) with sample size=1000.

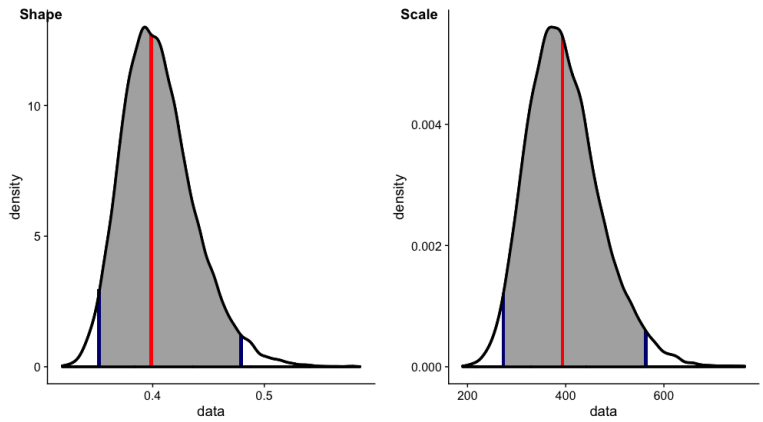


Figure 28: Confidence Interval for Weibull parameters

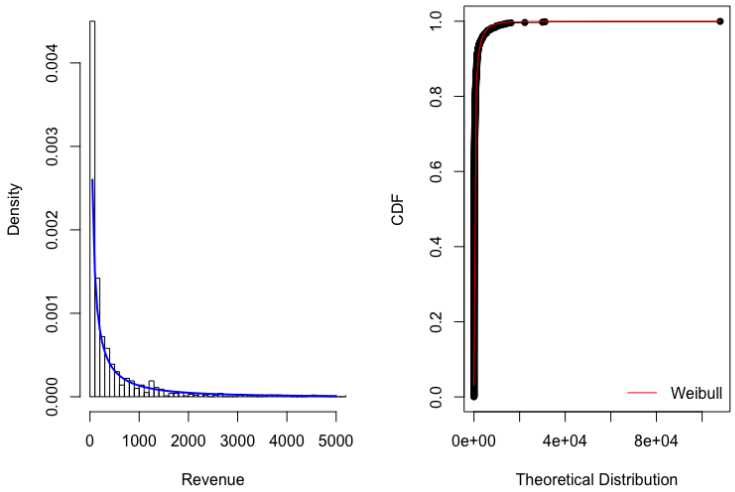


Figure 29: Density and CDFs of Weibull distribution fitted on Revenue

3.2.3 HO-RE-CA sector

HO-RE-CA sector originally contains 345,153 data from 73,211 distinct firms.

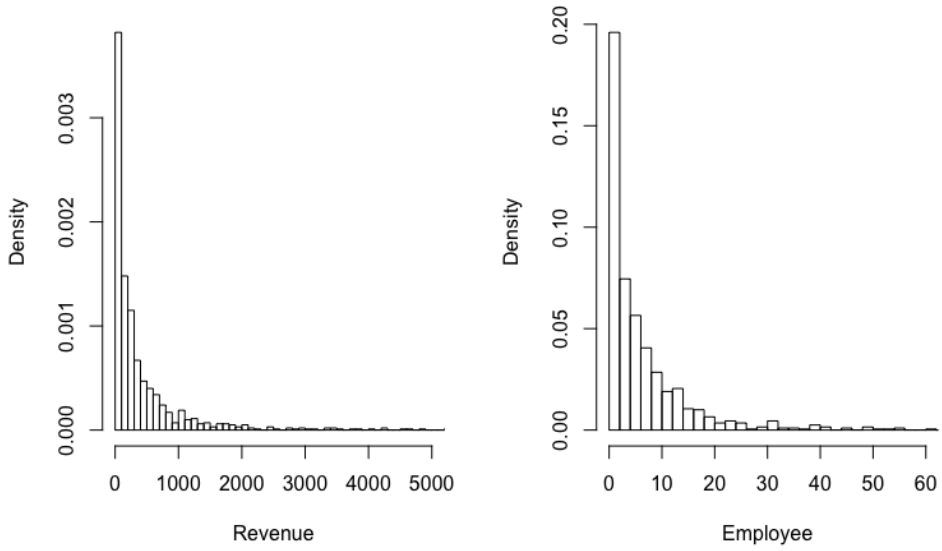


Figure 30: Histograms for Revenue and Employee

For what concerns kurtosis, its value is 11872.49 over Revenue and 5394.51 over Employees, referring a heavy significance of tails; while skewness values are respectively 92.511 and 68.979.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	1797366.2	871098.2	961995.1	924919.2	987195.8	997686.0	**779927.3**
	R	2983052	1987419	1952268	**1941339**	3177624	2149664	2068310
10000	E	134260.57	58066.81	66783.59	62356.18	70505.53	66379.27	**51824.08**
	R	219442.0	131991.5	131160.7	**129107**.2	249195.4	142566.3	137071.4
1000	E	11545.437	5887.523	6528.182	6261.586	6723.897	6747.819	**5312.105**
	R	20657.21	13451.84	13301.67	**13143.72**	23647.81	14599.15	14076.00
500	E	6930.719	3007.771	3609.581	3282.891	4079.497	3391.730	**2659.973**
	R	11192.155	6716.349	6794.289	**6594.764**	18926.559	7259.258	6987.295
300	E	2667.906	1781.565	1894.457	1863.555	1916.992	2033.982	**1598.626**
	R	5709.371	3956.972	3904.409	**3882.404**	7142.723	4252.081	4080.461
200	E	1430.649	1182.941	1208.338	1207.367	1206.358	1388.543	**1116.079**
	R	3115.977	2802.326	**2685.065**	2709.806	3609.381	3070.528	2973.951

Table 10: Ho.Re.Ca AIC values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
1000	E	0	0	0	0	0	0	0
	R	0	0	**1.923939e-11**	1.687539e-14	0	0	0
500	E	0	0	0	0	0	0	0
	R	0	0	0 **4.019149e-08**	0	0	0	
300	E	0	6.681322e-13	**1.365505e-08**	1.431419e-10	5.551115e-16	0	0
	R	0	4.006131e-10	**3.376111e-04**	6.334591e-06	0	0	0
200	E	3.207944e-09	6.417430e-05	**1.823888e-03**	7.041753e-04	1.408814e-03	0	3.611667e-12
	R	4.638472e-09	1.550878e-08	**1.556512e-02**	7.625658e-03	0	0	0

Table 11: Ho.Re.Ca P-values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	0.4828456	**0.2160859**	0.2538080	0.2935548	0.4020074	0.5000000	0.4004067
	R	0.48926591	0.13914077	0.22658549	**0.09016046**	0.29706436	0.51977372	0.35039180
10000	E	0.4553837	**0.2152657**	0.2491370	0.2911236	0.3805473	0.5000000	0.4007000
	R	0.46932933	0.13930311	0.19340924	**0.08692204**	0.29566869	0.52215430	0.35265954
1000	E	0.3719394	0.2359307	**0.2262489**	0.2656703	0.2946543	0.5000000	0.4240000
	R	0.41973621	0.14529912	0.11873977	**0.07816148**	0.29216772	0.50438163	0.33539822
500	E	0.4397224	**0.2132556**	0.2579852	0.2994338	0.4099030	0.5000000	0.4020000
	R	0.44826025	0.14789128	0.20152426	**0.09901236**	0.27738750	0.53477305	0.36398227
300	E	0.3630902	0.2653904	**0.2310444**	0.2706819	0.2930147	0.5000000	0.4600000
	R	0.3910738	0.1467948	**0.1070054**	0.1120202	0.2920201	0.4869634	0.3325420
200	E	0.4297650	**0.2073474**	0.2840897	0.3021988	0.4994924	0.5000000	0.4100000
	R	0.4377964	0.1661171	0.2276606	**0.1239872**	0.3304284	0.5563692	0.3831279

Table 12: Ho.Re.Ca D values

About R, Weibull distribution gives the best AIC values for every sample size, but once more these values are very close to the ones yielded by Gamma, Lognormal and Pareto distributions. In terms of p-values Gamma distribution seems to fit better than the others, but still we cannot assume that our data are consistent with being drawn from any distribution we've tested. Pareto seems to be the model that better describes our Employees data, at least in terms of AIC; as seen in the previous sectors, p-values are very low for E.

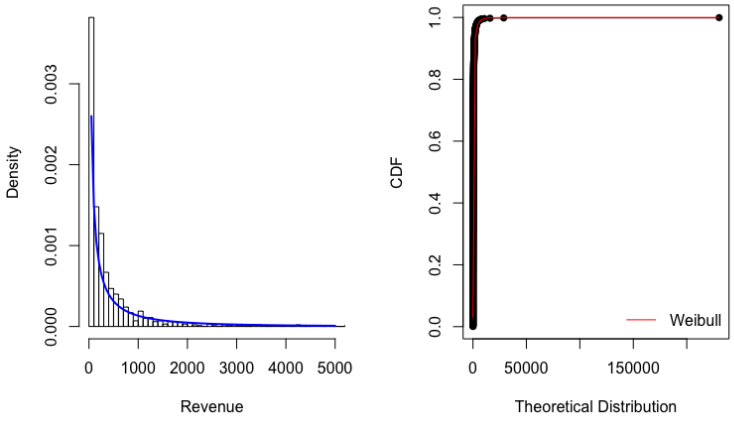


Figure 31: Histograms and CDF for Revenue and corresponding Weibull fit

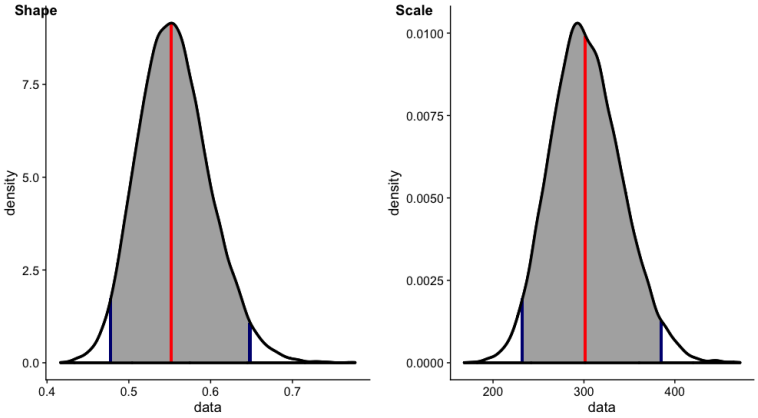


Figure 32: Confidence Interval for Weibull parameters

In (31, 32) we show the fitted Weibull(scale=301.3974 -95 %CI [231.8853, 385.0528], shape=0.5521 - 95%CI [0.4777169, 0.6478326]) distribution on R that we've seen with sample size=1000

3.2.4 Hypotesis testing over distinct Sectors

Since the results obtained by hypotesis testing over the fits of the distinct distributions are too weak to draw solid conclusions over the differences(if there) between the Sectors distributions, here we try to hypothesize that the Revenues and Employees over the three Sectors have been drawn from the same model:

- H0: the Revenues(Employees) of two distinct Sectors are drawn from the same model;
- H1: the Revenues(Employees) of two distinct sizes are not drawn from the same model.

  For both Revenue and Employee, clearly comes out that we can reject any null hypotesis (14): i.e. Manufacturing, Media and HO.RE.CA. sizes distributions come out from different population distribution function.

Sectors	D-statistic	p-value	Null Hypothesis
Manufacturing v. Media	0.35	< 2.2e-16	Rejected
Manufacturing v. HO.RE.CA	0.35838	< 2.2e-16	Rejected
Media v. HO.RE.CA	0.065501	< 2.2e-16	Rejected

Table 13: Hypotesis Testing results for R over distinct Sectors

Sectors	D-statistic	p-value	Null Hypothesis
Manufacturing v. Media	0.30148	< 2.2e-16	Rejected
Manufacturing v. HO.RE.CA	0.1746	< 2.2e-16	Rejected
Media v. HO.RE.CA	0.15363	< 2.2e-16	Rejected

Table 14: Hypotesis Testing results for E over distinct Sectors

3.3 Firm Size Distribution By Different Sizes

As for Sectors, we further analyze how R and E distributions behaves for Small, Medium and Large firms. After testing the fits of the various distribution families on each dataset, we test the hypotesis that Small, Medium and Large firms follow the same distribution.

3.3.1 Small Firms Size Distributions

These firms represent 87% of our whole dataset: their R ranges from 0 to 10,416,876, with mean=1060.789, median=159.13, 25%Quantile=14.835 and 75%Quantile=660.208, revealing a very sparse distribution with the majority of small values(in fact, its skewness=274.103, that shows very strong asimmetry), and a lot of very sparse tails, as reveals its kurtosis=169570.2. Talking about Employees, its values of course range in [0,49], with mean=3.894, median=1, 25%Quantile=0 and 75%Quantile=5, and a kurtosis=10.81.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	2387482	1676026	1806717	1790892	1812178	1851482	**1264449**
	R	7892933	4701105	4661133	**4631100**	10698331	5035368	4822318
10000	E	68214.65	48127.08	51797.44	51351.86	51957.00	53180.67	**36424.03**
	R	206855.8	133568.3	132390.6	**131637.1**	300746.4	142903.7	136751.6
1000	E	6874.720	4739.414	5138.364	5087.167	5154.176	5216.789	**3530.310**
	R	90002.55	13712.63	13538.50	**13503.82**	29701.05	14735.09	14148.84
500	E	3464.113	2422.134	2614.995	2588.095	2626.129	2659.257	**1813.564**
	R	10721.121	6677.632	6706.957	**6605.949**	19925.634	7138.565	6834.796
300	E	2094.391	1358.555	1506.222	1482.681	1515.019	1458.375	**921.347**
	R	6036.776	3908.576	3887.462	**3860.590**	9565.051	4153.021	3955.722
200	E	1377.2975	898.6057	994.3593	981.5001	996.0031	974.3693	**628.1592**
	R	9131.651	2669.765	2655.462	**2637.952**	6477.293	2851.113	2729.509

Table 15: Small AIC

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
1000	E	0	0	0	0	0	0	0
	R	0	0	**5.899129e-09**	1.432188e-14	0	0	0
500	E	0	0	0	0	0	0	0
	R	0	9.871484e-10	**7.222226e-07**	1.973601e-07	0	0	0
300	E	0	0	0	0	0	0	0
	R	0	2.980222e-08	**1.677416e-04**	4.355806e-07	0	0	0
200	E	0	0	3.774758e-15	7.982504e-14	0	0	0
	R	0	2.179901e-04	**3.208034e-02**	1.486302e-03	0	0	1.110223e-16

Table 16: Small P-values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
350000	E	0.2973952	0.2909645	0.2722711	**0.2486431**	0.3054929	0.5000000	0.4901657
	R	0.4776931	0.1525526	**0.1061919**	0.1392511	0.3416736	0.4474322	0.3027618
10000	E	0.2952949	0.2901632	0.2717650	**0.2470318**	0.3055079	0.5000000	0.4887000
	R	0.4447477	0.1558787	**0.1098319**	0.1424452	0.3356472	0.4431229	0.3002862
1000	E	0.3047839	0.2938986	0.2747265	**0.2548149**	0.3101965	0.5000000	0.4970000
	R	0.37558402	0.14663894	**0.09909997**	0.12760546	0.36143742	0.46364632	0.31216698
500	E	60.2983627	0.2959556	0.2783395	**0.2520838**	0.3211223	0.5000000	0.5000000
	R	0.4442141	0.1463877	**0.1217952**	0.1270094	0.3419182	0.4408236	0.2972653
300	E	0.3249185	0.3384150	0.3234251	**0.2852524**	0.3738516	0.5000000	0.5700000
	R	0.4202246	0.1733101	**0.1250749**	0.1598944	0.3474546	0.4221372	0.2906570
200	E	0.3243368	0.3103071	0.2910862	**0.2777204**	0.3273189	0.5000000	0.5300000
	R	0.3926654	0.1510315	**0.1016447**	0.1342070	0.3319464	0.4489915	0.3064523

Table 17: Small D

The results in Table 15 show that Pareto (for E) and Weibull(for R) always seems to fit in terms of AIC, but looking at p-values and D, Gamma seems to be the best fit for R. The results obtained, by the way, are very weak and basically not different from the ones seen previously. We can thus reject the null hypotesis for each distribution.

3.3.2 Medium Firms Size Distributions

Medium Firms are the 12.6% of the AIDA firms. Their Revenues range between 0 and 30,636,816 , with a mean of 27851.4 and a median of 11922.52 (25% Quantile=5086.34, 75% Quantile=24984.49): as for the entire dataset and for the Small firms, they are more concentrated on the top left of the Mean, and have a very high kurtosis(=8130.67). E is in [50,249], with a mean of 98.25 and a median of 82, and curiously a pretty low kurtosis(=0.775), somehow reporting a low weight of tails.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	1584252	1520527	1535511	1559450	1679299	2251147	**1480407**
	R	4096773	3323482	3339496	**3313197**	84044983	3888656	3803321
10000	E	105292.52	100979.76	102003.58	103667.26	111784.56	149923.84	**98350.82**
	R	265458.0	221652.3	222294.3	**220719.5**	5512864.7	258855.1	253164.0
1000	E	10548.329	10160.191	10248.500	10398.265	11224.940	15048.809	**9961.302**
	R	76021.79	22190.32	22510.76	**22211.92**	678134.93	25976.94	25409.55
500	E	5314.207	5101.543	5151.445	5227.086	5610.070	7513.197	**4954.984**
	R	13803.87	11161.01	11290.40	**11143.31**	354052.04	13013.35	12729.12
300	E	3124.396	2990.532	3022.700	3078.644	3331.710	4475.145	**2897.867**
	R	7267.572	6595.776	6533.157	**6517.993**	125380.541	7679.156	7511.195
200	E	2098.729	2014.445	2034.241	2067.131	2229.216	2990.489	**1948.046**
	R	5357.615	4530.250	4560.108	**4511.423**	171531.513	5214.343	5100.222

Table 18: Medium AIC

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
1000	E	0.000000e+00	**6.396838e-07**	7.403587e-10	0.000000e+00	0.000000e+00	0.000000e+00	2.121002e-08
	R	0.000000e+00	**5.548717e-06**	0.000000e+00	1.323733e-06	0.000000e+00	0.000000e+00	0.000000e+00
500	E	3.830969e-11	1.662003e-04	5.831168e-06	3.629630e-11	0.000000e+00	0.000000e+00	**1.728802e-04**
	R	0.000000e+00	**1.029844e-03**	3.336242e-11	6.311853e-04	0.000000e+00	0.000000e+00	0.000000e+00
300	E	3.996895e-07	1.890535e-03	2.051702e-04	9.654114e-08	0.000000e+00	0.000000e+00	**6.400089e-02**
	R	0.0000000000	0.0007451023	0.0410260347	**0.2539186350**	0.0000000000	0.0000000000	0.0000000000
200	E	7.644030e-05	9.785570e-03	4.795612e-03	4.280268e-05	0.000000e+00	0.000000e+00	**1.141498e-01**
	R	0.000000e+00	9.359276e-03	4.637959e-05	**3.887496e-02**	0.000000e+00	0.000000e+00	0.000000e+00

Table 19: Medium P-values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
150000	E	0.15513815	0.09221804	0.11021119	0.15682424	0.40181003	0.79350028	**0.07773013**
	R	0.44686659	0.07787607	0.11170416	**0.06411296**	0.96793190	0.71288705	0.50146462
10000	E	0.15543335	0.09131797	0.11125835	0.15824014	0.40443658	0.79381775	**0.08050215**
	R	0.42357544	0.08346756	0.11080563	**0.06570139**	0.96604090	0.71096977	0.50021589
1000	E	0.14623649	**0.08647381**	0.10420421	0.14721474	0.39950442	0.79271979	0.09581738
	R	0.44414466	**0.07998466**	0.14995483	0.08434513	0.97503941	0.71954545	0.50600188
500	E	0.15709381	0.09793020	0.11289571	0.15726554	0.39950442	0.79319230	**0.09736017**
	R	0.44168771	**0.08701434**	0.15753329	0.08978340	0.97296100	0.71831062	0.50837834
300	E	0.16034195	0.10773457	0.12372562	0.16756314	0.41677839	0.79546888	**0.07574049**
	R	0.32034088	0.11471077	0.08048485	**0.05862064**	0.97114633	0.72252515	0.51266787
200	E	0.15946902	0.11532556	0.12281288	0.16395165	0.41097857	0.79481703	**0.08460504**
	R	0.39361627	0.11580732	0.16333859	**0.09925403**	0.95727546	0.70043435	0.48963829

Table 20: Medium D

The results obtained on the hypotesis testing for fit are very similar to the Small firms ones in terms of AIC, instead they're quite interesting in terms of p-values and D statistic, since Pareto seems to be the best fit for E (yielding p-values $\geq$0.05 for sample sizes of 200 and 300).

3.3.3 Large Firms Size Distributions

Large firms just represent 0.4% of AIDA firms. Their Revenue ranges in [0; 47,813,192], with mean=265,404.9 and median=73,956.32, 25% quantile=28227.66 and 75% quantile=175,530.6: the latter is even below mean value, again showing that the values are condensed at the bottom. Its kurtosis is, again, pretty high(=550.633). Employees ranges from 250 to 223,352; their mean=946.715 and their median=434, with 25% quantile=314 and 75% quantile=755.

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
20000	E	380388.1	298754.4	313207.9	312486.4	555327.4	386275.7	**282953.1**
	R	617584.5	523683.0	527704.9	**522956.4**	104277316.4	599770.6	587057.9
10000	E	187087.7	149341.4	156230.8	156016.0	274498.0	193061.3	**141325.6**
	R	309038.3	261619.0	264024.3	**261606.0**	52939938.8	299936.9	293588.5
1000	E	18735.15	14892.27	15627.47	15590.72	27432.04	19275.08	**14071.57**
	R	31473.26	**26114.11**	26460.29	26152.50	5838736.01	29931.87	29299.58
500	E	8688.377	7508.607	7776.256	7790.451	13469.870	9687.293	**7132.805**
	R	14331.93	13083.52	12966.80	**12923.65**	1912210.70	14834.06	14516.19
300	E	5189.275	4466.281	4635.279	4644.881	7814.702	5781.770	**4220.010**
	R	9783.347	**7864.020**	8119.972	7971.348	2948754.290	9081.959	10111.433
200	E	3312.956	2928.507	3014.033	3040.770	4829.468	3865.908	**2836.754**
	R	6182.279	5295.631	5276.369	**5231.577**	1268704.007	5933.221	5805.081

Table 21: Large AIC

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
20000	E	0	0	0	0	0	0	**0.04214588**
	R	0	0	0	0	0	0	0
10000	E	0	0	0	0	0	0	**0.1846507**
	R	0	0	0	0	0	0	0
1000	E	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	**0.408663**
	R	0.000000e+00	**1.865666e-06**	0.000000e+00	5.206946e-14	0.000000e+00	0.000000e+00	0.000000e+00
500	E	0.000000e+00	2.607643e-10	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	**8.939452e-01**
	R	0.000000e+00	5.186120e-05	4.650820e-06	**1.374349e-02**	0.000000e+00	0.000000e+00	0.000000e+00
300	E	0.000000e+00	8.139564e-07	7.814083e-12	0.000000e+00	0.000000e+00	0.000000e+00	**9.892414e-01**
	R	0.000000e+00	**1.078144e-01**	0.000000e+00	3.204507e-05	0.000000e+00	0.000000e+00	0.000000e+00
200	E	4.440892e-16	1.188979e-03	4.598522e-05	1.021220e-09	0.000000e+00	0.000000e+00	**6.785279e-02**
	R	0.000000e+00	3.616685e-05	8.167230e-06	**1.358139e-02**	0.000000e+00	0.000000e+00	0.000000e+00

Table 22: Large P-values

n	size	norm	lnorm	gamma	weibull	expo	llogis	pareto
20000	E	0.417864024	0.152871512	0.216716598	0.280338561	0.918731761	0.793742338	**0.009823131**
	R	0.41605954	**0.07607228**	0.15257128	0.09595069	0.98469996	0.71944728	0.50652978
10000	E	0.40649867	0.15426419	0.21497760	0.28232921	0.91873176	0.79383173	**0.01091249**
	R	0.41580488	**0.06987113**	0.15225664	0.09878670	0.98562659	0.71972495	0.50620550
1000	E	0.40767271	0.15657757	0.22005876	0.29018897	0.91873176	0.79436458	**0.02810002**
	R	0.42984954	**0.08332179**	0.18728204	0.12506076	0.98301461	0.72257166	0.50820134
500	E	0.32843427	0.15086601	0.20221646	0.25511608	0.91873176	0.79284399	**0.02577588**
	R	0.3183700	0.1027623	0.1138930	**0.0705715**	0.9798799	0.7125773	0.5030042
300	E	0.33329677	0.15660197	0.20923769	0.26600809	0.91873176	0.79444191	**0.02562368**
	R	0.43279594	**0.06976549**	0.25740760	0.13565590	0.99648820	0.72940655	0.91361585
200	E	0.30048940	0.13627002	0.16340393	0.23127589	0.91873176	0.79376833	**0.09197177**
	R	0.3988194	0.1652310	0.1761287	**0.1117162**	0.9619463	0.6912700	0.4795413

Table 23: Large D

The results here are very interesting. Even if about Revenue they seem to be similar to the ones achieved previously, with Weibull and Gamma that as the best families, about Employees here we can see that Pareto distribution gives great results, being the best fit over any sample, for every measure: especially, if we assume a 99% CI, the null hypothesis that Employees data come from a Pareto distribution cannot be rejected. Here we show KDE, Pareto fitting and CIs(34, 33) for the sample(1000 E) that has yielded the highest p-value: its parameters are x_{min}=251 (95% CI [250.9873, 251.735]), α=1.30 (95% CI [1.22054, 1.388609]).

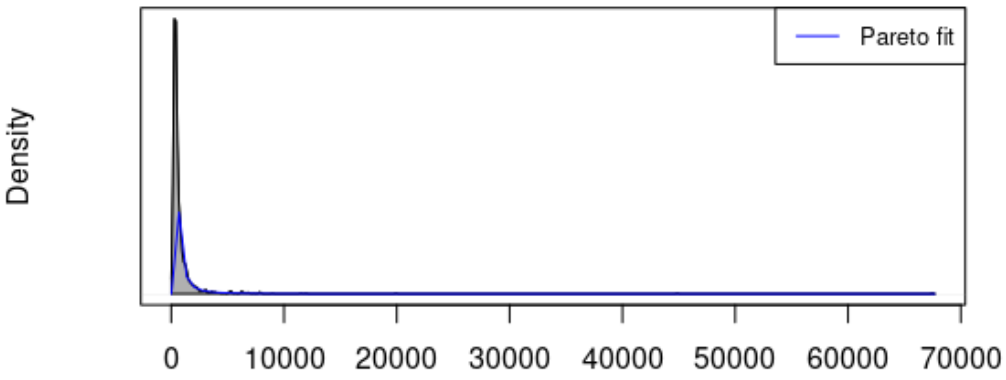


Figure 33: KDE of Large firms Employees and corresponding Pareto fit

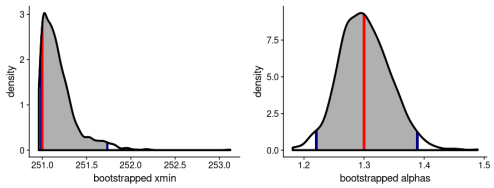


Figure 34: CIs for Pareto parameters over Large firms

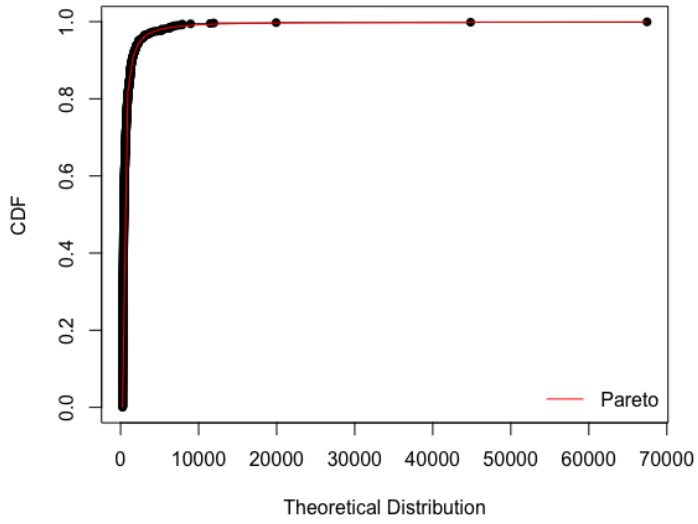


Figure 35: CDF of Large firms Employees and corresponding Pareto fit

3.3.4 Hypothesis testing over Revenue of distinct firm Sizes

Being Revenue another common measure for firm Size, it's interesting to see if the distributions of Revenues for Small, Medium and Large sizes(that we've

defined by considering their number of Employees) are drawn from the same model or not(as it should be). Thus, we've made an hypotesis testing over the Revenues for each pair of distinct sizes:

- H0: the Revenues of two distinct sizes are drawn from the same model;
- H1: the Revenues of two distinct sizes are not drawn from the same model.

We've used two sample KS-Test of the empirical data to yield the following results:

Firm Sizes	D-statistic	p-value	Null Hypothesis
Small v. Medium	0.78596	$<$ 2.2e-16	Rejected
Small v. Large	0.91146	$<$ 2.2e-16	Rejected
Medium v. Large	0.54124	$<$ 2.2e-16	Rejected

Table 24: Revenues of distinct Sizes - Hypotesis Testing

For every pair of sizes, the null hypothesis is rejected, since it has confidence$\approx$0.

4 Power Law Distribution of Firms Size

Power Law distribution seems to be a "natural" way to fix firms size data, in literature [2][3][4]; that is, with increasing firm Size(s), the number (n) of firms of Size s actually decreases (as power of s): this sort of "inverse proportion" relationship is described by Power Law probability distribution:

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{\alpha} \tag{1}$$

, where x is a continuous random variable, x_{min} is a positive number describing the lower bound for x, and α is a number(>1). If x is a discrete random variable, then the Power Law distribution is:

$$p(x) = \frac{x^{-\alpha}}{\sum_{n=0}^{\infty}(n + x_{min})^{-\alpha}} \tag{2}$$

In this section, we describe our deeper analysis made on Power Law distribution for firms Employees. First, we show different results obtained on a sample of the entire dataset, then we focus on how power law hypothesis eventually holds for distinct subsamples of data, and how maybe the alpha value changed during time. Again, here our two hypothesis are:

- H0: the (Employees) data are generated from a Power Law distribution function;
- H1: the (Employees) data don't come from a Power Law distribution function.

For each part, we clearly describe how we've handled the tricky choice of the parameters(x_{min} and α).

4.1 Results on AIDA

In contrast to what we have seen for the fits described in the previous chapter, the results (in terms of p-values for KS Test) reached by the Power Law distribution seem to be less linked to the sample Size; nonetheless, they are very related to the behaviour of the two parameters x_{min} and α. First of all, we've used the estimate of x_{min} by choosing the value of x_{min} that makes the probability distributions of the measured data and the best-fit power-law model as similar as possible above x_{min}, as described in [5]; this is done by minimizing the KS-statistic D. In our case, by using more and more data, we've seen that the estimated x_{min} parameter usually increases: this is probably due to the potentially higher number of big values for Size(Revenue or Employee) column (that is: by choosing larger sample sizes, it's more likely to have larger values with respect to smaller sample sizes). In this way, the number of potential tails is very high, and so the value for an appropriate x_{min} to fit a Power Law is increased. In the pictures of Figure 36 we show how x_{min}(and, thus, α) values change by increasing sample Size. For each size chosen, we've generated 2000 bootstrapped samples in a non-parametric way from the original dataset, to estimate the parameters for our distribution.

As emerges from Figure 36, the interval of values for x_{min} gets wider, as well as its most likely value(marked by a vertical line in blue). For what concerns alphas, they also seem to increase with increasing sample size; by the way, this

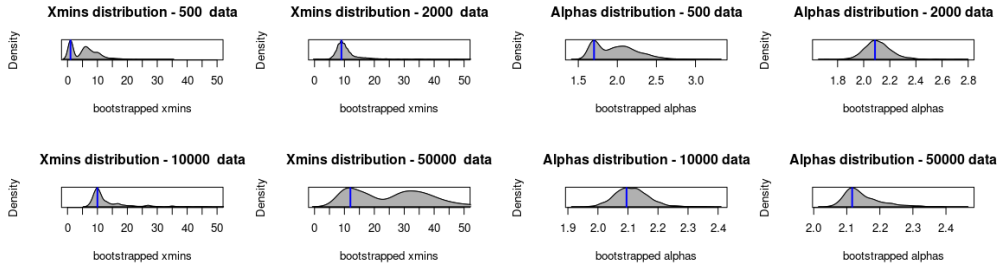


Figure 36: x_{min} and α distribution over bootstrapped samples of increasing Sizes

is more consequent to the x_{min} value: with lower x_{min}s, alpha is usually lower, as shown in Figure 37. Below, we also show a real example of such behaviour, obtained on a sample of 50k data. (Figure 38)

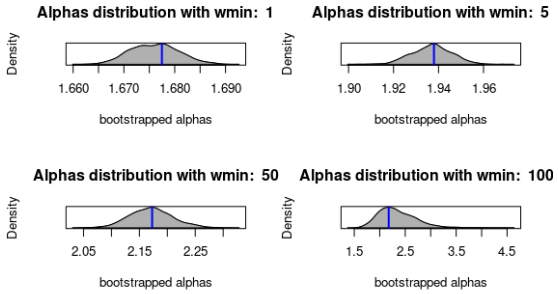


Figure 37: Different shapes by increasing x_{min}

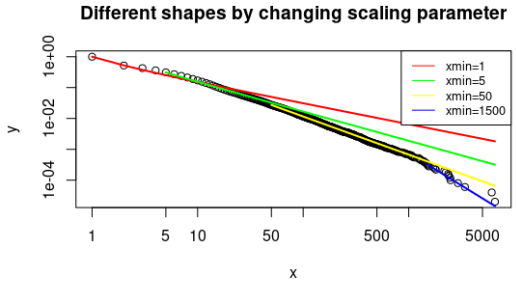


Figure 38: Different shapes by increasing x_{min} - an example

Finally, in Figure 39 we show the results reached on an AIDA sample (size=2000). Similarly to the steps of the previous chapter, we've first estimated the parameter x_{min} by minimizing KS-D statistic, then we've estimated α by MLE, finally we've tested the fit of the Power Law distribution by KS-Test, by using a bootstrapping procedure, as suggested in [5]. Here we've reached a p-value of 0.56, by using parameters x_{min}=11 (95% CI: [6, 24]) and α=2.088 (95% CI: [1.936, 2.339]). It

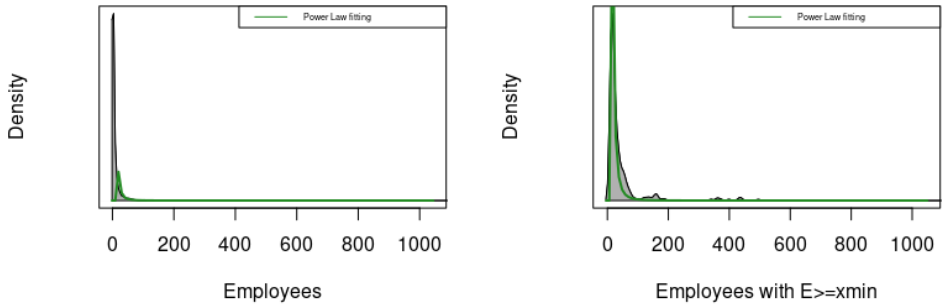


Figure 39: Employees distribution from minimum x and with x$\geq$ x$_{min}$ - an example

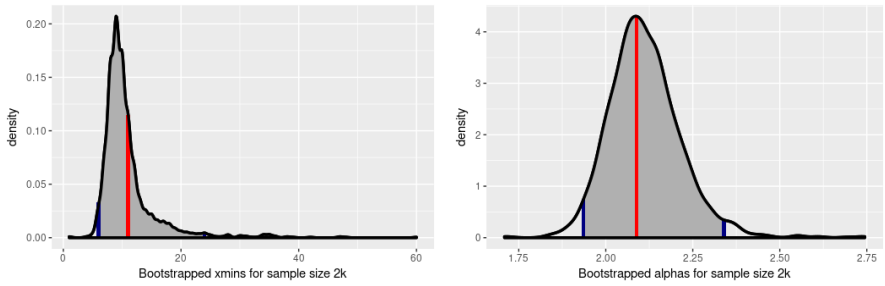


Figure 40: Confidence intervals for x$_{min}$ and α with sample size 2000

may look strange that p-value reaches such a high value, if we see picture one in Figure 39, but it seems very reasonable if we look at the second picture, that only takes into account Employees$\geq x_{min}$; this particular result seems to be common in our data, so that with higher values of x_{min} usually Power Law hypothesis seems (more) reasonable: it's surely caused by a (great) reduction in data size, but it's also linked to the fatness of tails. This is confirmed by high value in kurtosis(that is 296960 on the whole dataset), that describes a very high weight of tails in comparison to the normal distribution [6]; because of this, and starting from the results obtained in the previous chapter, we've decided to analyze the Employees distributions on the different firm Sizes: Small, Medium and Large.

4.1.1 Results on Small Firms

These firms represent 87% of our whole dataset. Because of the large difference in dimensions with respect to the Large firms(0.4% of the entire data), we've decided to use the entire sample of Large firms(29000), and we've chosen a sample of the same size for Small and Medium firms. To avoid bias due to random choice of subsamples from original data, we've generated confidence intervals for both the parameters and we've checked that the samples we've been working on have

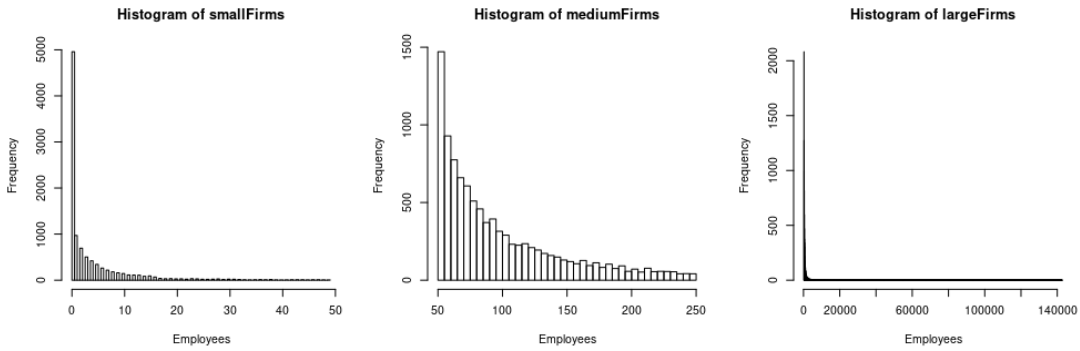


Figure 41: Employees histograms of Small, Medium and Large sizes

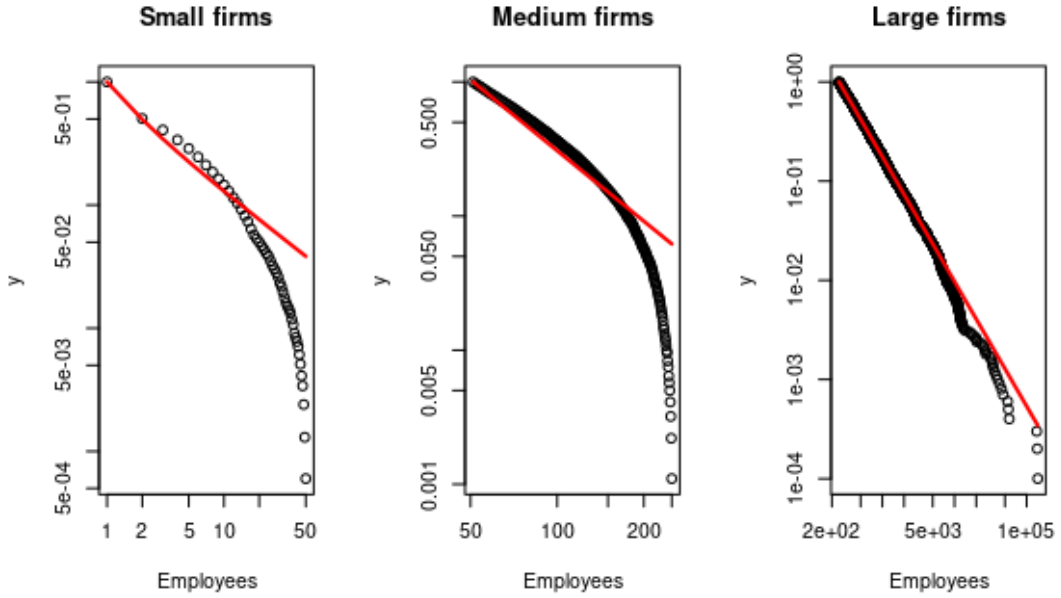


Figure 42: ECDF of Small, Medium, Large firms - Power Law fits for minimum x_{min}

parameters that fall into 95% CI (x_{min} - 95% CI [1, 14]; α - 95% CI [1.728, 3.1]). Unfortunately, by testing over distinct Small Firms samples, we've reached very poor results, that's probably due to pretty high x_{min} values (see Figure 43), cutting out the majority of original data(Figure 2b). By the way, even by further analyzing only the subsets of firms having Employees$\geq x_{min}$, the results are very poor.
This can be explained by the visible increase of α with increasing x_{min}s (Figure 44 and 45), i.e: the distribution of Employees values becomes more and more sparse as x_{min} rises up; in fact, similarly to what happens on the whole dataset, kurtosis value is quite high($\approx$10.82), describing a high weight of tails. Anyway, the only conclusion we can get for Small Firms is that they don't seem to follow a Power Law distribution.

4.1.2 Results on Medium Firms

Medium firms Employees are distributed with a kurtosis ≈ 0.775 and a skewness ≈ 1.25. The number of Medium firms , as for the Small ones, also decreases

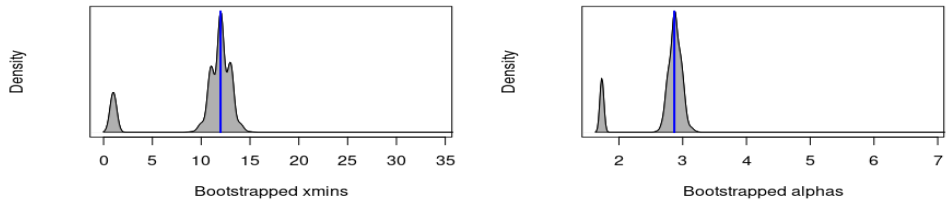


Figure 43: x_{min}s and α bootstrapped distributions for Small Size firms over sample size:29000

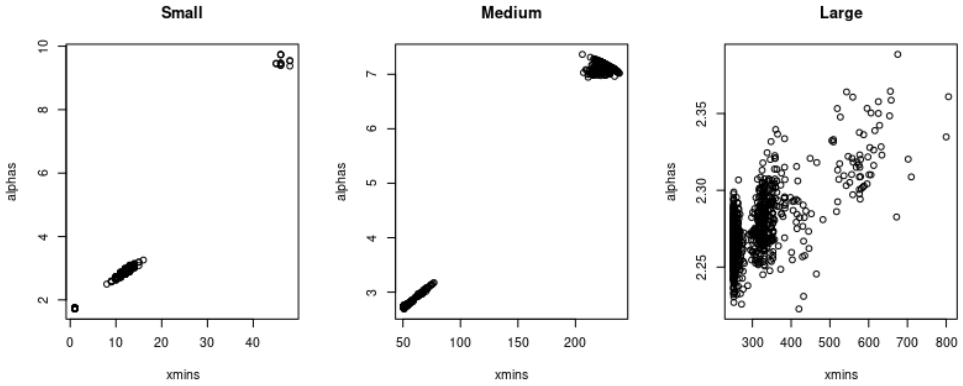


Figure 44: Distribution of x_{min} and α for Small, Medium and Large firms

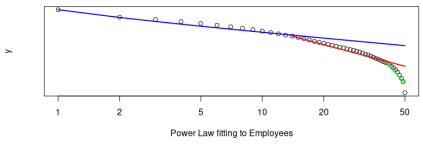


Figure 45: Increase of α by raising up x_{min} on a Small firm

with the growth of their number of Employees (Figure 41); it's therefore potentially confident with a Power Law distribution. But nevertheless due to its (small) kurtosis and skewness it is not as immediate to state that. Because of the usage of a subsample to analyze the Medium firms Employees distribution, as for the Small firms, we've first estimated the CIs of Power Law parameters through a non-parametric bootstrap, ensuring to avoid bias linked to the choice of sample, getting the following results: x_{min}=72 - 95% CI [51, 232.025]; α=3.089 - 95% CI [2.711, 7.237]; as you can see, they're quite large, since the feasible values for x_{min} range from 51 to 249 for Medium firms. This result is clearly shown in Figures 42 and 48: there are no xmins in [78, 205], while its value seems very likely to be around its minimum possible value(50) or in [200,249]. This is someway trivial if we look at Figure 42: there seems to be a "natural" split into two intervals: the first one starting at x_{min}=51, with lower values of α, the second one starting around 170, with quite larger values of α, and this is what really happens in Figure 44.

That said, it's not surprising that the results seen over the entire Medium firms sample are very poor: KS test, as with Small Firms, yields a p-value=0.

We've then decided to analyze separately the two intervals: Employees in [50, 170] and Employees in [171,249]. The results obtained are very good, with a p-value of 0.94 reached for the second interval - with estimated parameters: x_{min}=228 (95% CI [172.475, 233.625]), α=7.157 (95% CI [6.407, 7.502]; and a p-value of 0.2 reached over the first interval - with estimated parameters: x_{min}=117 (95% CI [50, 132.05]), α=7.891 (95% CI [3.328, 8.02] (Figures [**?**] and [**?**].

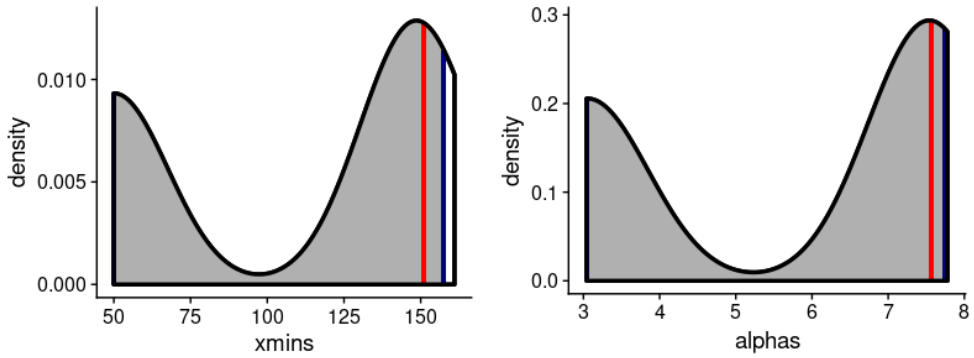


Figure 46: x_{min} and α CIs over Medium Firms having Employees in [50, 170]

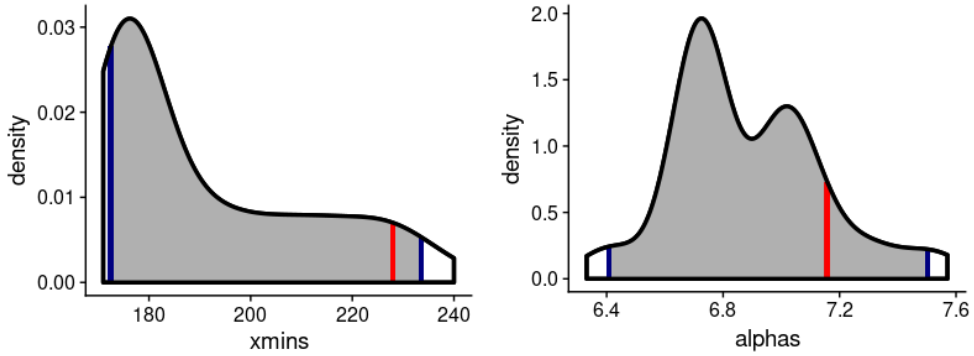


Figure 47: x_{min} and α CIs over Medium Firms having Employees in [171, 249]

As you may notice, the estimated x_{min}s are quite high in both the intervals, explaining that Power Law hypothesis just holds for tails in that interval. We've then further analyzed the firms having Employees in the intervals [50, 116] and firms with Employees in [171, 227]. Regarding the latter, we've seen a p-value of 0.4, i.e. we cannot reject the null hypothesis; in the first interval, instead, there seems to be a further good split into [50, 75] and [75,116], but still p-values are very low($\approx$0), so we can reject null hypothesis in the whole lowest interval [50, 170]. In Figure 49 we show ECDF and how Power Law fits in the whole intervals [50,170] and [171,249].

4.1.3 Results on Large Firms

Large firms represent a very tiny subset of our data, being $\approx$ 0.4% of the firms. Large firms Employees distribution shows a kurtosis of 1370.24 and a skewness=31.99. Below we show the distribution of bootstrapped CIs for both x_{min} and α. (Figure 50) The results obtained are very good, with a p-value of 0.45 on the whole sample, with parameters x_{min}=251 (95% CI [251, 587.025]), α=2.2517 (95% CI [2.242, 2.329]). Interestingly, the estimated x_{min} is even the lowest possible value, revealing that the null hypotesis cannot be rejected over the whole Large firms sample.

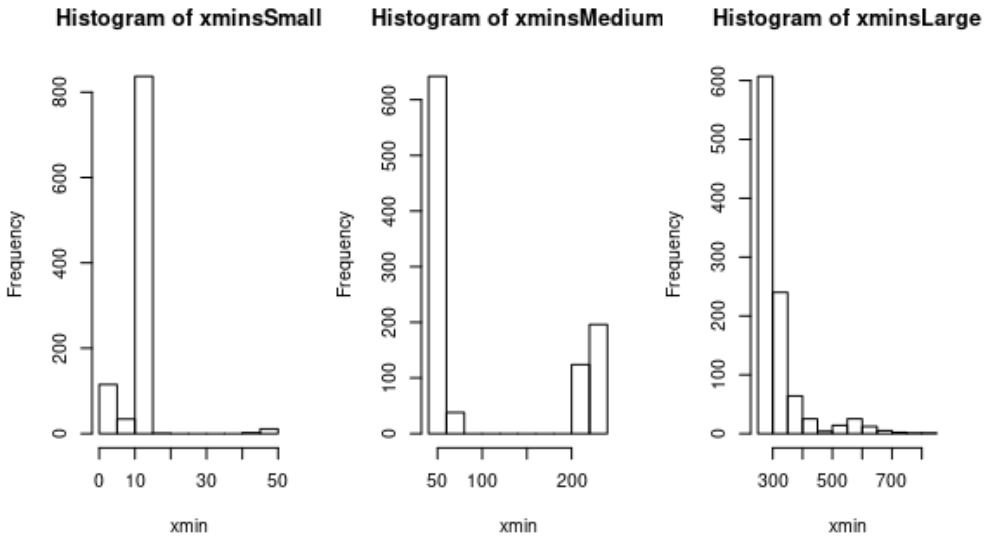


Figure 48: Distribution of estimated x_{min}s for Small, Medium and Large firms

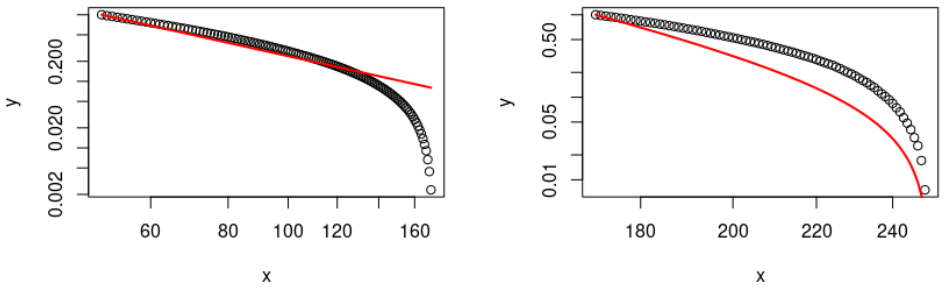


Figure 49: CDF and Power Law fit over Medium firms having Employees$\leq$170 and Employees$\geq$170

4.2 Results over years

One of the most discussed conclusions from the original Pareto distribution of incomes is the costancy of α: this seems to be untenable, indeed its value seems to vary over time and over different places. [4]
Here we try to answer this non-trivial question, by analyzing how eventually α has changed during the period we've analyzed. We've first tried to estimate the parameters for each year-sample, setting x_{min} in order to minimize KS-D statistic, then estimating α via MLE. The results for each year are very poor, as they always reach p-value ≈ 0. Then, we've tried to estimate how α has moved over the years, by using a subsample of 10k data.
First of all, as shown in Figure 37, α is strictly linked to the value we choose for x_{min}: if we decided to use random subsamples for each year, then we could've reached biased results, in terms of the randomness of the subsample, and in terms of the estimated x_{min} for that subsample. Due to this, we've fixed x_{min} to further analyse the behaviour of α along time: we've decided to fix it to the most likely value v for that sample size, given the empirical density function. To be sure that v is meaningful for each year, we've generated bootstrapped distributions

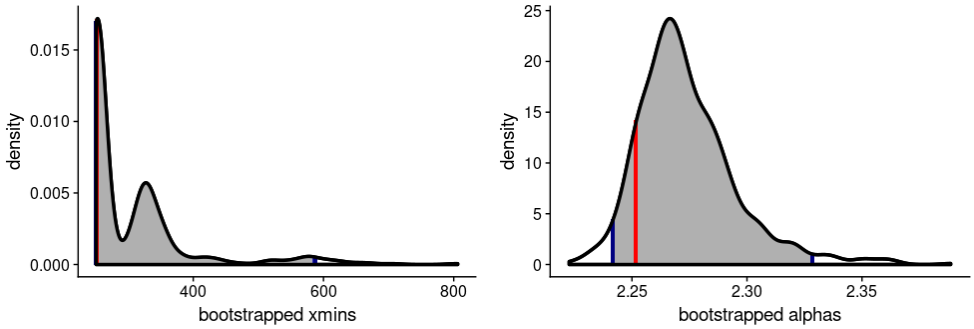


Figure 50: Distribution of bootstrapped x_{min}s and αs for Large firms

for x_{min} for each year, checking it's actually inside 95% CI.
The CIs we've found are the following: 2007 - 95% CI [11, 91]; 2008 - 95% CI [8, 84.025]; 2009 - 95% CI [8, 74.025]; 2010 - 95% CI [8, 68]; 2011 - 95% CI [9, 35]; 2012 - 95% CI [9, 30]; 2013 - 95% CI [9, 31]; 2014 - 95% CI [9, 32]; 2015 - 95% CI [9, 32]. We've then used v=12 (see Figure 36), that is consistent for each year. In Figure 51 we show interesting results obtained on Years 2011 and 2012,

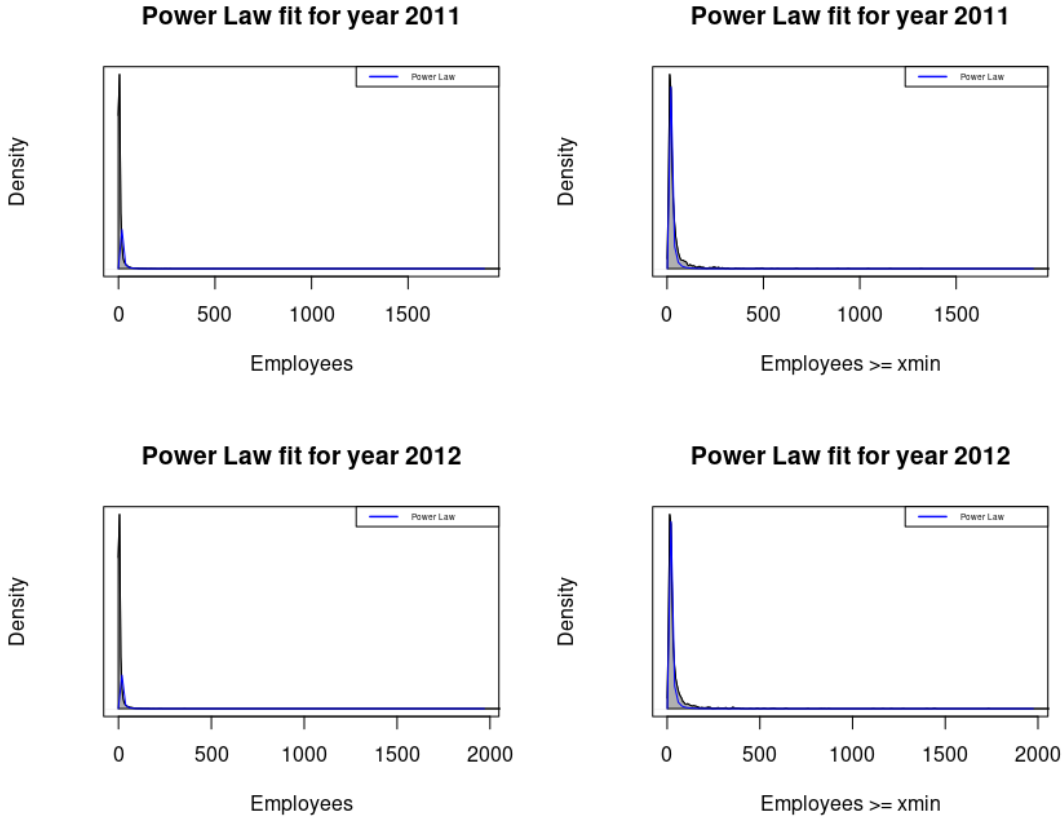


Figure 51: Fitting Power Law(x_{min}=12) on Employees Distribution over years

which also yields the best results in terms of p-values(0.85 and 0.95 respectively, for E≥xmin). For what concerns alphas, we've got the following results: 2007 - $\alpha \approx 2.021$ (95% CI: [1.974, 2.371]); 2008 - $\alpha \approx 2.136$ (95% CI: [2.007, 2.334]); 2009 - $\alpha \approx 2.067$ (95% CI: [1.98, 2.278]); 2010 - $\alpha \approx 2.074$ (95% CI: [1.927, 2.295]); 2011 - $\alpha \approx 2.186$ (95% CI: [2.104, 2.304]); 2012 - $\alpha \approx 2.145$ (95% CI:

[2.11, 2.299]); 2013 - $\alpha \approx 2.172$ (95% CI: [2.088, 2.26]); 2014 - $\alpha \approx 2.195$ (95% CI: [2.083, 2.253]); 2015 - $\alpha \approx 2.152$ (95% CI: [2.083, 2.239]). There seems to be a small growth in its values from 2011 on, with very similar values within the first 4 years(2007-2010) and the remaining ones(2011-2015), while there appear a quite large gap in 2010-2011. To answer the question regarding the change of α, we've made a hypothesis testing for each pair of subsequent years(y_1, y_2), over samples $((S_1 and S_2) generated by estimated Power Law distributions (D_1 and D_2) for y_1 and y_2$:

- H0: $S_1 and S_2$ are drawn from the same population distribution function;
- H1: $S_1 and S_2$ are not drawn from the same population distribution function.

Years	D-Statistic	P-Value	Null Hypothesis
2007-2008	0.023812	$<$ 2.2e-16	Rejected
2008-2009	0.17203	1.568e-07	Rejected
2009-2010	0.003	0.574	Accepted
2010-2011	0.022	$<$2.2e-16	Rejected
2011-2012	0.007	0.031	Rejected
2012-2013	0.005	0.034	Rejected
2013-2014	0.005	0.092	Accepted
2014-2015	0.008	0.005	Rejected

Table 25: α Changed over Year - Hypothesis Testing

From table 25 somehow appears a quite large change in the value of α during 2007-2009 and 2010-2011, while it seems to vary poorly during the periods 2011-2015 and, especially, 2009-2010. By considering a threshold value for p-value equal to 0.05, we cannot reject the null hypothesis for years: (2009-2010) and (2013-2014).

5 Growth rate analysis

Before we begin our analysis of the growth rate distribution across the AIDA firms, we define "growth" and make some assumptions in the context of this study.

First, the consecutive yearly growth of every firm is calculated using the formula:

$$ln(\frac{S(t)}{S(t-1)}) \tag{3}$$

Where **S(t)** represents the firm size (revenue) in the current year and S(t-1) represents the firm size in the previous year. Given that the revenue is used as a size metric, we replaced every row in the dataset for which *"R=0"* with *"R=1"* to properly reflect the (lack of) growth under the natural logarithm.
Second, we embrace the assumptions widely used in literature regarding the distribution of growth rate, namely:

1. The growth time series of each individual firm is a specific and independent realization of the same stochastic process.
2. All firms have the same specific functional form of the growth rate distribution, although respective parameters may differ from firm to firm.

5.1 Firm growth rate distribution in AIDA

In the beginning, we depict the growth rate distribution (GRT) of the entire AIDA database using non-parametric estimates such as the kernel density estimate and histogram plot in order to create an initial idea about the potential family of distributions that characterizes the empirical data:

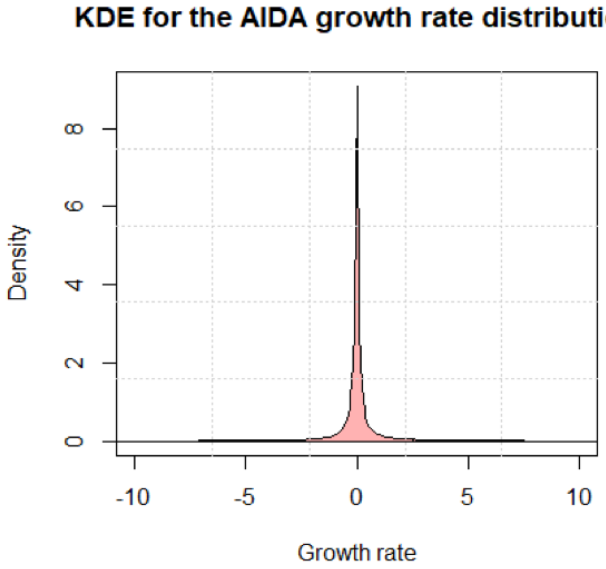


Figure 52: KDE for the AIDA growth rate distribution

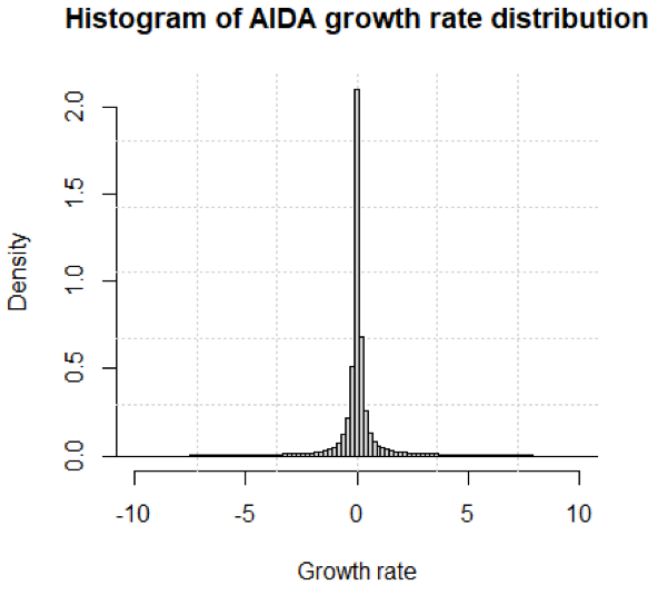
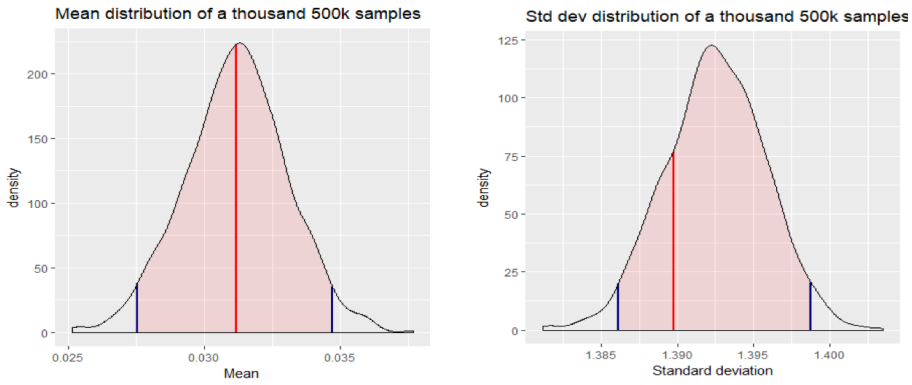


Figure 53: Histogram of AIDA growth rate distribution

One thing that should be emphasized is the presence of considerable firms that have a growth of zero. This is reflected in the high peak located at the center of the graph above. Some basic stats are shown below:

- Skewness ≈ 0.4
- Kurtosis ≈ 11.8
- Median $= 0$
- Mean ≈ 0.03

The distribution seems to be slightly positively skewed, indicating a lack of symmetry and that the tail on the right side is fatter that its counterpart on the left (i.e. there is more data on the right tail). As a rule of thumb, this also means that the mean is on the right of the median, as it is in fact the case. A positive value for the kurtosis indicates the presence of heavy tails. The typical normal distribution has a kurtosis of 3 and is known as a mesokurtic distribution. In our case, the kurtosis is much higher than that. This potentially hints at the presence of a non-normal leptokurtic distribution.
To determine which theoretical distribution fits the empirical data of AIDA best, we adopted the parametric maximum likelihood estimation (MLE) approach. The MLE functions used were fitdist() and mle2() of the fitdistrplus and bbmle packages respectively.
Normal, Laplace, Cauchy, Logistic, Lognormal, Exponential, Pareto, Gamma, Weibull and Beta were among the families tested. Random samples with varying sizes were extracted and fitted for each type of distribution. In order to avoid sampling bias, for every sample size we performed a thousand random re samplings of the same size, drawn from the dataset. The idea is to infer a 95% confidence interval for the parameters belonging to a certain family (for example, the mean and standard deviation for Gaussian) based on the empirical distribution, as we plotted below:



(a) Mean (b) Std. Dev.

Figure 54: Confidence Interval

Parameters for every selected sample are estimated via MLE. The thin red line indicates the value of the mean and standard deviation respectively for our 500k sample (i.e. the one that we actually analyze). The blue lines on both sides represent the confidence interval limits. Given that our sample parameters are well within the CIs, it is safe to assume that it is unbiased; hence, it is an appropriate choice for further analysis. We have followed the very same line of reasoning for all other sample sizes and parameters associated with the rest of theoretical distributions, selecting to work only on those samples that have parameters confined within the boundaries of their respective CIs and rejecting

those who do not fulfil this criterion.
It is only now that we are able to safely proceed with trials of fitting empirical data with different continuous distribution functions. The one-sample Kolmogorov-Smirnoff test along with the Chi-squared test were used to determine the goodness of fit for the maximum likelihood approximation. The KS test quantifies the maximal distance D between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, whereas in the Chi-squared test, the observations are classified into mutually exclusive bins, and the probability that any observation falls into the corresponding class is calculated. We describe the null and alternative hypotheses as follows:

- *Null Hypothesis* : the sample is drawn from the referenced distribution.
- *Alternative Hypothesis*: the sample is not drawn from the referenced distribution.

Given a significance level of α=5% (i.e. the largest acceptable probability of committing a type I error) and various sample sizes, we extracted the relevant stats from the KS-test and ranked them in the following table (for the top four fits):

Sample Size	Fit	D-Statistic	P-Value	AIC	Null Hypothesis
500,000	Cauchy	0.079343	< 2.2e-16	829940	Rejected
500,000	Laplace	0.17203	< 2.2e-16	1202469	Rejected
500,000	Logistic	0.1929	<2.2e-16	1467071	Rejected
500,000	Normal	0.25386	< 2.2e-16	1745741	Rejected
50,000	Cauchy	0.07903103	< 2.2e-16	82465.17	Rejected
50,000	Laplace	0.152071	< 2.2e-16	119999.9	Rejected
50,000	Logistic	0.1930132	< 2.2e-16	146519.4	Rejected
50,000	Normal	0.2535603	< 2.2e-16	174598.3	Rejected
5,000	Cauchy	0.08313392	< 2.2e-16	7939.525	Rejected
5,000	Laplace	0.1473177	< 2.2e-16	11523.29	Rejected
5,000	Logistic	0.1863629	< 2.2e-16	14128.57	Rejected
5,000	Normal	0.2509302	< 2.2e-16	17011.87	Rejected

Table 26: Growth Statistics

Given that the p-value is below the given significance level in every case, it is would wrong from a statistical point of view to claim that the Cauchy fit is a 'good' fit. Moreover, this issue was present throughout every test, independent form sample type or size, as we will see in the further experiments. However, based on the D-statistic and Akaike Information Criterion (AIC) score, we can conclude that the Cauchy fit consistently explains the empirical distribution of the growth considerably better compared to the other family of distributions taken into account. AIC estimates the quality of each model, relative to each of the other models. The lower it is the better. A similar thing can be said about the D-statistic, which is the lowest for Cauchy in every instance. Below, we graphically represent the four above-mentioned fits for an unbiased sample of 50,000 empirical data:

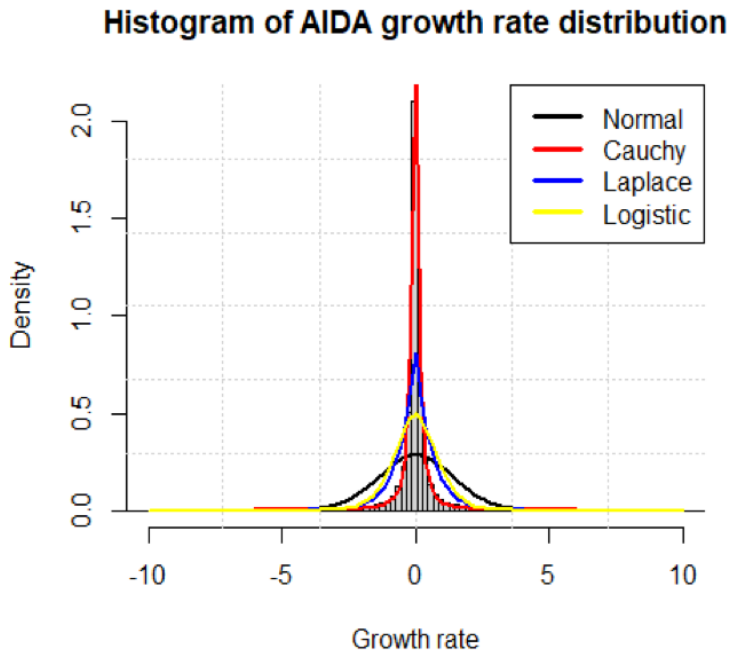
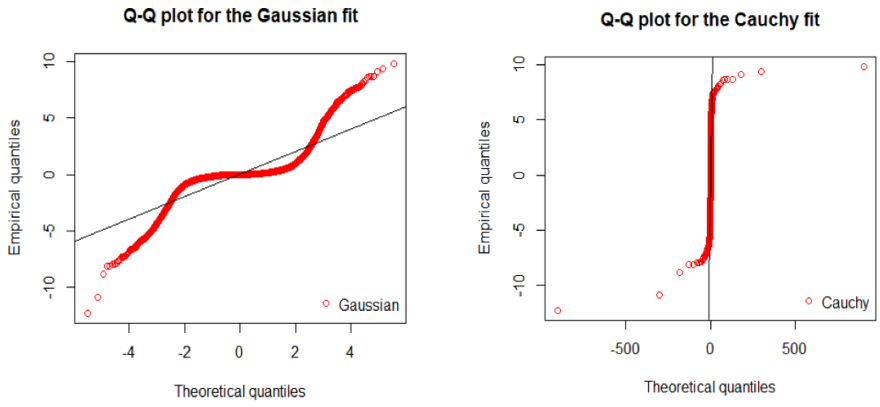


Figure 55: Growth Rate Distribution

As expected, also visually the Cauchy distribution provides the best fit for the growth distribution since it is characterized (like the empirical data) by the presence of heavy tails. Below, we compare the Q-Q and P-P plots between the Gaussian and Cauchy fit on the same sample:

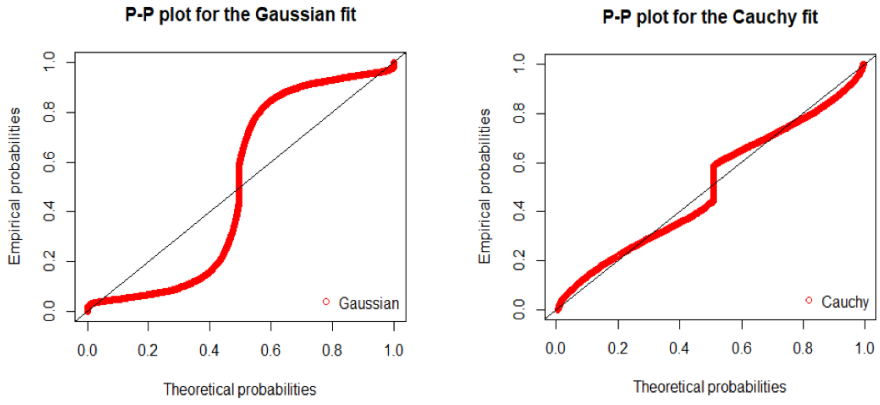


(a) Weibull

(b) Gamma

Figure 56: Confidence Interval

Ideally, in a Quantile-Quantile plot, the data should follow a diagonal-like direction if the referenced distribution is a good one. We can clearly see that on the right side graph, although far from perfect, explains the variation of the data much better, especially around the mean, than the left side (Gaussian) representation.



(a) Weibull (b) Gamma

Figure 57: Confidence Interval

A similar interpretation can be laid out over the Probability-Probability plots, where we see that Cauchy provides a more acceptable description of the empirical probabilities. Furthermore, we plotted the cumulative distribution functions of the empirical data and the data generated by the MLE parameters of the Gaussian and Cauchy distributions:

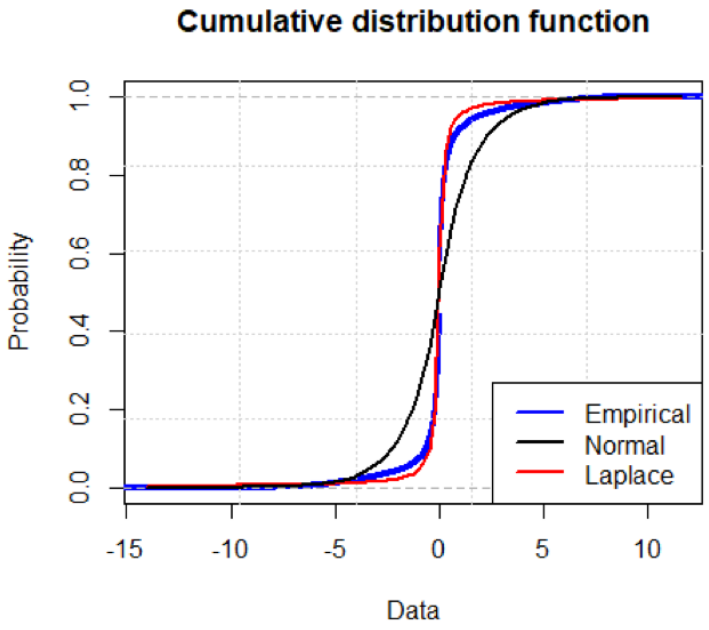


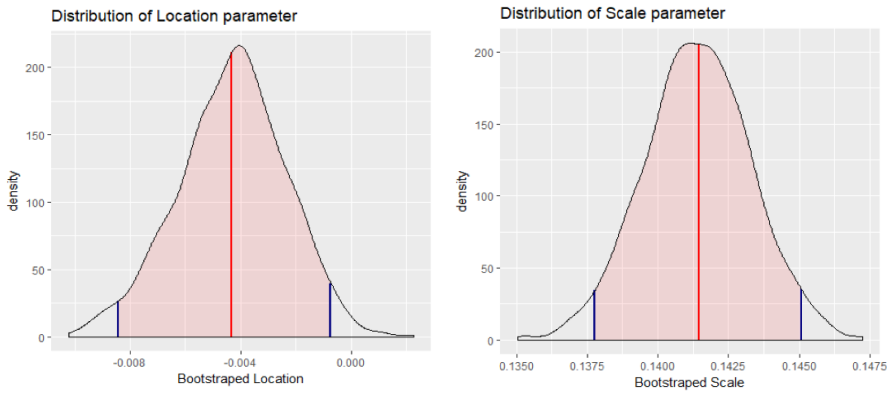
Figure 58: CDF Growth

The CDF of the Cauchy generated sample overlaps over a long stretch with the CDF of empirical data. The same observation cannot be made about the Gaussian generated sample CDF. The graphical representations reinforce our belief that the Cauchy fit is the most suitable in explaining the variation of the growth rate when compared to the others. In addition, the Gaussian fit consistently ranks as the worst from a statistical perspective. Hence, we can confidently say that the Gibrat Law does not hold on Italian firms.

5.2 Bootstrap confidence intervals

For our initial unbiased sample of 50,000 observations, we can also perform a parametric bootstrap by passing to the bootdist() function the Cauchy fit object with the MLE estimated parameters. The function generates a thousand equally sized samples (with similar values with respect to the empirical sample values), for which the location and scale parameters are calculated using MLE. Below we show the distribution of the computed parameters and draw up 95% confidence intervals.

- *Null Hypothesis*: The estimated parameters of the empirical data are equal to the true population parameters;
- *Alternative Hypothesis*: The estimated parameters of the empirical data are not equal to the true population parameters;



(a) Location (b) Scale

Figure 59: Confidence Interval of parameters

Again, the red lines represent the respective parameters of our sample, while the blue boundaries correspond to the CI limits. Clearly, the values of interest are situated between the 95% confidence intervals, so we accept the null hypothesis that the true parameters are equal to the parameters of our sample. As mentioned earlier, we also implemented the Chi-squared test to infer the goodness of fit associated to the referenced distribution. Testing was done following the instructions of Ricci et al[7].
However, we did not bother with displaying the computed p-values since they were very close to zero, even for small sample sizes. A snippet of the results is shown below:

```
        Chi-squared test for given probabilities

data:  f.os
X-squared = 1809.8, df = 112, p-value < 2.2e-16
```

Figure 60: Chi-Squared Test

5.3 Firm growth rate distribution in the Manufacturing sub-sector

Distribution of growth on the manufacturing sub-sector exhibits some of the same characteristics as we have seen in AIDA. However, the kurtosis has a higher value (25), indicating heavier tails. Again, the rank of the fits is consistent, with Cauchy taking the first place as the best fit, followed by Laplace, Logistic and the Normal distribution respectively. Below we plot the histogram:

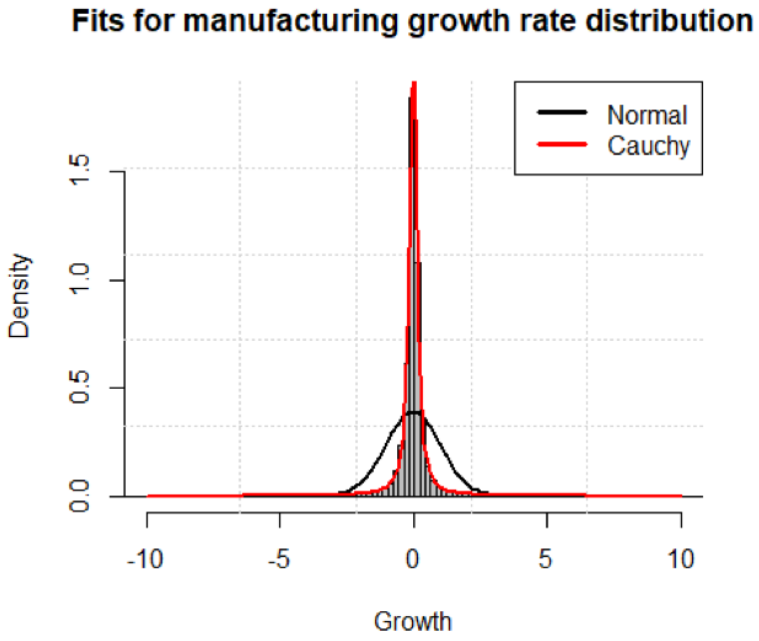


Figure 61: Fits Manufacturing Growth

What is notable about the manufacturing subset is that, for a small enough sample size, say 1000, we get a KS-test p-value of about 0.13 for the Cauchy fit, so we accept the null hypothesis that the sample was drawn from the Cauchy distribution family. Nevertheless, with sample being so small, we decided to conduct another test in order to reaffirm the validity of the obtained fit. Using parametric bootstrap, a thousand samples of size 1000 were generated starting from the location and scale parameters of our initial sample. For every sample, the respective D-statistic was extracted. In the following graph, we plotted the density of D using a log spline fit. Log spline is similar to kernel density estimation, with the exception that it allows us to easily derive the p-value of the test because in every case, the smallest value of the x-axis will be zero. By simply subtracting from 1 the CDF value of the empirical D-stat, we get the probability of observing values that are equal to or more extreme than the max distance we already have, or in other words, the p-value:

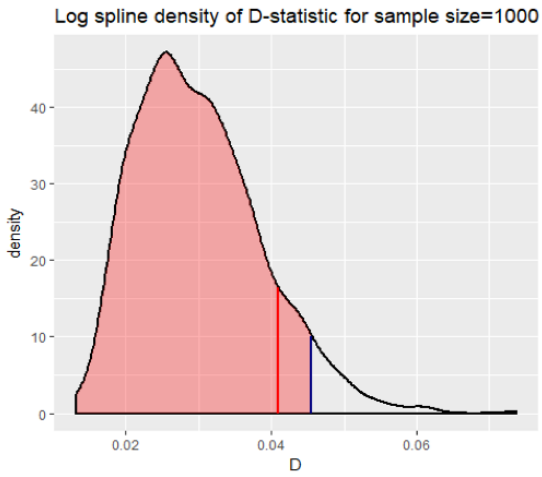
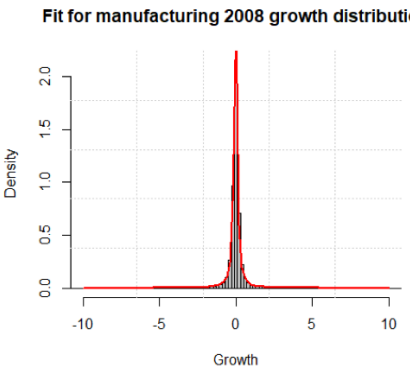


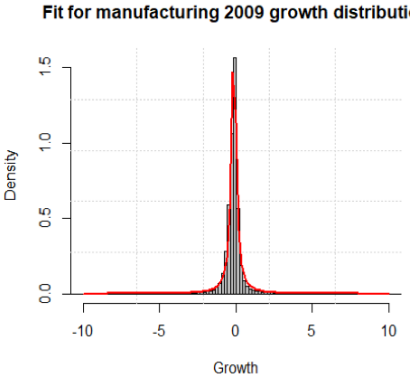
Figure 62: Log spline Density

- *Null Hypothesis*: The Cauchy distribution fits the empirical data well;
- *Alternative Hypothesis*: The Cauchy distribution does not fit the empirical data well;

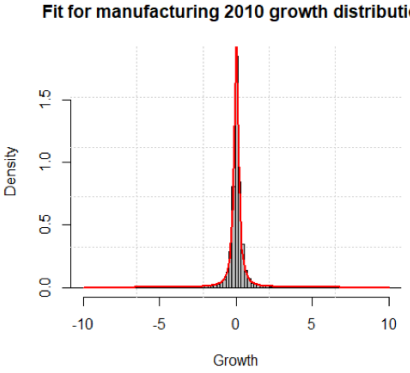
The p-value is 0.1, and we can see that the empirical D-stat is within the 95% confidence interval in the one-tailed test, so we accept the null hypothesis and conclude that the fit for our initial sample is a good one.
Now we proceed to see if there is any significant change for the distribution of the growth throughout each year (form 2008 to 2015). We have fitted Cauchy distributions for the entire dataset in each case.



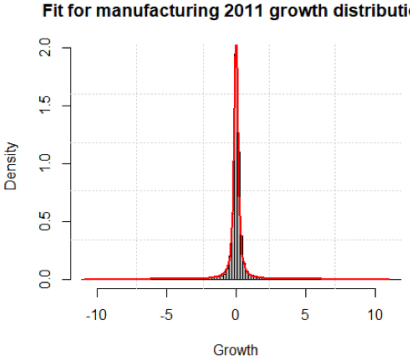
(a) 2008



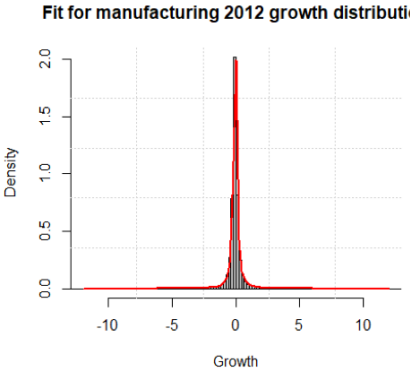
(b) 2009



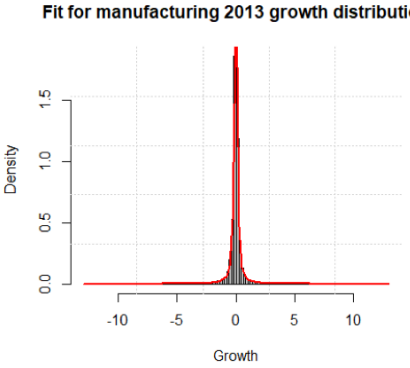
(c) 2010



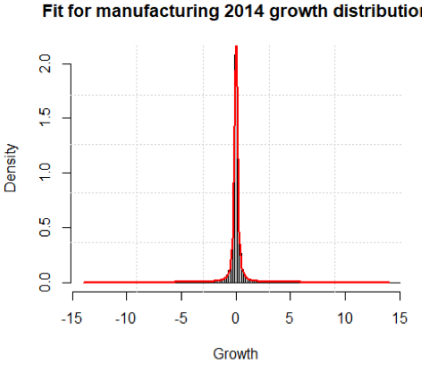
(d) 2011



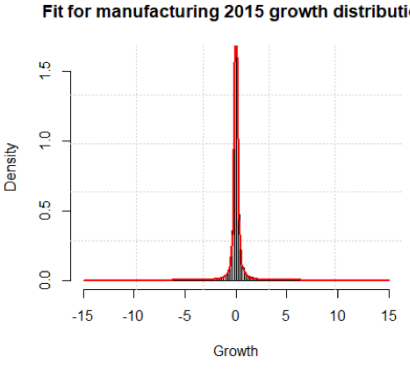
(e) 2012



(f) 2013



(g) 2014



(h) 2015

Figure 63: Fit by Year

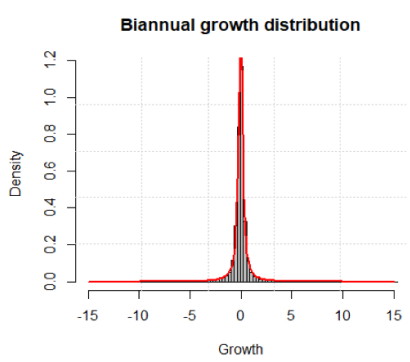
The best fit remains the same (Cauchy) across all years depicted above. At first glance, it may seem as though the distributions do not differ much among each other. This is not the case, however. We conducted two sample KS-tests of the empirical data for every pair of subsequent years in order to see if there is any statistical difference between the respective distributions:

- *Null Hypothesis*: Empirical distributions are drawn from the same family of distributions;
- *Alternative Hypothesis*: Empirical distributions are not drawn from the same family of distributions

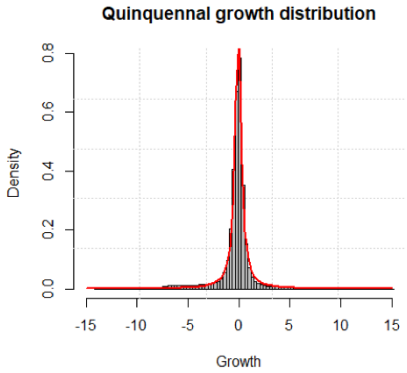
Years	D-Statistic	P-Value	Null Hypothesis
2008-2009	0.23812	$<$ 2.2e-16	Rejected
2009-2010	0.17203	$<$ 2.2e-16	Rejected
2010-2011	0.050075	$<$2.2e-16	Rejected
2011-2012	0.17576	$<$ 2.2e-16	Rejected
2012-2013	0.10108	$<$ 2.2e-16	Rejected
2013-2014	0.068017	$<$ 2.2e-16	Rejected
2014-2015	0.04359	$<$ 2.2e-16	Rejected

Table 27: Growth Statistics by Year

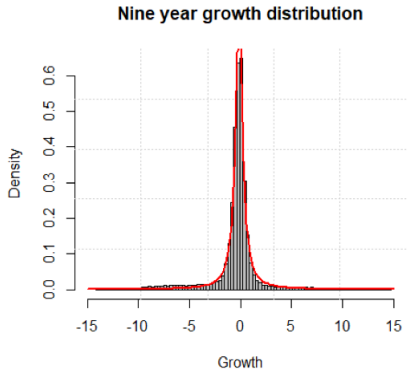
We can see that for every pair of years there is a statistically significant change in the "nature" of the distribution. If we assume that the family of distributions is Cauchy (it is in fact the best fit relative to the other families), then we could alternatively say that it is the parameters of the Cauchy fit that are statistically different. The change between 2008 and 2009 is notably the biggest, judging by the value of D-statistic. This may be due to the negative economic growth experienced during that period. The smallest change in distribution parameters seems to be between the years 2014 and 2015, indicating a possible stabilization of the economic growth.
Following the same line of reasoning, we check if there is any significant change in distribution for different time spans of firm growth (biannual, quinquennal and nine-year lag). The biannual growth for every firm was calculated on subsequent odd and even years while the quinquennal growth was computed for years with a difference of five units. We also estimated the growth by taking into account only the first and last years (2007 and 2015) for every firm. MLE parameters were estimated on the entire datasets, which we plot as follows.



(a) Biannual



(b) Quinquennal



(c) Nine Year

Figure 64: Fit by Group Year

Cauchy remains the best fit. The following table summarizes some of the most relevant metrics:

Time span	Kurtosis	Skewness	Mean	Variance
Biannual	16	-0.29	-0.08	1.85
Quinquennal	8.4	-0.78	-0.29	3.77
Nine years	7.1	-0.83	-0.41	5.1

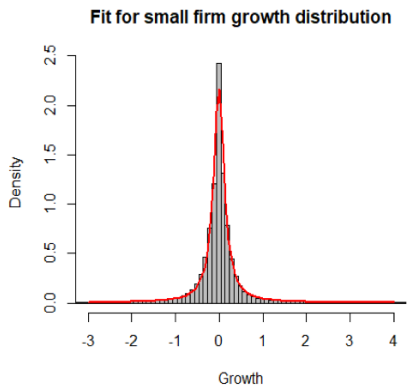
Table 28: Growth Statistics by Group Year

The distributions are leptokurtic and negatively asymmetric (judging by the negative skewness values). The kurtosis decreases as the time span increases. Due to the fact that the number observed data decreases, the tails get thinner. The variance increases for the same reason (i.e. less, more diverse observations). We show the stats of the two-sample KS-tests among pairs of different time spans. The hypotheses are the same as for the previous two-sample KS-test:

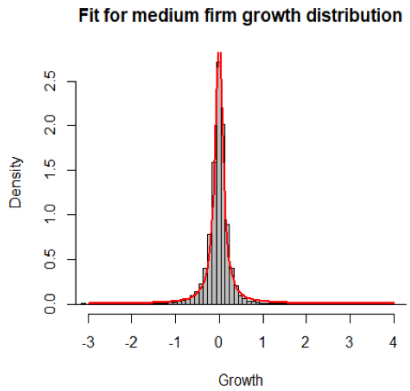
Time spans	D-statistic	p-value	Null Hypothesis
Biannual v. Quinquennal	0.12428	$<$ 2.2e-16	Rejected
Biannual v. Nine years	0.18198	$<$ 2.2e-16	Rejected
Quinquennal v. Nine years	0.057933	$<$ 2.2e-16	Rejected

Table 29: Growth Statistics by Group Year

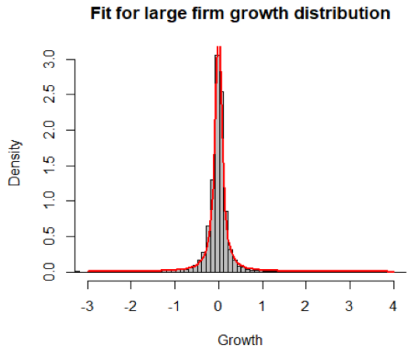
There is a statistically significant difference between the empirical distributions for every temporal span. As a result, we conclude that the empirical growth rate distribution changes across different lag intervals.
We now classify firms according to their size (small, medium, large) and follow the same procedure as previously by performing tests on the newly obtained datasets:



(a) Small



(b) Medium



(c) Large

Figure 65: Fit by Size

As expected, Cauchy distribution again outperforms the Laplace, Logistic and Gaussian ones.
The most important features are summarized below:

Firm Size	Kurtosis	Skewness	Mean	Variance
Small	21.5	0.27	-0.014	1.05
Medium	172.8	-1.78	-0.03	0.22
Large	341.6	7.4	-0.019	0.184

Table 30: Features by Firm Size

The empirical distributions are leptokurtic and positively asymmetric (i.e. skewness is greater than 0) with the exception of medium firms, which are negatively asymmetric. The kurtosis increases with increasing firm size. What this means is that the bigger the firm is, the more likely we are to see outliers or relatively large growth values in a random sample. Below we show the stats of the two sample KS-tests among pairs of different firm sizes.

Firms	D-statistic	p-value	Null Hypothesis
Small v. Medium	0.080428	< 2.2e-16	Rejected
Small v. Large	0.11289	< 2.2e-16	Rejected
Medium v. Large	0.045109	1.797e-13	Rejected

Table 31: Firm Size vs. Firm Size

There seems to be a significant difference in the kind of distribution between small and large firms, while this discrepancy is smaller between medium and large sized firms.
Finally, we classify the firms also according to their region of operation, intuitively obtaining three main categories: South, Center and North. The Cauchy fit explains the variation of the empirical data the best in every case. The table below sums up the obtained stats after conduction the two-sample KS-test.

Regions	D-statistic	p-value	Null Hypothesis
South v. Center	0.03658	< 2.2e-16	Rejected
South v. North	0.052095	< 2.2e-16	Rejected
Center v. North	0.023049	< 2.2e-16	Rejected

Table 32: Geo Area vs. Geo Area

The low p-values suggest a considerable difference in the nature of each empirical distribution. Growth in firms located in the northern region seem to be closer (distribution wise) to the firms operating in the central region of Italy.

5.4 Symmetry test on empirical distributions

To see if the yearly growth distribution in the manufacturing sub-sector is symmetric or not, we applied the two sided symmetry test by Miao, Gel, and Gastwirth (2006). Confidence intervals stand at 95%. Implementation is done by calling the symmetry.test() function of the lawstat library in R.

- *Null Hypothesis*: The distribution is symmetric;
- *Alternative Hypothesis*: The distribution is asymmetric.

We give a rundown of the results below.

Growth year	p-value	Null Hypothesis
2008	0.138	Accepted
2009	< 2.2e-16	Rejected
2010	< 2.2e-16	Rejected
2011	< 2.2e-16	Rejected
2012	0.09	Accepted
2013	< 2.2e-16	Rejected
2014	0.1	Accepted
2015	< 2.2e-16	Rejected

Table 33: Growth Distribution by years

Indeed, the growth distribution belonging to the years 2008, 2012 and 2014 is symmetric according to the p-values that are greater than the significance level of 5%.

5.5 Hypothesis testing on the mean of the growth

We want to test whether the mean of each yearly growth rate in the manufacturing sub-sector is statistically equal to zero. Given that we do not know the true variance of the population in each dataset, it would be logical to perform a one-sample t-test. However, the empirical distribution is not Gaussian, and as a result we cannot assume that our test statistic has a t(n-1) distribution under the null hypothesis. Nevertheless, if the sample size is large enough, the distribution of the t-test statistic under the null hypothesis can also be approximated by a standard normal distribution (Dekking et. al [8]). The respective sample sizes are greater than 80,000 at all times. Hence, we can perform a simple Z-test for every instance by passing as standard deviation argument the sample standard deviation.
In addition, we performed non-parametric bootstrap for the mean on each dataset, generating a thousand samples and computing the mean in each iteration in order to plot the bootstrapped distribution of the mean and extract the bootstrap confidence intervals. Parametric bootstrap was not preferred due to the large sample sizes. The Cauchy fit, being poor, would cause the obtained p-values to be unreliable. The hypotheses are listed below:

- *Null Hypothesis*: The true mean is equal to zero;
- *Alternative Hypothesis*: The true mean is not equal to zero;

The tests are two tailed, with 95% confidence intervals

Growth year	Z-test p-value	Bootstrap p-value	Z-test CIs	Bootstrap CIs	Null Hypothesiss
2008	0.192	0.194	[-0.013, 0.002]	[-0.012, 0.002]]	Accepted
2009	< 2.2e-16	0	[-0.166,-0.152]	[-0.165,-0.152]	Rejected
2010	< 2.2e-16	0	[0.043, 0.056]	[0.044, 0.056]	Rejected
2011	< 2.2e-16	0	[0.053, 0.066]	[0.053, 0.066]	Rejected
2012	< 2.2e-16	0	[-0.064,-0.053]	[-0.064,-0.053]	Rejected
2013	< 2.2e-16	0	[-0.060,-0.049]	[-0.061,-0.049]	Rejected
2014	0.5253	0.51	[-0.007, 0.003]	[-0.007, 0.003]	Accepted
2015	< 2.2e-16	0	[0.049, 0.060]	[0.049, 0.060]	Rejected

Table 34: Growth Year Statistics

The table above summarizes the results of both the Z-test and bootstrap approach. We accept the null hypothesis for the years 2008 and 2014. The distribution of the bootstrapped mean for 2008 is plotted as follows:

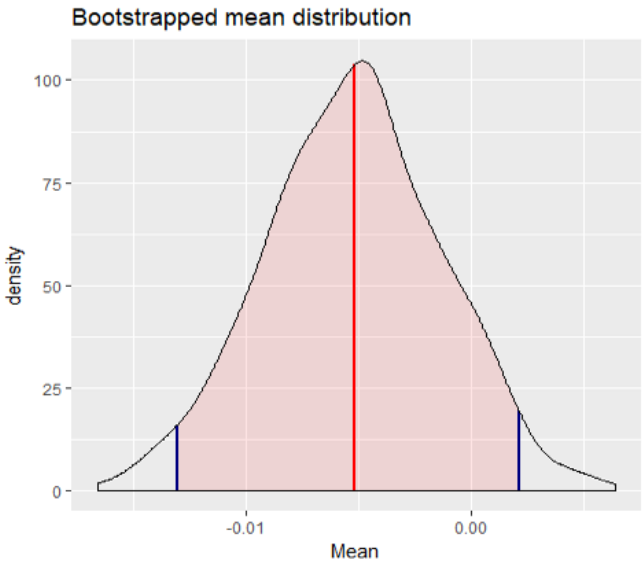


Figure 66: Bootstrap confidence interval

5.6 Hypothesis testing on the difference of the means for two populations

Similar to the previous section, we want to see if the difference of the growth means for every pair of consecutive years is equal to zero. Again, we do not know the true variances of the populations we are comparing. Therefore, we could choose to perform a two-sampled t-test. Yet again, we face the problem of a non-Normal distribution of the empirical data. However, since the sample sizes are large the distribution of the t-test statistic under the null hypothesis even in this case can be approximated by a standard normal distribution. This happens because, if we compute the difference between the same normally distributed statistics from two populations, the resulting statistic is still going to be normally distributed.
We can perform a simple Z-test by passing as standard deviation arguments the sample standard deviations of the two populations we are testing.
Non-parametric bootstrap for the mean differences was also carried out:

Years	Z-test p-value	Bootstrap p-value	Z-test CIs	Bootstrap CIs	Null Hypothesis
2008 v. 2009	< 2.2e-16	0	[0.143, 0.164]	[0.143 0.164]	Rejected
2009 v. 2010	< 2.2e-16	0	[-0.218,-0.200]	[-0.218,-0.200]	Rejected
2010 v. 2011	0.03496	0.0326	[-0.018,-0.001]	[-0.018,-0.001]	Rejected
2011v. 2012	< 2.2e-16	0	[0.110, 0.127]	[0.109, 0.127]	Rejected
2012 v. 2013	0.358	0.325	[-0.011, 0.004]	[-0.011, 0.004]	Accepted
2013 v. 2014	< 2.2e-16	0	[-0.062, -0.044]	[-0.061,-0.049]	Rejected
2014 v. 2015	< 2.2e-16	0	[-0.064,-0.048]	[-0.065,-0.0482]	Rejected

Table 35: Growth Year Statistics

The table above summarizes the results. We accept the null hypothesis only for the years 2012 v. 2013, so only for this pair, the difference of the means is statistically zero.

The graph below plots the distribution of the bootstrapped mean difference for

the years 2008-2009. The empirical mean difference is within the 95% confidence intervals and has a value of 0.154.

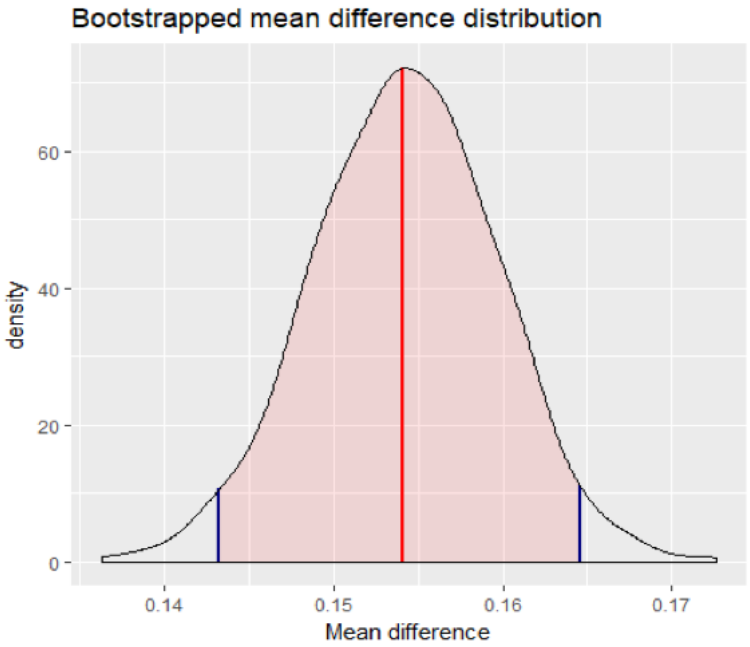


Figure 67: Mean Difference

5.7 Linear regression models for the growth rate distribution of subsequent years

In order to determine whether there is a dependence between growth in subsequent years, we can model a simple linear regression of the form "growth2009∼growth2008 ", where growth-2008 is the predicting variable and growth-2009 is the target variable.

- *Null Hypothesis*: Model is not significant (i.e. beta =0) ;
- *Alternative Hypothesis*: Model is significant (i.e. beta != 0);

The default confidence interval stands at 95%.

Below we have summed up some of the more interesting stats associated with the fit of the linear regression models.

It is interesting to note that most models are significant, or in other words, the growth of the previous year adequately predicts the growth in the current year. However, it is also noteworthy that the values of the R-squared statistic (that indicates the goodness of fit) are far from ideal. R-squared can take values between 0 (very bad fit) and 1 (ideal fit). A good fit in a significant model is typically associated with an R-squared value that is greater than 0.7, but in our case, we obtained significant models with horrendous R-squared values. So what is going on here? A possible interpretation could be formulated as follows: the independent variable is correlated with the dependent variable, but it do not explain much of the variability in the dependent variable. This is likely due to the

presence of heteroscedasticity around the fitted values.

Years	F-statistic	Model p-value	R-squared	Slope	Null Hypothesis
2009 $\sim$ 2008	40.43	2.056e-10	0.0007019	-0.024278	Rejected
2010 $\sim$ 2009	112.3	$<$ 2.2e-16	0.001407	-0.033502	Rejected
2011 $\sim$ 2010	0.7307	0.3927	8.362e-06	-0.002515	Accepted
2012 $\sim$ 2011	3.196	0.07384	3.262e-05	-0.004916	Rejected
2013 $\sim$ 2012	196.3	$<$ 2.2e-16	0.001738	0.039106	Accepted
2014 $\sim$ 2013	138.8	$<$ 2.2e-16	0.00121	0.03252	Rejected
2015 $\sim$ 2014	84.1	$<$ 2.2e-16	0.0007577	0.023250	Rejected

Table 36: Summary Linear Regression

Heteroscedasticity implies that there are sub populations in the dependent variable that have different variance from others. Let us take for instance the first model (2009 $\sim$ 2008) and plot the fitted values.

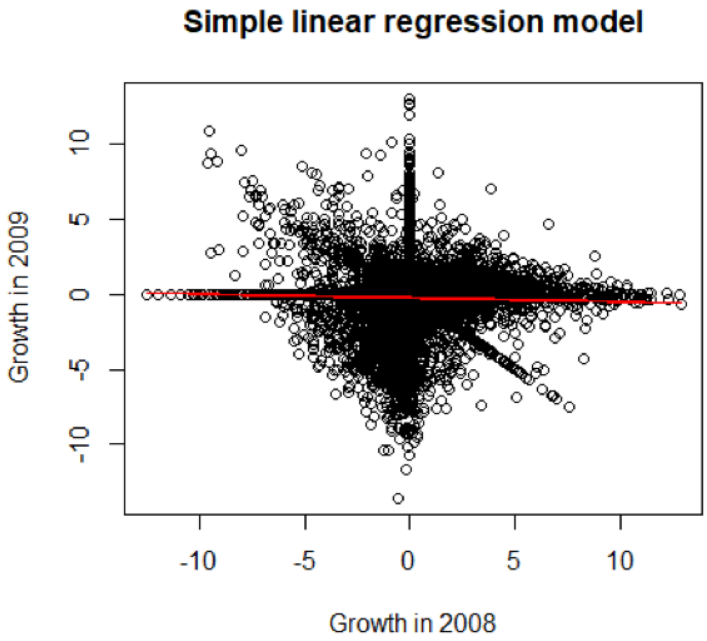


Figure 68: Simple Linear Regression

The red line across the graph represents the fitted values. Obviously, the fit is not very good, to say the least. On average, we can say that if the growth in 2008 increases with one unit, the growth of 2009 with decrease with -0.002 units, so there is a slight negative correlation between them. Still, actual values of the 2009 growth are not at all in alignment with the fitted values of the model.

The residuals (difference between actual y-values and predicted y-values) are plotted against the fitted valued in the graph below.

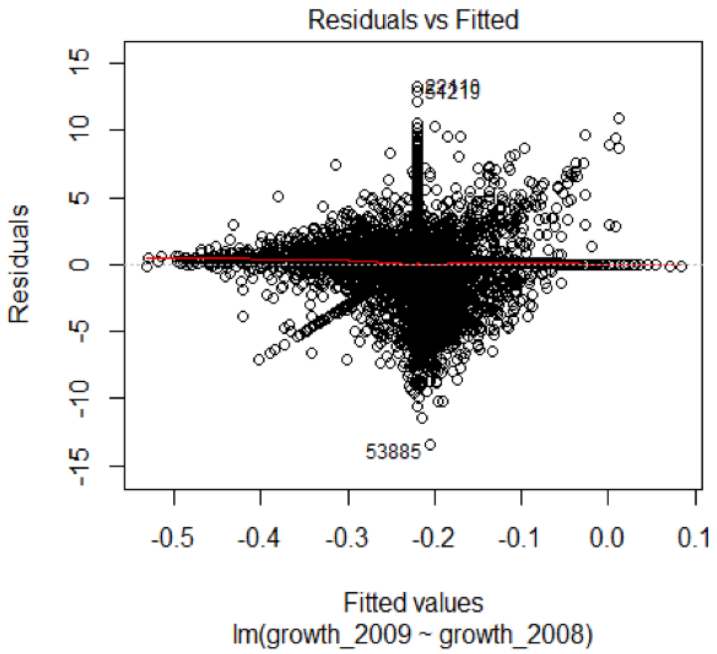


Figure 69: Residual from Linear Regression

Residuals seemingly bounce around in a random manner. To check for the presence of heteroscedasticity, we have used the Breush-Pagan test of the lmtest library. The hypotheses are:

- *Null Hypothesis*: Variance of residuals is constant;
- *Alternative Hypothesis*: Variance of residuals is not constant.

We display a screenshot of the test results:

```
        studentized Breusch-Pagan test

data:  fit_growth
BP = 296.24, df = 1, p-value < 2.2e-16
```

Figure 70: Breush-Pagan test

The p-value is much lower than the significance level of 5%; therefore, we reject the null hypothesis.

Overall, we can conclude that on average, there has been consistent negative growth in Italian manufacturing firms until 2011. This is reflected in the negative slopes for the preceding year growths. From 2012 and onwards, growth has been positive.

References

[1] https://en.wikipedia.org/wiki/Small_and_medium-sized_enterprises.

[2] Axtell, Robert L. "Zipf distribution of US firm sizes." science 293.5536 (2001): 1818-1820.

[3] Fujiwara, Yoshi, et al. "Do Pareto–Zipf and Gibrat laws hold true? An analysis with European firms." Physica A: Statistical Mechanics and its Applications 335.1-2 (2004): 197-216.

[4] Atkinson, Anthony Barnes. "Pareto and the upper tail of the income distribution in the UK: 1799 to the present." Economica 84.334 (2017): 129-156.

[5] Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." SIAM review 51.4 (2009): 661-703.

[6] Delignette-Muller, Marie Laure, and Christophe Dutang. "fitdistrplus: An R package for fitting distributions." Journal of Statistical Software 64.4 (2015): 1-34.

[7] Vito Ricci. "Fitting distributions with R" Free Software Foundation (2005): 17-18.

[8] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaa, L. E. Meester. "A Modern Introduction to Probability and Statistics." Springer (2005): 402-405.