Politecnico di Torino

Dipartimento di Ingegneria Gestionale e della Produzione
Collegio di Ingegneria Gestionale

Corso di Laurea Magistrale in Ingegneria Gestionale
A.a. 2021/2022
Luglio 2022

Power-law distributions and Venture Capital Investors' Behaviour

A Methodology for Using Heavy-Tailed Distribution Estimates to Describe Differences in European and American Entrepreneurial Ecosystems

Relatore:
Prof. Emilio Paolucci

Candidato:
Francesco Giuliani

Summary

The increasing coverage in the recent literature of comparison studies between the European and American entrepreneurial ecosystems indicates an ever-increasing need for tools that facilitate the understanding of entrepreneurial phenomena to facilitate the necessary economic and political changes in support of innovation.

Other interesting recent literature supports a shift in probability distributions assumed in entrepreneurship research from the Gaussian distributions to heavy-tailed ones.

This thesis tries to develop a method that combines the two research trends above.

Firstly, it tries to inquire about the distributions of three of the most common variables of interest in entrepreneurship used to describe the financial story of a startup: *total funding*, *exit value* and *multiple on invested capital* (MOIC). The distributions are calculated from different samples that control for the geographic origins of the investor that funded the startups composing the samples.

Secondly, by looking at the estimated distribution parameters, it tries to make a comparison between the investing behaviour of investors in the American and European ecosystems.

If the variables under scrutiny are plausibly distributed according to a power law, then the α parameter of the power law can be used to draw the desired comparisons across samples with different startup headquarters and different compositions of investors' geographic origins.

The data used in this work and the results obtained by the analyses performed suggest that it is not always possible to describe the variables of interest using a power law. However, from the cases in which this is possible, it seems that the American investors tend to have a higher α, which suggests a potentially higher likelihood of having larger values. Therefore, it seems that American investors have a higher propensity to invest high amounts of money in single startups, in case the observed variable was the *total funding*.

Finally, a simple descriptive analysis of the data available confirms the dominance of US cities as entrepreneurial ecosystems: San Francisco and New York seem to almost always end up first and second respectively, considering the sums of exit values and investments, in addition to having significantly high MOICS. London is the main European ecosystem which seems to perform at a similar level and compete at a similar scale to that of the American cities.

Table of Contents

Introduction

Stemming from the work of Andrea Scipione in his master thesis (Scipione, 2020), this thesis tries to develop a deeper understanding of the investment patterns of American (i.e. US-based) and European (in geographical terms) venture capital funds. It will also try to highlight potential differences between the behaviours of the funds and formulate possible explanations regarding their causes.

While Scipione's work mainly focuses on the general differences between the European and the American entrepreneurial ecosystems, this work follows more closely the behaviour of the investors within the ecosystem. Therefore, it attempts to connect the trends and differences observed in the entrepreneurial ecosystems with investor behaviour.

In the last decades, a particular branch of entrepreneurship research has focused on the differences between the European and the American ecosystems, identifying the differences between the outcomes obtained in the two markets and trying to link them to their causal factors. Another interesting line of research has been focusing on describing entrepreneurial phenomena using heavy-tailed distributions, instead of the normal distribution, generally used in the economics research field.

This thesis will therefore combine the two approaches, trying to understand if the observed differences in the ecosystems can be described using heavy-tailed distributions, with a particular focus on power-law distributions. This could not only help to better describe the differences between investment patterns of investors from different geographic locations but also relate those better described differences to their potential causal factors. A better description and understanding of these factors would in turn lead to a better understanding of the overall entrepreneurial ecosystem itself, allowing regulators to prepare reforms that allow overall performance improvements while helping startup founders optimize their investment-seeking strategies and venture capital funds to identify potential improvements to their investment strategies.

Even though possibly affected by missing information and unavoidable biases in the data, the results of this work are not the only important outcome. Rather, it can be argued that the method developed to answer the research questions is in itself a potentially useful outcome that can be re-applied to more numerous and complete datasets to produce even more reliable results.

Research questions

Combining the two lines of research mentioned above, this thesis will observe the values and the distributions of some of the most widely used startup performance indicators, calculated on different samples, and try to observe and describe the potential differences between the results. This work expects to find differences in the selected indicators and their distribution based on the startup's headquarters location and the geographic origin composition of its investor.

This work investigates samples composed of startups that: (1) already had an exit, (2) were founded between January 2005 and December 2020 and (3) were headquartered in Europe or the US, trying to identify differences based on their geographic location and the composition of the geographic origins of their investors. The measures investigated are the total funding (i.e. the total amount of capital invested in each startup), the exit value, the multiple on the invested capital, and the *time to exit* (i.e. the amount of time elapsed between the founding date of the startup and its exit date). The main questions, measure by measure, are therefore the following:

1. Total funding:
    a. is the total funding of startups distributed following a power-law distribution?
    b. Are there particular differences between the total funding amounts of European and American startups?
    c. Does a different geographic origin composition of investors determine differences in the total amount and distribution of startup funding?
2. Exit value:
    a. Is the total exit value of startups distributed following a power-law distribution?
    b. Are there particular differences between the distribution of exit values of European and American startups?
    c. Does a different geographic origin composition of investors determine differences in the total amount and distribution of exit values?
3. Multiple on invested capital (MOIC or Multiple):
    a. Is the MOIC of startups distributed following a power-law distribution?
    b. Are there particular differences between the distribution of the MOIC of European and American startups?
    c. Does a different geographic origin composition of investors determine differences in the distribution of the MOIC?

To answer these questions, this thesis uses a dataset based on the one used in Scipione's work. However, this was modified and refined based on the specificities of the research questions above.

Thesis structure

This work is organized in chapters that follow the typical structure of a research paper. Even if, for reasons that will be explained in the next sections, it was not possible to collect new data and samples from scratch, this thesis emulates a research paper, as it tries to give its contribution, how small as it may be, to further the understanding of entrepreneurial phenomena.

In doing so, for sake of conciseness, the thesis will assume the reader is familiar with the basic concepts and definitions of entrepreneurship.

Chapter 1 gives a brief overview of the main developments of entrepreneurship research to paint the general picture of the scientific literary context from which this research stems and on which it finds its foundation.

Chapter 2 describes the original dataset used for this research, how it influenced the methodological choices, the cleaning operations performed on the original dataset to prepare it for the analysis stage, and the analysis methodologies applied.

Chapter 3 describes the analysis performed and the results it generated, starting from some general descriptive statistics, and then going to the core results ought to answer the main research questions.

Chapter 4 contains the conclusive arguments identifying which research questions did find answers and it underlines what they entail for the general literary context. Moreover, it provides a section summarizing the shortcomings of the research described in this thesis, their potential solutions, and future developments of the line of research presented in this work.

Finally, Chapter 4 is followed by the Bibliography, the Statement of Independent Work, and the Appendix, of which each different section contains additional demonstrative information, like additional tables, the code used in different phases of the thesis, and the different plots generated.

Notation, naming conventions and abbreviations

In this work, some terms will be used with a specific meaning, unless stated otherwise in the text, on a case-by-case basis. A list of the most used terms and their intended meaning in this thesis is here provided.

- **European or Europe**: both terms are used to indicate Europe as a geographic region, not to indicate the political organization.
- **American**: the adjective is used to indicate someone or something from the United States of America, not generally coming from the American continent.
- **EU**: Used as an abbreviation for Europe with a geographic meaning. It does *not* indicate the European Union as a political or diplomatic entity.
- **05-20 startups:** startups that were founded between 2005 and 2020, one of the main filters for Crunchbase extraction in Scipione's work.
- **Deals and funding rounds**: the terms will be used as synonyms unless otherwise stated. Similarly, when referring to tables in *italics*, *Deals* and *Deal* tables are used interchangeably to indicate tables that have on each row the specific funding found of a specific startup.
- **Startup and Company**: the terms are often used interchangeably. Scipione's work tends to use more the *Company* term. This work will mostly use the *Startup* or *Startups* terms to indicate tables containing on each row information on a specific startup.
- **Invested capital and total invested amount**: the terms are used as synonyms to indicate the total amount of funding collected by a startup in all its funding rounds.

Chapter 1
Theoretical Framework

This chapter will describe and mention some of the main themes developed by the entrepreneurship literature on the disparity between the European and American startup ecosystems, the power-law-like behaviour of some of the typical quantities of entrepreneurial phenomena, and on the methodologies that should be used when inquiring about potentially power-law-distributed quantities. Moreover, it will give a general overview of theoretical concepts used by this thesis such as an introduction to the power-law and the methodologies used to fit the data which this work inquires upon.

1.1 Differences between the US and European ecosystems

The literature on the topic has produced quite an amount of works on the matter. It shows quite clearly that the US, being the birthplace of venture capital (VC) markets and the fast-growing tech-based startup model, has a significant lead across a wide variety of VC market performance indicators. One of the most important works in this field, performed on a sample of companies from the period between 1997 and 2003, highlighted a significant gap between the value generated by US venture capital investments and European investments (Hege, Palomino, & Schwienbacher, 2009). The authors also reported finding significant differences between US and European VC firms with respect to their behaviour, indicating a more active role of the former.

The study focused on performing several statistical tests, mainly focusing on IRRs as a performance benchmark. The authors based their analysis on each round total startup valuation, rather than just on the round's invested amount (i.e. the startup's cash inflow). This allowed them to better focus on the economic value, as perceived by the market, created by the startup until that moment, rather than merely on the performance of the startup as a financial asset. Unfortunately, as discussed in Chapter 2 (Paragraph 2.1.3), this approach is not feasible with the data available for this thesis. Therefore, the analysis performed in this work must focus on the amount of cash invested in each round to calculate total invested amounts, and on the final exit values as the only data point that can be used to estimate the economic value created by startups and VCs (venture capitalists).

Moreover, as stated in the Research section of the Introduction chapter, it should be noted that the main goal of this study is an inquiry regarding the nature of the distributions of entrepreneurship performance measures. Therefore, the data-richness problem highlighted above is not that important. Given the different focus of this thesis, a comparison among the calculated distributions is still possible and is performed in Chapter 3 of this thesis. The comparison analyses samples of startups grouped by different combinations of their headquarter location, geographic origin composition of their investors, and, potentially, their industries.

1.2 Differences between the US and European ecosystems

The issue of the distribution of entrepreneurial phenomena has only relatively recently been focused on by the literature. Very often, the assumption of Gaussian distribution for a very wide range of metrics was made, either implicitly or explicitly, without checking whether the shape of the available data was compatible with a normal distribution. This very common behaviour of entrepreneurship research opened the way to a variety of practices, such as eliminating the outliers, that exposed researchers to potentially wrong or misleading conclusions in their studies.

However, new literature has been produced in the last decade which suggests that very often entrepreneurial phenomena cannot be described by a normal distribution, as they behave more similarly to heavy-tailed distributions, such as the power-law (Crawford, Aguinis, Lichtenstein, Davidsson, & McKelvey, 2015). In their study, Crawford and colleagues argue that by using Gaussian distributions to describe entrepreneurial phenomena researchers run the risk of excluding very important outliers that generate a significant portion of the outcomes, therefore reaching biased and potentially misleading results.

Subsequently, the researchers proceeded to test that many entrepreneurial phenomena are compatible with a power-law distribution. They limit themselves to showing that the data are more compatible with a power-law distribution than a normal one, but do not perform a *p*-value test of the obtained distributions, nor do they test for other heavy-tailed distributions. These couple of steps are quite important to corroborate that the data are distributed according to a power law – more on this in the next paragraph (**Error! Reference source not found.**).

One example of a suspected power-law distributed quantity is the MOIC of EIF-backed startups (Prencipe, 2017). In his work Prencipe analyses liquidity events and returns of startups backed by funds in which the EIF is a partner. Among the findings of his report, he shows that the MOICs of the startups in his sample are compatible with a power-law distribution.

Principe's work, however, performed on a large sample, does not provide a comparison between startups coming from different regions (Europe vs. the US) or among different geographic investor origin combinations. This thesis will try to precisely do that, not only to verify if entrepreneurial metrics are compatible with power-law distributions but also to understand if the mentioned geographic components imply different distributions.

1.3 Fitting power-law distributions to data

Ever since the idea of power-law or, more generally, heavy-tailed distributed quantities came about in the scientific literature, the trend which tried to identify and describe phenomena with power-law distributions significantly increased. Many fields of research started trying to adapt power-law distributions to their data. However, very often this was simply done by plotting these quantities on a log-log scale with the investigated quantity (e.g. random variable X) on the x-axis and its complementary cumulative distribution function (CCDF) on the y-axis (i.e. $P[X \geq x]$). If their data appeared as a straight line on the log-log chart, the authors often concluded that it was power-law distributed, other times they used least-squares fitting methods, which can often produce inaccurate estimates for the parameters of the distribution, but which at the same time often happen not to indicate whether the data obey a power-law or not. Therefore, the need for a rigorous method for estimating power-law parameters and for testing whether the data obeyed a power law was developed (Clauset, Shaliza, & Newman, 2009) (Clauset, Shaliza, & Newman, 2009).

1.3.1 Main characteristics of heavy-tailed and power-law distribution

Heavy-tailed distributions are described in probability theory, as probability distributions (PDs) with not exponentially bounded tails (i.e. their tails are heavier than the exponential distribution). They are of significant interest in many fields as they help in describing rare inputs that often produce a significant portion of the outcomes of a phenomenon or process.

1.3.1.1 A general definition of heavy-tailed distributions

A random variable X, of distribution function F, has a heavy (right) tail distribution if its moment generation function, $M_X(t)$, is infinite for all $t > 0$ (Rolski, Schmidli, Schmidt, & Teugels, 1999).

Using the tail distribution function, or complementary distribution function

$$\bar{F}(x) \equiv P[X > x] = 1 - P[X \leq x], \tag{1.1}$$

then a probability distribution (PD) is said to be heavy-tailed if:

$$\lim_{x \to \infty} e^{tx} \bar{F}(x) = \infty, \forall\, t > 0. \tag{1.2}$$

At a more intuitive level, if a random variable X is heavy-tailed distributed, then, if it ever surpasses a given large value x, it is also likely to exceed any value larger than OBJ.

1.3.1.2 The power-law functional form and its properties

Power-law distributions are a specific type of heavy-tailed distributions. They have the general form $f(x) = Cx^{-\alpha}$, where α defines the shape of the power-law and C is a normalization constant used to make the total under the curve equal to 1.

Among their properties, power laws are scale-invariant and lack well-defined average values. The first property implies that scaling the argument does not change the fact that the underlying variable is also power-law distributed, and the two distributions

only differ by a scaling factor: $f(kx) = C(kx)^{-\alpha} = Ck^{-\alpha}x^{-\alpha} = k^{-\alpha}f(x)$. This property is important for this thesis when preparing the samples (v. paragraph 2.3.2).

The second property is related to the fact that power-laws have a well-defined average over $x \in [1; \infty[$ only if $\alpha > 2$, and have a finite variance only if $\alpha > 3$ (Newman, 2005). The not-so-common property which requires $\alpha > 3$ for the power-law to have finite variance is the cause of the very common black swan behaviour (Taleb, 2007) of phenomena that are power-law distributed. Black swans, rare events that have a significant impact, on entrepreneurship are also known as unicorn startups, which have valuations higher than 1 billion dollars.

1.3.2 Methodology for fitting power-law distributions to empirical data

Following Clauset's and colleagues' work (Clauset, Shaliza, & Newman, 2009), this small paragraph gives an overview of the method they developed. It is up to the reader reading the article to dive deeper into its inner workings.

The mathematical equation for the probability density function of a power-law distribution is:

$$p(x) = \frac{\alpha-1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha}. \tag{1.3}$$

Where the lower bound x_{min} is needed to prevent the probability density function (PDF) from diverging for $x \to 0$, and α is the exponent of the distribution. The CCDF for this functional form is defined as

$$\bar{F}(x) = \int_x^{\infty} p(v)\,dv = \left(\frac{x}{x_{min}}\right)^{-\alpha+1}. \tag{1.4}$$

Using the two formulations above, the authors proceed to detail the following steps for fitting power-law distributions to empirical data.

1.3.2.1 First step: estimating the power-law distribution parameters

The first step requires estimating the parameters of the power-law distribution. Using the *method of maximum likelihood*, the authors first assume x_{min} to be known to get the first estimate *maximum likelihood estimator* (MLE) for the parameter α. Then they proceed to estimate x_{min}. The choice of x_{min} is quite important as picking too low a value will result in a biased distribution while choosing too high a value will reduce the usefulness of the resulting distribution, which will be able to explain a smaller than desired part of the data.

To better estimate the value of x_{min}, the authors suggest the use of the Kolmogorov-Smirnov or KS statistic and try to minimize it by varying the value of x_{min}. Selected x_{min}, they then proceed to re-estimate the parameter α. For more details on the rigorous mathematical steps, it is suggested to read Clauset's paper.

The method provided is suitable, with some adaptations that can be found in the original paper, to both continuous and discrete distributions.

It should be noted that the authors mention that it can be useful to look at the value n_{tail}, the number of observations included in the tail (i.e. the values of $X > x_{min}$ described by the estimated power-law). This value can be useful in evaluating if the obtained distribution is potentially a good fit: $n_{tail} \geq 1000$ usually indicates that the distribution could represent the data rather well. A similar concept is mentioned in a later paper co-authored by Clauset, which states that $n_{tail} \geq 300$ can also lead to good results (Virkar & Clauset, 2014). This concept should be kept in mind, as it is found to play a part in the results of this work (v. par. 3.4).

1.3.2.2 Second step: calculating the goodness of fit of the power-law

Considering the objective is evaluating if the empirical data can plausibly come from the calculated power-law, Clauset and colleagues build a hypothesis test, with the following hypothesis:

- H_0: the empirical data come from the estimated power law
- H_1: the empirical data do not come from the estimated power law.

To build the test and generate a *p*-value, they start with the parameters estimated for the power-law distribution fitting the empirical data (v. subpar. 1.3.2.1). Using those values, they generate a large number, N, of power-law distributed synthetic data sets. For each data set $i, i \in [1, \dots, N]$, they estimate the parameters of the power-law distribution that best fits those data points, applying the methodology discussed above.

For each synthetic data set i and its calculated power-law distribution, they calculate the Kolmogorov-Smirnov distance, KS_i.

The fraction of times that $KS_i > KS$, where KS is the Kolmogorov-Smirnov statistic calculated for the estimated power-law distribution in the first step of the process (v. subpar. 1.3.2.1), is the *p*-value for the test hypothesis. If its value is sufficiently large, then the null hypothesis cannot be refused, and the empirical data could be generated by the power-law estimated in the first step.

1.3.2.3 Third step: compare the power-law hypothesis with alternative distributions

The checks to be carried out to reasonably affirm that the empirical data could be power-law distributed do not end with the non-refusal of the null hypothesis. Other distributions could fit the data. It is, therefore, necessary to try to fit different distributions (e.g. the log-normal distribution) to the data, carry out the *goodness of fit* test described above for them too, and, if their null hypothesis cannot be rejected, then evaluate whether they are a better fit for the empirical data than the estimated power-law.

In case the fit of the estimated power-law is ruled out by the refusal of the null hypothesis for the *goodness of fit* test, the step described here is not needed, as the hypothesis that a power-law fits the empirical data has already been refused.

When the null hypothesis for the power-law *goodness of fit* test cannot be rejected, then it is necessary to perform the comparison mentioned above. Clauset and colleagues suggest using a *likelihood-ratio* test to try to understand which distribution type better fits

the data. The idea behind this test requires computing the *likelihood* of the data under the two competing distributions. The method looks at the sign of the log of the *likelihood ratio*, $\mathcal{R}$, and, based on its sign, infers which of the two distributions is better than the other, or, if $\mathcal{R} = 0$, it indicates equivalent *goodness of fit* to the empirical data of the competing distributions. However, given that the sign of $\mathcal{R}$ can easily be influenced by statistical fluctuations, the method suggests the calculation of a *p*-value based on synthetic data in a similar manner to that used in estimating the *p*-value for the *goodness of fit* test. If said *p*-value is sufficiently small, then a winning distribution can be selected. Moreover, the described method allows determining if the provided data is insufficient to determine a winning distribution. For a deeper description of the method, this thesis recommends Clauset's (Clauset, Shaliza, & Newman, 2009) and Vuong's (Vuong, 1989) papers.

1.3.3 R statistical programming software implementation

When looking for a statistical package able to implement the methods described above, there are different options available. Concerning the mathematical and statistical software environments, the main choices are MATLAB, Python and R. The choice for this thesis was to use the R language as it is open-source software with simple plotting functionalities and is more tailored to the kind of statistical analysis performed here than Python. From some general searches, it seemed R was better suited to fast implementation and usage for this thesis. Moreover, it gives users easier access to plotting functionalities.

Please note that all the analyses performed with R were performed with R Statistical Software (v4.2.0; R Core Team 2021).

Chosen R, there is a need for an R library able to implement the power-law fitting methodology developed in Clauset's paper. The chosen package should allow for easy implementation of the methods previously described (v. par. 1.3.2), allowing the focus of this thesis to remain on the test and interpretation of power-law distributions to entrepreneurial phenomena across different geographic settings.

One of the most appreciated and widely used R packages for such analysis is the R *poweRlaw* package (Gillespie, 2015), which is the package chosen to perform the analysis in this thesis. A brief overview of the main functions provided by this package is provided in the paragraphs below, however, it is suggested that the interested readers deepen their knowledge of the package by consulting Gillespie's paper and the poweRlaw package documentation[1].

The following subsections mirror the structure of the subsections from the previous paragraph (par. 1.3.2).

1.3.3.1 Preliminary step: setup and data import

The first step is to import the R package. On the chosen development environment, RStudio, this is achieved by clicking on *tools and* then on *install packages* to open a menu

[1] The documentation for the poweRlaw package is available at the following web address: https://www.rdocumentation.org/packages/poweRlaw/versions/0.70.6.

in which in the CRAN directory it is possible to type *poweRlaw*, once the package is selected, clicking on *Install* will complete the step. This can also be obtained, by typing:

```
R> install.packages("poweRlaw")
R> library("poweRlaw")
```

where the first line installs the package and the second imports it into the workspace.

Before running any function of the package it is necessary to import the data on which to perform the analysis. This can be done in different ways. However, the chosen method for the work described in this thesis is importing data from MS Excel sheets, which is where the single datasets were prepared (v. par. 2.3.2). Data can be imported from Excel into the R workspace using the following code:

```
R> install.packages("readxl")
R> library("readxl")
R> db <- read_excel("C:/Users/96fra/chosenFolder/chosenFile.xlsx",
sheet = "chosenSheet")
```

which installs and imports the *readxl* package and imports data, preferably from a sheet which contains the data as a two-column table with headers. The variable `db` is an R object that contains the imported data, and it is also possible to reference only one of its columns by typing `db$chosenColumnName`.

1.3.3.2 First step: fitting a power-law distribution

Once the data is imported and the *poweRlaw* package is installed, it is possible to estimate the power lad distribution that better fits the data available using the code below.

```
R> pl_m <- displ$new(db)
R> pl_m$setPars(estimate_pars(pl_m))
R> (est_pl <- estimate_xmin(pl_m))
R> pl_m$setXmin(est_pl)
```

The code sample first creates a *power-law distribution* object, then it sets a starting value of the parameter α, which it then uses to calculate the MLEs for the power-law parameters, x_{min} and α (v. par. 1.3.2.1). Finally, the code sets the parameter values of the *power-law distribution* object equal to the estimate results (`est_pl`)[2].

Subsequently, it is possible to plot the data points and overlap the estimated distribution using the code below.

```
R> plot(pl_m)
R> lines(pl_m, col = 2)
```

The software will then visualize the images of the plot and allow the user to export them in different formats.

[2] It is possible to visualize the variables by simply typing their names (e.g. `R> est_pl`). The `est_pl` object also contains information on the number of data points included in the estimate, (v. n_{tail} from par. 1.3.2.1).

1.3.3.3 Second step: bootstrapping

The second step tries to estimate the *goodness of fit* of the *distribution* object. The function

```
R> bs <- bootstrap_p(pl_m, no_of_sims = 5000, threads = 4)
```

implements the *goodness of fit* test procedure to produce a *p*-value described in par. 1.3.2.2.

The inputs, together with the *distribution* object (`pl_m`), can include the number of simulated sets, the number of threads used to run the simulation and other parameters.

It should be noted that this is the part of the code that requires the longest amount of time to run, as it performs MLE calculations on a very large number of simulations.

The results, saved in the `bs` variable, can be accessed simply by typing the variable name as an R command in a new line (e.g. `R> bs`). The software will visualize the information contained in the *bootstrap* object, among which there are the *p*-value and the *KS* measure of *goodness of fit*. It is therefore straightforward to rule out the compatibility of the data with the estimated distribution if the *p*-value is low enough (e.g. less than 0.1).

1.3.3.4 Third step: comparing the estimated power-law to other distributions

If the hypothesis of the data being described by a power-law distribution cannot be rejected, then it is necessary to verify whether other distributions fit the data as well, and then compare their *goodness of fit* with that of the estimated power-law (v. par. 1.3.2.3).

The first step requires replicating similar steps to those presented in the previous paragraph, creating a new *distribution* object of a different distribution type (e.g. log-normal). Subsequently, it is necessary to estimate the parameters of the new distribution and to test its *goodness of fit* and its *p*-value. This is showcased below[3].

```
R> ln <- conlnorm$new(db$MultipleVal)
R> ln$setXmin(pl$xmin)
R> ln$setPars(estimate_pars(ln))
R> bs_p_ln <- bootstrap_p(ln, no_of_sims = 500, threads = 6)
```

If the resulting *p*-value for the alternative distribution does not allow to reject the null hypothesis that the alternative distribution fits the data, it is then necessary to compare it with the estimated power-law.

The procedure to perform the comparison in the *poweRlaw* package is implemented by the following line of code:

```
R> comp <- compare_distributions(pl_m, ln_m)[4]
```

Which implements in the R language (Gillespie, 2015) Vuong's method (Vuong, 1989) to determine which distribution better fits the data among two competing distributions.

[3] The `conlnorm` distribution type at line 1 corresponds to a continuous log-normal distribution.

[4] The arguments are the two distribution objects, the first, `pl_m`, representing the power law, and the second, `ln_m`, representing the competing log-normal distribution. The order of the arguments is important as it determines the direction of the comparison.

The hypothesis tested by the function are:

- H_0: the two distributions are equally far from the true distribution
- H_1: one of the two competing distributions is closer to the true distribution.

Hence, a sufficiently small p-value is needed to reject H_0 and state that one of the two better fits the data (e.g. $p < 0.1$).

On the contrary, if H_0 cannot be rejected, it is not possible to state that the data is, for example, likely to come from the estimated power-law distribution.

Finally, it should be noted that even if H_0 is rejected in favour of the power-law distribution, this does not mean that the best possible distribution to fit the data is a power-law. On the contrary, it only means that the power law better fits the data than the competing distribution. There could be other distributions that would fit the data just as well as, or even better than, the estimated power law. Therefore, to have the most accurate possible results, one should keep testing the estimated power-law against other competing distribution types theoretically compatible with the data until either the power-law is rejected or there are no more distributions left to test.

Performing such a high number of competing distribution tests is not practical and, in most cases, knowing that the power-law could fit the data better than a competing distribution could be enough to state that it is plausible that the data is power-law distributed, accepting the risk that another – untested – distribution better fits the data.

Chapter 2
Dataset and Methodology

This chapter describes how the available data shaped the development of the methodology for this work. The most significant limiting factor for this work was given by the restricted available data. Due to budget, data availability and time constraint, it was not feasible to collect a new dataset from scratch. Therefore, it was necessary to utilize already available data to build a dataset. It was therefore the data available that determined what this work could inquire. As shown in the next paragraphs, missing information, distortion biases, data quality and sample dimension are determinant factors in what could be extrapolated from the available information.

Hence, the research questions that could be answered, or at least inquired upon, have been quite significantly influenced by the quality and quantity of the dataset itself. One of the leading questions of the preliminary stages of this work was in fact *what hypotheses can be formulated and tested with the available data?*

2.1 Description of the original dataset

The data used in this work come from the dataset collected by Scipione[5] in his master's degree thesis (Scipione, 2020). His data, mainly divided into five MS Excel workbooks, contained information retrieved from the commercial database service Crunchbase. Table 2.1 shows how the information was organized among Scipione's files.

The type of data can be easily conceptualized by thinking of it in terms of just three object classes: *Startup*, *Deal*, and *Investor*. The tables *EU Startups* and *US Startups* contain information on single startups (i.e. each one of their rows corresponds to a single startup), while the tables *EU Delas* and *US Deals* contain information on single financing rounds (i.e. each one of their rows represents a financing-round for a specific startup), and finally, the table *Investors* contains information on single investors (i.e. each one of its rows contains information on a specific and unique investor).

The exact Crunchbase queries used by Scipione are not reported in his work. Therefore, it was not possible to establish exactly all the filters applied to the data, nor if it was somehow manipulated in Excel after extraction. A meeting with Scipione revealed that in some cases some arbitrary data truncation was performed when the number of rows returned by Crunchbase was too large.

The main filters for the tables containing information on startups and funding rounds are not too restrictive, as they only required the startups to be headquartered in the EU or US, to have been founded between 2005 and 2020 and to have already had an exit of M&A or IPO type. Therefore, it does not seem like the data on startups and deals have significant distortions related to how the query has been performed.

[5] The author of this work would like to reserve a special thanks to Andrea Scipione for sharing his original dataset and for dedicating time for an introductory call about his work.

Table 2.1. *The subdivision of data in Scipione's work.*

File	Table	Has Exit[6]	Sample Description	Sample Dimension
1	EU Startups	Yes	05-20 startups with HQ in the EU	8817
1	EU Deals	Yes	Funding rounds of 05-20 startups with HQ in the EU	5312
2	Investors	-	Investors with HQ in all regions that invested in 05-20 startups with HQ in EU or US	9658
3	US Startups	Yes	05-20 startups with HQ in the US	15653
4	US Deals	Yes	Funding rounds of 05-20 startups with HQ in the US	15055
5	All startups	No	05-20 startups with HQ in the EU or US	9982

Some attention must be given to the query used to download the investor data. The applied filters require investors to be headquartered either in Europe or in the US. However, in his work, Scipione mentions that the result was ordered by Crunchbase's ranking score, the number of lead investments and the number of IPO exits to select only the top investors. Therefore, it seems that the investor table was truncated by only selecting top-performing investors, which could introduce biases in the data for certain types of analysis. This element will be dealt with in the later paragraphs when discussing what type of information can be extrapolated from the available data.

More detail on the columns in the original tables is provided in Columns of the original dataset tables in Appendix A.1.

2.1.1 Crunchbase's shortcomings

To be able to formulate a sensible methodology, it appears necessary to examine the possible distortions, biases and problems that could arise from Crunchbase. It is well known that commercial datasets are often plagued by many problems that should be considered when designing experiments or extracting statistics on data coming from these platforms.

Some of the most well-known problems are the *reporting bias*, the *bias toward US investments*, the *lack of complete information on financing rounds*, the *under-representation of the Biotech sector*, and the *difficulty in knowing the status of investments whose outcome is other than IPO or acquisition* (Prencipe, 2017)

Even if Crunchbase is among the three most accurate commercial databases (Retterath & Braun, 2020), it is still plagued with the above-mentioned problems. Particular attention should be paid to the *reporting bias* which tends to create distortions in the reported financing round information as, given that, usually, the information on rounds is self-reported by startups or venture capitalists (VCs), there is a tendency to report larger financing rounds more than smaller ones. This distortion could impact results relating to the total funding received by startups, and other indicators.

[6] For the tables containing startups as rows, one of Scipione's selection requirements was that startups had an exit. File 5, for which he did not require startups to have had an exit, is an exception. This extraction is not going to be used in this work and will not be explained further. Please, consult Scipione's thesis to find out more about this.

Moreover, the *bias towards US investments* could lead to an overrepresentation of the American ecosystem compared to the European one. This must be considered when using Crunchbase to compare the dimensions of the American and European ecosystems.

The *lack of complete information on financing rounds* is potentially a significant factor that could heavily impact the availability of complete data and therefore reduce the sizes of complete and representative samples and therefore the accuracy of the inferences on the populations that are *object-of-interest*[7]. Hence, a potentially biased *object-of-study* population[8] (i.e. the dataset available for this work) could produce biased results.

2.1.2 Data-related problems in previous work

Before starting to simply export and use Scipione's datasets, a few checks on his data reveal some potential problems with the way his data were encoded. This short paragraph names a few of the most important ones.

In all the six tables mentioned in Table 2.1. The subdivision of data in Scipione's work. Table 2.1, there were character encoding problems. In fact, many names of startups, cities or other fields, presented out-of-context characters that had to come from the file data being imported into excel using the wrong encoding. This encoding problem could lead to problems in the search formulas, the readability of results and their aggregation.

In addition to this, it seemed that the Startup table column *Funding Status,* used in Scipione's work to determine if the exit type had been IPO or M&A, referred to the current funding status of the startup at the time of the Crunchbase query. Therefore, its use to determine the exit type could have led to distorted outcomes in determining the frequencies of one exit type compared to the other.

Another important aspect to consider is that it seemed that Scipione used the *Name* field as the primary keys for the tables: *Organization Name* as the primary key for the *Startup* tables, *Transaction Name* as the primary key for the *Deals* tables, and *Organization/Person Name* as the primary key for the *Investors* table. Given the extensive use of search and comparison formulas, having duplicate values in these fields could generate distortions in the results.

In his extraction, Scipione considers only companies that have either an M&A or an IPO exit. This could introduce some distortion in the data. Nonetheless, usually, those two are the only exit types considered in the relevant literature.

Paragraph 0 deals with how the criticalities highlighted here are handled when creating the data model used in this thesis.

[7] The *object-of-interest population* is the population on which a statistical study tries to produce inferences. In the case of this work, the *object-of-interest population* is the population of all startups, deals and investors belonging to the European and American entrepreneurial ecosystems.

[8] The *object-of-study population* is the population from which the samples are drawn to produce inferences on the *object-of-interest population*. In this case, the *object-of-study population* is the population of all startups, deal, and investor records in Crunchbase and extracted by Scipione in his work.

2.1.3 The research potential of the available data

Some of the most concerning aspects regarding the data available are the *representation bias* and the *US-centricity bias* coming from Crunchbase's dataset. Results coming from the analyses performed on the available data trying to compare the European and American ecosystems in their entirety are at very significant risk. Crunchbase is an American company based in the US and it seems therefore reasonable to think that a bias towards a better coverage of the American ecosystem is to be expected.

Hence, this work tends to avoid comparing the two ecosystems (US vs EU) by using absolute measurements such as total amounts (e.g. the total number of startups in the two ecosystems, the total funding amount in the observed period, and the total exit values, etc).

The bias towards a higher likelihood of representation that larger funding rounds have compared to smaller financing rounds can be problematic for some types of analysis that try to inquire about quantities like the number of funding rounds performed by startups. However, the impact of missing information on small financing rounds on the total amounts invested in each startup does not constitute an issue as the relative error on the total will be small. For example, if a startup had 5 rounds for a total invested amount of 5 million dollars, but on Crunchbase the 1st round of 50 thousand dollars was missing, the calculated total invested amount from Crunchbase data would be 4.5 million dollars, with a relative error on the total invested amount of only 1%. Therefore, when performing analysis on quantities like the investment amounts on single startups this kind of bias is not particularly important.

Considering all the above, it seems reasonable to think that trying to fit power-law distributions to the investments, the exit values, and the multiples should not be too heavily affected by the discussed biases. In fact, having a higher percentage of American Startups compared to European ones, should not be a problem when fitting distributions to region-based samples (e.g. fitting a power-law to the exit values of 05-20 startups headquartered in the EU, and *then* comparing the obtained distribution to that obtained from a similar sample of startups headquartered in the EU). The US overrepresentation bias could be problematic when trying to fit samples made of startups from both regions as the composition of the sample would not be representative of the population of interest – i.e. the population of startups headquartered in Europe or the US. The sample would be biased with a higher percentage composition of US startups compared to the population that is the *object of study*. Therefore, inferences made on the population of interest could be distorted.

In conclusion, it seems acceptable to use Scipione's original data to perform an analysis on the distribution of certain quantities that are not heavily impacted by the *representation bias*, such as exit values, invested amounts and multiples, and on samples that are not affected by the *overrepresentation of the US ecosystem bias*. Inferences on other quantities, such as total dimensional comparisons requiring summing up quantities over all the companies belonging to the samples should be avoided. For example, comparing the total funding for all European vs American companies to infer how much bigger one market is than the other could be affected by the *US overrepresentation bias*.

2.2 Methodology

In the previous paragraph, it was established that the research questions this thesis inquires upon can theoretically be answered with the data available, without having too large a distortion on the inferred results because of Crunchbase's inherent biases.

This paragraph discusses how the research methodology should be structured to prevent introducing distortions and to obtain the most accurate possible results. Moreover, some assumptions must be made about unknown behaviours of Crunchbase and the distribution of missing company information.

Reasons for the choice of which quantities to inquire upon are provided (e.g. why this work inquires about the distribution of multiples instead of that of IRRs).

Finally, the steps of the analysis are planned.

2.2.1 Assumptions

The assumptions to be made for the planned analysis span across different levels: those to be made about Crunchbase, those about Scipione's extracted data, and those about the distribution of the errors controlled for by the logical consistency filters applied to Scipione's data.

2.2.1.1 Assumptions on Crunchbase

Some of Crunchbase's biases and why they should not be too problematic for the analysis performed in this work have already been discussed in par. 2.1.3. However, a few assumptions must be made about the information collected by Crunchbase.

Firstly, to be able to infer about general American and European populations using the information available on Crunchbase (CB), it must be assumed that, even if not all startups present in the two ecosystems are present in CB, those present in CB are as close as possible to a random sample – i.e. the *object-of-study population* of all EU startups in CB must be a representative sample of the *population of interest* of all European startups. For this to be possible, overall, all startups, founded in the 2005-2020 period, must have an equal likelihood of being inserted in CB. This is not likely for all startups, as when Scipione performed the extraction in 2020, the startups that had been recently founded had less time to be inserted in CB. However, since this analysis focuses on startups founded between 2005 and 2020 that *had already had* an exit, it seems reasonable to think that startups that had an exit, must have been founded at least 1 or 2 years before the date of data extraction.

The dataset contains some companies with a very short *time to exit* (i.e. time elapsed between the founding date and the exit date) which would undermine the argument for a period of 1 or 2 years needed by a startup before having an exit. However, it seems that these companies are vehicle companies, which therefore don't represent the definition of a startup that is here inquired, and, hence, they do not create distortion.

Given that startups founded in the 2005-2020 period that had already had an exit, must have been founded at least about 1 or 2 years before the day the data was extracted

by Scipione, it seems acceptable to think that startups in CB, and Scipione's dataset, represent fairly the population inquired upon by this research. It can be argued that startups that were founded at least 1 or 2 years prior to the data extraction should have a similar probability, at least timewise, of being inserted in CB. The same argument made here for the European startups can be made for the American ones.

There could be other factors that influence the likelihood that CB is a representative sample of the *object-of-interest population*, however, these cannot be fully controlled in this study and could be explored by future research.

Even if it was determined that CB's startup population (the *object-of-study*) was not representative of the *population of interest*, the inferences made by this research can at least be applied to CB's startup population, provided that the assumptions in the following paragraphs below hold true.

Secondly, the *Deals* tables should respect the conditions mentioned above by having deal information about startups that obey the abovementioned criteria. The main problem with CB's *deals population* is missing information on financing rounds for many of the startups in the *Startup* population (v. par. 2.2.1.3).

Observing the number of entries in each of the *Deals* tables, 15055 for US startups and 5312 for EU ones (v. Table 2.1), it is possible to notice that they have fewer rows than the *Startup* tables. Given that Scipione imposed that the startups in the *Deal* tables respected the same filters imposed on the *Startup* tables if CB had complete information and that every startup had to have at least one deal, the number of entries in the *Deal* tables should equal at least that of the *Startup* tables. This showcases the fact that CB has significant missing information on deals.

It should be assumed that the subset of startups that do have deal information is still representative of the *population of interest* and of the *population* that is *object-of-study*.

2.2.1.2 Assumptions on Scipione's dataset

The assumptions about Scipione's dataset regard the number of entries present in his tables. It must be assumed that all the tables in his dataset contain all the startups in CB obeying the constraints imposed in the CB queries described above and in his work.

Similarly, it must be assumed that the *Deal* tables downloaded by Scipione contain all the deals available in CB respecting the filters he imposed.

Moreover, it should be assumed that, when he mentions that duplicate entry removal operations were performed on his datasets, his operations did not compromise the data. The potential problem, identified in par. 2.1.2, arises from him using the *Organization Name*, *Transaction Name*, and *Person/Transaction Name* fields as primary keys for the tables *Startups*, *Deals*, and *Investors* respectively, which could lead to the elimination of entries. However, some duplicate values in each table were found in these fields, so this problem can probably be discarded and simply related to his use of those fields in his *VLOOKUP* Excel formulas, but not in his preliminary data cleaning operations.

The *population of interest* representativeness for the population of European and American investors cannot be assumed for Scipione's datasets, as he mentions truncating

his CB query results to only select the biggest investors, according to CB ranking and other criteria (v. par. 2.1). This produces an *Investors* table that can be non-representative of the general population of investors. However, this is not directly a problem for this thesis as it does not directly use investors from the downloaded dataset. The focus is on the startups and deals.

What this work needs to assume is that requiring the samples of startups and deals to have at least one of the investors from the *Investors* table (to derive investor origin information) does not introduce an unbearable distortion. Unfortunately, due to the significant lack of information present in CB's data and due to Scipione's truncated Investor data, this distortion risk cannot be avoided. However, as discussed in par. 2.3.2, It is partially mitigated by not requiring every single investor reported in each round to be present in the *Investors* table. The requirement is for at least one investor in each round to be present in the *Investors* table. The requirement arises from a data scarcity problem, and it leaves room for uncertainties about the true investors' geographic origin composition, but, at the same time, it mitigates the unrepresentativeness problem of the *Investor* table.

2.2.1.3 Assumptions on single CB columns meanings on the applied logical filters

The information incompleteness problem plaguing CB must be addressed to create logically consistent samples. However, it must be assumed that the logical filters imposed upon the tables when creating the samples did not create distorted samples themselves. Therefore, the general assumption must be that information incompleteness is randomly distributed in CB's dataset.

Hence, when, for example, filtering out companies that have no exit value information, the sample of companies that do have an exit value must remain representative of the population of companies with exits.

This assumption cannot be entirely verified for all fields, however, from preliminary data analysis and dataset exploration, it seems that the missing information problems apply to all almost all table entries. Almost all of them present at least one field missing information and it seems that there is not a clear pattern determining a higher probability for certain entries to have specific missing information.

Finally, it is not always possible to determine with certainty the meaning and the way that the content of certain CB columns was gathered. The most significant one, is, for example, the *number of rounds*. It is assumed that, if the field is not empty, the number inserted there represents the real total number of rounds, and not just the number of deals reported in Crunchbase. This assumption is partially corroborated by the fact that for some startups in the *Startup* table that have rounds in the *Deals* table the value indicated in the *Number of rounds* field is greater than the deals reported in the *Deals* table. Under the assumption that Scipione's queries retrieved all deals available on CB for the startups in the *Startup* table, this observation suggests that the reported *Number of rounds* indicates the actual number of rounds.

2.2.2 Analysed quantities

In this paragraph, detail is provided on the research perimeter and how it is sectioned: discussing the choices of the variables to analyse and the sample sectioning criteria and dimensions.

2.2.2.1 Variable choices

Once described the datasets, their potential distortions and the assumptions, this paragraph describes the chosen variables to be studied and justifies their choice.

Under the established assumptions, it is not possible to compare the US and EU ecosystems in terms of total values, due to the *US overrepresentation bias*, however, it is possible to compare the behaviours and distributions of some variables in the American and the European ecosystems.

As mentioned in the Research questions paragraph, the chosen variables in this work are, for a generic startup, the *total funding amount*, the *exit value*, and the *multiple on the invested capital* (MOIC).

A list of some of the main candidates as research variables together with the main reasons for their inclusion or exclusion from the research perimeter is provided below.

Exit value: it indicates the total economic value generated by the startup until the exit. It is chosen as it is one of the most important descriptors of a startup and, when looking at its probability distribution, it can help understand how the ecosystems behave. The data available allowed this variable to be used in this work.

Invested capital: it indicates how much is invested in startups and how capital intensive they can be. Moreover, looking at its distribution in an ecosystem can give information regarding the level of concentration and the investing capacities of the above-mentioned ecosystem.

Number of rounds: it describes the number of rounds which startups normally have. It could be useful if used together with the *time-to-exit* to determine how involved investors are and how closely the funding process is monitored. This is not explored in this work as the focus is more on the financial aspect.

Time-to-exit: it describes the time elapsed between the founding date and the exit date of a startup. It can be useful in determining and comparing how long startups normally take before exiting. It can be used in conjunction with the number of rounds as described above. The data available do not allow a precise calculation of this value as Crunchbase's precision of the founding dates and exit dates varies significantly, as it can be daily, monthly, or yearly. For a startup with a TTE of 4 years calculated starting from founding and exit dates with yearly precision, the relative error could be as much as 25%, rendering the analysis subject to errors. A more precise dataset is needed.

IRR: the IRR is a powerful instrument to evaluate a startup's economic value generation capacity and great performance indicator. However, in the case of this thesis, the IRR is not used as it requires a level of precision and uniformity in the data that is not available. In particular, the high variability in the precision of the dates provided for each

financing round makes it impossible to calculate IRRs homogeneously and coherently. In addition to this, power-law fitting does not allow for negative values, and many startups have a negative IRR. These would have to be discarded from the samples used to calculate the estimated distributions, an operation which would further reduce the sample sizes, distort the results, and reduce their descriptive power on the ecosystems.

MOIC: the choice of the MOIC as a performance measurement allows to include startups with non-successful exits in the samples used for estimating the distributions. If estimated using MOICs, non-successful exits are positive numbers smaller than 1 and can therefore be included in the samples when fitting power-law distributions. Moreover, MOICs are not hugely sensitive to missing data or *underreporting of small rounds bias* because they are calculated starting from exit values and investment values, which are only marginally affected by *underreporting of small rounds bias*.

2.2.2.2 Dataset sectioning dimensions for sample creation

When estimating the distribution parameters, the sample on which they are calculated is important, not only because its representativeness influences the significance of the results, but also because, by being compared to different starting samples, it allows to compare different populations. Each sample can be produced by sectioning the cleaned dataset (v. par. 0) along different dimensions.

Startup geographic dimension

Since this work tries to spot differences among the European and American ecosystems, the main sectioning dimension is the *headquarters region* of the startup. Therefore this work identifies the following possible values for the *headquarters region*:

- *All*: used to include all regions (as discussed above samples composed in this way could be affected by the *US overrepresentation bias*)
- *US*: used to select companies headquartered in the US
- *EU*: used to only select companies headquartered in Europe

Optional – startup industry

The *industry group* field is not used in this thesis as a sample sectioning dimension, as all industries were included to have the highest possible sample numerosity. However, this could lead to an *underrepresentation of the biotech sector bias* in the samples and the results. When sampling along this dimension, a similar startup may belong to multiple industries. This is because the *Industry* and *Industry Group* fields in Crunchbase contain a list of keywords that create sets with non-null intersections. This is not necessarily a problem as it could help with having higher sample dimensions, besides a startup can compete in different industries. Some of the mentioned industries may not be industry sectors, but rather technologies used by the companies to deliver their product, rather than industries per se (e.g. AI is often used as an industry marker, but it is a technology used to deliver a product or service in a specific industry). However, if paired with a more precise dataset, this could allow sectioning samples along two new dimensions: the industry in

which the company competes and the technologies it works with to deliver its product/service.

Investor origin geographic dimension

An interesting dimension to use when sectioning samples is the geographic composition of the investors that invested in the startup. This dimension can be affected by the problems discussed in the Assumptions paragraph. However, it presents a remarkably interesting possibility to compare how different investors behave by looking at the investor's geographic origin composition of each startup. The created dimensions are:

- *InvAll*: the investors that invested in the startup are from all regions
- *InvEU*: all the investors (for which the geographic origin is known) that invested in the startup are based in Europe
- *InvUS*: all the investors (for which the geographic origin is known) that invested in the startup are based in the US
- *InvEUandUS*: the startup had at least one American and one European investor.

It should be noted that when evaluating the geographic origin composition of investors of a startup the investors from all investing rounds are considered. Hence, to have this information, the startups considered in the analysis must be taken from the *Deals* tables, as only those contain information on the investors that were present in each round.

Sample selection based on the variable of the study

Depending on the variable under study, *exit value*, *investment amount*, or *MOIC*, the starting population on which to apply the sectioning dimensions varies. Depending on the dimension and the variable along which one wants to section the data, the requirements change. The goal is to have the most numerous possible samples.

For example, if one wants to find the distribution of *exit values*, for *US* startups, independent of the investor's geographic origin composition, one can take startups that have exit values but do not necessarily have complete information on invested amounts, from the *US Startup* tables.

On the contrary, if someone wants to find the distribution of the *MOICs* of EU startups that only have EU investors, they must take startups headquartered in Europe from the *Deal* table that only have investors headquartered in Europe across all their rounds. Given the starting population of startups in the *Deals* table is smaller than that in the *Startups* table, the selected sample is going to be smaller.

To summarize, Table 2.2 provides an idea of how the sampling would work on the starting datasets.

On the rows, it shows the *startup headquarter region* dimension values, on the columns the *investor geographic composition* dimension values, while in the cells it contains three starting populations options, based on the desired variable under inquiry, on which to apply the dimensional filters.

Table 2.2. *Sampling dimensions. Starting dataset table from which each sample is formed depending on the chosen dimension values.*

Inv Origin → **↓ Startup HQ**	InvAll	InvEU	InvUS	InvEUandUS
All	Exit: *Startup table with exit values (STwEV)* Inv: *Startup table with tot investments (STwI)* MOIC: *STwEV&I (STwEV&I)*	Exit*: Deal table with exit values (DTwEV)* Inv*: Deal table with tot investments (DTwI)* MOIC*: Deals Table with exit values and tot investments (DTwEV&I)*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*
US	Exit: *STwEV* Inv: *STwI* MOIC: *STwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*
EU	Exit: *STwEV* Inv: *STwI* MOIC: *STwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*	Exit*: DTwEV* Inv*: DTwI* MOIC*: DTwEV&I*

2.2.3 Analysis steps

The last aspect of the methodology to define remains the definition of the steps to be taken to reach the final goal of this thesis, which is to determine the plausibility that the variables *exit values*, *investment amount*, and *MOIC* are power-law distributed.

2.2.3.1 General step outline

Firstly, Scipione's datasets need to be refined and prepared to extract all the different samples needed for the analysis. The precise way in which this is accomplished by this thesis is described in the next paragraph,

Dataset construction.

Secondly, once the data has been prepared and the data model created, the samples, for each combination of sample sectioning dimensions and variables (v. Table 2.2), need to be extracted from the new data model and prepared to be read by the R script used in the distribution estimations (v. par. 2.3.3).

Thirdly, the simulation must be run for each of the prepared samples, its results must be saved in a summary Excel workbook, and the plots for each distribution must be exported as images and saved in corresponding folders. Information on how the R script for fitting distributions works is provided in paragraphs 1.3.2 and 1.3.3.

Finally, reviewing the obtained plots and the results summary table, which contains the summary information on each distribution fitting estimate, allows this work to make the desired comparisons and draw the conclusions for this thesis.

2.2.3.2 Focus on the R script distribution parameter estimation step

This section briefly describes what the script sample in Appendix B.1 does. The first operation is importing the desired sample, after which the power-law distribution is estimated. A p-value is estimated for the power law and both the distribution, and the results of the bootstrapping procedure are plotted.

Secondly, the code prepares the competing distribution, by taking a log-normal distribution with the same cut-off parameter, x_{min}, as the previously estimated power-law distribution. The code then estimates the MLEs for the log-normal parameters performs the bootstrapping test for it as well and plots all the results.

It should be noted that only the lognormal distribution is chosen as a competing distribution for time and calculating power constraints. However, by looking at the shape of the data in a plot, it seems that the shape it follows is very similar to that of a log-normal distribution.

Thirdly, a comparison test is performed between the two estimated distributions.

Finally, the data is fitted to another log-normal distribution, however, its cut-off parameter, x_{min}, is estimated autonomously in this case. The reason behind this choice is that, as mentioned above, by looking at the plots during a few trial runs of the R script, it seems that the plotted points have a very similar shape to that of a log-normal distribution, which could therefore have a lower cut-off value, explaining a higher portion of data with the chosen distribution (v. Figure 2.1).

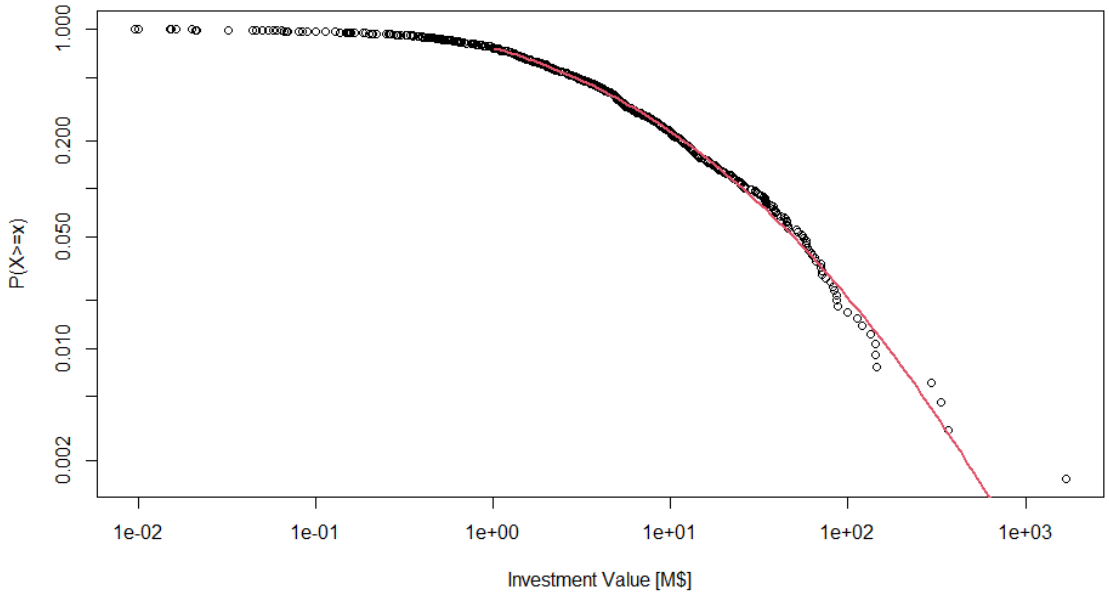

Figure 2.1 *Example of how well a log-normal with a low cut-off value seems to fit the data point.*

2.3 Dataset construction

The steps to take in preparing the dataset can be subdivided into three main categories: taking the data from Scipione's dataset, cleaning it, and making it available for analysis, creating the data model and those taken to create the data samples.

Each category is explained in one of the paragraphs below.

2.3.1 Data cleaning and preparation

The first step to undertake is the creation of original data backups. Therefore, Scipione's Excel worksheets are copied and stored in a dedicated backup folder. Subsequently, the encoding problem, mentioned in par. 2.1.2, needs to be tackled.

2.3.1.1 Solving the encoding problem

Each of the original 5 tables (v. Table 2.1) that need to be used in this work is exported in a CSV file that uses the ";" character to separate column fields and that wraps each column field in quotation marks. This is achieved through a series of concatenation Excel formulas that prepare the whole table to be exported by copying and pasting all the rows into a *.txt* file. Then the file is saved as a *.txt* file, making sure that the encoding option is set to ASCII. The ASCII encoding ensures that all the previous encoding problems are solved.

Subsequently, the data needs to be imported into the Excel sheet that will serve as the basis for the new clean dataset. To do this, the saved ASCII encoded *.txt* file is opened again, its content copied and pasted back into Excel. The *text import wizard* option must be opened and the *column separating character* option must be set to ";", the *column value identifier* option must be set to the quotation mark. Finally, the column types must be selected.

The steps to take are those just mentioned. Exporting in CSV and re-importing with a different encoding was found not to work, as within column values there are often values separated by commas, that once the *.csv* is imported back into Excel are mistaken for separate fields and compromise the whole dataset.

2.3.1.2 Joining tables

It is quite cumbersome to keep the 5 new encoded tables, *US Startups*, *EU Startups*, *US Deals*, *EU Deals*, and *Investors* in 5 different worksheets, especially those two pairs (*Startups* and *Deals*) which refer to the same object class. Hence, the tables *US Startups* and *EU Startups* are joined in a new table called *Startups*, after making sure the columns coincided. The same is done for the *US Deals* and *EU Deals* tables, joined in a single *Deals* table.

The three new resulting tables, *Startups*, *Deal*, and *Investors* are saved into three separate workbooks and backed up. The next steps are performed on these new joined tables.

It should be noted that joining the tables does not represent a problem as the field *Headquarters Location* still contains all the information needed to be able to distinguish between EU and US startups and deals.

2.3.1.3 Column selection

The number of columns in the original dataset is quite cumbersome and many fields are either not useful or report redundant information. Therefore, some columns have been eliminated from the newly created tables. Moreover, the name of some of the columns was changed to a shorter format to save screen space.

Some of the eliminated columns are, for example, all the columns referring to currency amounts. By looking at the original column names, it can be noticed how each currency amount has three columns: an original currency amount, an original currency label, and a USD value. The first two column types are redundant for this work, as the same currency is needed to perform analysis on homogeneous quantities, measured in USD in this case.

On top of deleting unimportant columns, some are hidden and new ones, with calculated fields, are created. More on the created columns in the next paragraph. A comprehensive list of all the columns in the new tables is provided in Appendix A.2.

2.3.2 Creation of the data model

The data model creation step is one of the most important ones. In fact, it enables later steps to be performed much more easily and with more precision. This activity can be subdivided into separate pieces of work, which are described below.

2.3.2.1 Primary key column definition and modification

The first step is to define adequate primary keys. It can easily be achieved for the three tables *Startups*, *Deal*, and *Investors* by looking at the columns *Organization Name URL*, *Transaction Name URL*, and *Organization/Person Name URL* respectively. They contain the Crunchbase URL corresponding to Crunchbase's web page related to that table entry. This seems to be a unique value. By performing a quick check on Excel it becomes evident that it is indeed a unique value for each row, and it can then be used as the primary key.

To reduce the amount of memory required, the URLs are shortened by removing the first part of the URL address that is the same for all entries.

For example, given a URL like "https://www.crunchbase.com/organization/absd-45", the part "https://www.crunchbase.com/organization" is removed, and the new primary key is "absd-45".

2.3.2.2 Import all tables into a single Excel workbook

All the tables are taken from the separated workbooks, which remain as backups and imported into a single workbook, which will be the central piece of the new data model.

It should be mentioned that the new tables have the following number of entries:

- *Startups:* 24470 (unique startups)
- *Deal:* 20367 (unique rounds)
- *Investors:* 9658 (unique investors)

2.3.2.3 New calculated field columns creation

Some new calculated fields to be used for different purposes can now be created.

Firstly, for each of the three tables *Startups*, *Deal*, and *Investors* the columns *Region, Country, State* and *City* are created. They are used to split the *Headquarter Location* field by the combined use of Excel's MATCH, FIND, LEFT, MID and RIGHT formulas to locate the comma characters that are used to separate the fields in the original column. The entries from the US are assigned "US" in the *Region* field, "USA" in the *Country* field, the reported State in the *State* field and the reported city in the *City* field. A similar procedure is followed for EU entries, with the difference that the *State* field is left empty. In the *Deals* table, the only reported fields are *Region* and *Country*.

The geographic data fields calculated from the original column *Headquarter Location* need additional cleaning steps as many cities are reported with more than one name. Given CB is a US-based company, its main language is English, which means that British, American, and Irish cities are reported correctly, but many cities from non-English-speaking countries are often reported with multiple names, generally in different languages. This type of data inconsistency can be problematic for several reasons, but the main one is related to the fact that, when performing analyses at a city level, the functions that need to group certain quantities (e.g. the number of startups, total invested capital, total exits, etc.) by city the results for the cities that have a multiple-naming problem are distorted. Given the large number of cities in the dataset (3700+), checking manually each city is not feasible and is highly prone to errors. Solving this problem fast and in an automated manner would be possible with some commercial services, which are however out of budget. Therefore, this thesis uses some Excel filters, based on the count of table entries (in the table containing all the cities in the sample) that are from the same country, start with the same three letters, and are of similar length. After the filters are applied, a manual inspection is performed on the remaining cities. The ones that have multiple entries with names from different languages are marked and corrected in the dataset by using Excel's *replace all* tool.

Finally, some logical check columns need to be created. The checks are mainly of logical consistency and are applied to the *Startups* and *Deals* tables.

New columns in the Startups table

The *Startups* table presents the following date-consistency fields: *Founded < Exit, Founded < Closed, Founded < Announced, Founded < IPO, Founded < Delist, IPO < Delist*. They check if the dates inserted are logically consistent. The main one is the

requirement *Founded* < *Exit* which requires that the founding date of that single startup is chronologically before its exit date.

The fields *Exit Type* and *Exit Value [USD]* are also created. They are used to determine the right exit vehicle used by the startup and the right exit value. Sometimes startups exit through a vehicle, but then undergo other financing rounds and the *Funding status* field may be misleading, as highlighted in the previous paragraphs. The new fields compare the *Exit Date* field with the *Announced Date* (date of M&A, which is empty if the startup did not have one) and with the *IPO Date* (empty if no IPO) fields, depending on which one is equal to the exit date, it is possible to determine which exit type the startup had. Similarly, for the *Exit Value [USD]* field, based on the determined exit date, the *Valuation at IPO [USD]* or the *Price [USD]* fields are used as exit value.

New columns in the Deals table

Similar fields to the ones above are created also for the *Deals* table. However, this table has more calculated fields needed for integrity and data completeness tests that are needed in later stages.

The *Tot Rounds [USD]* field contains a sum of the *Money Raised [USD]* fields. It groups rows (i.e. funding rounds) that have the same *Organization Name URL* field (i.e. the same startup) and calculates the total funding for each startup.

The *Max Same Date* and the *Num Same Date* columns are used together to determine a startup's number of rounds performed on the same day. The fact that two rounds are performed on the same day might be an overreporting or a data consistency problem and it is therefore highlighted.

The *Num of Rounds Missing Money Raised* field counts for each startup how many rounds have an empty cell in the *Money Raised [USD]* field.

The *Exit Amount [USD]* imports, using a VLOOKUP on the *Startups* table the exit value of the deal's startup.

The *Round Dopo Exit* field checks if the round has been performed after the startup's exit. The rounds performed after the exit, are excluded from the analysis.

The *Num of Missing Rounds (Round count-num fund rounds)* column counts how many rounds are missing by counting for each startup the number of rows (deals) in the *Deals* table and comparing it to the number of rounds indicated in the *Startups* table.

The *Num Of Rounds Missing Investor Names* field is used to count, for each startup, the number of rounds that miss completely information on the *Investor Names* field.

The *Startup Inv Geog Composition* and *Count comp rounds missing geo info* fields are used to monitor the composition of the startup's investors' geographic origins and to count how many of the startup's rounds are missing this information. These fields are described in more detail in the next paragraph, as determining the composition of the startup's investors' geographic origins requires some additional steps.

The fields *Use* and *Use Exit* are used to combine logical data integrity checks and are called into action when creating the samples.

The other calculated fields, reported in the appendix, are created to perform additional checks and to explore the dataset but are not used in the following steps.

2.3.2.4 Data model creation

Once all the calculated columns have been created, the three tables are uploaded to Excel's Power Pivot Data Model and relationships are created between primary keys. Additionally, some Pivot Tables, based on the new data model, are created to facilitate the extraction of the samples.

In particular, the *Organization Name URL* field in the *Startups* table is in a 1-to-many relationship with its homonym in the *Deals* table.

Unfortunately, it is not possible to create direct relationships between the *Deals* table and the *Investors* table, as the *Deals* table carries every round's investors' information in the *Investor Names* field as a list of comma-separated investor names, which, one by one, correspond to the *Organization/Person Name* in the *Investors* table.

To be able to create relationships between rounds in the *Deals* table and geographic origin information on the investors that took part in those financing rounds, it is necessary to split in different rows the information connecting every single round with each one of the investor names listed in its *Investor Names* field.

Creating the RoundInvestor table

A new table must be created, called *RoundInvestor*, that has in each row a primary key given by the combination of the *Transaction Name URL* and a single investor name.

This cannot be done simply by manually splitting columns for a table containing more than 20 thousand entries, each of which contains a variable number of investor names. Consequently, a custom VBA code is created to automatically create the new *RoundInvestor* table. An Excel VBA script is created for this purpose. It is reported in Appendix B.2.

Generates the *RoundInvestor* table (47600 entries), which has some additional calculated columns allowing it to draw information on the investor's headquarters location from the *Investors* table and the round from the *Deals* table and the *Startups* table.

Generating the field containing the composition of a startup's investors' geographic origins

The *RoundInvestor* table, pulling the headquarter location of the investor in each row, counts the number of investors coming from the *EU*, *US,* and *Other* regions. Combining these counts, the column *Startup Inv Geog Composition*, then determines the composition of the startup's investors' geographic origins assigning the following values:

- *EU Only*: the startup, across all its rounds, only has European investors
- *US Only*: the startup, across all its rounds, only has American investors
- *Other Only*: the startup, across all its rounds, only has investors headquartered in regions different than Europe or the US
- *EU and US*: the startup, across all its rounds, has at least one European and one American investor (no *Other* region investors were detected)

- *EU and Other*: the startup, across all its rounds, has at least one European and one *Other* region investor
- *US and Other*: the startup, across all its rounds, has at least one American and one *Other* region investor
- *All*: the startup, across all its rounds, has at least one investor from each of the three established regions.

It should be noticed that the existence of the *Other* region value is because the *Investors* table contains investors from all around the world that have invested in EU or US startups.

Finally, the *Deals* table can pull the *Startup Inv Geog Composition* field with a VLOOKUP query on the *RoundInvestor* table, matching the *Organization Name URL* fields of the two tables.

2.3.3 Creation of the samples

Once the data model has been created, it is possible to create the samples to be fed to the R script. Thanks to how the data model has been structured, this can be done with simple filters, depending on the combination of variables and sample sectioning dimensions required (v. Table 2.2).

As already mentioned, each sample is extracted from a different starting population to allow the maximum possible numerosity of each sample. Some filters on the new calculated columns are imposed simultaneously by using the *Use* (for samples to be used to fit the *investment amount* variable) and *Use Exit* (for samples to be used to fit the *exit values* variable) columns, depending on the sample.

The samples are then saved in different worksheets of three different workbooks as mentioned in the Analysis steps paragraph. An overview of the created samples is provided in Table Appendix.0.1 – Table showcasing the samples created.

An important aspect to keep in mind is that the columns exported in each worksheet of each workbook contain different information from the data model, depending on the variable it is prepared for. Namely, for the variables:

- *Exit values*: the data exported is the column *Exit Value [USD]* from either the *Startups table* or the *Deals* table, depending on the selected value on the investor geographic origin composition dimension
- *Invested amounts*: the data exported is the column *Total Equity Funding [USD] or the Tot Rounds [USD]* from either the *Startups table* or the *Deals* table, depending on the selected value on the investor geographic origin composition dimension
- *MOICs*: the data exported are either taken directly from the column *Equity Multiplier* (*Startups* table) *or* a combination of both the columns *Tot Rounds [USD]* and *Exit Value [USD]* from the *Deals* table.

For the variables *invested amounts* and *exit values,* the exported values are divided by 1 million to obtain their value in million dollars. This scaling operation is why the scale

invariance property of power-laws, mentioned in par. 1.3.1.2, is important. By dividing the variables of interest by 1 million, there is a scaling of the variables and therefore of the power law. However, the results provided for the scaled variables regarding the existence of a fit hold true also for the non-scaled variables. The *MOIC* variable is not scaled.

Chapter 3
Analysis of the results

This chapter will analyse the results of the analyses performed, by giving some comments on some general descriptive statistics on the data, but especially by commenting on the results of the power-law R script estimates. To do so, it will explore the results variable by variable.

It should be noted that not all the plots are provided in the Appendix. Some plots referring to estimates that run into bootstrapping errors are omitted. The attached zip folder of this thesis contains the complete results.

3.1 Descriptive statistics

This paragraph provides a general description of the populations, trying to avoid comparisons in absolute terms, as, for the reasons previously discussed, using the totals of certain quantities (e.g. exit values) to describe the populations in analysis makes little sense. Therefore, some quantitative data, including some totals, are here provided, but they should not be used to make a direct comparison between the American and European ecosystems.

3.1.1 Numbers at play in the dataset

By looking at the number of startups in the dataset, it is possible to better grasp what is present in it and the quantities at play.

3.1.1.1 Startups table

The new data model merges US and EU startups in a single table, *Startups*, which contains 15653 US startups and 8817 European ones. It should be remembered that the startups present in the dataset are those founded between 2005 and 2020, with HQ in either EU or the US, that, at the time of data extraction – v. (Scipione, 2020) –, had already had an exit.

Exploring their distribution by country (v. Figure 3.1), it is possible to notice that the number of startups by country seems to follow an extremely skewed distribution, which shows the first country, the United States, to have almost double the startups as all the other countries combined. As previously mentioned, this distribution needs to be taken with a grain of salt, as its precision and representativeness of the true proportions at play are probably not reliable. However, it might be still useful to understand what quantities are at play and how they might be distributed.

A country's population, its GDP and other macroeconomic indicators are also clearly for how the number of startups is distributed. Being the US a much larger economy than all the other ones, it is reasonable that it has a much larger number of startups. The same holds with regards to the population. However, an interesting fact to point out is that the sum of the GDPs of the countries included in the EU sample should be of a comparable dimension to that of the US. Therefore, it seems reasonable to highlight that, while having

a similar GDP size, the European sample still has almost half the number of startups, remaining at a large disadvantage. It should be noted part of this difference is probably accounted for by CB's *bias of overrepresentation of the US ecosystem*. However, it is unlikely that said bias can account for the whole difference.

It might be interesting to perform an analysis on the number of startups by country, correcting it by dividing it by the GDP per capita. Such an analysis could give some

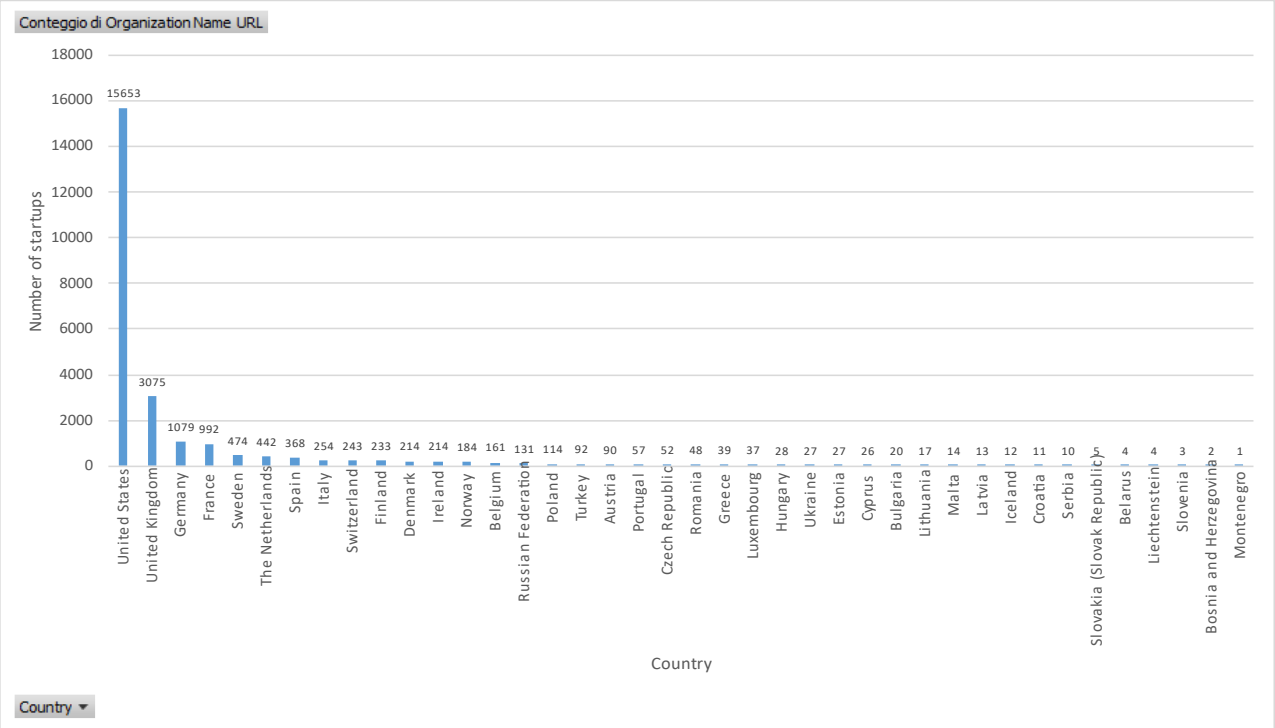

***Figure 3.1** - Number of startups by country*

indication of how the economic well-being of a nation influences the number of startups it produces. This type of analysis is however out of the scope of this work and could be explored by future research.

3.1.1.2 Deals table

The deals table, contains information on 20367 deals, referring to a total of 8668 startups, 25 of which are not present in the companies table, meaning there is no information available on their headquarters location. Of the remaining 8643 startups, 2657 are European and 5986 American.

The number of rounds per company represented in the *Deals* table is reported in Table 3.1. Most companies have 5 or fewer rounds, while the overall average number of rounds per company is 3.55. For European startups the average number of rounds is 3.05, while for American ones it is 3.74, indicating that American startups tend to receive more money per round. However, in the *Deals* table, 1704 companies are missing at least one round compared to the number of rounds declared on the *Number of rounds* column in the *Startups* table. These companies cannot be included in the samples.

***Table 3.1 – N**umber of startups by the number of rounds and by region in the Deals table.*

Number of Rounds	Region not known	EU	US	Grand Total
1	16	1385	2233	3634
2	1	595	1419	2015
3	5	327	993	1325
4	2	169	586	757
5		89	356	445
6	1	54	173	228
7		23	112	135
8		6	58	64
9		4	31	35
10			14	14
11		1	5	6
12		2	3	5
13		1		1
14		1	3	4
Grand Total	**25**	**2657**	**5986**	**8668**

3.1.1.3 Investors table

The I*nvestors* table contains 9658 records, of which 5915 are headquartered in Europe, 3041 in the US, and 702 in other regions. In the *Investor Type* column, the table also contains 260 investors which are labelled as *Corporate Venture Capital*, 138 of which are European, 104 American, and 18 from other regions. The column also contains 692 empty values.

It should be noted that the *Investor Type* column, does not contain single values, meaning that every investor can have more than one *investor type* value. The *Investor Type* column values, in case they are multiple, are indicated as a comma-separated list.

The *Investment Stage* column contains comma-separated lists of values indicating at which stages the investors in the *Investors* table generally invest. The column contains 879 empty values (i.e. for 879 investors there is no information regarding their typical investment stage).

To provide the reader with an idea of what the composition of the Investors sample by *Investor Type* and *Investment Stage* is, a quick analysis is performed for each of the two columns. The first step requires extracting the single values that these fields can take. Given that they are comma-separated lists, each value needs to be split using the combination of a comma and a space (“, “) as a marker. The operation described can be performed with a simple VBA macro, reported in Appendix B.3.

Once the list of single row-by-row values is obtained, using Excel *remove duplicates* functionality, the true list of possible values is obtained, and it can be used in combination with Excel’s COUNTIFS function to create a table with the count of the number of investors that have each value in their comma-separated lists.

Table 3.2 showcases the single investment stage values and the number of investors who report them in their *Investment Stage* field. It shows how the *seed* stage is the most common, however, if the *venture*, *early-stage venture* and *late-stage venture* values are combined, their number surpasses that of the seed stages. Moreover, it is interesting to notice that, even if the number of EU investors in the sample described in the *Investors* table is higher than that of American investors (EU 5000 circa vs US 3000 circa), the number of investors containing the *late-stage venture* field is higher than that of their European counterparts (814 US vs. 715 EU). This difference could be a clue pointing to the fact that US investors tend to have higher capital and to invest more in later stages. However, the reliability of this last observation must be contextualised, considering that the sample might not be representative of the original populations.

Table 3.2 *– Number of investors for each investment stage value in the Investors table.*

Investment Stage	US	EU	Other	Tot
Seed	1690	4339	296	6325
Venture	1630	2423	295	4348
Early Stage Venture	1496	2160	278	3934
Late Stage Venture	814	715	140	1669
Private Equity	429	512	58	999
Debt	157	31	24	212
Convertible Note	49	97	5	151
Secondary Market	34	38	7	79
Post-Ipo	37	14	16	67
Crowdfunding	11	49	6	66
Grant	26	9	8	43
Non Equity Assistance	15	8	6	29
Initial Coin Offering	15	11	1	27

Concerning the *Investor Type* column, the analysis indicates that the *Individual/Angel* investor type is the most common, followed by the V*enture Capital* type. As shown in Table 3.3, there is another interesting remark to make about the composition of the sample, when comparing the US and EU.

The European case has a much higher number of *Individual/Angel* type of investors, than the American one, both in absolute terms (2884 vs 744) and relative terms with respect to the percentage of investor types compared to the total (42% of EU investor types are *Individual/Angel*, while only 22% of the US ones are). This characteristic in turn translates to the US having a higher percentage of *Venture Capital* investor types compared to Europe (36% US vs 27% EU).

The last remark could be again taken as a clue pointing to the different approach of the US ecosystem, which probably tends to invest larger sums at later stages.

Table 3.3 – *Count of investor type values by regions in the Investors table.*

Investor Type	US	EU	Other	Tot
Individual/Angel	744	2884	105	3733
Venture Capital	1181	1859	248	3288
Private Equity Firm	302	512	47	861
Micro VC	302	370	45	717
Investment Partner	199	308	14	521
Accelerator	166	231	40	437
Corporate Venture Capital	104	138	18	260
Angel Group	60	162	17	239
Incubator	36	110	22	168
Investment Bank	46	56	20	122
Government Office	30	62	12	104
Family Investment Office	28	58	2	88
Entrepreneurship Program	16	20	6	42
Hedge Fund	34	7	0	41
University Program	16	7	6	29
Fund Of Funds	13	14	2	29
Venture Debt	11	10	8	29
Syndicate	6	13	0	19
Co-Working Space	7	7	2	16
Secondary Purchaser	4	6	1	11
Startup Competition	3	4	0	7
Pension Funds	0	1	0	1

It could also be interesting to analyse the distribution of the total funding amount of the investors to understand how much funds they have available to invest. However, a quick analysis of the number of investors for which the *Total Funding Amount [USD]* field is known (i.e. > 0), shows that the numerosity of the sample is too small (less than 400 values overall). The results of such an analysis would therefore be very unlikely to be of any significance.

The next analysis that can be performed relates to the number of portfolio organizations. This type of analysis could help to grasp an understanding of the investor size differences between the regions. Not all records report the number of portfolio organizations, however, most records seem to report this information, therefore, it seems reasonable to perform the comparison, at least for the most represented *investor type*s and *investment stages*. The easiest way to perform this comparison is by looking at the average number of portfolio organizations. Clearly, when performing this kind of comparison, it should be kept in mind that the distribution of this kind of quantity is very likely to be heavy-tailed. Therefore, using average values could be misleading. However, in this analysis, the average will only be used as a descriptive measure and no inferences are going to be made on the *populations of interest*.

Table 3.4 – *Average number of portfolio organizations by investor's Investment Stage. The table has been truncated to only show Investment Stage values for which there are at least 100 data points.*

Investment Stage	US	EU
Seed	45.87	10.13
Venture	60.24	18.71
Early Stage Venture	63.18	19.41
Late Stage Venture	67.71	28.61
Private Equity	53.1	21.32

As shown in Table 3.4, the average number of portfolio organizations for US investors is much higher than that of their European counterparts. The difference is the highest in the *venture* and *early-stage venture* cases and could hint at the fact that American investors are larger. However, it should be noted that, without knowing the total amount of capital invested by the investors, it is not possible to deduce any insight into the level concentration of capital. It is therefore not possible to assert that American investors tend to invest in a more concentrated manner in proportion to the available capital.

Table 3.5 confirms the trend discussed above: in this case, the differences for the *Individual/Angel* case are greatly increased.

Table 3.5 - *Average number of portfolio organizations by investor's Investor Type. The table has been truncated to only show values for which there are at least 100 data points.*

Investor Type	US	EU
Individual/Angel	11.54	2.503
Venture Capital	60.25	19.28
Private Equity Firm	35.28	16.84
Micro VC	59.53	26.11
Investment Partner	21.48	5.672
Accelerator	116.7	30.69
Corporate Venture Capital	60.41	21.58
Angel Group	51.12	15.88
Incubator	35.71	23.86

3.1.2 A focus on the *Deals* table

Given the *Deals* table is the table from which the samples accounting for startups' investors' geographic origin composition are retrieved, it seems useful to deepen its description in this separate paragraph.

Starting from its numerosity (v. par. 3.1.1.2), it seems interesting to describe how the sample numerosity varies when increasing the logical and informational rigour of the sample-selection criteria.

3.1.2.1 Number of missing rounds

One of the first steps when imposing logical requirements is to check that there are no missing rounds. To do this, an additional calculated column is created in the Excel table that allows comparing the counted number of records with the same *Organization Name URL* field in the *Deals* table to the reported number of rounds in the *Startups* table. The difference between these numbers represents the number of missing rounds.

Table 3.6 *– Number of startups in the Deals table by the number of missing rounds (rows), and headquarter regions (columns).*

Number of missing rounds	No Region Info	EU	US	Row Total
0	20	2310	4634	**6964**
1	4	260	871	**1135**
2		58	267	**325**
3	1	19	102	**122**
4		5	57	**62**
5		2	24	**26**
6		2	14	**16**
7			6	**6**
8			3	**3**
9		1	2	**3**
10			2	**2**
11			2	**2**
12			2	**2**
Column Total	**25**	**2657**	**5986**	**8668**

As Table 3.6 showcases, the number of startups not missing any rounds is not too low, remaining at 6964 in total, 2310 for European startups and 4634 for American ones. The first informational requirement, therefore, reduces the size of the population available for sampling by 1704 units, but still maintains a decently large starting sample dimension.

3.1.2.2 Temporal consistency requirements

The next requirements are of a logical consistency type and are based on the relationship between the startup's founding date, exit date and the round's date of the announcement. They are calculated using the columns *Ann* > *Founded* and *Founded* <= *Exit*. The column names are self-explanatory, and they indicate that the startups must have a founding date preceding the exit date and that the *Announced Date* field in a round is after the founding date of the startup. When imposing this couple of requirements, the sample numerosity goes from 6954 to 6864 (2276 EU and 4588 US).

The next constraint imposed involves the number of rounds performed by a startup on the same day. It seems quite unlikely that a startup can have multiple rounds performed on the same date, and, if this happens in the dataset, it is quite likely due to the poor information quality of the record. Therefore the *Max Same Date* column calculates, for every startup, the maximum number of rounds that have the same *Announced Date* field

and discards those startups that have a number higher than 1 (i.e. at least one of their rounds has more than one record in the *Deals* table). The number of startups available for sampling drops to 6791 (2257 EU and 4534 US).

3.1.2.3 Information completeness requirements

Depending on the different samples to be built, the structure of the information completeness requirements will change.

Investor's geographic origin composition

If the analysis is performed on samples that are not sectioned by the investor's geographic origin composition, no requirement on information completeness of investors' information is needed.

However, in case the samples need to include investors' geographic origin composition, then the column *Count comp rounds missing geo info* comes into play. Its value must be equal to 0 (i.e. the number of a startup's rounds missing investors' geographic information must be 0). When applying this requirement the number of startups available for sampling drops to 4705 (1731 EU and 2974 US).

It should be noted that, as previously explained (v. par. 2.3.2.4), the way that the values describing a startup's investors' geographic origin composition have been created, could not represent all the startup's investors. Many of the investor names in the *Deals* table are not present in the *Investors* table, therefore for many rounds, it was not possible to determine the geographic origin of all the investors. This was unavoidable, as requiring full information completeness for every single round would have given a total sampling pool of fewer than 100 startups, without introducing any other requirement. The clear problem that this incompleteness introduces is partially mitigated by the fact that the *Investors* table contains the most important investors, as stated by Scipione's work (v. Chapter 4 for more on this).

Information completeness accounting for the investigated variable

Depending on the variable under study (*Exit Value*, *Tot Equity Funding* or *Tot Invested Amount*, *MOIC*), the sample requires additional information completeness requirements.

If the object of analysis is the distribution of *exit values*, then the additional information completeness requirement involves the *Exit Amount [USD]* column and excludes all startups that have no Exit Amount information or have a value less than or equal to 1. Applying this filter, together with the one in the previous paragraph (v. Investor's geographic origin composition), dramatically reduces the sample size to 816 startups (262 EU and 554 US).

If the object of analysis is the *Tot Equity Funding*, then information completeness on the *exit values* is not required, and the *Num of Rounds Missing Money Raised* column comes into play. The *Exit Amount [USD]* filter is removed and the *Num of Rounds Missing Money Raised* field is required to be zero. The number of startups available in this case is 2583 (816 EU and 1767 US). The reduction compared to the 4705 available after the

Investor's geographic origin composition requirement is quite significant, but still less than that caused by the *exit value* information completeness constraint.

In case the variable under scrutiny is the *MOIC* since it is a calculated field that requires both the total invested amount and the exit value, the filters described above need to be applied simultaneously. The number of European startups available drops to 150, and that of American startups drops to 393 for a total of 543 startups.

The reduction is quite significant and, as already repeatedly mentioned, leads to extremely low sample numerosity, especially when the aim is to section the starting samples by startup investors' geographic origin composition. Such additional sectioning would lead to very low sample numerosity. In this work, precedence is given to the development of a structured methodology for the comparison between EU and US distributions, therefore the low sample numerosity is tolerated introducing a caveat on their statistical significance. A comparison is made among the values obtained in the results, but for the reasons repeatedly explained, it should be taken into consideration more on a methodological level than a quantitative one.

Finally, it should be noted that after the filters were applied there were no companies that had rounds in the samples that were after the startups' exit. This would have been an important requirement to satisfy otherwise, as their *Money Raised [USD]* field would have contributed to total funding, distorting possibly quite significantly the results for the *tot investment amount* and the *MOIC* variables.

3.1.3 An overview of city-level entrepreneurial ecosystems

Given the importance of geographic proximity in entrepreneurial phenomena and ecosystems, it seems useful to provide some descriptive statistics from a single city perspective.

The next paragraphs show the distributions and average values for different cities, starting with the number of startups, and continuing with the total invested amounts, the total exit amounts, and, finally, the MOICs.

To analyse the *Startups* table with a focus on single cities, attention must be paid to the *City* column when trying to use its field alone to aggregate the variables under study. Some cities belonging to different countries have the same name (e.g. Venice in the US vs Venice in Italy). When trying to aggregate using only the *City* field of the *Startups* table the values of both cities would erroneously be aggregated under a single city named *Venice*. To correct for this possible distortion, an additional column, *City_CountryISO2*, is added to the *Startups* table, joining the city name, the "_" character, and the 2-characters country ISO code, creating unique values for cities[9] (e.g. Venice_US).

[9] It is not necessary to perform a similar operation at a single country level as in the *Startups* table there are not different cities with the same name in the same country, even though in reality this is the case for some countries.

3.1.3.1 Number of startups

Created the appropriate field to use when aggregating values at a single city level, it is possible to use an Excel Data Model Power Pivot table to aggregate the quantities to be studied. This paragraph deals with the total number of startups by city. Given that to obtain the number of startups by city only a rather simple count of rows in the *Startups* table is needed, there is no need for particularly complicated filters on logical consistency or information completeness. Therefore, the analyses presented below include all rows in the *Startups* table.

Figure 3.2 shows the forty European cities with the highest number of startups. It appears quite evident that the distribution of the number of startups seems quite skewed towards the highest values, with significant differences between the values in the first positions. London, with 1384 startups, is the first European city by the number of startups and it makes up about 15% of the total number of European startups. To overtake the number of London-based startups, the next six top cities (Paris, Berlin, Stockholm, Madrid, Amsterdam, and Dublin) need to combine their number of startups. This behaviour is an additional clue suggesting that the number of startups might be power-law, or at least heavy-tailed, distributed.

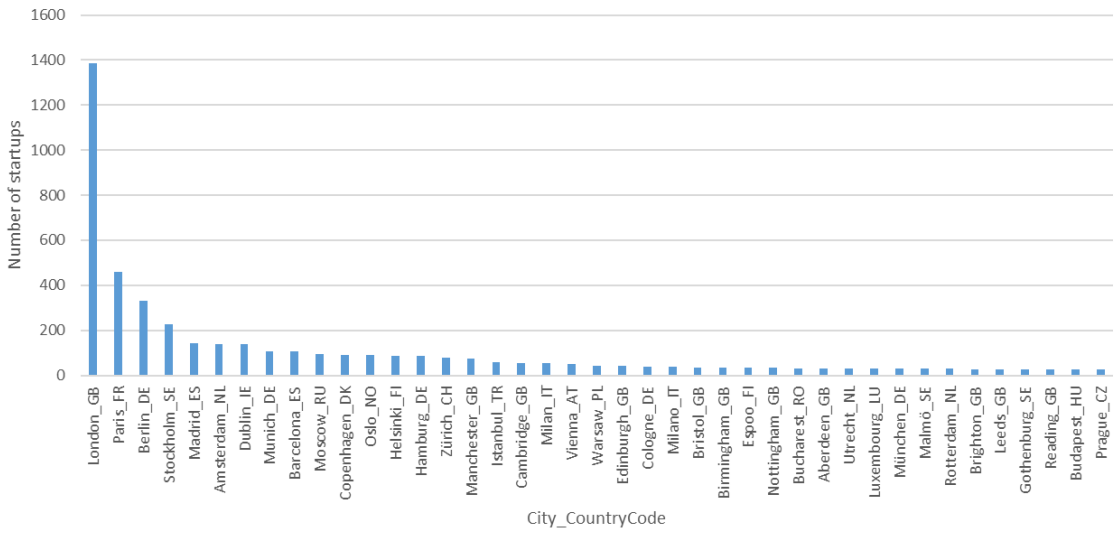

Figure 3.2 - *Top 40 European cities by number of startups.*

Figure 3.3 shows the distribution of the number of startups in the top 40 American cities. While still significantly skewed, the distribution appears to be much more concentrated around the two top cities, San Francisco (1526 startups) and New York (1464 startups). Visually, unlike the European case, which shows a somewhat progressive "curve" leading from the top city to the bottom one, the American case shows a net step from the two top cities to the rest.

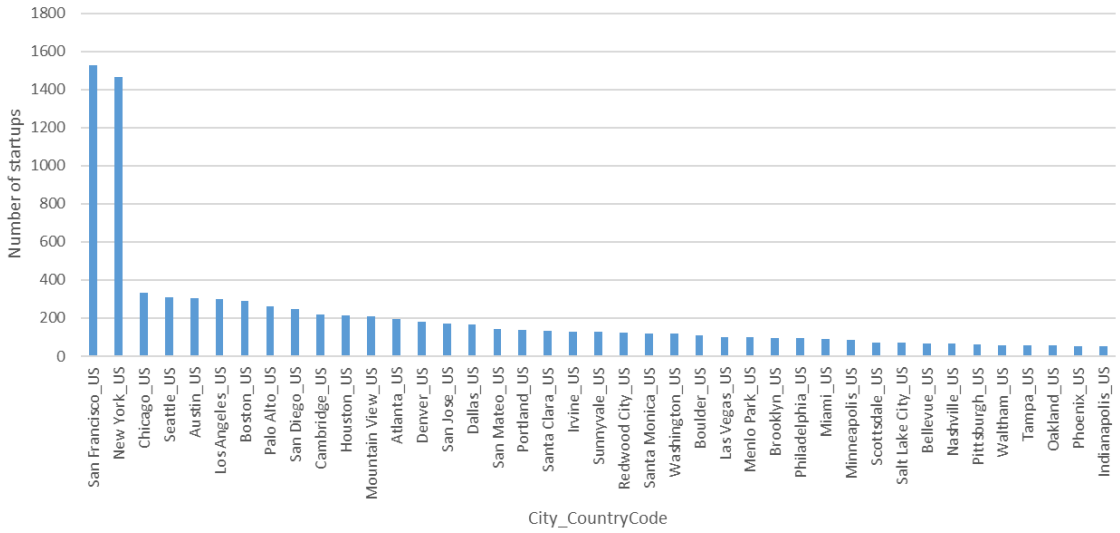

Figure 3.3 *- Top 40 American cities by number of startups.*

Figure 3.4 shows the top 42 cities by number of startups for both the US and EU. Visually, it is quite clear how the US dominates this category, as indicated by the prevalence of the orange columns in the histogram.

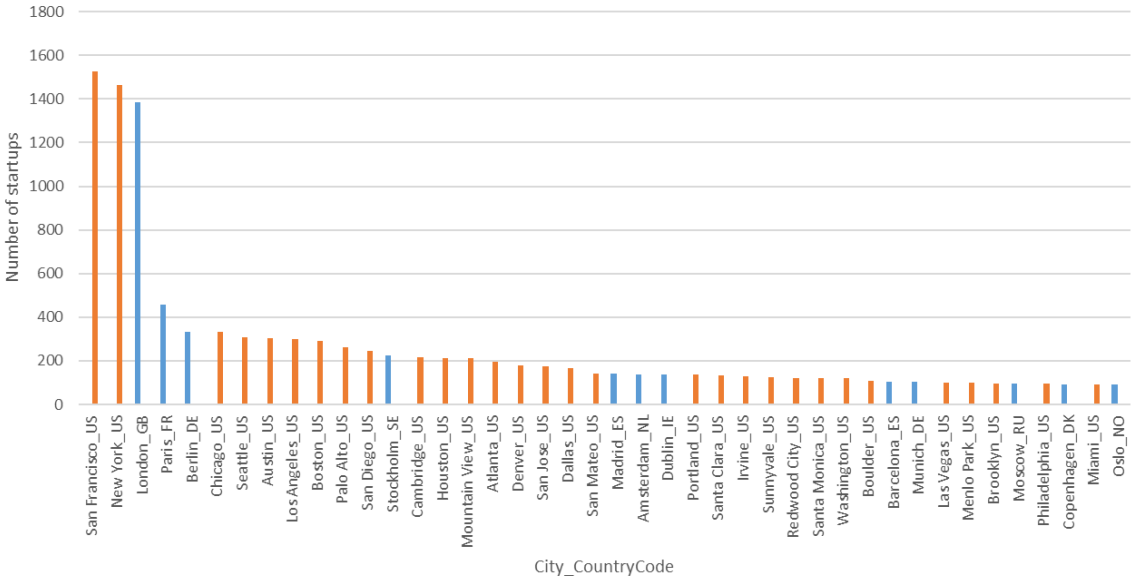

Figure 3.4 - *Top 42 cities in the Startups table by number of startups. US startups are in orange, while EU ones are in blue.*

Even though the evidence provided does not constitute statistically significant proof of the distribution of the number of startups in the US and the EU, or of the difference in dimensions between the two ecosystems, it still points out how concentrated entrepreneurial phenomena are: just a few cities are responsible for most of the effects.

More in detail, the most important cities worldwide by number of startups are San Francisco, New York, London, Paris, and Berlin. Overall, they account for 5164 startups in the table or 21.1% of the total amount of startups.

3.1.3.2 Investment Amounts

To find how the investment amounts are distributed by city in the *Startups* table, a new filter needs to be created. A new column called *Use_Inv*, controlling for several logical and informational completeness constraints on individual startups, is created. Its main requirements involve logical consistency of founding and exit dates and information completeness on the *Total Equity Founding Amount [USD]* column. When requiring the *Use_Inv* field to be *TRUE*, the remaining number of startups is 6997, 1573 European and 5424 American.

***Table 3.7** – Top 42 cities by Number of startups before and after the use of the Use_Inv filter.*

After *Use_Inv* filter			**Before *Use_Inv* filter**		
Top 42 cities	**N. of startups**	**% of startups**	**Top 42 cities**	**N. of startups**	**% of startups**
San Francisco_US	890	12.72%	San Francisco_US	1526	6.24%
New York_US	621	8.88%	New York_US	1464	5.98%
London_GB	287	4.10%	London_GB	1384	5.66%
Palo Alto_US	159	2.27%	Paris_FR	458	1.87%
Paris_FR	153	2.19%	Berlin_DE	332	1.36%
Boston_US	149	2.13%	Chicago_US	331	1.35%
Seattle_US	147	2.10%	Seattle_US	309	1.26%
Mountain View_US	140	2.00%	Austin_US	302	1.23%
Austin_US	140	2.00%	Los Angeles_US	301	1.23%
Cambridge_US	136	1.94%	Boston_US	291	1.19%
San Mateo_US	101	1.44%	Palo Alto_US	261	1.07%
Chicago_US	91	1.30%	San Diego_US	247	1.01%
Los Angeles_US	83	1.19%	Stockholm_SE	226	0.92%
Berlin_DE	82	1.17%	Cambridge_US	218	0.89%
San Diego_US	81	1.16%	Houston_US	213	0.87%
San Jose_US	75	1.07%	Mountain View_US	211	0.86%
Redwood City_US	74	1.06%	Atlanta_US	197	0.81%
Santa Clara_US	69	0.99%	Denver_US	179	0.73%
Sunnyvale_US	68	0.97%	San Jose_US	173	0.71%
Atlanta_US	61	0.87%	Dallas_US	166	0.68%
Santa Monica_US	59	0.84%	San Mateo_US	142	0.58%
Portland_US	57	0.81%	Madrid_ES	142	0.58%
Menlo Park_US	56	0.80%	Amsterdam_NL	139	0.57%
Boulder_US	46	0.66%	Dublin_IE	138	0.56%
Denver_US	41	0.59%	Portland_US	137	0.56%
Barcelona_ES	41	0.59%	Santa Clara_US	131	0.54%
Stockholm_SE	40	0.57%	Irvine_US	128	0.52%
Washington_US	39	0.56%	Sunnyvale_US	126	0.51%
Madrid_ES	38	0.54%	Redwood City_US	122	0.50%
Dublin_IE	37	0.53%	Santa Monica_US	121	0.49%
Brooklyn_US	35	0.50%	Washington_US	120	0.49%
Waltham_US	32	0.46%	Boulder_US	110	0.45%
Amsterdam_NL	32	0.46%	Munich_DE	105	0.43%
Philadelphia_US	30	0.43%	Barcelona_ES	105	0.43%
Helsinki_FI	30	0.43%	Menlo Park_US	99	0.40%
Copenhagen_DK	30	0.43%	Las Vegas_US	99	0.40%
South San Francisco_US	29	0.41%	Moscow_RU	95	0.39%
Moscow_RU	29	0.41%	Brooklyn_US	95	0.39%
Munich_DE	29	0.41%	Philadelphia_US	94	0.38%
Irvine_US	29	0.41%	Copenhagen_DK	92	0.38%
Tot. startups: 6997	**4366**	**62.39%**	**Tot. startups: 24470**	**11129**	**45.47%**

The variable of interest is the total equity funding amount in Million of US dollars, which is calculated city by city. Before analysing the distribution and sums of total equity funding amounts, a quick check on the change in numerosity of the startups available when applying the *Use_Inv* filter. Table 3.7 shows how the list of the top 40 cities by number of startups changes after introducing the *Use_Inv* filter.

The top 3 elements of the list do not change; however, it is possible to notice how the top 5 cities do change with Paris ending up 5th in the list and Berlin exiting the list, overtaken by Palo Alto which goes in 4th place. Moreover, the concentration of the list increases after the introduction of the filter: before the *Use_Inv* filter, the top 5 list contained about 21% of startups, while after the use of the filter it contained about 30% of them. The top 40 cities go from representing about 45% of all startups to about 62% after the introduction of the *Use_*Inv filter. The fact that the concentration increases is additional proof that CB's dataset is distorted, and it might point out that there is in fact a bias towards having more complete information for certain locations. In addition to this, the shift towards a higher US representation in the composition of the top 40 cities by number of startups, is yet another clue pointing towards the *US over (and better – from an informational completeness point of view) representation bias.*

However, even with the highlighted biases, a descriptive analysis can be performed to get a general grasp of the scale and dimensions of the ecosystems, even if its results are not completely statistically correct.

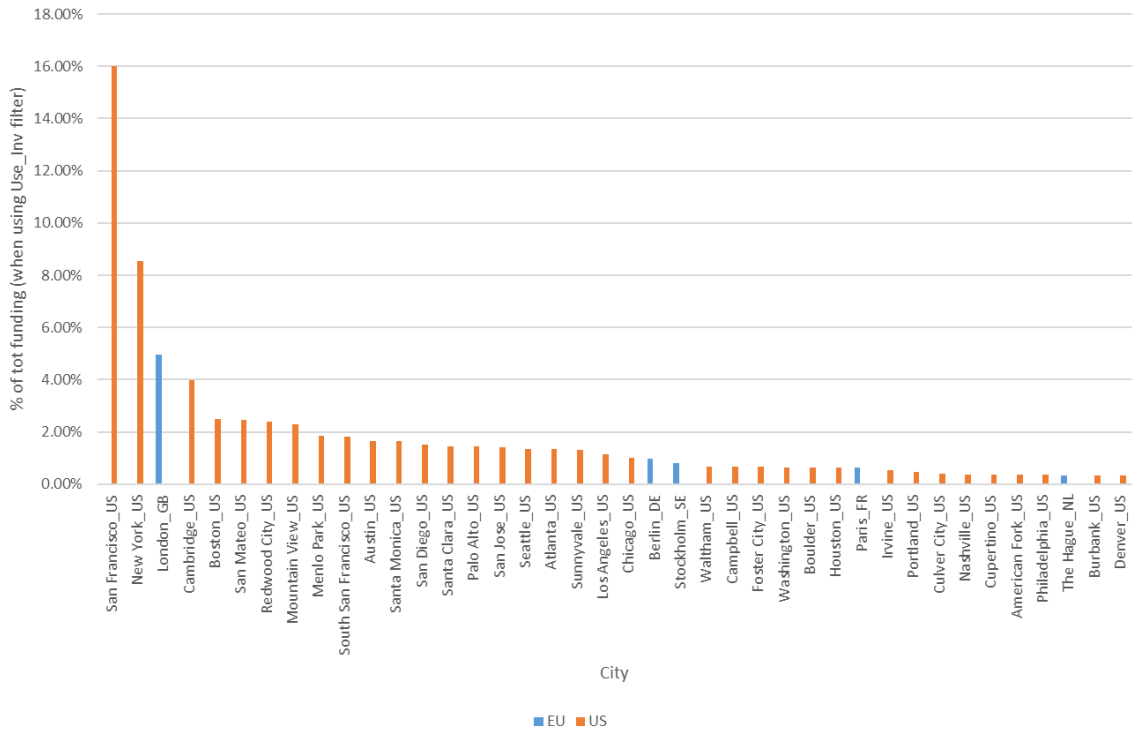

Figure 3.5 *- Distribution of the percentage of the total investment amounts generated by the top 40 cities.*

Considering the distribution of the total investment values by city in the *Use_Inv* filter case, Figure 3.5 shows how the percentage of the total invested amounts is distributed

by city. It clearly shows a net dominance of US cities, with San Francisco and New York dominating all other ecosystems. London is third, controlling about 5% of the investment values.

Table 3.8 illustrates how the list of the top 20 cities by tot funding amount is dominated by American cities, both when applying the *Use_Inv* filter and when not. In both cases, San Francisco and New York remain the leaders of the list, with a percentage of total funding that remains quite similar at 17% and 7% respectively in case the *Use_Inv* filter is not applied, and at 16% and 8.5% respectively otherwise. London (UK) and Cambridge (US) are very close at 4th and 3rd place, which they swap depending on the application of the filter, representing between 4% and 5% of the total funding depending on the case.

Table 3.8 *– Top 20 cities by tot funding when using the Use_Inv filter, and when not.*

Without *Use_Inv* filter			**With *Use_Inv* filter**		
City	**Tot funding [M$]**	**% of tot funding**	**City**	**Tot funding [M$]**	**% of tot funding**
San Francisco_US	54,729.15	17.08%	San Francisco_US	32,312.59	16.02%
New York_US	22,267.73	6.95%	New York_US	17,214.20	8.53%
Cambridge_US	14,327.22	4.47%	London_GB	10,010.82	4.96%
London_GB	13,707.50	4.28%	Cambridge_US	8,025.80	3.98%
Houston_US	9,599.47	2.99%	Boston_US	5,002.41	2.48%
Berlin_DE	6,975.78	2.18%	San Mateo_US	4,991.88	2.47%
San Mateo_US	6,585.73	2.05%	Redwood City_US	4,835.73	2.40%
San Diego_US	6,458.45	2.01%	Mountain View_US	4,609.23	2.28%
Boston_US	6,457.09	2.01%	Menlo Park_US	3,739.45	1.85%
Redwood City_US	5,667.38	1.77%	South San Francisco_US	3,672.90	1.82%
Venice_US	5,025.87	1.57%	Austin_US	3,321.71	1.65%
Mountain View_US	4,862.03	1.52%	Santa Monica_US	3,306.98	1.64%
Palo Alto_US	4,781.88	1.49%	San Diego_US	3,038.89	1.51%
South San Francisco_US	4,388.26	1.37%	Santa Clara_US	2,949.21	1.46%
Menlo Park_US	4,272.11	1.33%	Palo Alto_US	2,904.13	1.44%
Chicago_US	4,089.79	1.28%	San Jose_US	2,856.23	1.42%
Austin_US	3,987.80	1.24%	Seattle_US	2,694.66	1.34%
Seattle_US	3,713.68	1.16%	Atlanta_US	2,689.02	1.33%
San Jose_US	3,561.75	1.11%	Sunnyvale_US	2,629.63	1.30%
Santa Monica_US	3,536.28	1.10%	Los Angeles_US	2,328.66	1.15%

3.1.3.3 Exit Values

To find how the exit values are distributed by city in the *Startups* table, a new filter needs to be created. A new column called *Use_Exit*, controlling for several logical and informational completeness constraints on individual startups, is created. Its main requirements involve logical consistency of founding and exit dates and information completeness on the *Exit Value [USD]* column. When requiring the *Use_Exit* field to be

TRUE, the remaining number of startups is 3612 (down from the 24000+ startups in the *Startups* table), 1103 European and 2509 American.

Table 3.9 shows the top 20 cities by total exit values. In this case, San Francisco and London, 2nd place, lead the list. Comparing the lists before and after applying the filter, the distributions remain quite similar. Even in this case, the US dominates the list with only 2 European cities making the list.

Table 3.9 *– Top 20 cities by tot exit value when using the Use_Exit filter, and when not.*

Without *Use_Exit* filter			**With *Use_Exit* filter**		
City	**Tot Exit Values [M$]**	**% of tot Exits**	**City**	**Tot Exit Values [M$]**	**% of tot Exits**
San Francisco_US	90,481.91	8.70%	San Francisco_US	90,481.26	8.93%
London_GB	81,943.74	7.87%	London_GB	81,899.84	8.08%
New York_US	54,142.15	5.20%	New York_US	54,052.78	5.33%
Houston_US	35,445.97	3.41%	Houston_US	30,801.47	3.04%
Chicago_US	25,013.98	2.40%	Santa Clara_US	22,406.90	2.21%
Santa Clara_US	22,406.90	2.15%	Palo Alto_US	21,411.74	2.11%
Palo Alto_US	21,411.74	2.06%	Chicago_US	21,313.98	2.10%
Cambridge_US	18,812.09	1.81%	Cambridge_US	18,812.09	1.86%
Menlo Park_US	15,598.04	1.50%	Menlo Park_US	15,598.04	1.54%
Dallas_US	14,751.03	1.42%	Boston_US	14,237.28	1.40%
Boston_US	14,237.28	1.37%	Dallas_US	13,751.03	1.36%
Los Angeles_US	12,955.96	1.25%	Los Angeles_US	12,555.96	1.24%
Denver_US	12,244.92	1.18%	Denver_US	12,244.92	1.21%
Helsinki_FI	11,665.58	1.12%	Helsinki_FI	11,665.58	1.15%
South San Francisco_US	11,508.75	1.11%	South San Francisco_US	11,508.75	1.14%
San Diego_US	11,094.15	1.07%	San Diego_US	11,035.15	1.09%
Atlanta_US	9,444.35	0.91%	Atlanta_US	9,444.35	0.93%
Stratford_US	9,000.00	0.86%	Stratford_US	9,000.00	0.89%
San Jose_US	8,608.23	0.83%	San Jose_US	8,608.23	0.85%
Santa Monica_US	8,542.20	0.82%	Santa Monica_US	8,542.20	0.84%

As Figure 3.6 shows, the distribution of the percentage of the total exit values for each of the top 40 cities is quite skewed, showing very similar behaviour to that of the other observed variables.

Even in this case, the US ecosystem dominates the EU one with only 7 out of 40 cities making it into the top 40 list. However, London is quite close to San Francisco considering the Exit Values metric, which could indicate that UK-based startups have in general higher multiples: in fact, while only representing about 5% of total investments,

London represents about 8% of total exit values. On the contrary, San Francisco represents about 16% of the total investments, while only representing about 9% of total exit values.

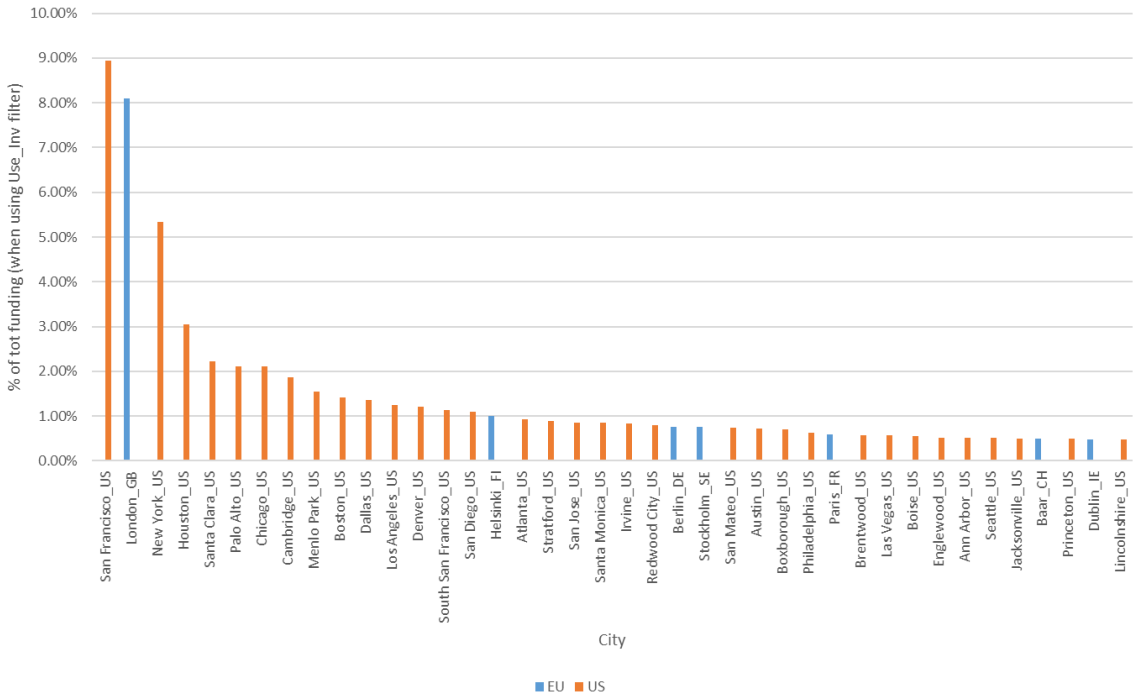

***Figure 3.6** – Distribution of the percentage of the exit values generated by the top 40 cities – Use_Exit filter applied.*

3.1.3.4 Multiples

To study how the MOICs are distributed among cities it is necessary to consider some aspects of the MOIC as a measurement. It is calculated on a single startup, dividing it's the cash received by the startup and its previous investors at the exit, by the cash invested in the startup throughout its lifetime (total equity invested). Therefore, it is not possible to simply sum the MOICs of all the startups in a city to compare different cities.

An approach could focus on comparing, for each city, the maximum, the minimum, and the average MOIC values calculated from the startups headquartered in the city. Another approach, more focused on the city's overall entrepreneurial ecosystem to generate value, could divide the sum of the total exit values of the startups in the city by the sum of the startup's total funding. This section tries to compare the results coming from the two approaches.

Similarly to the filters imposed on the previous measurements, the startups to be included in the analysis must satisfy the filter *Use_MOIC*, which demands startups in the *Startups* table to have both an Exit value and a total equity founding amount. Moreover, these startups must respect founding and exit date logical constraints and the requirement stating that the last funding date of the startups must be earlier than the exit date. The last constraint is needed to make sure that the total equity funding value provided is only made of investments performed before the exit so that the right MOIC can be calculated.

Given that there is a need to calculate aggregate values such as averages, medians, max and min, a numerosity constraint is needed. Therefore the cities to be analysed must

have a minimum number of startups that satisfy the required filters. To reach a compromise between having a small number of cities to compare with higher startup numerosity and many cities with low startup numerosity, which could lead to lower significance, the cities to be included in this analysis must have more than ten startups that satisfy the *Use_MOIC* filter.

Table 3.10 *– Descriptive statistics of the MOICS of the startups headquartered in the top 27 cities. The list is in descending order of the median MOIC of the startups in each city.*

City	Median	Max	Avg	Min	City MOIC	Num of Startups
Barcelona_ES	6.41	116.33	18.34	1.89	3.04	12
San Francisco_US	5.22	3273.91	37.67	0.06	3.30	177
Paris_FR	5.15	86.18	9.99	0.16	6.42	30
Palo Alto_US	5.14	669.18	30.17	0.39	6.92	37
London_GB	5.12	1274.67	36.48	0.05	1.71	49
Atlanta_US	4.82	14.40	5.39	0.14	2.74	12
Santa Monica_US	4.78	35.66	7.88	0.93	3.38	15
Chicago_US	4.73	2040.82	123.31	1.34	6.13	18
Irvine_US	4.71	50.61	9.27	0.77	4.46	13
Boston_US	4.55	431.31	18.34	0.14	3.97	41
Los Angeles_US	4.49	125.79	15.14	0.15	8.23	20
New York_US	4.39	500.00	11.99	0.02	2.78	118
Berlin_DE	4.31	63.49	9.82	1.02	3.00	14
San Diego_US	4.06	22.86	5.79	0.26	4.17	30
Portland_US	4.03	44.94	7.90	0.09	4.33	12
Santa Clara_US	4.01	315.35	29.40	0.40	18.69	13
Sunnyvale_US	3.96	17.78	5.93	0.10	2.46	18
Seattle_US	3.86	38.46	10.02	0.16	3.00	19
Mountain View_US	3.19	12.90	3.89	0.00	1.89	35
San Jose_US	3.02	26.67	4.69	0.07	3.54	22
Austin_US	2.98	27.50	6.15	0.25	1.80	29
Menlo Park_US	2.06	41.67	6.46	0.39	3.91	21
San Mateo_US	1.89	48.72	5.67	0.51	2.49	18
Redwood City_US	1.68	200.00	13.66	0.19	1.98	23
Waltham_US	1.32	13.21	3.84	0.34	2.27	15
Cambridge_US	1.17	1315.00	29.92	0.33	2.36	65
South San Francisco_US	0.96	186.00	11.56	0.20	3.06	23
Overall	**4.09**	**3273.91**	**21.59**	**0.00**	**3.32**	**899**

As Table 3.10 shows, the top city by median is Barcelona in Europe, however, the number of startups included is not very high, so the results must be taken with a grain of salt. However, it is interesting to notice that the second highest median value is San Francisco's, which once again reaffirms itself to be among the top ecosystems. Paris, Palo Alto and London complete the top 5. It is interesting to notice that South San Francisco is the only ecosystem for which the median MOIC is less than one, indicating that the lower two quartiles all have startups with unsuccessful exits (i.e. their exit amount is lower than their

total equity funding). However, the city's MOIC is 3.06, indicating that overall the startup has very successful startups in the top quantiles that make up for the losses of the other startups.

If one compares the median values to the average values of the MOICs, it is quite apparent that these present quite a significant difference, caused by the type of distribution typical of the MOICs: a distribution that is quite probably heavy-tailed. These kinds of distributions, as previously mentioned, have extreme events that significantly influence the average, in the case of the MOICs they increase it significantly. The median value, instead, is not as sensitive as the average to extreme values, which explains the significant differences observed between the values.

When ordering the list of the top 27 cities by the aggregated city MOIC, the top 5 cities appear to be Santa Clara (18.7), Los Angeles (8.2), Palo Alto (6.9), Paris (6.4), and Chicago (6.1). They appear to be the cities that have the best ability to generate value. However, it should be noted that these cities have quite a small number of startups included in the sample. Therefore, they might not be the most representative ones. If one considers the list of the top 5 cities with at least 40 startups that satisfy the required logical constraints by aggregated MOIC (MOIC = sum of exits in the city divided by the sum of tot equity funding in the city), the familiar list made of Boston, San Francisco, New York, Cambridge, and London comes out. However, it should be noted that these are the only five cities which have at least 40 startups. Therefore, the results which seem to indicate the supremacy of US cities should be taken with significant caution, as the results are heavily influenced by sample numerosity, and given the biases previously discussed, there is a certain level of uncertainty to consider.

Looking at the maximum values of the MOICs, there are several cities with max MOICs above 100. These values are incredibly high and indicate startups that have achieved incredible success and generated extraordinary economic value, for the ecosystems as a whole and their investors.

Table 3.11 presents an overview of the top startups by MOIC from the sample in Table 3.10. The multiples of these startups are incredibly high, but most of them still have quite large exits. For example, seven out of the sixteen in the table, have exits larger than 500 Million dollars.

Table 3.11 *– Startups, in the top 27 cities by MOIC, that have a MOIC greater than 100.*

Startup	City	MOIC	Exit [M$]	Funding [M$]
looksery	San Francisco_US	3273.91	150.60	0.05
citadel-plastics	Chicago_US	2040.82	800.00	0.39
txn	San Francisco_US	1835.54	91.78	0.05
boston-biomedical	Cambridge_US	1315.00	2630.00	2.00
hungryhouse-co-uk	London_GB	1274.67	248.48	0.19
blockseer	Palo Alto_US	669.18	12.71	0.02
hopstop-com	New York_US	500.00	1000.00	2.00
careport-health	Boston_US	431.31	1350.00	3.13
whatsapp	Santa Clara_US	315.35	19000.00	60.25
able-health	San Francisco_US	225.00	27.00	0.12
tilos-therapeutics	Cambridge_US	203.42	773.00	3.80
gliimpse	Redwood City_US	200.00	200.00	1.00
spitfire-pharma	South San Francisco_US	186.00	93.00	0.50
qlika	Palo Alto_US	150.00	3.00	0.02
honey-science	Los Angeles_US	125.79	4000.00	31.80
trovit	Barcelona_ES	116.33	101.32	0.87

3.2 Investment stages of trans-regional investors

To better understand the investment patterns, on a descriptive level it might be interesting to inquire about when cross-regional investors tend to invest in startups. To do this, this work uses the *RoundInvestor* table, selects a region (US or EU) and takes the first round (before exit) in which a cross-regional investor invested in every startup.

For example, considering startups headquartered in Europe, this work looks at how the number of US investors is distributed relative to the first entrance round represented as the percentage of completion between the startup's first round and the exit.

3.2.1 US investors' first entrance into European startups

It is necessary to remember that the results of this analysis are purely descriptive and risk not being representative of the original population. It should also be noted that the results of this analysis only involve the startups in which US investors invested before the exit, not the ones after that.

Table 3.12 shows that in about 40% of the cases US investors tend to invest before or at the 50%th round. In 13% of the cases, they tend to invest after the median round and right before the last round (percentage of completion = 100%) before the exit. In about 47% of cases, US investors invest in the company's last round before the exit.

Table 3.12 *– Entrance of US investors in European startups. The percentage value is reported in decimals.*

Percentage completion to exit	Num of Startups	Percentage of Startups
0.077	1	0.20%
0.111	1	0.20%
0.125	1	0.20%
0.143	2	0.41%
0.167	6	1.22%
0.182	1	0.20%
0.200	15	3.05%
0.250	25	5.08%
0.286	3	0.61%
0.333	41	8.33%
0.375	1	0.20%
0.400	11	2.24%
0.429	1	0.20%
0.500	86	17.48%
0.571	2	0.41%
0.600	5	1.02%
0.667	28	5.69%
0.714	3	0.61%
0.750	15	3.05%
0.800	7	1.42%
0.833	3	0.61%
0.857	1	0.20%
1.000	233	47.36%
Grand Total	**492**	**100.00%**

3.2.2 EU investors' first entrance into American startups

Table 3.13 shows when European investors tend to invest in US startups. In about 43% of the cases, EU investors tend to invest before or at the 50%th round. In 19% of the cases, they tend to invest after the median round and right before the last round (percentage of completion = 100%) before the exit. In about 38% of cases, EU investors invest in the company's last round before the exit.

By comparing the results with the previous case, it seems there are no significant differences between the investment patterns before the median round. The main difference is that American investors tend to invest in the last round before the exit compared to European investors (47% vs 38%).

This would support the idea that American investors tend to invest at later stages, particularly at the last stage, which very often is related to the scale-up phase and is likely to require more money. American investors, having usually larger amounts of money to invest, invest in the European startups at later stages.

Table 3.13 *- Entrance of EU investors in American startups. The percentage value is reported in decimals.*

Percentage completion to exit	Num of Startups	Percentage of Startups
0.091	1	0.13%
0.100	1	0.13%
0.125	2	0.26%
0.143	8	1.02%
0.167	15	1.92%
0.200	21	2.68%
0.250	45	5.75%
0.286	6	0.77%
0.333	71	9.07%
0.375	3	0.38%
0.400	25	3.19%
0.429	5	0.64%
0.444	1	0.13%
0.500	133	16.99%
0.545	1	0.13%
0.571	9	1.15%
0.600	16	2.04%
0.625	1	0.13%
0.667	56	7.15%
0.700	2	0.26%
0.714	6	0.77%
0.750	27	3.45%
0.800	15	1.92%
0.833	8	1.02%
0.857	4	0.51%
0.875	1	0.13%
0.900	1	0.13%
0.909	1	0.13%
1.000	298	38.06%
Grand Total	**783**	**100.00%**

3.3 Distribution estimates

This section discusses the results for the distribution estimates obtained in the R part of the analysis performed in this work. It is divided into three main subsections, one for each of the three variables.

3.3.1 Total investment values

The full table containing the results of the R estimates of the *total investment* variable across the different samples is available in Appendix D.1, together with the exported distribution plots.

As the table in Appendix D.1 reports, among the 36 tested cases (12 samples, 3 different distribution types), 26 cannot refuse the distribution estimates (p-value ≥ 0.1). Of these 26, none of them can be referred to as a single sample, meaning that all the samples have at least two compatible distributions. After a closer inspection, only two samples have two compatible distributions of the same type: *All_InvEUandUS* and *All_InvEU*. They are the samples composed of startups with HQ in all regions the investors' geographic origin compositions of which are EU and US investors and EU only investors respectively.

The interesting aspect about these two samples is that the distribution type that fits them is the log normal. Therefore, it seems that the hypothesis that these two samples fit a power-law distribution is rejected.

To verify that they are best fitted by a log-normal distribution, however, they should be tested against other distributions.

In the remaining 22 cases, the results are inconclusive and the hypothesis that the power law is a good fit for the samples involved cannot be rejected. However, the comparison tests performed against the log-normal distributions, that also fit the samples, are not able to significantly determine a winning distribution.

Therefore, what can be rejected is that the samples previously mentioned, together with *EU_InvAll*, are fitted by a power law.

The fact that in most samples the power-law distribution hypothesis cannot be rejected indicates that there is a possibility they might be distributed accordingly to a power law and more investigation is required.

For the cases in which the power law distribution hypothesis cannot be rejected, the estimated parameters can be used to compare the US and EU ecosystems. The results of this, given that the power law is not the winning distribution, might not have the desired statistical significance, but the comparison can be at least used as an example for the methodology developed in this thesis.

In the case of the *EU_InvEU* sample vs the *US_InvUS* one, the cut-off value for the EU sample is much lower than the American one (7.83 M$ vs 90.0 M$). However, more observation could be on the alpha values: 1.99 for the EU case vs 2.58 for the American case. The American parameter alpha is greater than that of the European case, indicating that the underlying distribution is more likely to have extreme outlier values. Therefore, it could be said that American investors tend to support higher rounds with a higher probability than EU ones.

3.3.2 Exit values

The full table containing the results of the R estimates of the *exit values* variable across the different samples is available in Appendix 0, together with the exported distribution plots.

The first elements to address are the samples for which the power-law distribution hypothesis can be rejected. These are: *All, US_InvEU, All_InvUS, All_InvEUandUS, US_InvEUandUS.*

The *EU_InvAll* sample cannot reject the power-law distribution, while it rejects all competing distributions. The remaining samples cannot reject the competing distributions.

For the *Exit Values* variables, there only is one sample which has US investors: *EU_InvUS*. However, this sample only has 49 data points, which makes its results very unlikely to be reliable.

Being there no sample with US investors, it is not possible to compare the power-law estimated parameters to get information on the differences between US and EU investors in this case.

3.3.3 Multiples

The full table containing the results of the R estimates of the *MOIC* variable across the different samples is available in Appendix D.2.1, together with the exported distribution plots.

In the case of the multiples, it seems that the power-law distribution has a better fit than the previous variables. The *US_InvUS*, *All_InvUS*, *EU_InvAll*, *All* samples are reasonably likely to follow a power-law distribution, as the competing distributions are rejected. The numerosity of the *EU_InvUS* and *US_InvEU* samples is too low for their results to be used.

Given the only two comparable samples with US and EU investors that cannot reject the power-law distribution hypothesis are *All_InvEU* and *All_InvUS*, these are the ones that are compared in this work.

The samples considered involve startups headquartered in all regions, but with European and American investors respectively. In this case, the European investors' MOICs power-law distribution has a 7.7 cut-off value vs the 5.85 cut-off of the Americans.

Regarding the alpha values, the European investors' alpha is 2.33 circa vs the Americans' 2.22. This example would seem to indicate that the Europeans tend to have a higher likelihood of having more extreme MOICs.

3.4 General comments on the fit quality

From the analysis of the results, it appears that a quite recurring problem is that, as mentioned in par. 1.3.1.2, the tail numerosity is rather scarce. This could be a problem for the accuracy and the credibility of the results, but, unfortunately, with the available data, it is not possible to reach more statistically significant results.

Overall, there is a rather small number of estimates that pass the *p*-value test, and many of the sample distribution estimates that do pass it cannot establish a winner among the two competing distributions (i.e. the power law and the log-normal).

It should be noted that Appendix D does not contain the plots for the estimated log-normal distribution that are labelled as *Log Norm Tot* in the results tables in Appendix sections D.1, D.2, and D.3. This is because the cases worth mentioning are reported and shown in the previous paragraphs and the *Log Norm Tot* estimates are not the main objective of this thesis. However, it seems interesting to remark that quite a few of the

samples, for each of the variables under scrutiny, are compatible with a log-normal distribution with a cut-off value, x_{min}, much lower than that allowed by the power law. Therefore, a much larger portion of the data can be described by a log-normal distribution.

Finally, it should probably be stated clearly that the estimated distribution parameters, and the resulting distributions they generate, do not describe the behaviour of all the sample points. Instead, they only describe the distribution of the data points belonging to the tail of the sample. As previously mentioned, the number of datapoints described by the estimated distributions is indicated as *ntail* in R, or as n_{tail} in the theoretical section of this thesis.

Chapter 4
Conclusions

Overall, this work showcases the application of a strong methodology to inquire about the distribution of variables deriving from entrepreneurial phenomena and to use the estimated distribution parameters to compare the behaviour of investors in different ecosystems. This method can easily be replicated with more complete datasets to estimate the distribution parameters of more entrepreneurship variables across several sampling criteria and sampling sectioning dimensions.

Even if, due to the lacking data quality previously discussed, the results are not of the best quality on a statistical significance level, they still enable some comparisons to be made about the distribution of the variables of interest and the differences in investment practices among EU and US investors.

Overall, the hypothesis that much of the data is power-law distributed cannot be rejected, especially for the MOICs and investment values. However, it was not possible to decisively prove the superiority of the power-law compared to the log normal.

In addition to this, in paragraph 3.1 some level of evidence has been collected suggesting that American investors, compared to European ones, tend to invest at later stages when investing cross-border.

4.1 Criticalities

Regrettably, the major issue faced by this work regards the data. Undoubtedly, low data availability, numerosity and quality are the main detractors of this work, as they generate significant flaws and bias in the data and expose its results to lower statistical significance.

Another element that could potentially distort the results is that many investing organizations are international branches of a single organization. Therefore, the way they invest could not be dictated by their investment strategies, but by the decision of the parent organisation. An example could be the existence of an EU-based Black Rock fund. Such a fund would be controlled by the parent organization, Black Rock, which is American. Therefore, considering the EU branch as a European investor might be misleading.

4.2 Advantages and strengths

The advantages of this methodology are mainly twofold. Firstly, it allows comparing the two ecosystems while avoiding a large part of the distortions in CB. This is discussed in the introduction part of this work.

In addition to this, comparing the ecosystems based on the distributions of some of their main descriptive and performance variables allows testing at the same time the type of distributions that describe the observed quantities, while using their parameters to compare the characteristics of the original populations.

As already mentioned, the method allows to bypass some of the biases in CB, however, it needs higher sample numerosity and information completeness to express its full potential.

4.3 Future research potential

Further research, applying the structured methodology developed in this thesis can obtain results from more numerous and more complete data, obtaining a higher reliability and inference potential from the results.

Another interesting aspect that can be inquired upon by future research is the addition of a sample-sectioning dimension controlling for the industries. Such an approach could give more sector-specific insights and reduce the risks associated with *industry over (or under) representation biases* in Crunchbase or other commercial datasets.

As mentioned in paragraph 3.4, it seems interesting that quite a few samples are compatible with a log-normal distribution with much lower cut-off values than those allowed for the power-law distributions estimated on the same samples. Future research might try to investigate the reason behind this to test whether it is due to the relatively low numerosity and quality of the data available for this thesis, or due to some other reason related to the behaviour of entrepreneurial phenomena (i.e. why does the log-normal fit well the data and describes a higher portion of it than the power law?). Might it be that different parts of the data are better described by different distributions? The author of this work hypothesises that the log-normal could be used to describe the lower values of the data and then a power-law may better describe the final part.

Another interesting research topic might be related to the distribution of the number of startups by country. In paragraph 3.1.1.1, the number of startups by country seems to follow a very skewed distribution. It might be interesting to perform an analysis on the number of startups by country, correcting it by dividing it by the GDP per capita. Such an analysis could give some indication of how the economic well-being of a nation influences the number of startups it produces.

Bibliography

Clauset, A., Shaliza, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *Physics*. doi:https://doi.org/10.48550/arXiv.0706.1062

Crawford, G., Aguinis, H., Lichtenstein, B., Davidsson, P., & McKelvey, B. (2015). Power law distributions in entrepreneurship: Implications for theory and research. *Journal of Business Venturing, 30*(5), 696-713. doi:https://doi.org/10.1016/j.jbusvent.2015.01.001

Gillespie, C. S. (2015, February). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software, LXIV*(2), 1-16. doi:http://dx.doi.org/10.18637/jss.v000.i00

Hege, U., Palomino, F., & Schwienbacher, A. (2009). Venture Capital Performance: The Disparity Between Europe and the United States. *Finance, 30*(1), 7-50. Retrieved June 20, 2022, from https://www.cairn-int.info/journal--2009-1-page-7.htm

Lin, X. (2000). Stochastic Processes for Insurance and Finance. By T. Rolski, H. Schmidli, V. Schmidt and J. Teugels (John Wiley, Chichester, 1999). *British Actuarial Journal, VI*(4), 876-877. doi:10.1017/S135732170000204X

Neumann, J. (2015, June 25). *Power Laws in Venture*. Retrieved June 20, 2022, from Reaction Wheel: Jerry Neumann's Blog: http://reactionwheel.net/2015/06/power-laws-in-venture.html

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, XLVI*(5), 323-351. doi:10.1080/00107510500052444

Prencipe, D. (2017, April 10). Liquidity events and returns of EIF-backed VC investments. (H. Kraemer-Eis, Ed.) *The European venture capital landscape: an EIF perspective, III*. Retrieved June 20, 2022, from https://www.eif.org/news_centre/publications/eif_wp_41.pdf

Retterath, A., & Braun, R. (2020, September). Benchmarking Venture Capital Databases. *SSRN*. doi:http://dx.doi.org/10.2139/ssrn.3706108

Rolski, T., Schmidli, H., Schmidt, V., & Teugels, J. L. (1999). *Stochastic Processes for Insurance and Finance.* Chichester: Wiley.

Scipione, A. (2020). *Differenze fra investitori europei e statunitensi: Analisi del mercato europeo delle start-up.* [Master's Degree Thesis, Politecnico di Torino]. Turin: Politecnico di Torino. Retrieved June 20, 2022, from https://webthesis.biblio.polito.it/16455/1/tesi.pdf

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable* (2nd ed.). Random House.

Virkar, Y., & Clauset, A. (2014). Power-law distributions in binned empirical data. *Power-law distributions in binned empirical data, VIII*(1), 89-119.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses. *Econometrica, LVII*(2), 307-333. doi:https://doi.org/10.2307/1912557

Statement of Independent Work

I hereby confirm that this thesis was written independently by myself without the use of any sources beyond those cited, and all passages and ideas taken from other sources are cited accordingly.

Appendix

This appendix contains the complete details, lists and figures of the work performed in this thesis. Appendix A contains the lists of the table columns in the original dataset and the modified data model prepared for this work, Appendix B contains the full detail of the scripts used for this thesis, and Appendix C showcases the characteristics of the samples created for the analysis, and Appendix D contains all the estimates for the distribution parameters and the plots of the calculated curves.

Appendix A
Details on the used datasets

The column names present in this appendix section are self-explanatory. Comments will be provided in case there are specific topics to be discussed.

A.1 Columns of the original dataset tables

The column names for 4 of the tables in the original dataset could be grouped in pairs. The two tables containing startups headquartered in the EU and US had the same columns, similarly, the two tables, *Deals EU* and *Deals US*, containing the information on financing rounds have the same columns. Hence, only one list of columns is provided here for each table type (*Deals, Startups, and Investors*).

A.1.1 Startup tables column names

Below is a list of the column names (each one separated by a comma) in Scipione's original *Startup* tables (the EU and US ones) his work, excluding his calculated columns:

Organization Name, Organization Name URL, Headquarters Location, Headquarters Regions, Diversity Spotlight (US Only), Estimated Revenue Range, Description, Operating Status, Founded Date, Founded Date Precision, Exit Date, Exit Date Precision, Closed Date, Closed Date Precision, Company Type, Org Name URL, Investor Type, Investment Stage, Number of Portfolio Organizations, Number of Investments, Number of Lead Investments, Number of Exits, Number of Exits (IPO), Industry Groups, Industries, Number of Founders, Founders, Number of Employees, Number of Funding Rounds, Funding Status, Last Funding Date, Last Funding Amount, Last Funding Amount Currency, Last Funding Amount Currency (in USD), Last Funding Type, Last Equity Funding Amount, Last Equity Funding Amount Currency, Last Equity Funding Amount Currency (in USD), Last Equity Funding Type, Total Equity Funding Amount, Total Equity Funding Amount Currency, Total Equity Funding Amount Currency (in USD), Total Funding Amount, Total Funding Amount Currency, Total Funding Amount Currency (in USD), Top 5 Investors, Number of Lead Investors, Number of Investors, Number of Acquisitions, Acquisition Status, Transaction Name, Transaction Name URL, Acquired by, Acquired by URL, Announced Date, Announced Date Precision, Price, Price Currency, Price Currency (in USD), Acquisition Type, Acquisition Terms, IPO Status, IPO Date, Delisted Date, Delisted Date Precision, Money Raised at IPO, Money Raised at IPO Currency, Money Raised at IPO Currency (in USD), Valuation at IPO, Valuation at IPO Currency, Valuation at IPO Currency (in USD), Stock Symbol, Stock Symbol URL, CB Rank (Organization), CB Rank (Company), CB Rank (School).

A.1.2 Deal tables column names

Below is a list of the column names (each one separated by a comma) in Scipione's original *Deals* tables (the EU and US ones) from his work, excluding his calculated columns:

> *Transaction Name, Transaction Name URL, Funding Type, Funding Stage, Money Raised, Money Raised Currency, Money Raised Currency (in USD), Announced Date, Pre-Money Valuation, Pre-Money Valuation Currency, Pre-Money Valuation Currency (in USD), Equity Only Funding, Organization Name URL, Organization Description, Organization Industries, Diversity Spotlight (US Only), Organization Location, Organization Website, Organization Revenue Range, Total Funding Amount, Total Funding Amount Currency, Total Funding Amount Currency (in USD), Funding Status, Number of Funding Rounds, Lead Investors, Investor Names, Number of Investors, Number of Partner Investors, CB Rank (Funding Round).*

A.1.3 Investor table column names

Below is a list of the column names (each one separated by a comma) in Scipione's original *Investors* table from his work, excluding his calculated columns:

> *Organization/Person Name, Organization/Person Name URL, Location, Regions, Description, Founded Date, Founded Date Precision, Exit Date, Exit Date Precision, Company Type, Estimated Revenue Range, Industry Groups, Industries, Number of Funding Rounds, Funding Status, Last Funding Date, Last Funding Amount, Last Funding Amount Currency, Last Funding Amount Currency (in USD), Last Funding Type, Last Equity Funding Amount, Last Equity Funding Amount Currency, Last Equity Funding Amount Currency (in USD), Last Equity Funding Type, Total Equity Funding Amount, Total Equity Funding Amount Currency, Total Equity Funding Amount Currency (in USD), Total Funding Amount, Total Funding Amount Currency, Total Funding Amount Currency (in USD), Investor Type, Investment Stage, Number of Portfolio Organizations, Number of Investments, Number of Partner Investments, Number of Lead Investments, Number of Diversity Investments, Number of Exits, Number of Exits (IPO).*

A.2 Columns of the new data model

The columns of the new data model created for this thesis are mostly similar to the original dataset columns. However, some columns, contain redundant information (e.g. the original currency amount columns and the original currency columns used for the total funding values of the startups). Moreover, as the reader will notice, there are new columns. Some of the new columns are calculated fields used to explore the dataset in the preliminary steps and have been fundamental in spotting important incongruences in the dataset, some other columns are calculated fields used as data consistency filters, logical filters, sample sectioning dimension filters and as evaluation variable filters (e.g. the field *Exit Value [USD]* is a calculated field combining logical criteria to determine the exit value of the startup, by looking at the exit date, the acquisition date, the IPO date, the acquisition price and the money raised with the IPO).

A.2.1 Startup tables column names

Below is a list of the column names (each one separated by a comma) in the new data model's *Startups* tables, including calculated columns:

Organization Name URL, Organization Name, Region, Country, State, Province, City, Headquarters Location, Description, Operating Status, Founded < Exit, Founded < Closed, Founded < Announced, Founded < IPO, Founded < Delist, IPO < Delist, Data Filter, Has Funding and Exit values2, Equity Multiplier, Founded Date, Founded Date Precision, Exit Type, Exit Value [USD], TTE [days/365], Exit Date, Exit Date Precision, Closed Date, Closed Date Precision, Company Type, Investor Type, Investment Stage, Number of Portfolio Organizations, Number of Investments, Number of Lead Investments, Number of Exits, Number of Exits (IPO), Industry Groups, Industries, Number of Funding Rounds, Funding Status, Last Funding Date, Last Funding Amount [USD], Last Funding Type, Last Equity Funding Amount [USD], Last Equity Funding Type, Total Equity Funding Amount [USD], Total Funding Amount [USD], Top 5 Investors, Number of Lead Investors, Number of Investors, Number of Acquisitions, Acquisition Status, Transaction Name, Transaction Name URL, Acquired by, Acquired by URL, Announced Date, Announced Date Precision, Price [USD], Acquisition Type, Acquisition Terms, IPO Status, IPO Date, Delisted Date, Delisted Date Precision, Money Raised at IPO [USD], Valuation at IPO [USD], USE?.

A.2.2 Deal tables column names

Below is a list of the column names (each one separated by a comma) in the new data model's *Deals* tables, including calculated columns:

Organization Name URL, Region, Country, Tot Rounds [USD], Transaction Name URL, Organization Name ID, Max Same Date, Num Same Date, Round Index, Funding Type, Funding Stage, Num of Rounds Missing Money Raised, Money Raised [USD], Exit Amount [USD], Num Exit-Ann < 30, Exit-Ann < 30, Num of Exit = Ann Date, Exit = Announced Date, |Exit - Announced Date|, Raised Money = Exit, Round dopo Exit, AnnDate_Year, AnnDate_Month, MonthsFromFounded, Announced Date, Pre-Money Valuation [USD], Num of Missing Rounds (Round count-num fund rounds), Declared Num Rounds Check (Num Rounds in Comp DB - Num Fund Rounds), Num Round Count, Num Rounds in Companies DB, Number of Funding Rounds, Lead Investors, Inv Name Count Check, Investor Count Check, Num Of Rounds Missing Investor

Names, Inv Name Len, Investor Names, Number of Investors, Number of Partner Investors, Founded Date, Exit Date, Founded <= Exit, Ann > Founded, Num Rounds with Ann date < Founded, Announced € [Founded; Exit], All Ann in Date Range, Is unique, Is in Companies, Startup Inv Geog Composition, Count comp rounds missing geo info, Startup Lead Inv Geog Composition.

A.2.3 Investor table column names

Below is a list of the column names (each one separated by a comma) in the new data model's *Investors* tables, including calculated columns:

Organization/Person Name, Organization/Person Name URL, Region, Country, City, Description, Industry Groups, Industries, Total Equity Funding Amount [USD], Total Funding Amount [USD], Investor Type, Investment Stage, Number of Portfolio Organizations, Number of Investments, Number of Partner Investments, Number of Lead Investments, Number of Exits, Number of Exits (IPO), Is Corporate, Equal Name Count, Present in Deals.

A.2.4 RoundInvestor table column names

Below is a list of the column names (each one separated by a comma) in the new data model's *Investors* tables, including calculated columns:

Transaction Name URL, Organization Name URL, Round Index, Investor Index, Investor Name, Is in Investors Table, Is in Companies Table, Investor Region, Num Inv in Round, Num Inv in DB in Round, Comp Round with Min inv in DB num, Num of Geo info of Round, Comp Round with min num of geo info, Round Num EU Inv, Round Num US Inv, Round Num Other Inv, Startup Inv Geog Composition.

Appendix B
Scripts

B.1 R script sample to generate the estimated distributions

A template of the R script used to generate distribution estimates for each sample is provided below. The script was the same for all samples, it only needed to change two fields, which are highlighted in bold, to change the analyzed sample. The origin Excel workbook (one for each variable) controls the chosen variable, and the Excel worksheet controls the sample to be used for the said variable.

```
R> library(poweRlaw)
R> library(readxl)
R> db <- read_excel("C:/Users/96fra/OneDrive - Politecnico di R>
Torino/Magistrale/Tesi/Databases/Giuliani/selected_variable_workbook.xlsx", sheet =
"selected_sample_worksheet")

# Power-law
R> pl<- conpl$new(db$MultipleVal)
R> pl$setPars(estimate_pars(pl))
R> est_pl <- estimate_xmin(pl)
R> est_pl
R> pl$xmin <- est_pl
R> pl
R> bs_p_pl <- bootstrap_p(pl, no_of_sims = 500, threads = 6)
R> bs_p_pl$p
R> bs_p_pl$gof
R> plot(pl, main="All Region - All Investors", sub="Plotting a power-law
probability distribution to startup multiples", xlab="Multiple", ylab="P(X>=x)")
R> lines(pl, col = 2, lwd =2)
R> plot(bs_p_pl)

# Log Norm
R> ln <- conlnorm$new(db$MultipleVal)
R> ln$setXmin(pl$xmin)
R> ln$setPars(estimate_pars(ln))
R> ln
R> bs_p_ln <- bootstrap_p(ln, no_of_sims = 500, threads = 6)
R> bs_p_ln$p
R> bs_p_ln$gof
R> plot(ln, main="All Region - All Investors", sub="Plotting a log normal
probability distribution to startup multiples - xmin same as power-law's xmin",
xlab="Multiple", ylab="P(X>=x)")
lines(ln, col = 2, lwd =2)
R> plot(bs_p_ln)

#compare distributions
R> comp <- compare_distributions(pl_m, ln_m)
R> comp

# Log Norm Tot
R> lntot <- conlnorm$new(db$MultipleVal)
R> lntot$setPars(estimate_pars(lntot))
R> est_lntot <- estimate_xmin(lntot)
R> est_lntot
R> lntot$xmin <- est_lntot
R> lntot
R> bs_p_lntot <- bootstrap_p(lntot, no_of_sims = 500, threads = 6)
```

```
R> bs_p_lntot$p
R> bs_p_lntot$gof
R> plot(lntot, main="All Region - All Investors", sub="Plotting a log normal
probability distribution to startup multiples", xlab="Multiple", ylab="P(X>=x)")
R> lines(lntot, col = 2, lwd =2)
R> plot(bs_p_lntot)
```

B.2 VBA script to split the investor columns

To obtain the *RoundInvestor* table the code must loop through the entries *Deals* table, split the *Investor Names* field, separating it into a variable-length array containing each name in a separate position. Then the VBA code needs to create a new table and paste all the newly created *Transaction Name URL – Investor Name* couples in it.

```
Sub SplitTableDimensionColumnToRows()
Application.ScreenUpdating = False

'1 - SETTING ORIGIN TABLE
Dim objWorkbook As Workbook
Dim objSheet As Worksheet
Dim orgTable As ListObject

Set objWorkbook = ActiveWorkbook
Set objSheet = objWorkbook.Sheets(InputBox("Insert Origin Sheet Name", "Insert Text
_ Input", "Sheet1"))
Set orgTable = objSheet.Range(InputBox("Insert Origin Table Name", "Insert Text _
Input", "Table1")).ListObject

'2 - SETTING DEST TABLE
Dim destSheet, destTblName As String
Dim destTable As ListObject

destSheet = InputBox("Insert New Table Destination Sheet Name", "New Table Sheet",
_ "TestSheet")
Sheets.Add.Name = destSheet

destTblName = InputBox("Insert New Table Name", "New Table Name", "TestTable")
Sheets(destSheet).ListObjects.Add().Name = destTblName

Set destTable = Sheets(destSheet).Range(destTblName).ListObject

destTable.ListColumns(1).Name = "Transaction Name URL"
destTable.ListColumns.Add(2).Name = "Organization Name URL"
destTable.ListColumns.Add(3).Name = "Investor Index"
destTable.ListColumns.Add(4).Name = "Investor Name"
destTable.ListColumns.Add(5).Name = "Round ID"
destTable.ListColumns.Add(6).Name = "Round Index"
destTable.ListColumns.Add(7).Name = "Announced Date"

'3 - PREPARING LOOP VARIABLES
MsgBox ("Origin Table Selected: " & orgTable.Name)
Dim orgTblRow As ListRow
Dim arrInvestorNames() As String
Dim invNameColIndex As Integer

'Setting index of investor names column
invNameColIndex = 6

For Each orgTblRow In orgTable.ListRows

    arrInvestorNames = Split(orgTblRow.Range(invNameColIndex), "; ")
    Dim InvCounter As Integer
    Dim Investor As Variant
    Dim newRow As ListRow

    InvCounter = 1

    For Each Investor In arrInvestorNames
```

```
        'CREATE TABLE ROW
        Set newRow = destTable.ListRows.Add
        newRow.Range(, 1) = orgTblRow.Range(, 1)
        newRow.Range(, 2) = orgTblRow.Range(, 2)
        newRow.Range(, 3) = InvCounter

        If (Investor <> 0 And Investor <> "0") Then
            newRow.Range(, 4) = Investor
            Else: newRow.Range(, 4) = ""
        End If

        newRow.Range(, 5) = ""
        newRow.Range(, 6) = orgTblRow.Range(, 4)
        newRow.Range(, 7) = orgTblRow.Range(, 5)

        'REMINDER to delete table empty row
        'REMINDER to insert condition for null inv names
        InvCounter = InvCounter + 1

    Next Investor

    Erase arrInvestorNames

    'EXIT CONDITION FOR orgTblRow LOOP
    'If (InputBox("Finire?", "Terminacript", "Yes") <> "Yes") Then

    '    Else
    '        Exit For
    '    End If

Next orgTblRow

Application.ScreenUpdating = True
End Sub
```

B.3 VBA script to split comma-separated lists

```
Sub separateWithComma()

Dim savedStr(1 To 700[10]) As String
Dim c As Range
Dim tempStrArr As Variant
Dim myRange As Range
Dim tmpStr As String
Dim cellVal As String
Dim j As Integer
Set myRange = Selection

j = 0

For k = 1 To myRange.Count Step 1

    tmpStr = myRange.Item(k)
    tempStrArr = Split(tmpStr, ", ")

    For i = LBound(tempStrArr) To UBound(tempStrArr)

        j = j + 1
        savedStr(j) = tempStrArr(i)

    Next i

Next k

myRange.Offset(, 2).Resize(UBound(savedStr)) =
WorksheetFunction.Transpose(savedStr)

End Sub
```

[10] The maximum size of the array depends on the number of single values contained in total in the table. This should be calculated and inserted manually before running the macro.

Appendix C
Created samples

The table reported below showcases all the samples created. They are created, for each measure, or *variable*, under study grouping the startups in the data model by origin HQ and by the startups' investors' geographic origin composition. The sample name reports how the sample is composed: the "_" character separates the first part of the name, indicating the region of the startup's headquarters, from the second part of the sample name, indicating the startup's investors' geographic origin composition.

It should be noted that the samples that have a *Sample Dim* field empty are samples for which there were no startups available that satisfied the additional logical and informational criteria, of which the R script was not able to find a solution and it threw an error.

Table Appendix.0.1 – Table showcasing the samples created and their numerosity

Measure	Sample [Region_Investors]	Sample Dim
Exit Values [M$]	All	2774
Exit Values [M$]	EU_InvAll	949
Exit Values [M$]	EU_InvEU	236
Exit Values [M$]	EU_InvUS	49
Exit Values [M$]	US_InvAll	1824
Exit Values [M$]	US_InvUS	613
Exit Values [M$]	US_InvEU	50
Exit Values [M$]	All_InvEU	286
Exit Values [M$]	All_InvUS	662
Exit Values [M$]	All_InvEUandUS	2743
Exit Values [M$]	EU_InvEUandUS	598
Investment Values [M$]	All	2482
Investment Values [M$]	All_InvEU	651
Investment Values [M$]	All_InvUS	1397
Investment Values [M$]	EU_InvEU	583
Investment Values [M$]	EU_InvUS	60
Investment Values [M$]	EU_InvAll	
Investment Values [M$]	US_InvAll	1725
Investment Values [M$]	US_InvUS	1337
Investment Values [M$]	US_InvEU	68
Investment Values [M$]	All_InvEUandUS	292
Investment Values [M$]	EU_InvEUandUS	112
Investment Values [M$]	US_InvEUandUS	180
Multiples	All	523
Multiples	All_InvEU	124
Multiples	All_InvUS	321
Multiples	EU_InvAll	139

Measure	Sample [Region_Investors]	Sample Dim
Multiples	US_InvAll	384
Multiples	EU_InvEU	104
Multiples	US_InvUS	308
Multiples	EU_InvUS	13
Multiples	US_InvEU	20
Multiples	All_InvEUandUS	
Multiples	EU_InvEUandUS	
Multiples	US_InvEUandUS	

Appendix D
Results

This appendix section illustrates the result estimates for the distributions fitted to the created samples. It is separated into three sections, one for each variable studied. At the start of each section, there is a table summarizing the parameters for each sample and distribution type that has been tested. It should be noted that not all samples were tested against the three distribution types considered. In general, the distributions that are not reported are those for which the R script returned an error. The column *Sample [Region_Investors]* indicated the sample that is being examined, while the column *Distribution* indicates the distribution that is being tested on the selected sample. The values of the field *Distribution* are:

- *Power-lax*: the distribution being tested is a power law.
- *Log Norm*: the distribution being tested is log-normal and it has the same cut-off parameter as the power law tested on the sample.
- *Log Norm Tot*: it is not available for all samples, as often the script could not run the test on the goodness of fit. The distribution type tested is log-normal, however, there are no constraints on the parameters and the script is free to explore a larger solution space.

D.1 Investment values [M$]

Table Appendix.0.2 - *Results for the investment values distribution estimates*

Sample [Region_Investors]	Sample Dim	Distribution	PL: xmin [M$]	PL: alpha	LN: par1	LN: par2	gof	p	p no_of_sims	ntail	Fit?
All	2774	Power-law	845.00	2			0.05045067	0.045	200	240	No
All	2774	Log Norm	845.00		2.913	1.919206	0.01604436	0.96	100	240	Yes
EU_InvAll	949	Power-law	1113.74	2.30368			0.04591232	0.926 6667	300	42	Yes
EU_InvAll	949	Log Norm	1113.74		-39.881149	6.16996				42	Bs Error
EU_InvEU	236	Power-law	188.70	2.244453			0.0630709	0.69	1000	49	Yes
EU_InvEU	236	Log Norm	188.70		-5	3	0.02708322	0.995	1000	49	Yes
EU_InvUS	49	Power-law	550.00	2.062334			0.1211432	0.546	1000	10	Yes
EU_InvUS	49	Log Norm	550.00		-5.840807	4	0.05106913	0.991	1000	10	Yes
EU_InvUS	49	Log Norm Tot	4.18		4.979406	1.706739	0.05106913	0.788	1000	47	Yes

Sample [Region_Investors]	Sample Dim	Distribution	PL: xmin [M$]	PL: alpha	LN: par1	LN: par2	gof	p	p no_of_sims	ntail	Fit?
US_InvAll	1824	Power-law	240.00	1.9843440			0.06639699	0	500	510	No
US_InvAll	1824	Power-law	799.67	2.394138			0.06849002	0.002	500	187	No
US_InvUS	613	Power-law	500.00	2.319953			0.06923154	0.028	1000	107	No
US_InvUS	613	Log Norm	500.00		4.477036	1.573682	0.02852301	0.908	500	107	Yes
US_InvUS	613	Log Norm Tot	101.00		4.523448	1.563834	0.02852301	0.36	500	346	Yes
US_InvEU	50	Power-law	147.00	2			0.1272477	0.095	1000	22	No
US_InvEU	50	Log Norm	147.00		5.403796	1.110474	0.06557694	0.744	1000	22	Yes
US_InvEU	50	Log Norm Tot	72.00		4.63729	1.417935	0.06557694	0.53	1000	33	Yes
All_InvEU	286	Power-law	175.00	2.153679			0.06580983	0.32	500	71	Yes
All_InvEU	286	Log Norm	175.00		3.151493	1.782664	0.03039123	0.93	500	71	Yes
All_InvEU	286	Log Norm Tot	13.50		4.175567	1.528958	0.03039123	0.396	500	221	Yes
All_InvUS	662	Power-law	499.33	2.309392			0.06388263	0.082	1000	120	No
All_InvUS	662	Log Norm	499.33		-1.594979	2.669603	0.02832578	0.878	500	120	Yes
All_InvUS	662	Log Norm Tot	115.00		4.23823	1.683145	0.02832578	0.226	500	345	Yes
All_InvEUandUS	2743	Power-law	169.13	1.969667			0.06394502	0	500	1261	No
All_InvEUandUS	2743	Log Norm	169.13		3.75919	1.793655	0.03232967	0	500	1261	No
All_InvEUandUS	2743	Log Norm Tot	69.00		4.589471	1.562302	0.03232967		500		Bs Error
EU_InvEUandUS	598	Power-law	145.38	1.944784			0.08549033	0	500	241	No
EU_InvEUandUS	598	Log Norm	145.38		4.090199	1.690159	0.03637801	0.148	500	241	Yes
EU_InvEUandUS	598	Log Norm Tot	3.00		4.48846	1.650041	0.03637801	0.002	500		No
US_InvEUandUS		Power-law	170.00	1.979613			0.06593744	0	500	1052	No
US_InvEUandUS		Log Norm	170.00		3.559048	1.847737	0.03493414	0	500	1052	No
US_InvEUandUS		Log Norm Tot	74.00		4.45767	1.610212	0.03493414	0	500		No

D.1.1 All regions and all investors

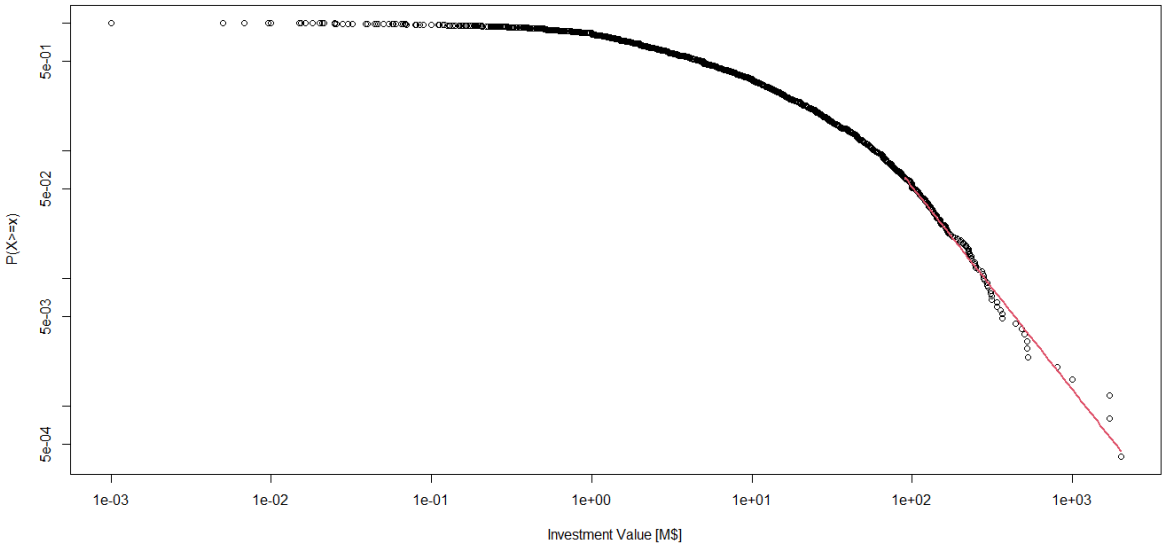

Figure 0.1 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from all regions.*

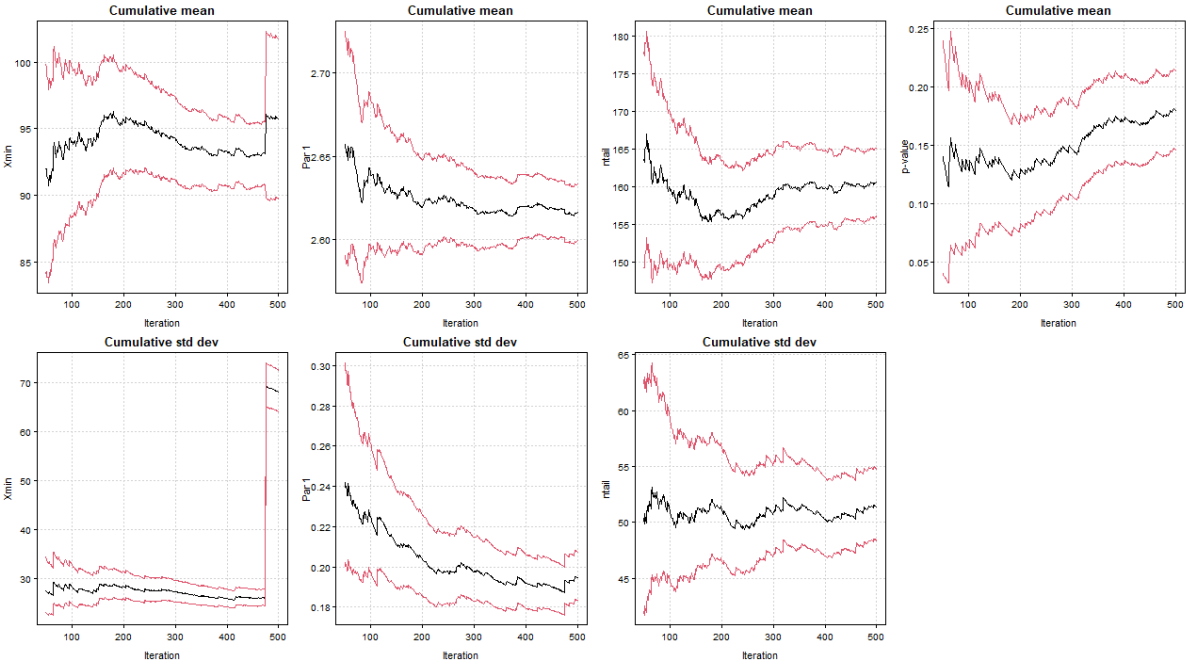

Figure 0.2 - *Bootstrapping results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from all regions.*

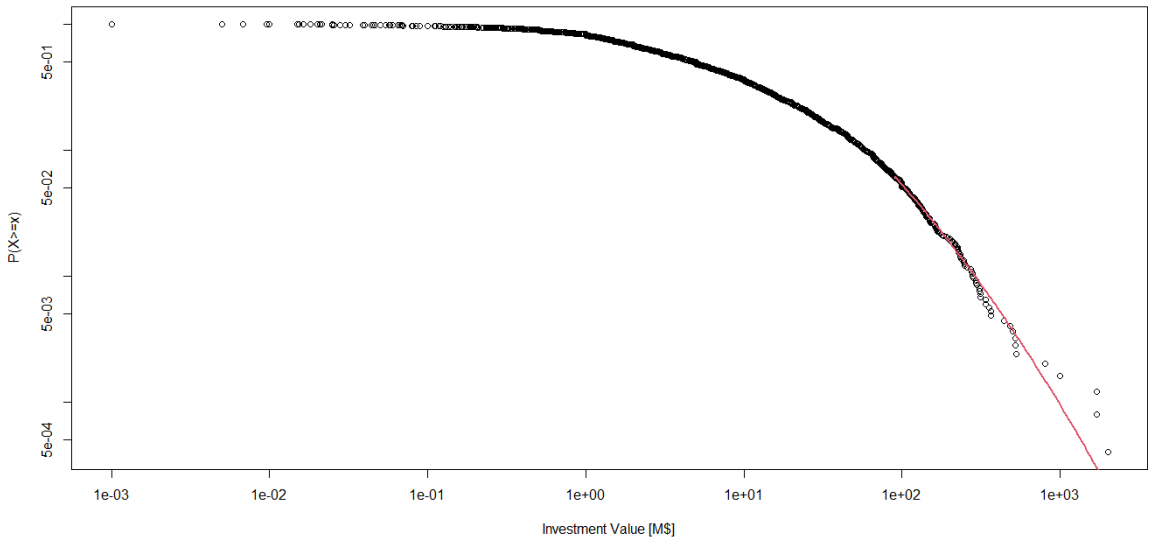

Figure 0.3 - *Estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from all regions.*

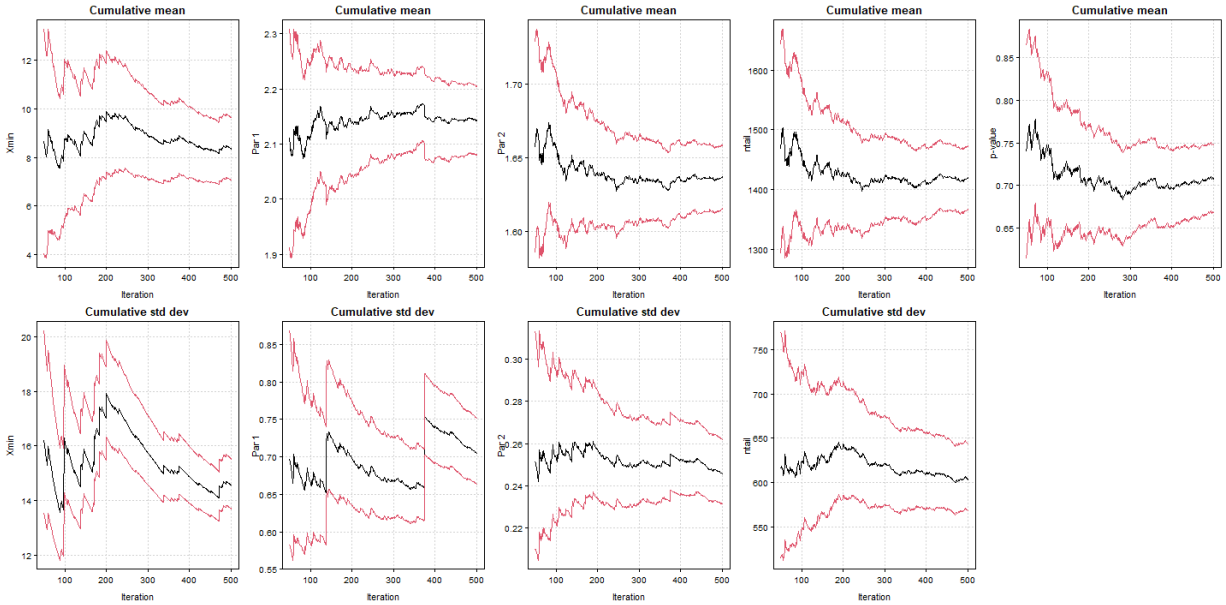

Figure 0.4 - *Bootstrapping results for the estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from all regions.*

D.1.2 All regions – EU investors

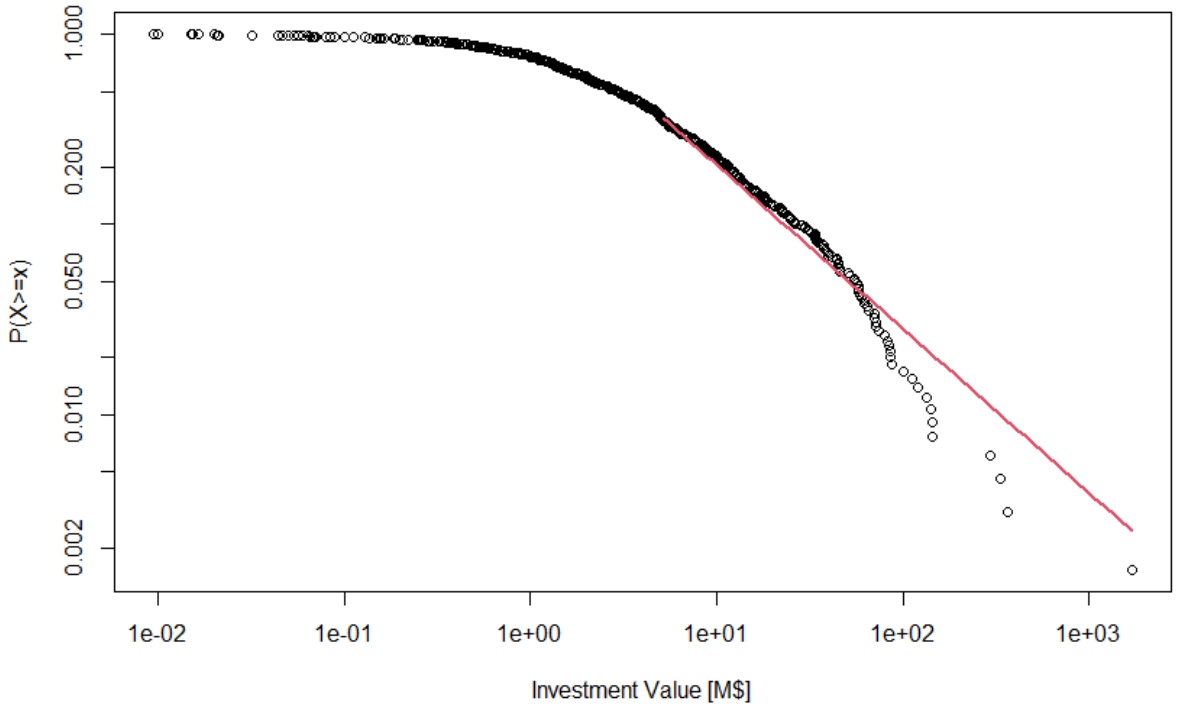

Figure 0.5 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the EU region.*

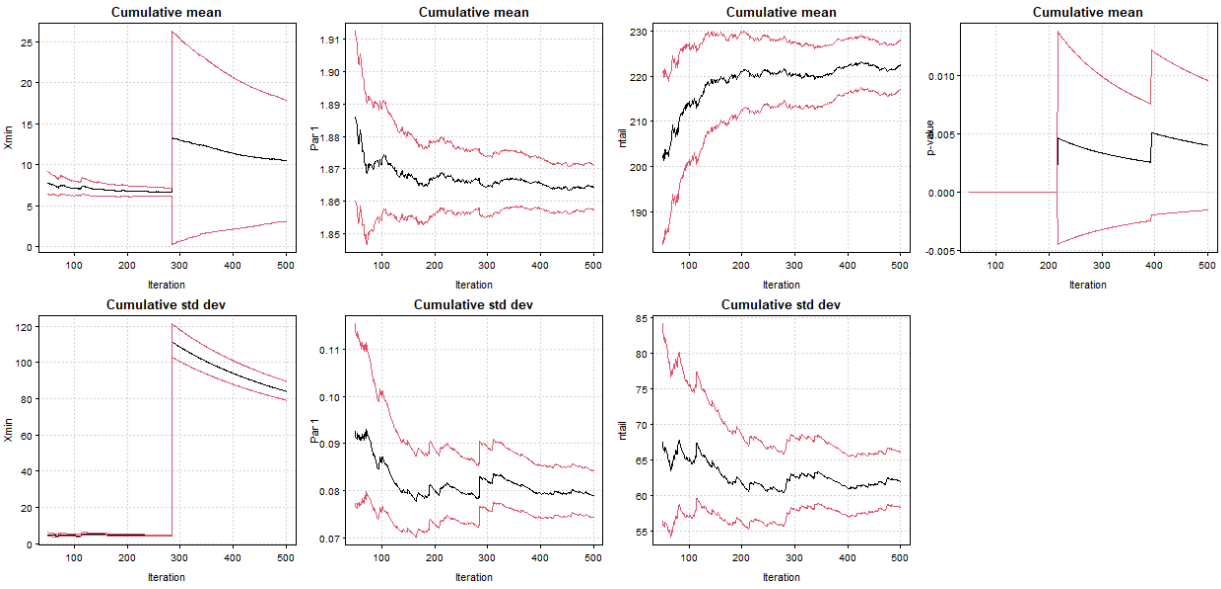

Figure 0.6 - *Bootstrapping results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the EU region.*

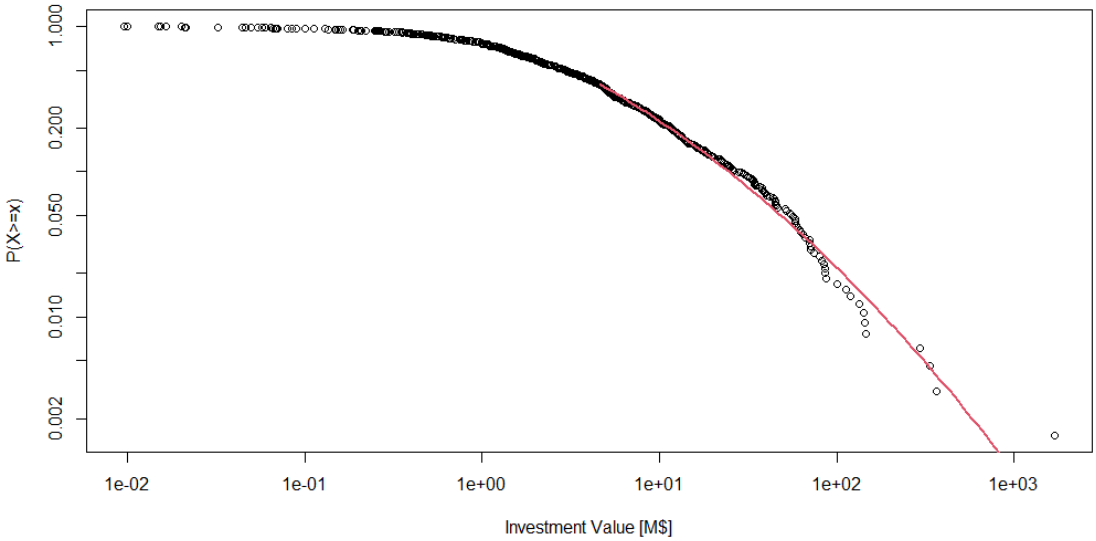

Figure 0.7 - *Estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the EU region.*

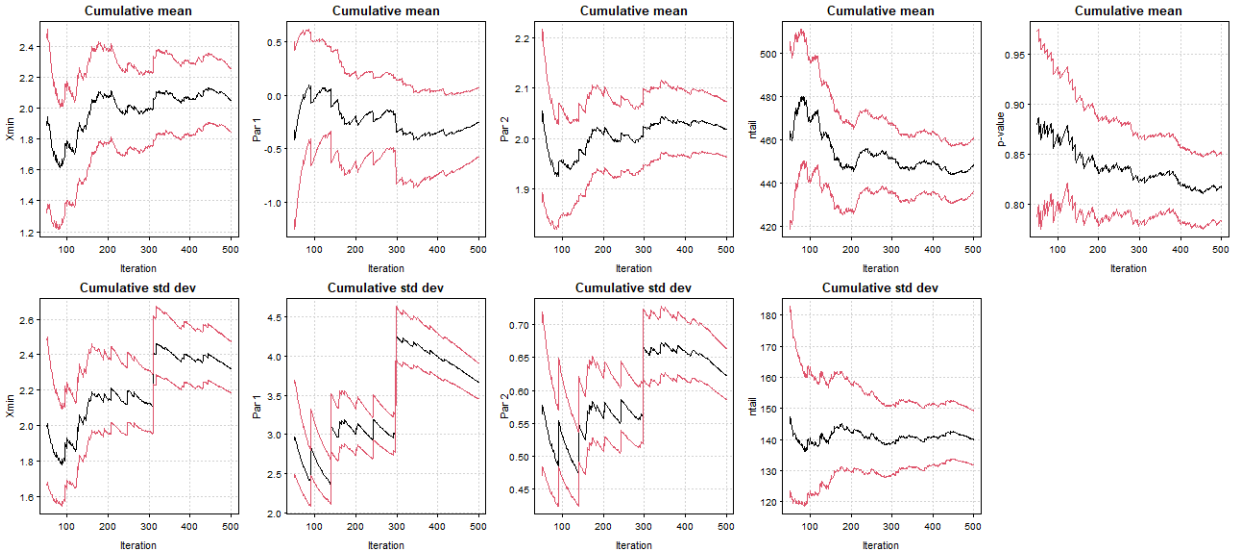

Figure 0.8 - *Bootstrap results for the estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the EU region.*

D.1.3 All regions – US Investors

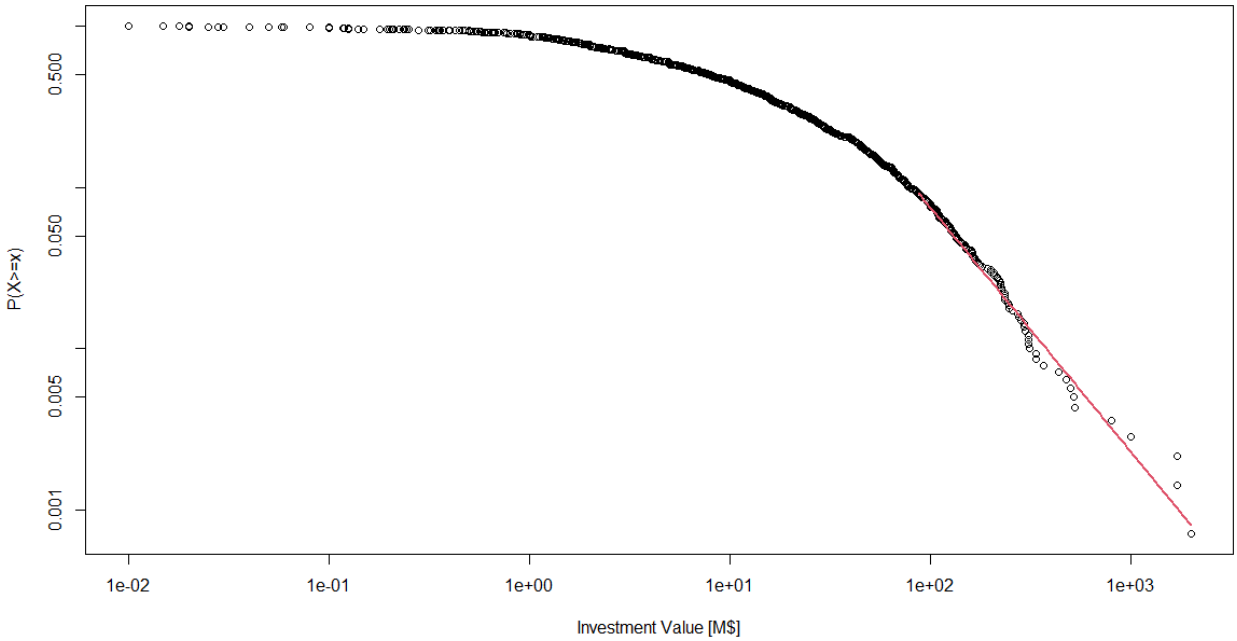

Figure 0.9 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the US region.*

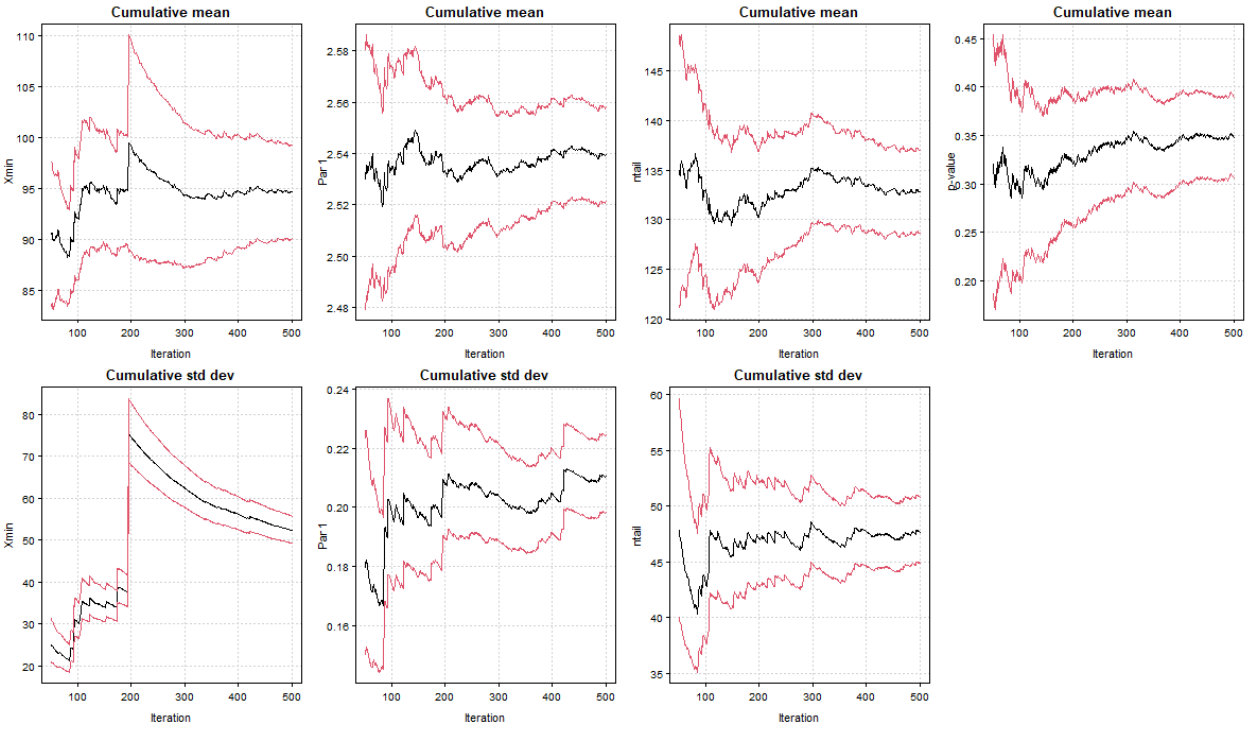

Figure 0.10 - *Bootstrap results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the US region.*

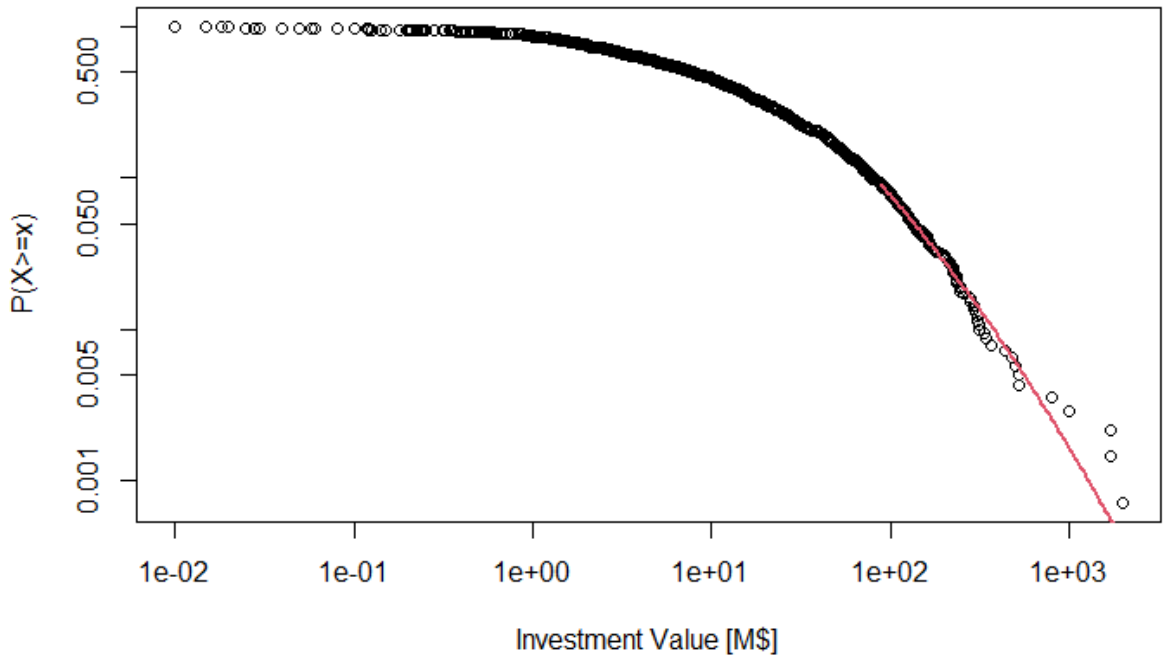

Figure 0.11 - *Estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the US region.*

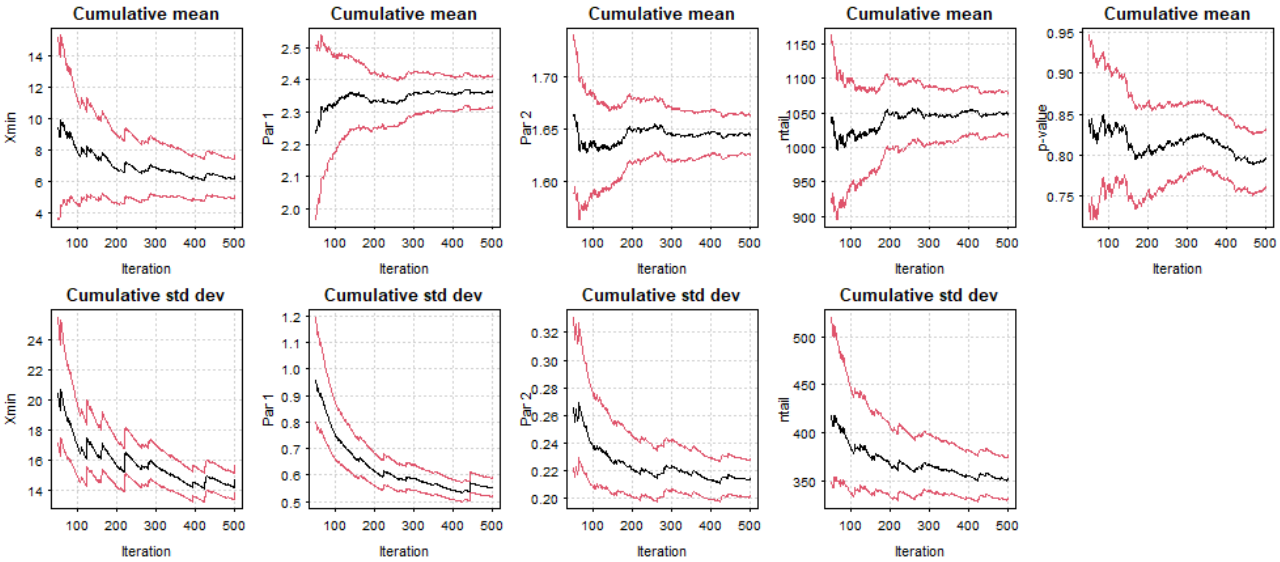

Figure 0.12 - *Bootstrap results for the estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in all regions and investors from the US region.*

D.1.4 EU region – all investors

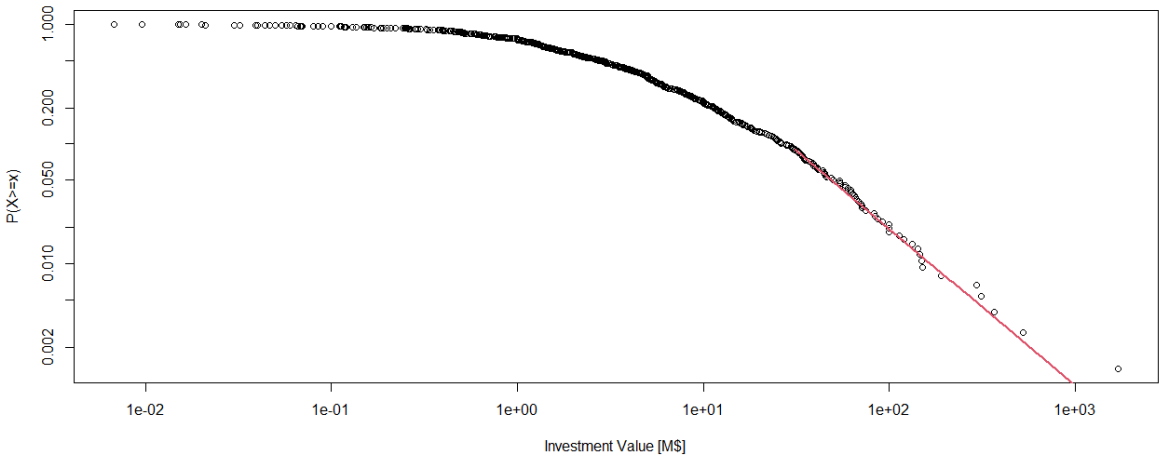

Figure 0.13 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the EU region and investors from all regions.*

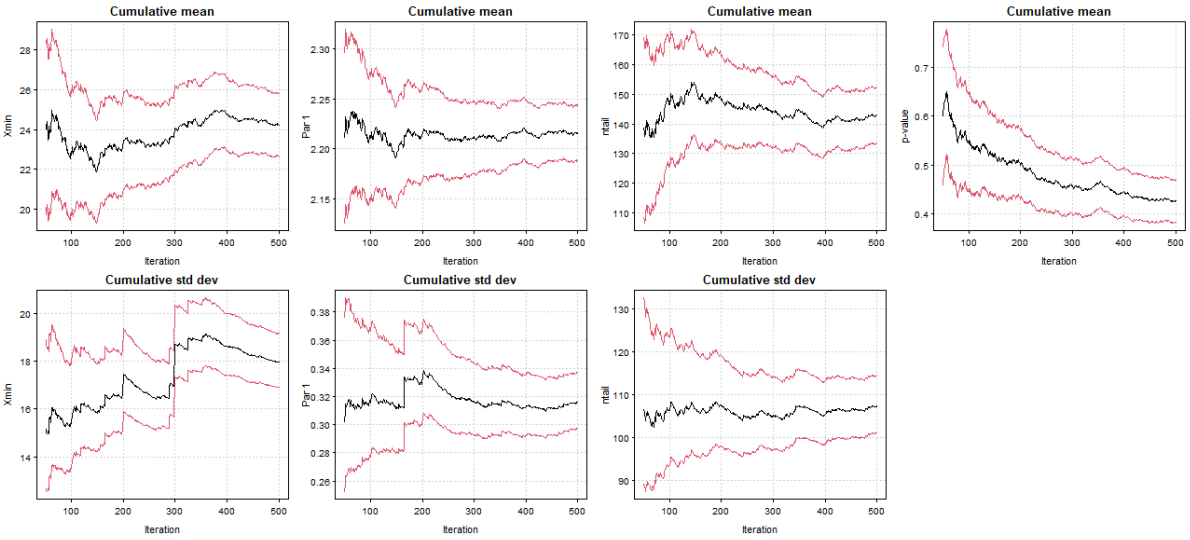

Figure 0.14 – *Bootstrap results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the EU region and investors from all regions.*

D.1.5 EU region – EU Investors

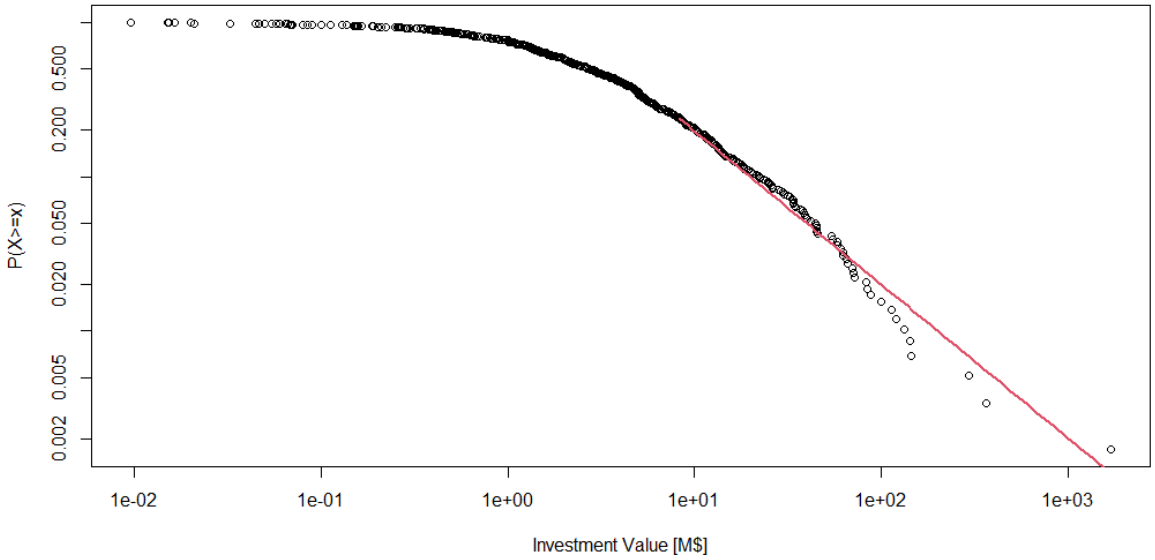

Figure 0.15 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

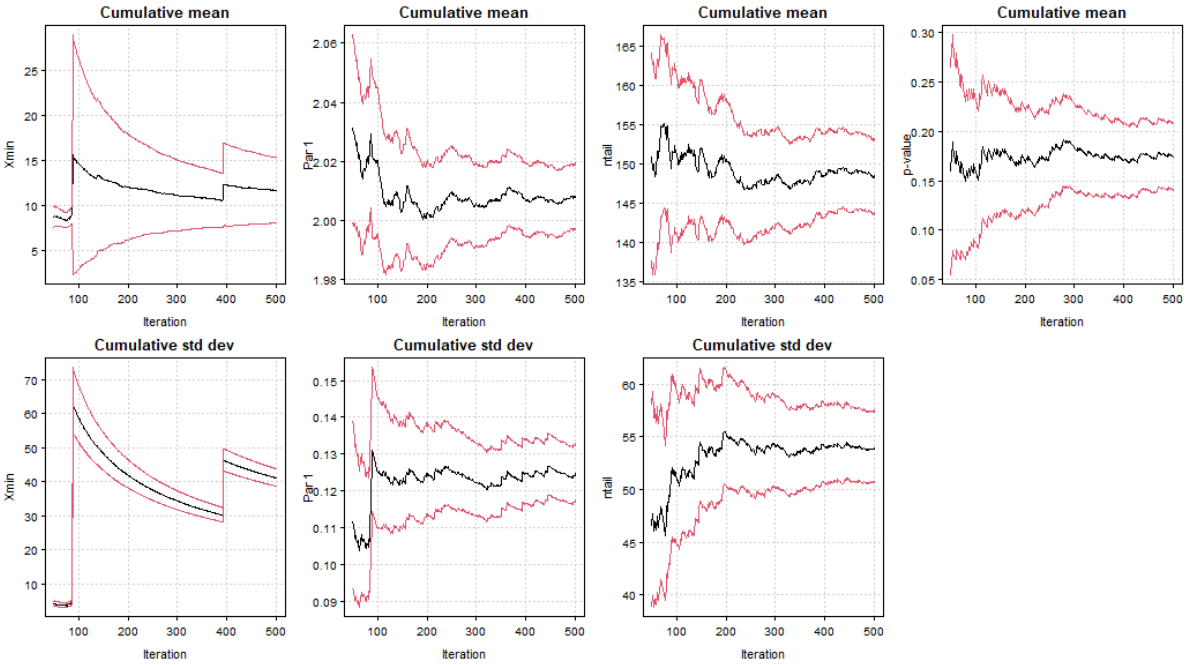

Figure 0.16 - *Bootstrap results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

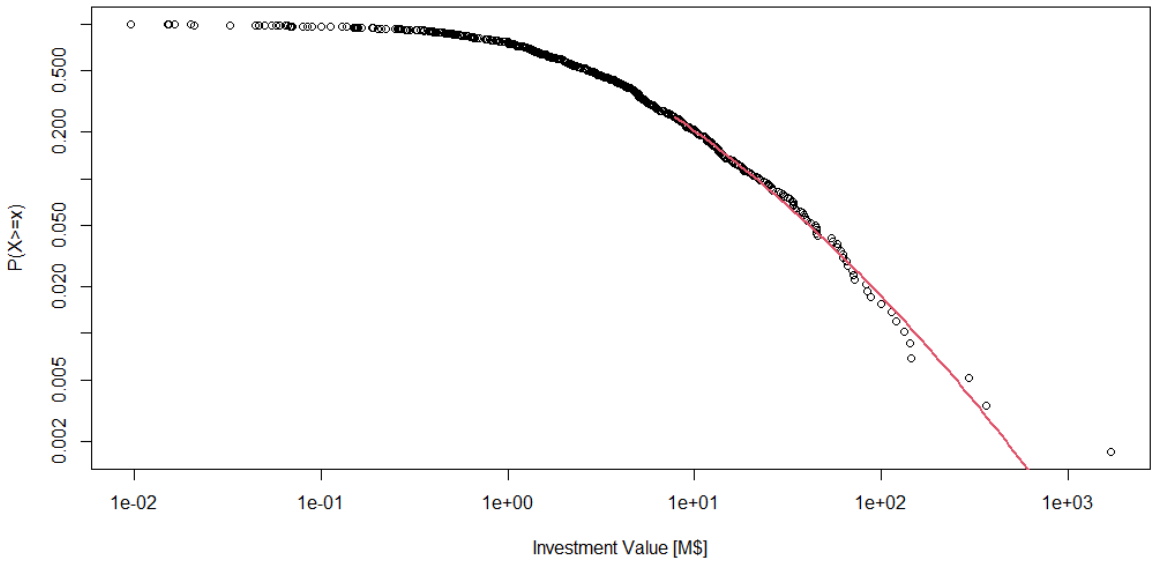

Figure 0.17 - *Estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

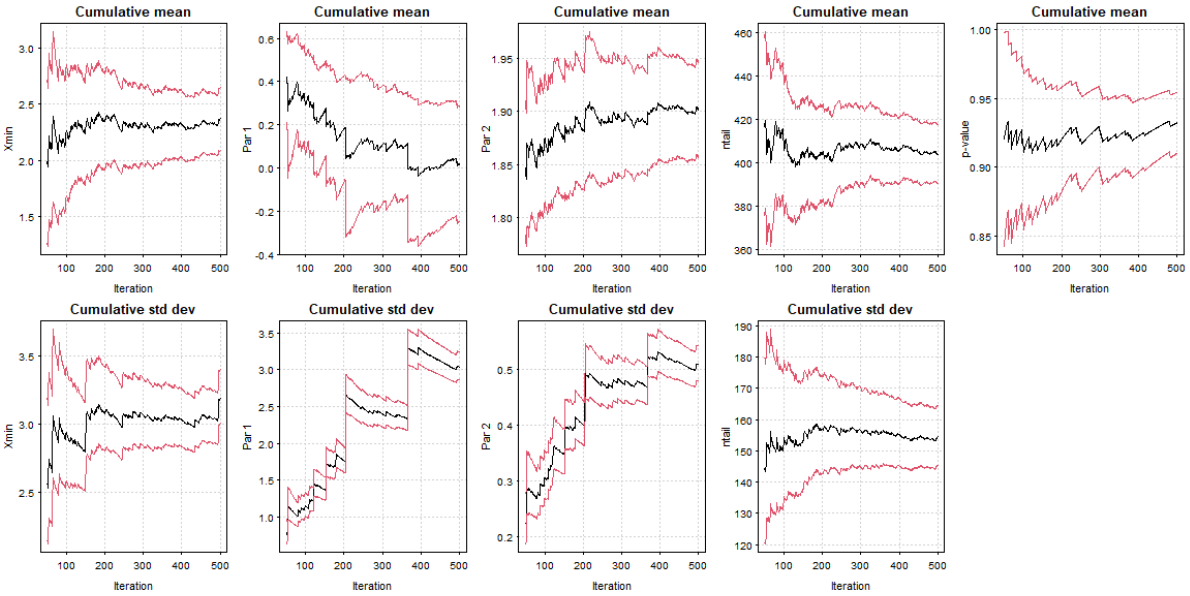

Figure 0.18 - *Bootstrap results for the estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

D.1.6 US region – All investors

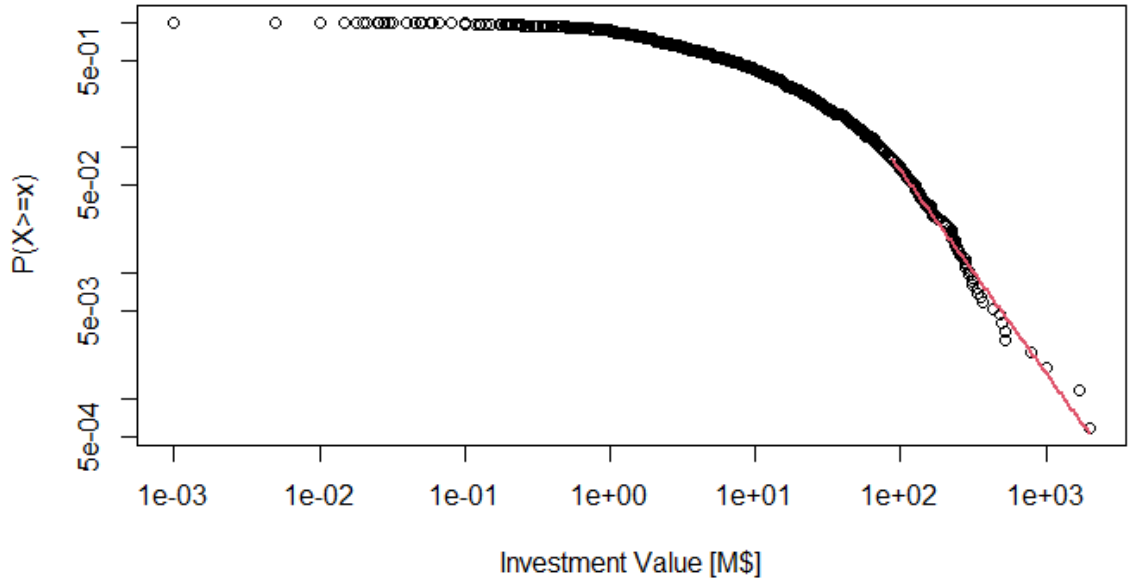

Figure 0.19 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from all regions.*

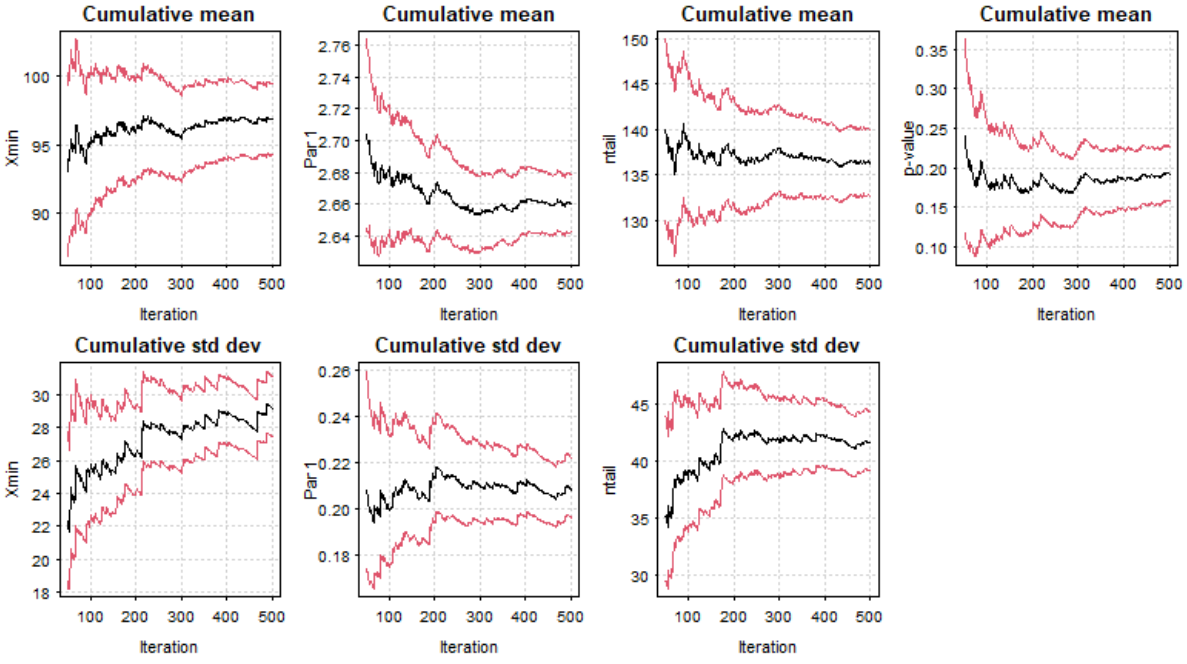

Figure 0.20 *– Bootstrapping results estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from all regions.*

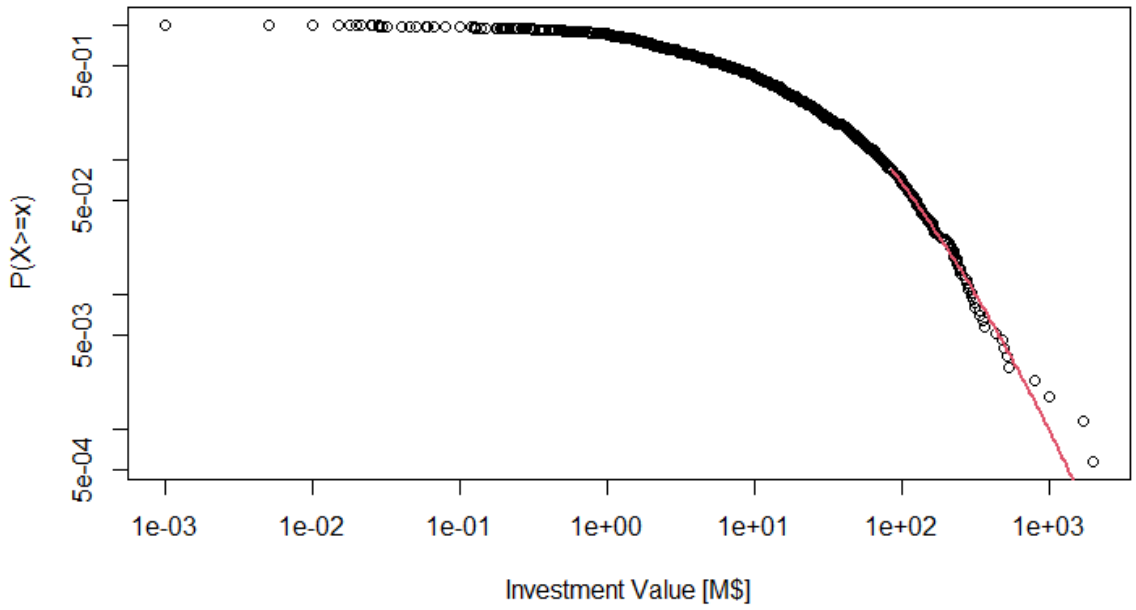

Figure 0.21 - *Estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from all regions.*

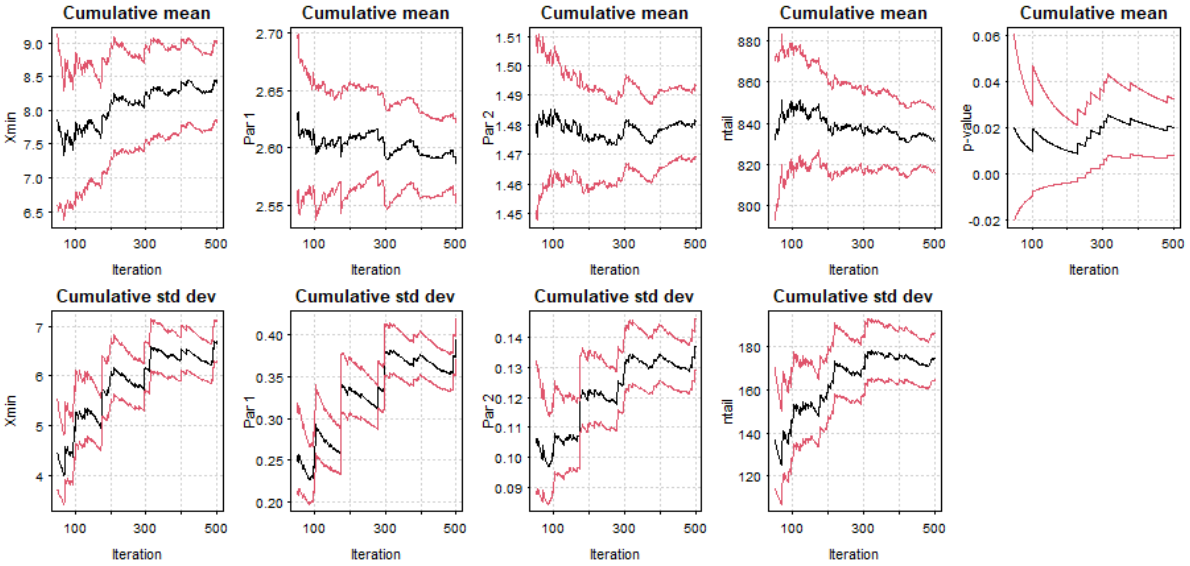

Figure 0.22 – *Bootstrapping results for the estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from all regions.*

D.1.7 US region – EU investors

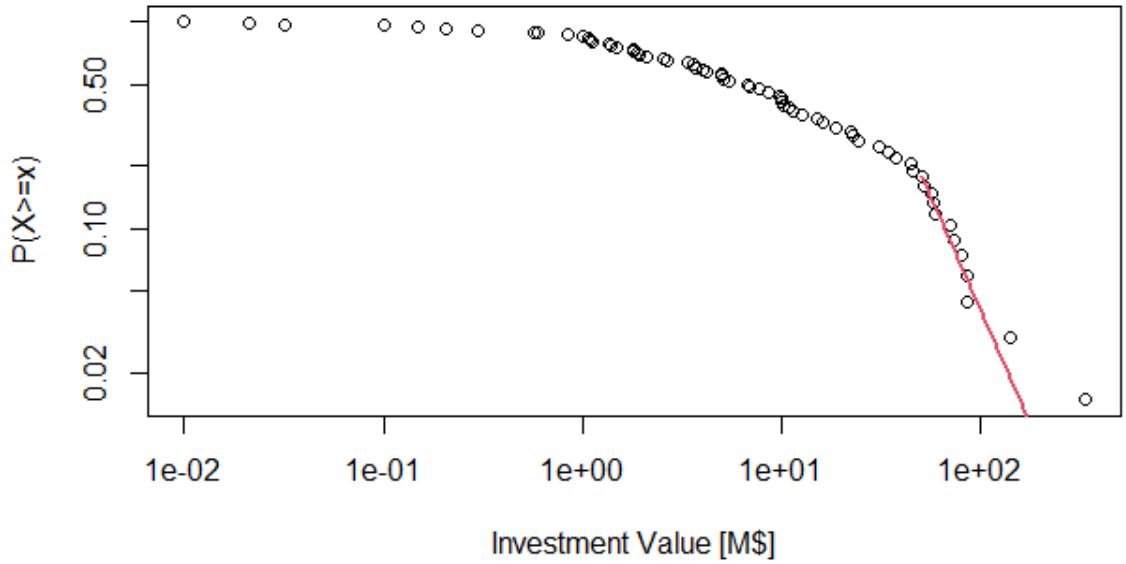

Figure 0.23 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from the EU region.*

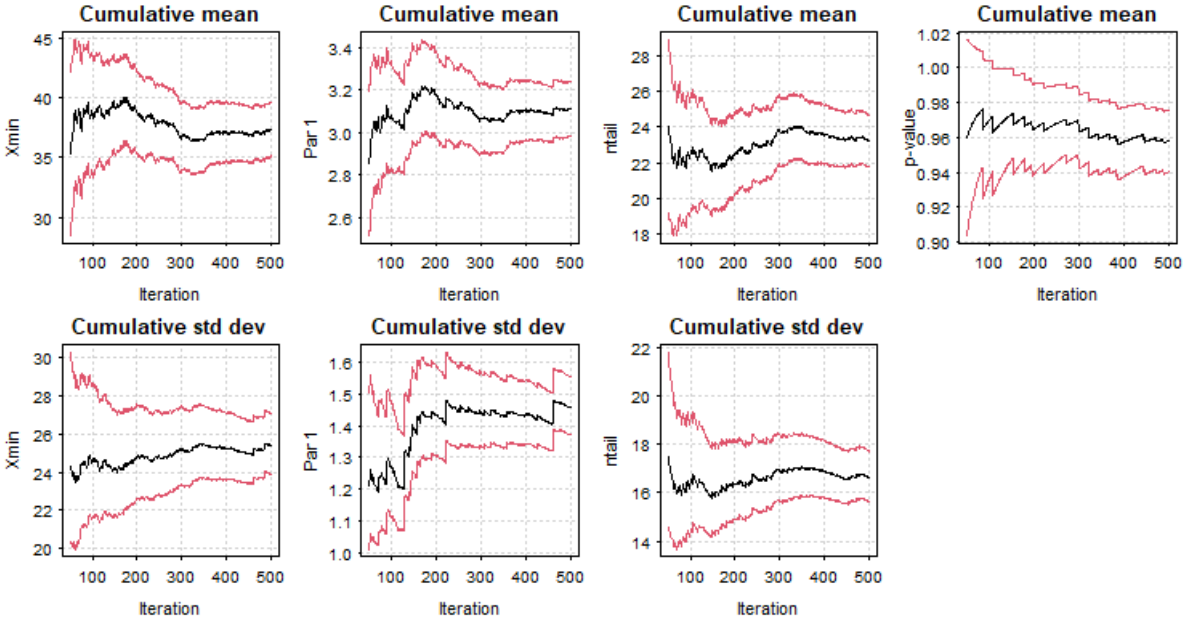

Figure 0.24 *– Bootstrapping results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from the EU region.*

D.1.8 US region – US investors

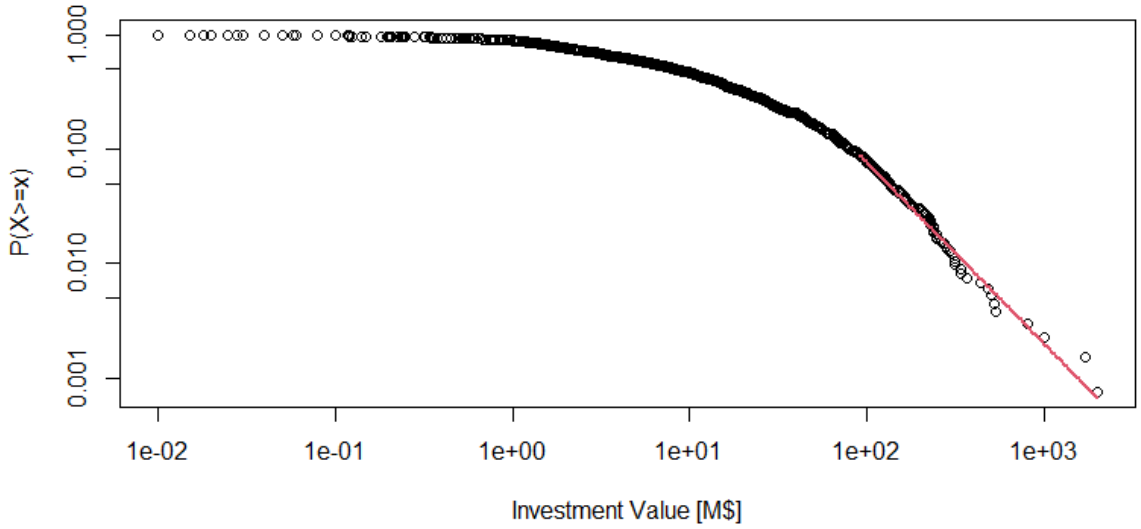

Figure 0.25 - *Estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from the US region.*

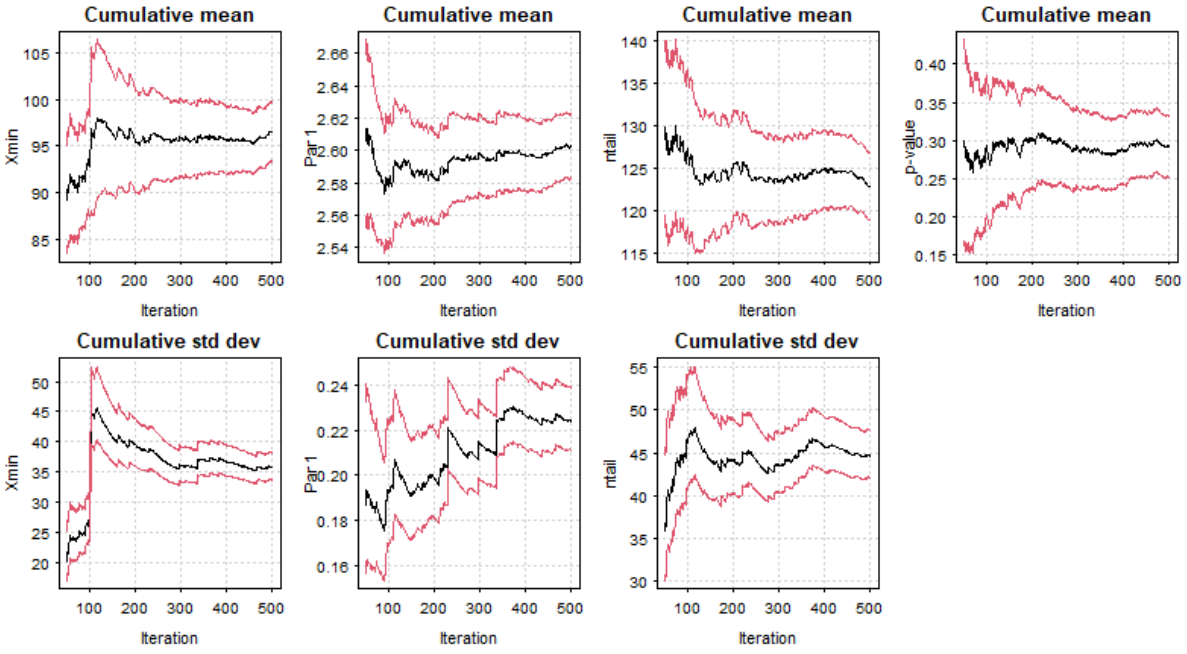

Figure 0.26 *- Bootstrap results for the estimated power-law distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from the US region.*

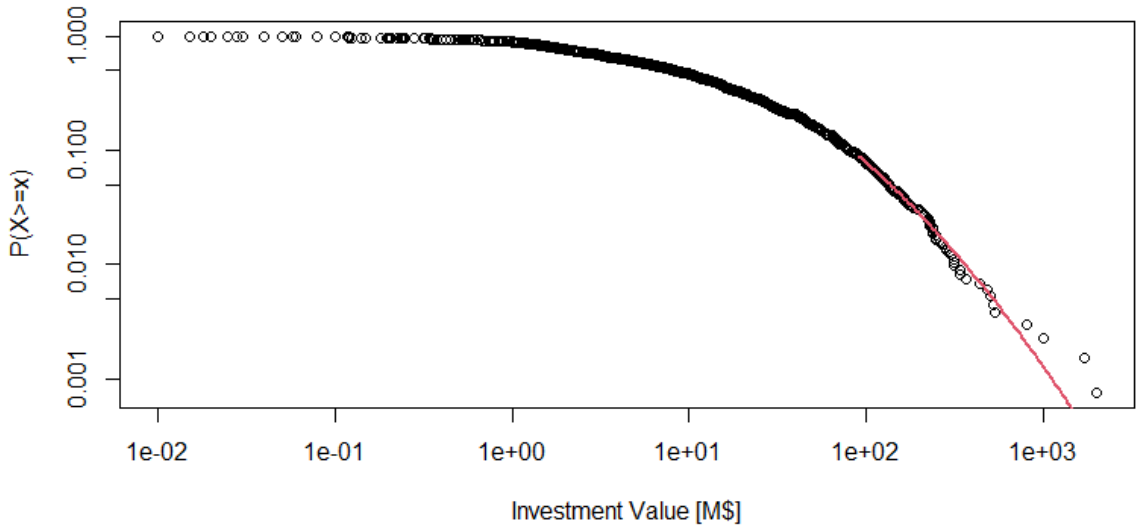

Figure 0.27 - *Estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from the US region.*

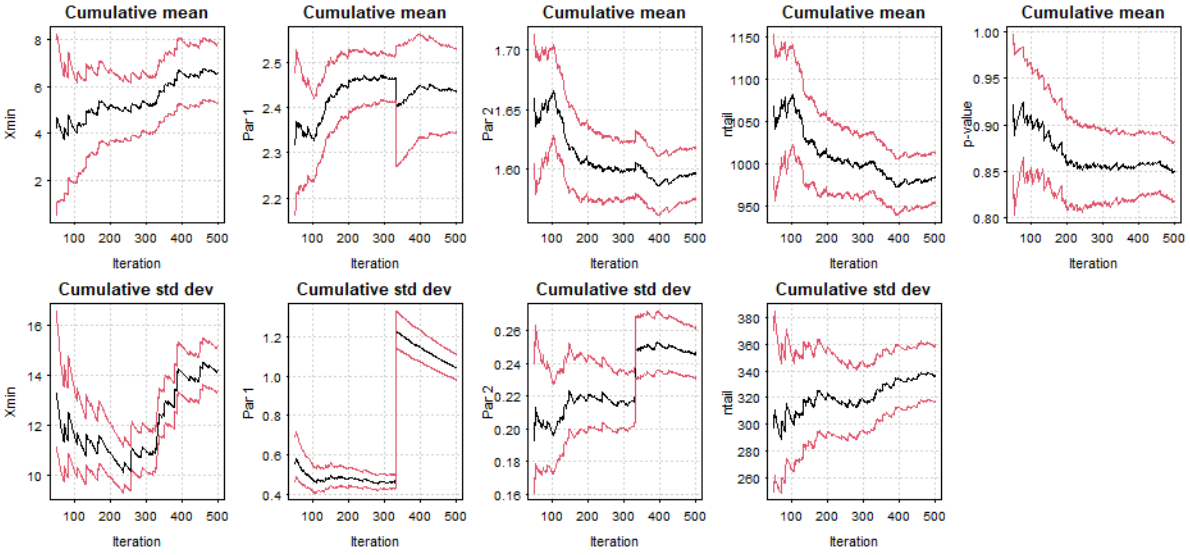

Figure 0.28 - *Bootstrap results for the estimated log-normal distribution of startups' total equity funding. Sample composition: startups with headquarters in the US region and investors from the US region.*

D.2 Exit values [M$]

Table Appendix.0.3 - *Results for the exit values distribution estimates*

Sample [Region_Investors]	Sample Dim	Distribution	PL: xmin [M$]	PL: alpha	LN: par1	LN: par2	gof	p	p no_of_sims	ntail	Fit?
All	2482	Power-law	90.00	2.587836			0.04947502	0.18	500	153	Yes
All	2482	Log Norm	90.00		0.5733729	1.8106869	0.02046975	0.708	500	153	Yes
All	2482	Log Norm Tot	5.30		2.307938	1.566301	0.02046975	0.034	500	1174	No
All_InvEU	651	Power-law	4.61	1.861759			0.05822278	0.004	500	261	No
All_InvEU	651	Log Norm	4.61		-0.9587419	2.2905372	0.02090581	0.816	500	261	Yes
All_InvEU	651	Log Norm Tot	1.01		0.6989886	1.8550472	0.02090581	0.456	500	495	Yes
All_InvUS	1397	Power-law	88.00	2.513946			0.04788165	0.348	500	128	Yes
All_InvUS	1397	Log Norm	88.00		0.6217243	1.8465578	0.021589	0.796	500	128	Yes
All_InvUS	1397	Log Norm Tot	42.00		2.770083	1.418334	0.021589	0.704	500	271	Yes
EU_InvEU	583	Power-law	7.83	1.994733			0.0519716	0.174	500	147	Yes
EU_InvEU	583	Log Norm	7.83		-1.353647	2.296045	0.02155227	0.932	500	147	Yes
EU_InvEU	583	Log Norm Tot	0.87		0.7002221	1.7650088	0.02155227	0.476	500	458	Yes
EU_InvUS	60	Power-law	3.00	1.622069			0.08119525	0.556	500	35	Yes
EU_InvUS	60	Log Norm	3.00		-7.76564	4.516281	0.06136831	0.674	500	35	Yes
EU_InvUS	60	Log Norm Tot	8.00		2.369312	2.192184	0.06136831	0.872	500	18	Yes
EU_InvAll									500		Bs Error
EU_InvAll									500		Bs Error
EU_InvAll									500		Bs Error
US_InvAll	1725	Power-law	85.70	2.61045			0.05069603	0.192	500	146	Yes
US_InvAll	1725	Log Norm	85.70		0.9416341	1.7050659	0.0235885	0.722	500	146	Yes
US_InvAll	1725	Log Norm Tot	5.40		2.605927	1.483696	0.0235885	0.02	500	914	No
US_InvUS	1337	Power-law	90.00	2.584359			0.05112865	0.292	500	120	Yes

Sample [Region_Investors]	Sample Dim	Distribution	PL: xmin [M$]	PL: alpha	LN: par1	LN: par2	gof	p	p no_of_sims	ntail	Fit?
US_InvUS	1337	Log Norm	90.00		1.825849	1.575886	0.02185357	0.85	500	120	Yes
US_InvUS	1337	Log Norm Tot	0.03		2.0926	1.792806	0.02185357	0	500	1323	No
US_InvEU	68	Power-law	51.11	3.190406			0.0842489	0.958	500	12	Yes
US_InvEU	68	Log Norm			-1017.90156	22.57197					Bs Error
US_InvEU	68	Log Norm Tot	0.01		1.750494	2.005883	0.05172677	0.526	500	68	Yes
All_InvEUandUS	292	Power-law	32.00	2.049198			0.1123064	0	500	116	No
All_InvEUandUS	292	Log Norm	32.00		3.915676	0.9912625	0.03919227	0.404	500	116	Yes
All_InvEUandUS	292	Log Norm Tot	8.25		3.592215	1.183139	0.03919227	0.228	500	181	Yes
EU_InvEUandUS	112	Power-law	34.06	2.448731			0.1117337	0.288	500	21	Yes
EU_InvEUandUS	112	Log Norm	34.06		-4.434081	2.600025	0.07100319	0.386	500	21	Yes
EU_InvEUandUS	112	Log Norm Tot	0.50		2.015011	1.585448	0.07100319	0.012	500	111	No
US_InvEUandUS	180	Power-law	207.45	5.025885			0.1163638	0.448	500	14	Yes
US_InvEUandUS	180	Log Norm	207.45		3.2971793	0.8232026	0.0481666	0.83	500	14	Yes
US_InvEUandUS	180	Log Norm Tot	58.00		4.6686303	0.6422162	0.0481666	0.472	500	62	Yes

D.2.1 All regions and all investors

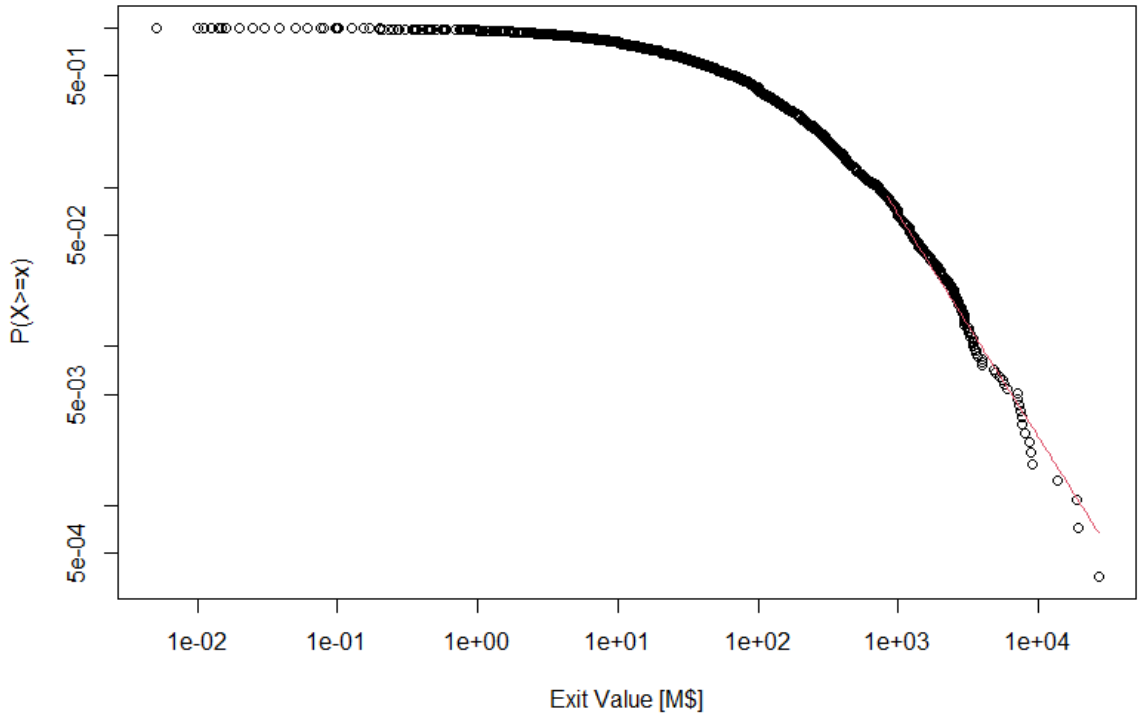

Figure 0.29 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from all regions.*

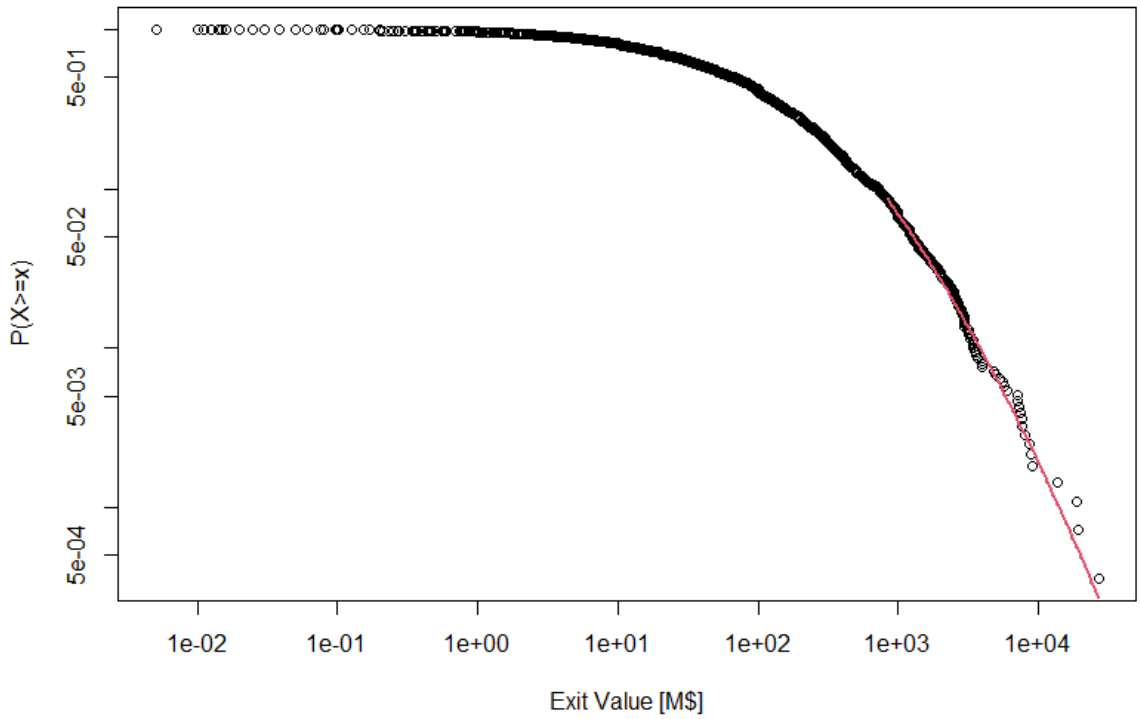

Figure 0.30 *- Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from all regions.*

D.2.2 All regions – EU investors

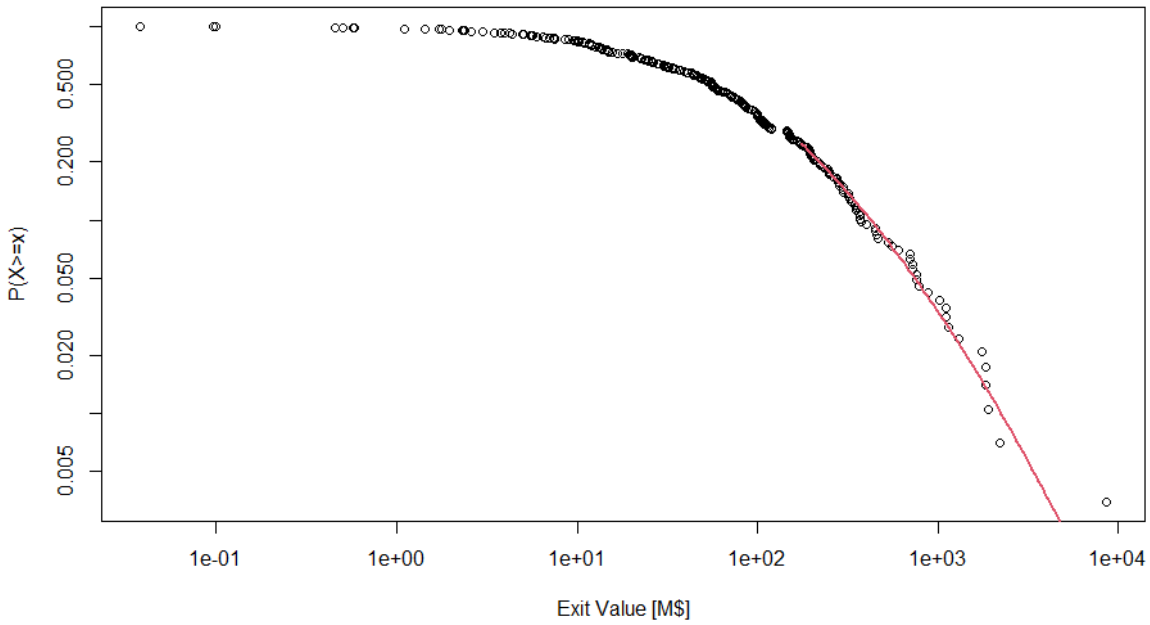

Figure 0.31 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the EU region.*

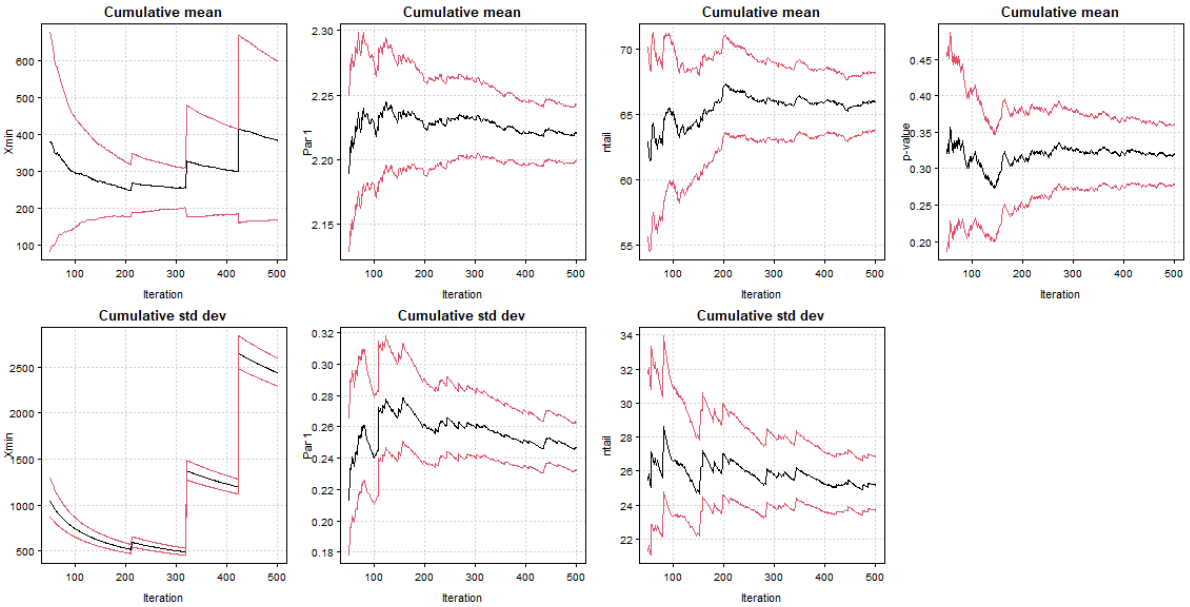

Figure 0.32 – *Bootstrapping results for the estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the EU region.*

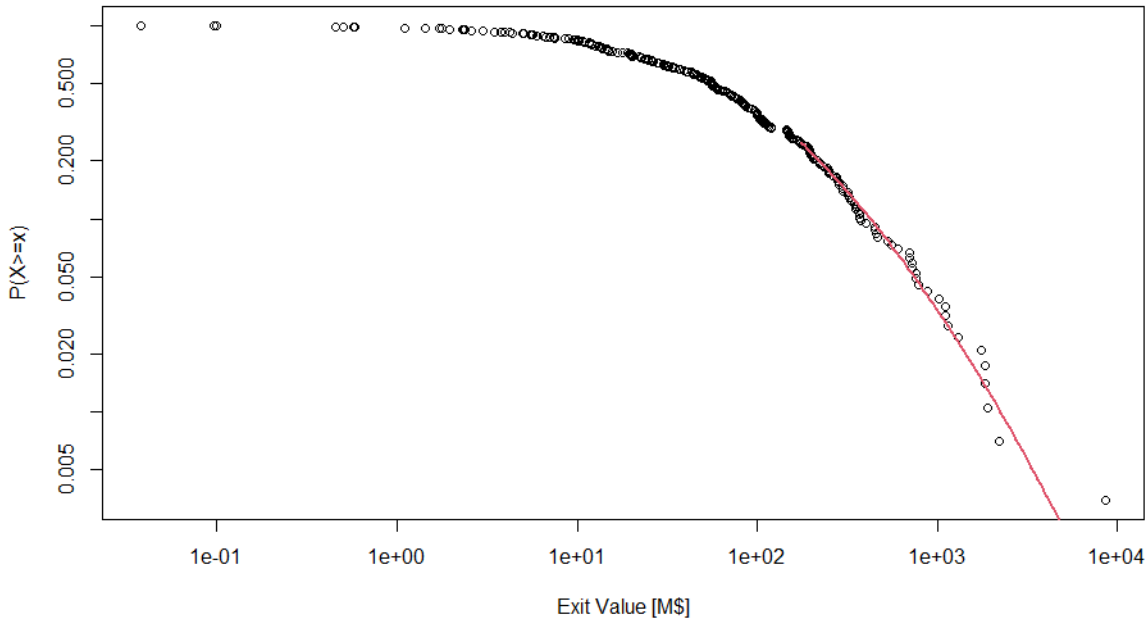

Figure 0.33 - *Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the EU region.*

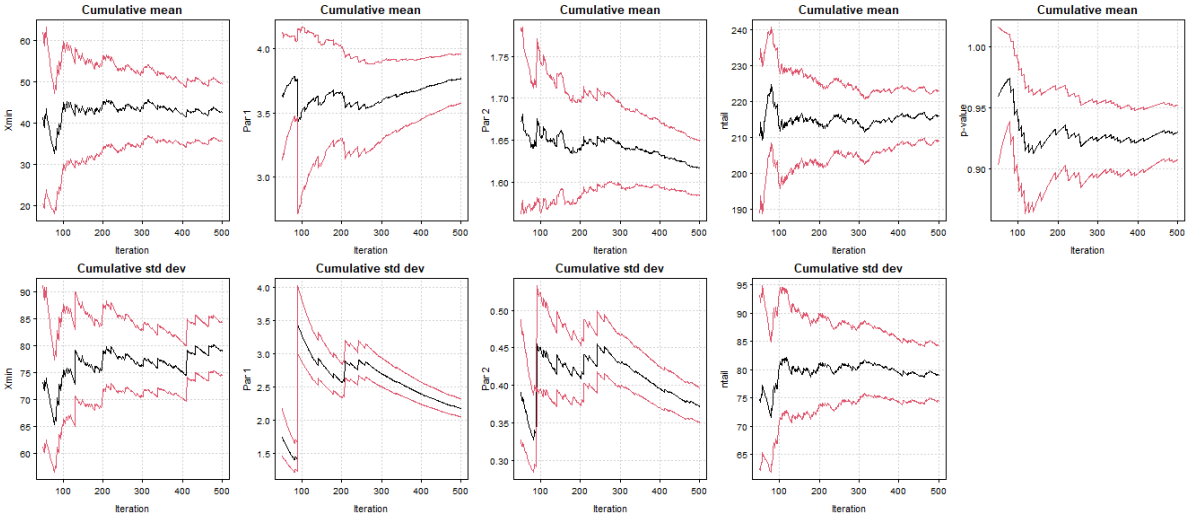

Figure 0.34 – *Bootstrapping results for the estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the EU region.*

D.2.3 All regions – US Investors

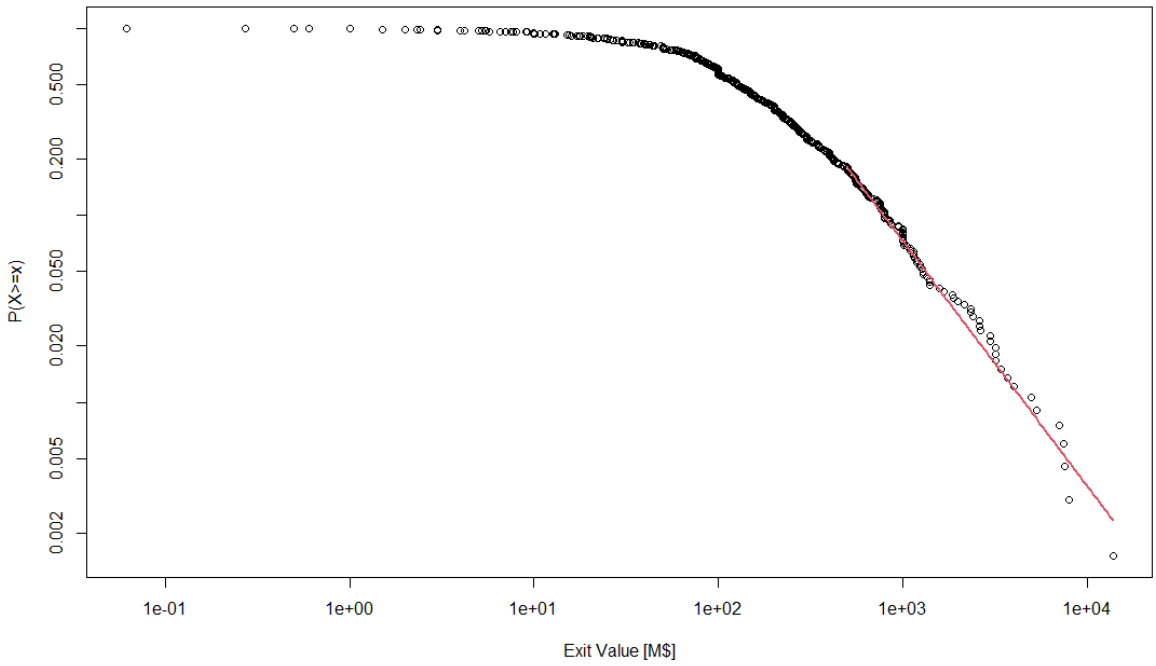

Figure 0.35 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the US region.*

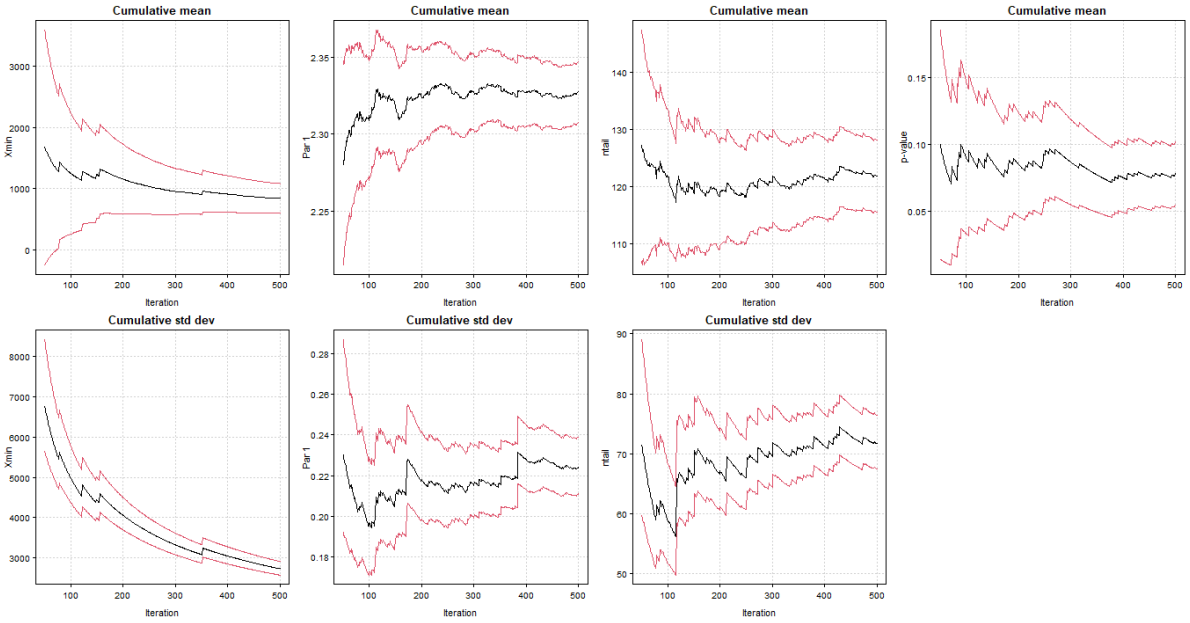

Figure 0.36 - *Bootstrapping results for the estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the US region.*

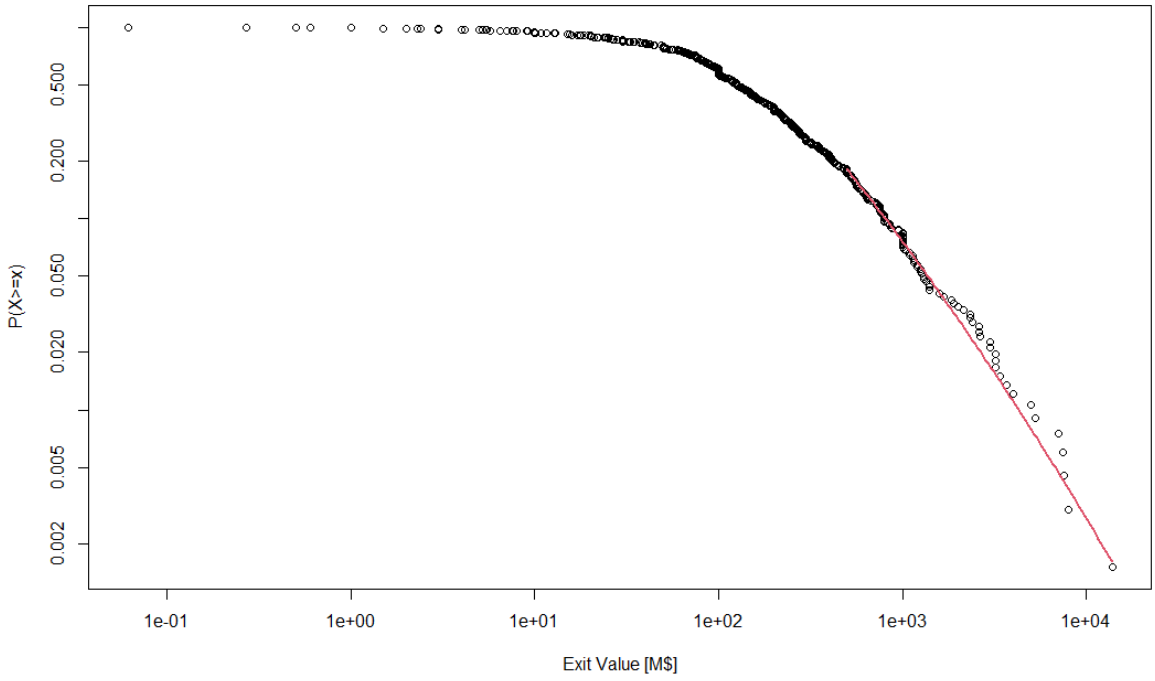

Figure 0.37 - *Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the US region.*

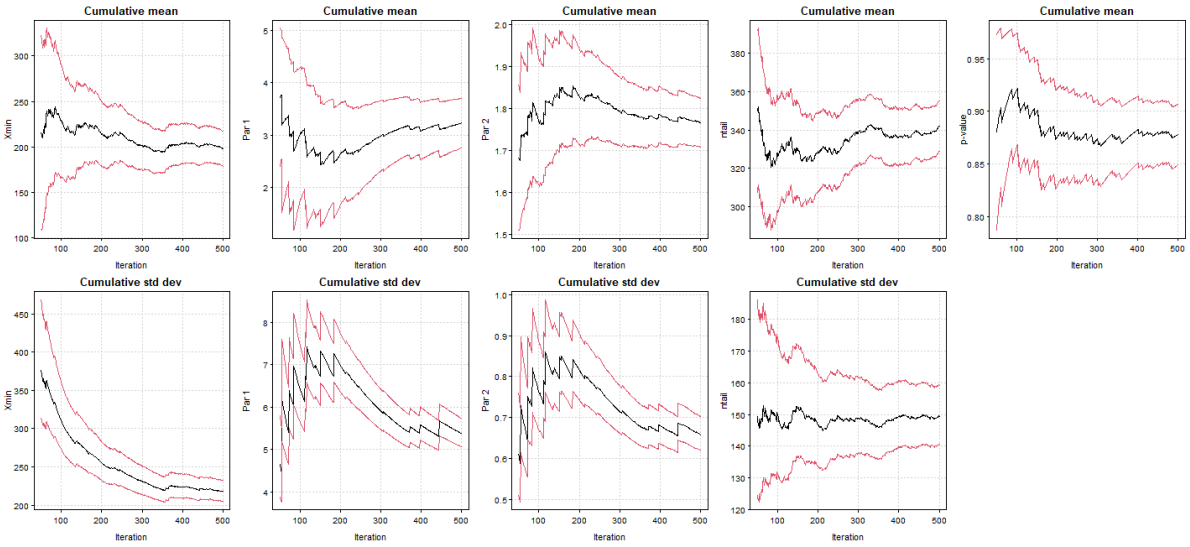

Figure 0.38 - *Bootstrapping results for the estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in all regions and investors from the US region.*

D.2.4 EU region – All investors

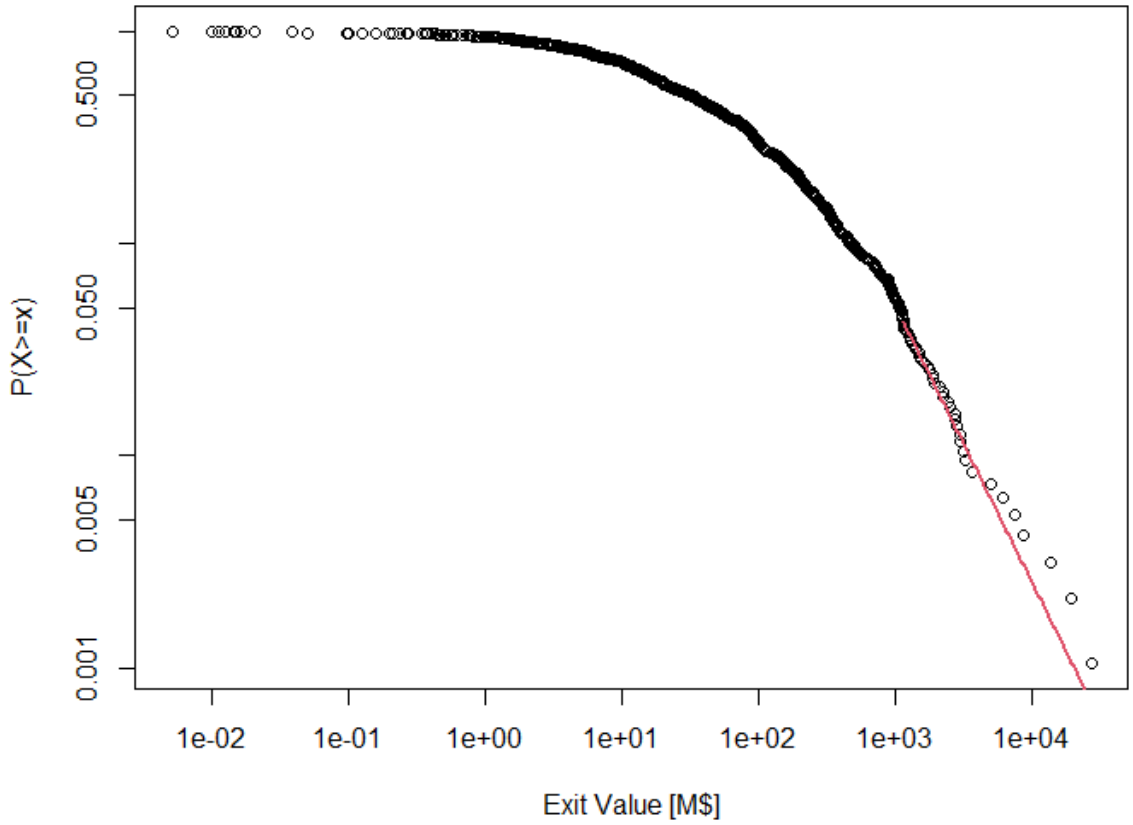

Figure 0.39 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the EU region and investors from all regions.*

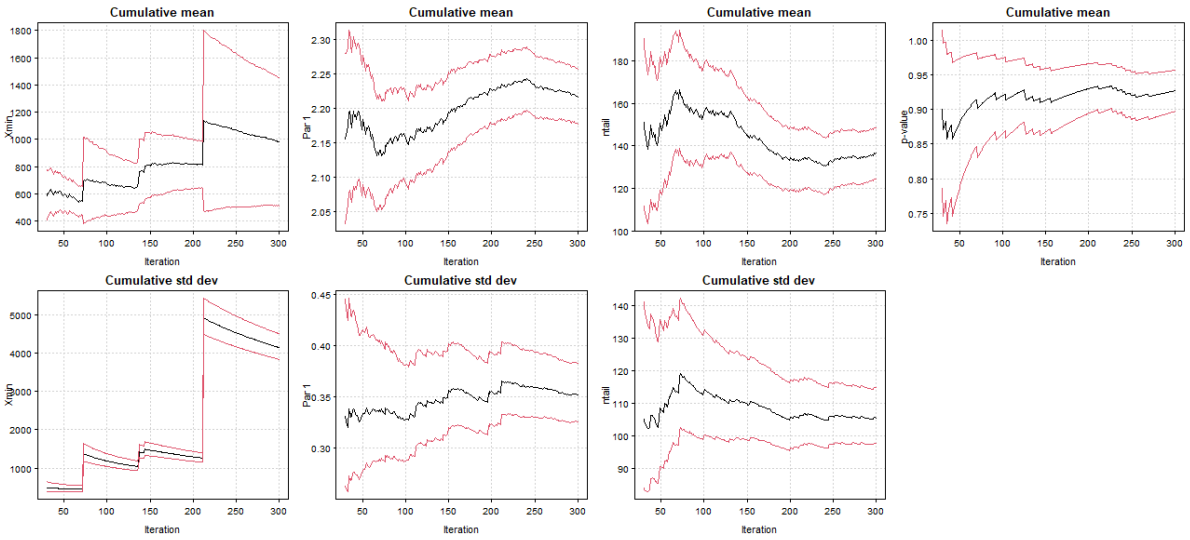

Figure 0.40 - *Bootstrapping results for the estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the EU region and investors from all regions.*

D.2.5 EU region – EU investors

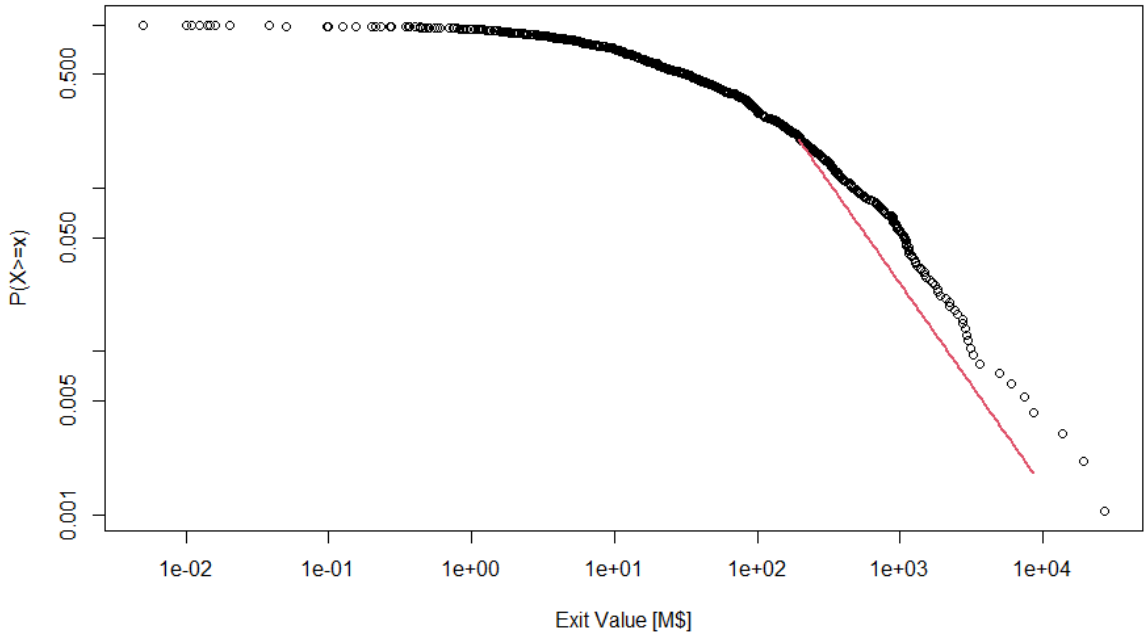

Figure 0.41 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

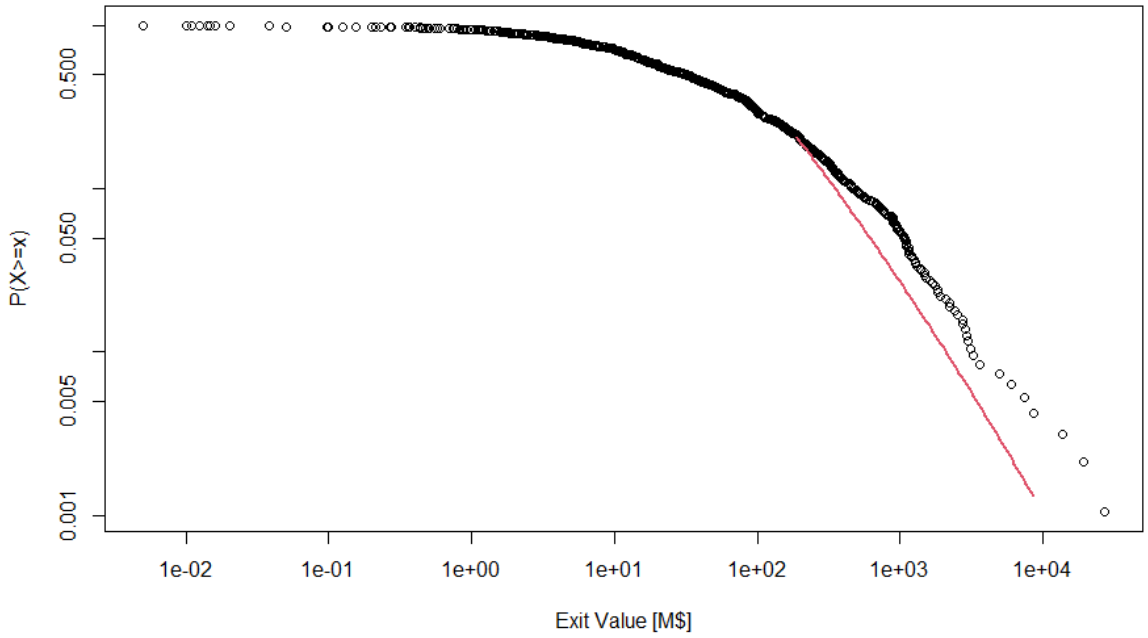

Figure 0.42 - *Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

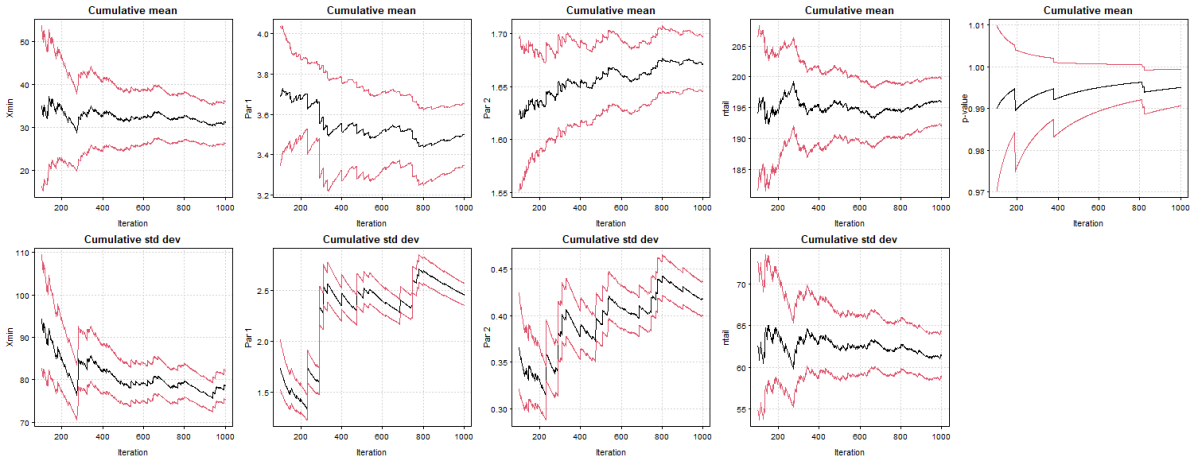

Figure 0.43 – *Bootstrapping results for the estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

D.2.6 US region – All investors

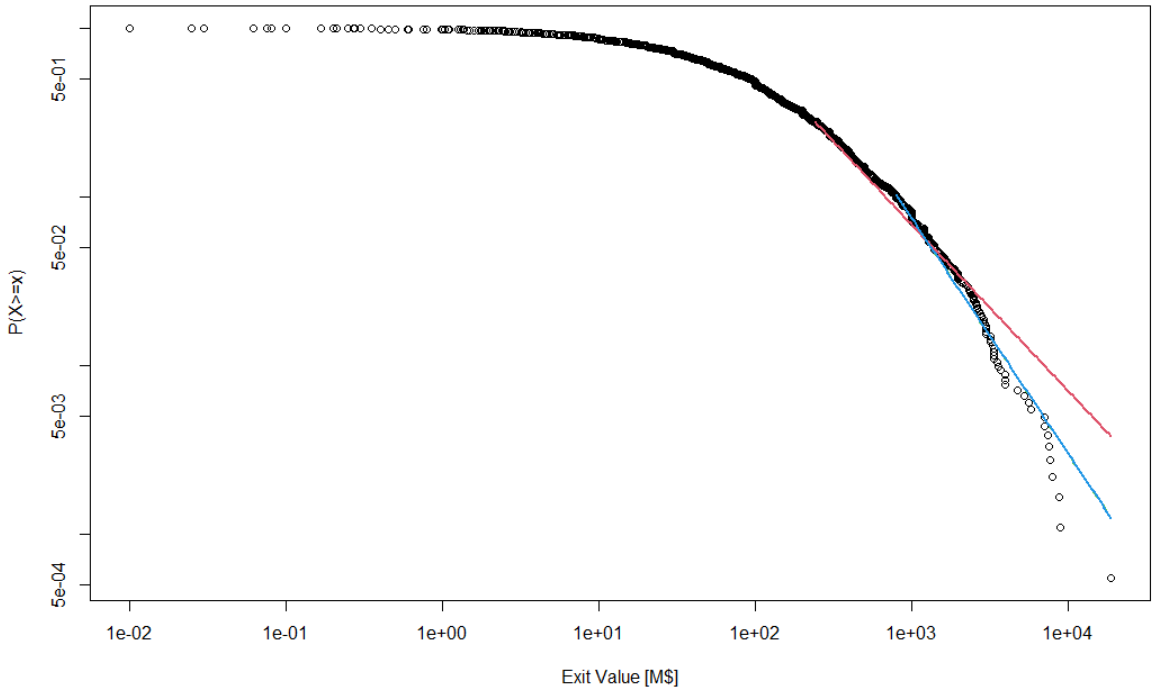

Figure 0.44 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from all regions.*

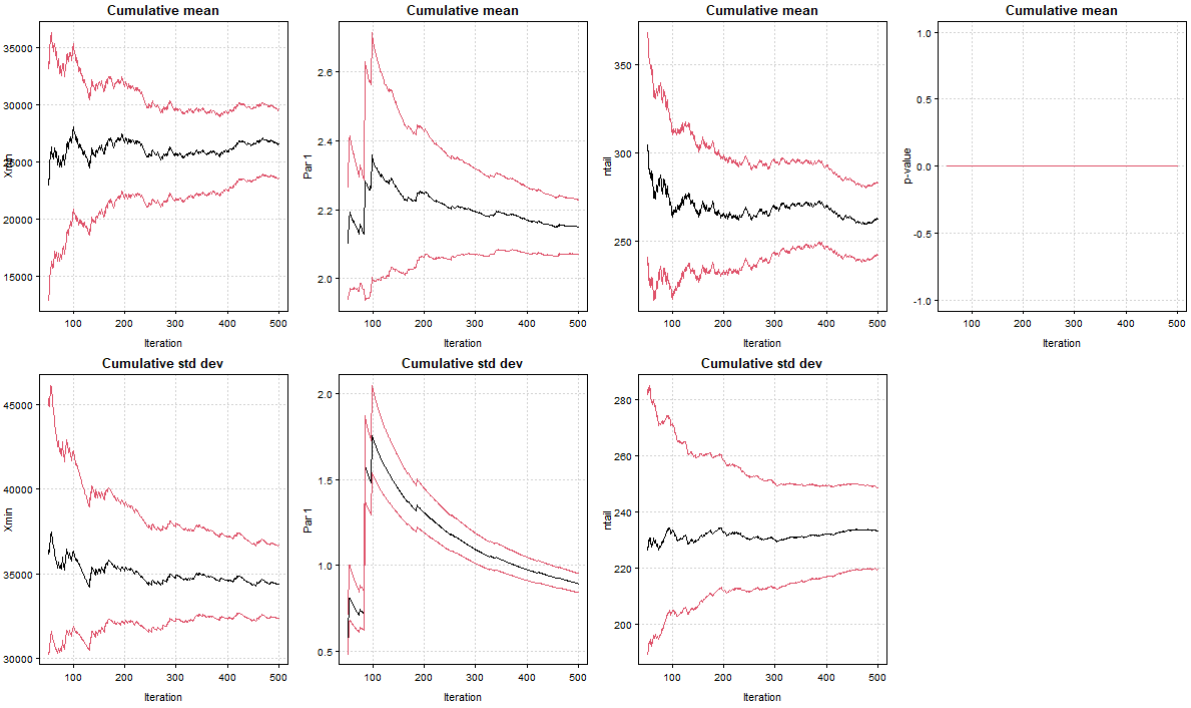

Figure 0.45 - *Bootstrapping results for the estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from all regions*

D.2.7 US region – US investor

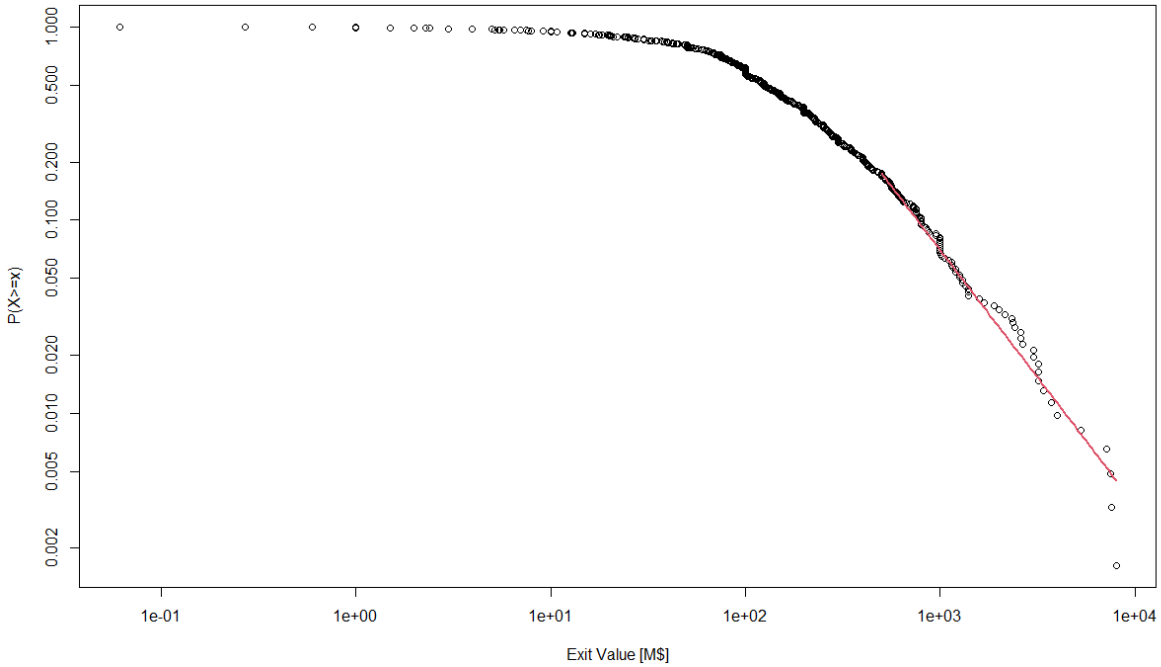

Figure 0.46 - *Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the US region.*

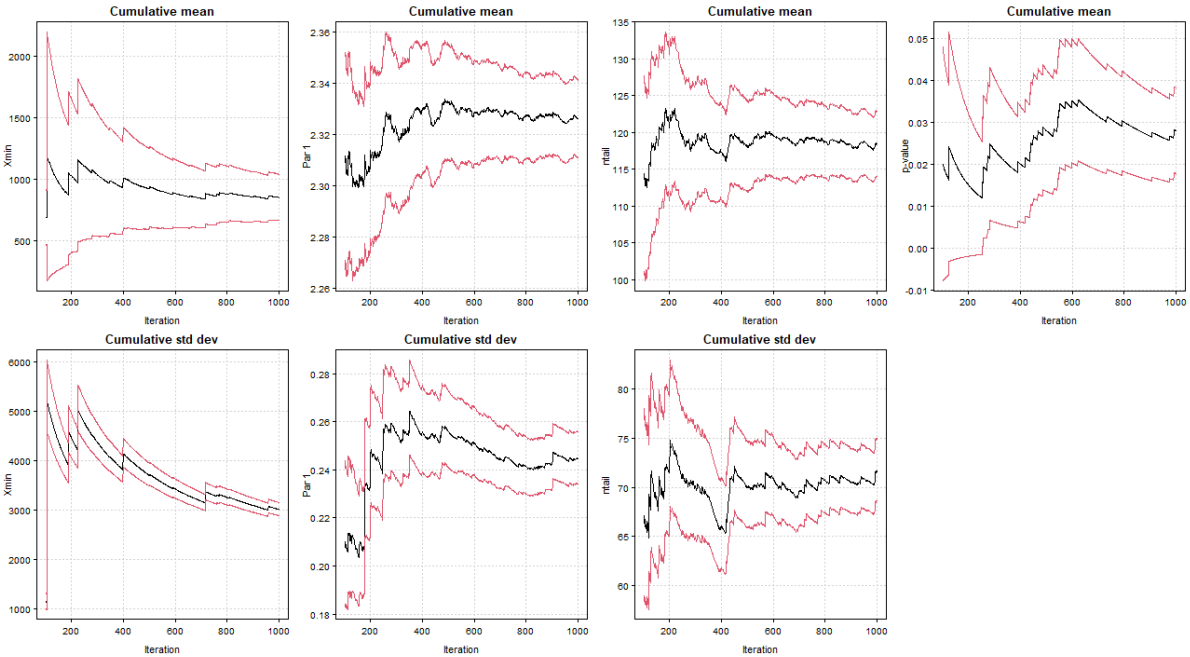

Figure 0.47 *– Bootstrapping results for the estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the US region.*

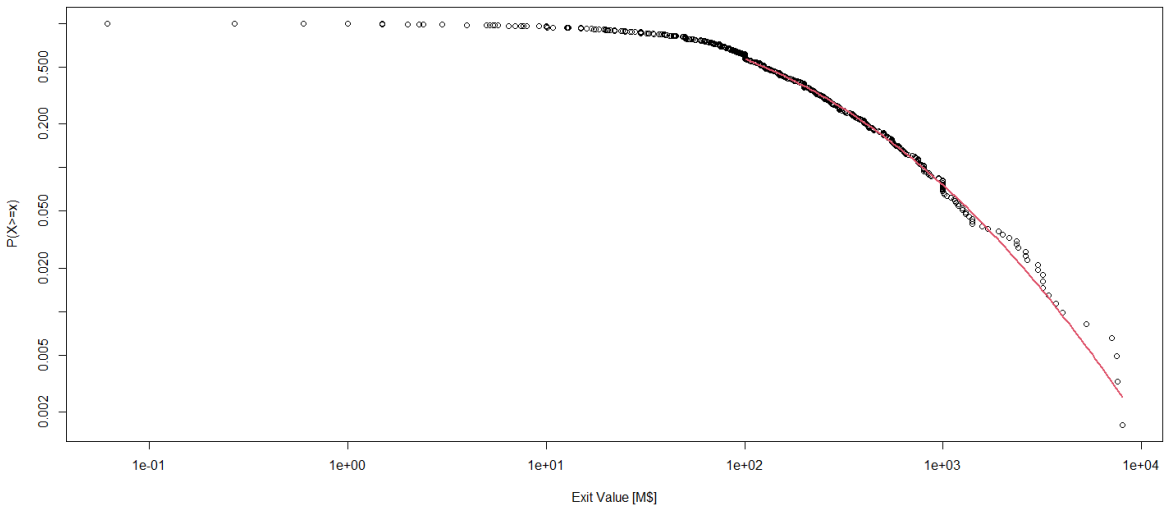

Figure 0.48 - *Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the US region.*

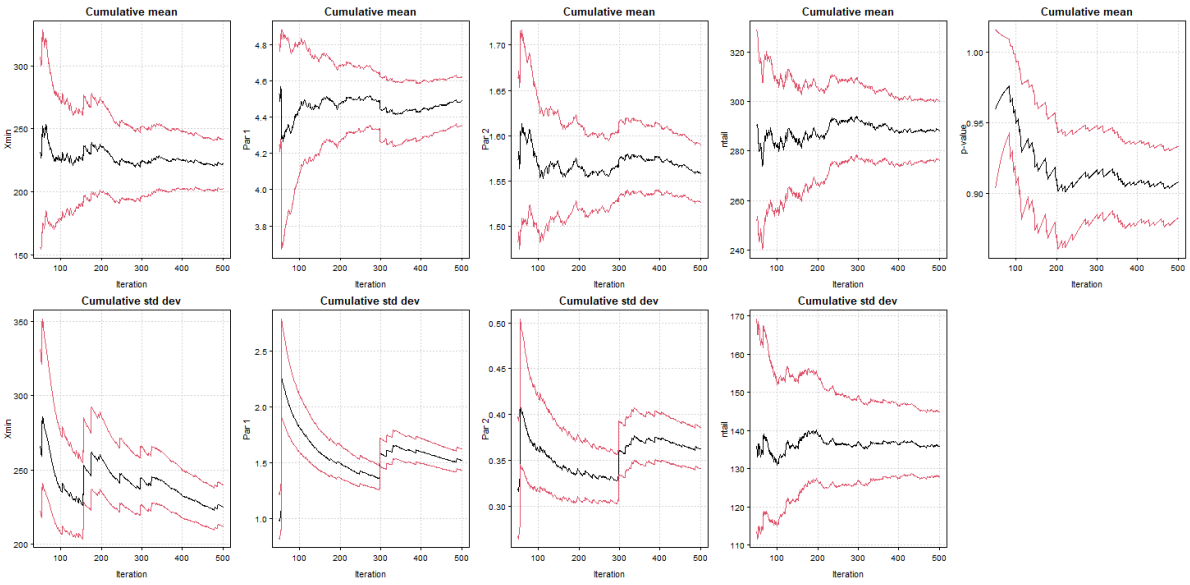

Figure 0.49 - *Bootstrapping results for the estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the US region.*

D.2.8 US region – EU investors

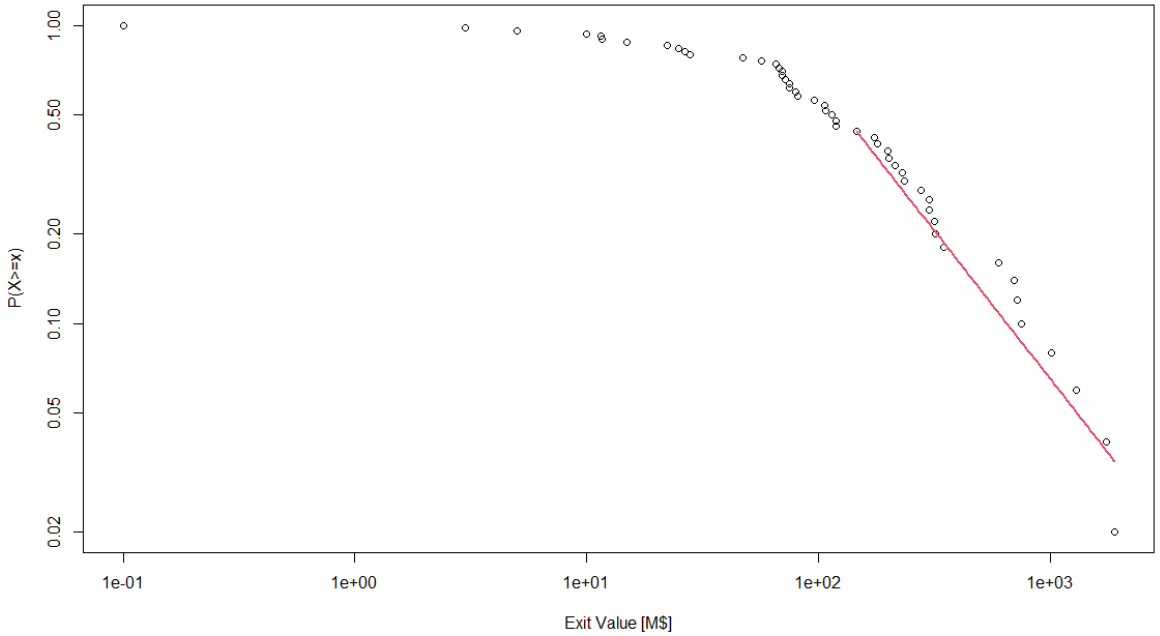

Figure 0.50 *- Estimated power-law distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the EU region.*

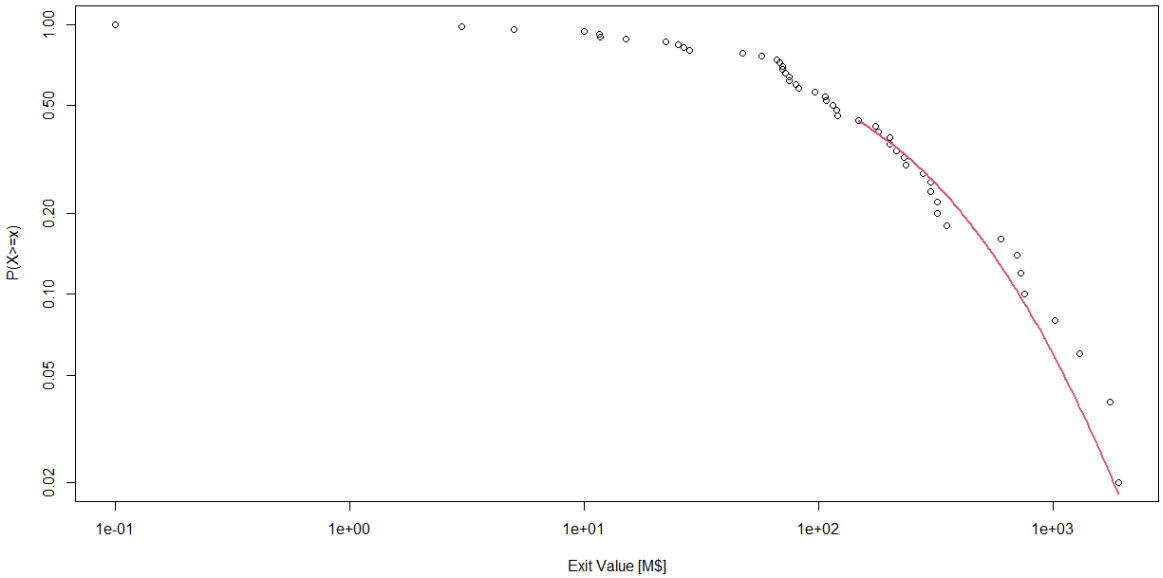

Figure 0.51 *- Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the EU region.*

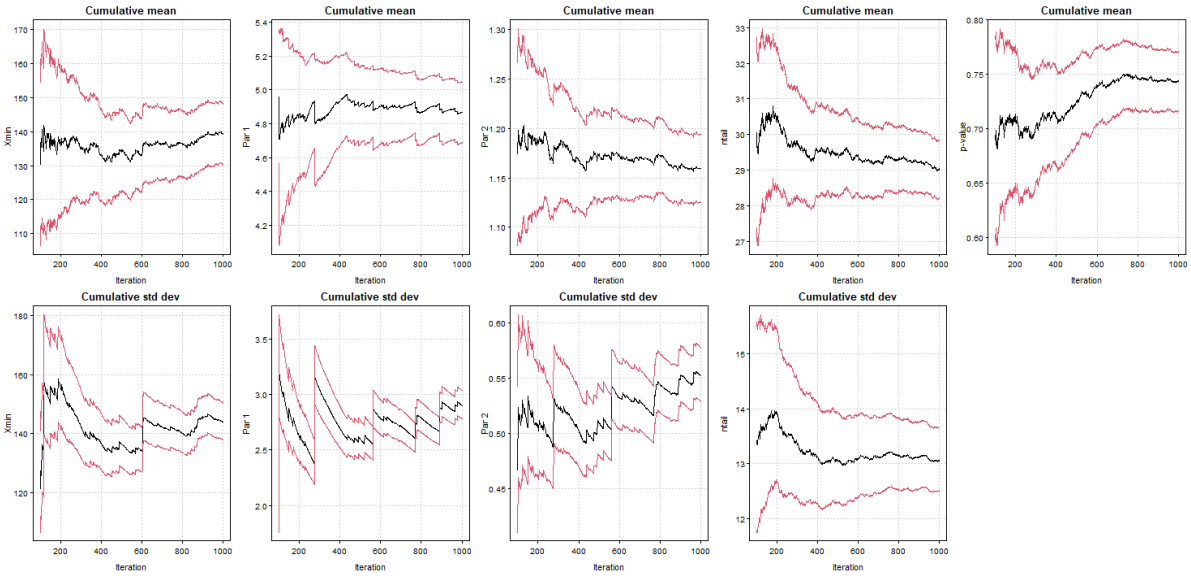

Figure 0.52 - *Bootstrapping results for the Estimated log-normal distribution of startups' exit values. Sample composition: startups with headquarters in the US region and investors from the EU region.*

D.3 MOICs

Table Appendix.0.4 - *Results of the MOICs distribution estimates*

Sample [Region_Investors]	Sample Dim	Distribution	PL: xmin [M$]	PL: alpha	LN: par1	LN: par2	gof	p	p no_of_sims	ntail	Fit?
All	523	Power-law	7.91	2.226264			0.0497996	0.12	500	188	Yes
All	523	Log Norm	7.91		-5053.0513	64.41533			500	188	Bs Error
All	523	Log Norm Tot	0.15		1.507504	1.402034	0.04689326	0	500		No
All_InvEU	124	Power-law	7.70	2.329504			0.05261546	0.888	500	51	Yes
All_InvEU	124	Log Norm	7.70		-1489.7494	33.85728					Bs Error
All_InvEU	124	Log Norm Tot	0.10		1.726852	1.183847	0.04662785	0.254	500	122	Yes
All_InvUS	321	Power-law	5.85	2.222136			0.04929488	0.358	500	139	Yes
All_InvUS	321	Log Norm							500		Bs Error
All_InvUS	321	Log Norm Tot	5.71		-23.372586	4.710227	0.04068985			142	Bs Error
EU_InvAll	139	Power-law	7.70	2.186098			0.07660799	0.276	500	61	Yes
EU_InvAll	139	Log Norm	7.70		-61.255126	7.463675					Bs Error
EU_InvAll	139	Log Norm Tot	0.15		1.845628	1.218263	0.06452479	0.006	500	137	No
US_InvAll	384	Power-law	5.72	2.147509			0.06172525	0.034	500	0.034	No
US_InvAll	384	Log Norm	5.72		-2390.8307	45.80763					Bs Error
US_InvAll	384	Log Norm Tot	0.08		1.38588	1.444079	0.0545788	0	500	382	No
EU_InvEU	104	Power-law	7.70	2.261716			0.06160864	0.78	500	44	Yes
EU_InvEU	104	Log Norm	7.70		-1290.0701	32.40056					Bs Error
EU_InvEU	104	Log Norm Tot	2.59		0.9512624	1.5634706	0.05724673	0.166		82	Yes
US_InvUS	308	Power-law	5.85	2.227627			0.05104992	0.3	500	134	Yes
US_InvUS	308	Log Norm									Bs Error
US_InvUS	308	Log Norm Tot									Bs Error
EU_InvUS	13	Power-law	4.0540540	2.181163			0.09725677	0.978	500	8	Yes
EU_InvUS	13	Log Norm	4.0540540		-2.422694	2.318469	0.08133307	0.918	500	8	Yes

Sample [Region_Investors]	Sample Dim	Distribution	PL: xmin [M$]	PL: alpha	LN: par1	LN: par2	gof	p	p no_of_sims	ntail	Fit?
EU_InvUS	13	Log Norm Tot	4.0540540		-2.418144	2.31753	0.08133307	0.93	500	8	Yes
US_InvEU	20	Power-law	5.72	2.753184			0.1930749	0.13	500	10	Yes
US_InvEU	20	Log Norm	5.72		1.8996766	0.7187923	0.07559356	0.932	500	10	Yes
US_InvEU	20	Log Norm Tot	1.95		2.0660343	0.6951417	0.07559356	0.856	500	13	Yes
All_InvEUandUS		Power-law									Bs Error
All_InvEUandUS		Log Norm									Bs Error
All_InvEUandUS		Log Norm Tot									Bs Error
EU_InvEUandUS		Power-law									Bs Error
EU_InvEUandUS		Log Norm									Bs Error
EU_InvEUandUS		Log Norm Tot									Bs Error
US_InvEUandUS		Power-law									Bs Error
US_InvEUandUS		Log Norm									Bs Error
US_InvEUandUS		Log Norm Tot									Bs Error

D.3.1 All regions and all investors

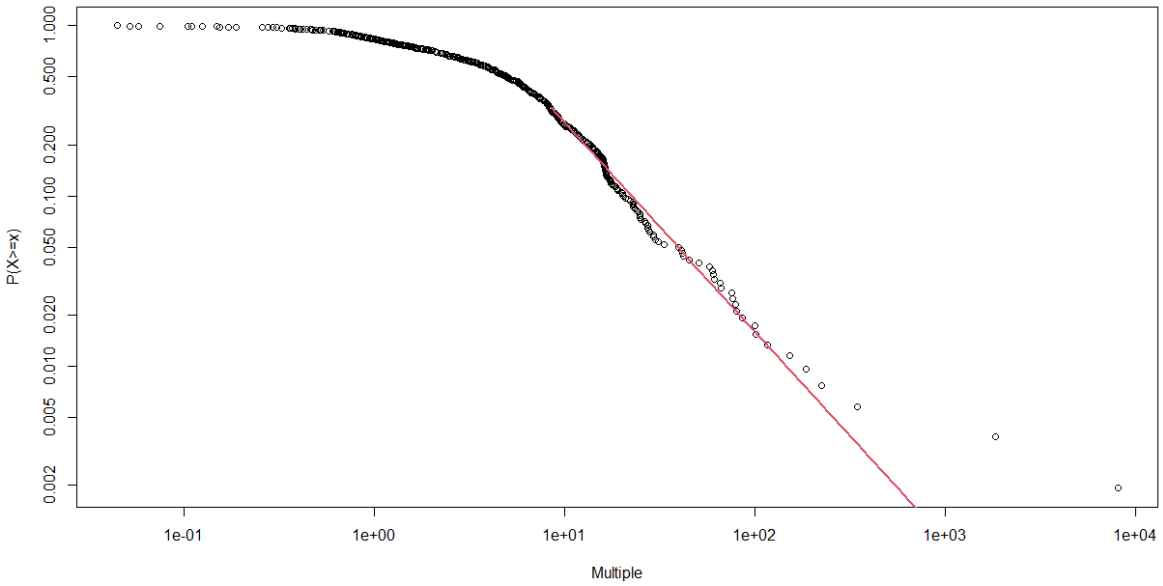

Figure 0.53 *- Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from all regions.*

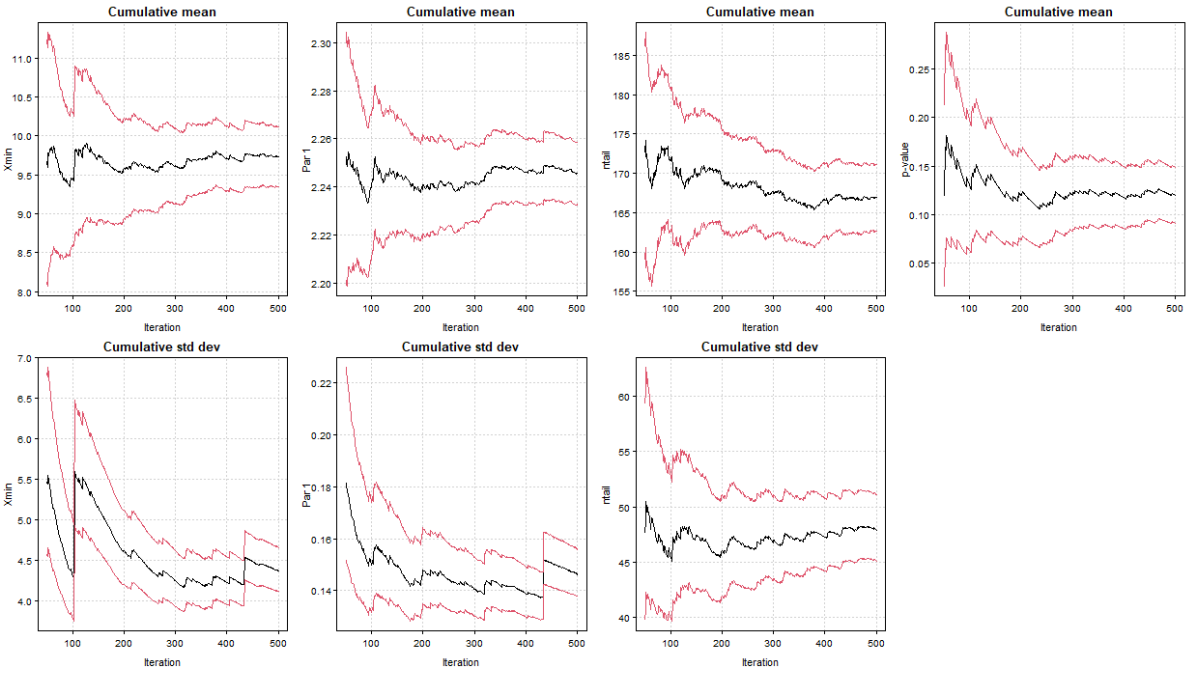

Figure 0.54 *– Bootstrapping results for the estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from all regions.*

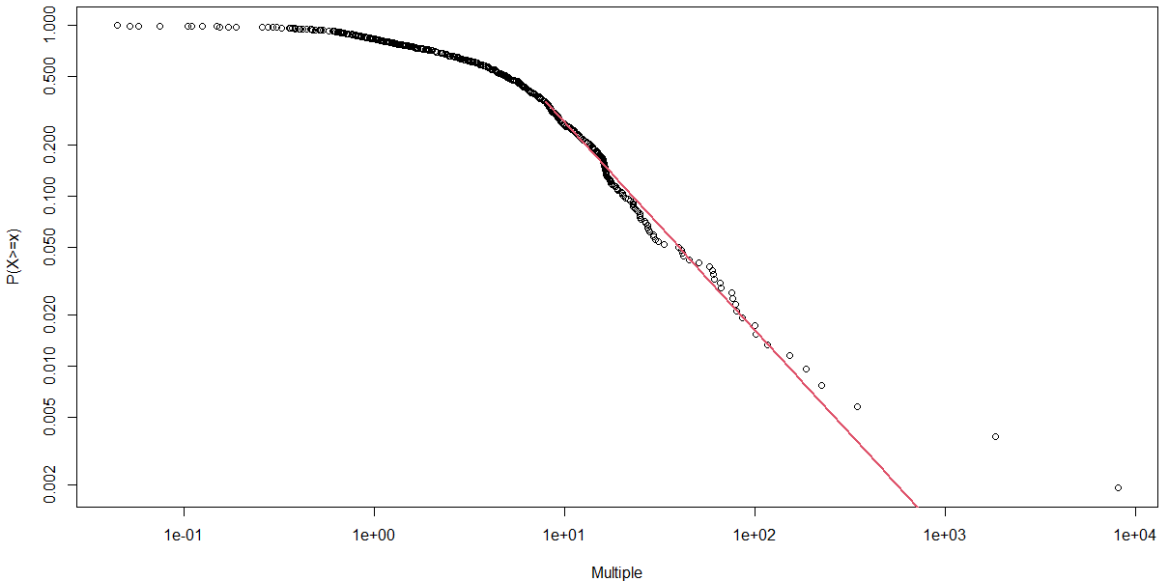

Figure 0.55 - *Estimated log-normal distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from all regions.*

D.3.2 All regions – EU investors

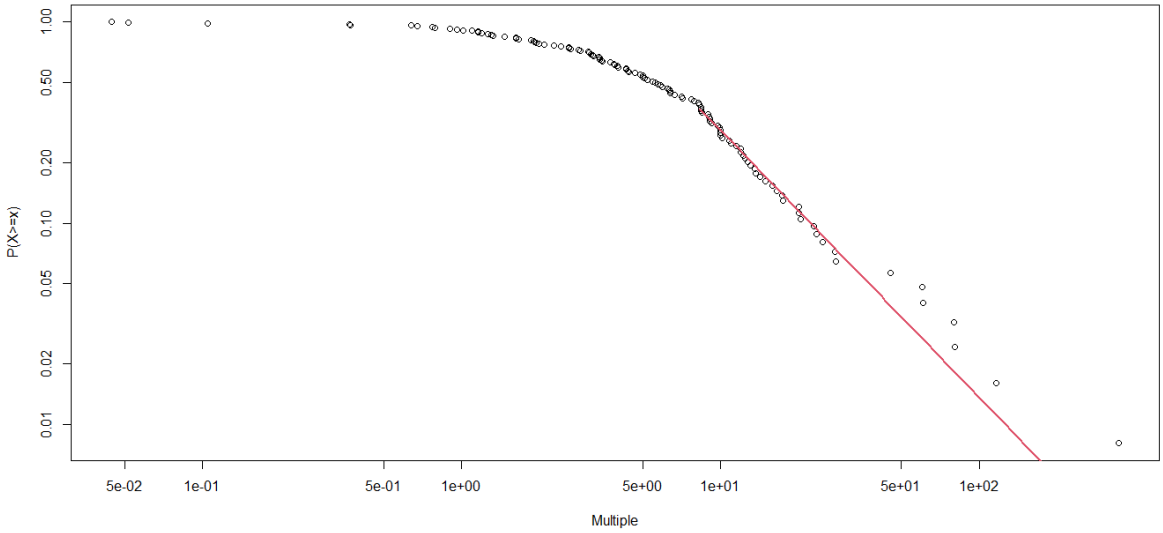

Figure 0.56 - *Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from the EU region.*

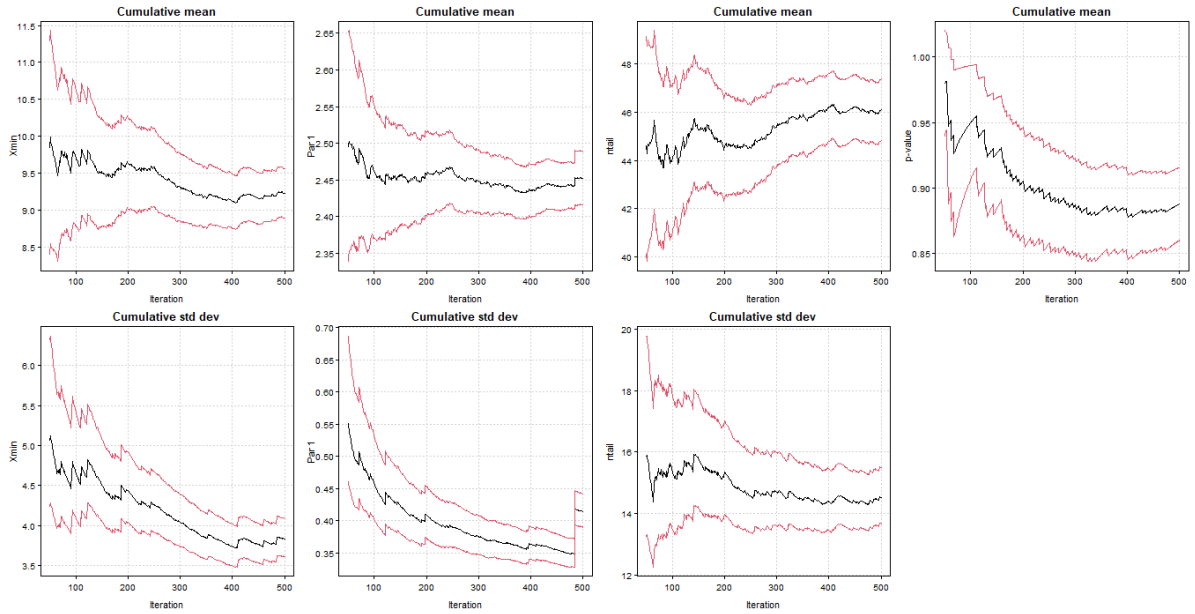

Figure 0.57 - *Bootstrapping results for the estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from the EU region.*

D.3.3 All regions – US Investors

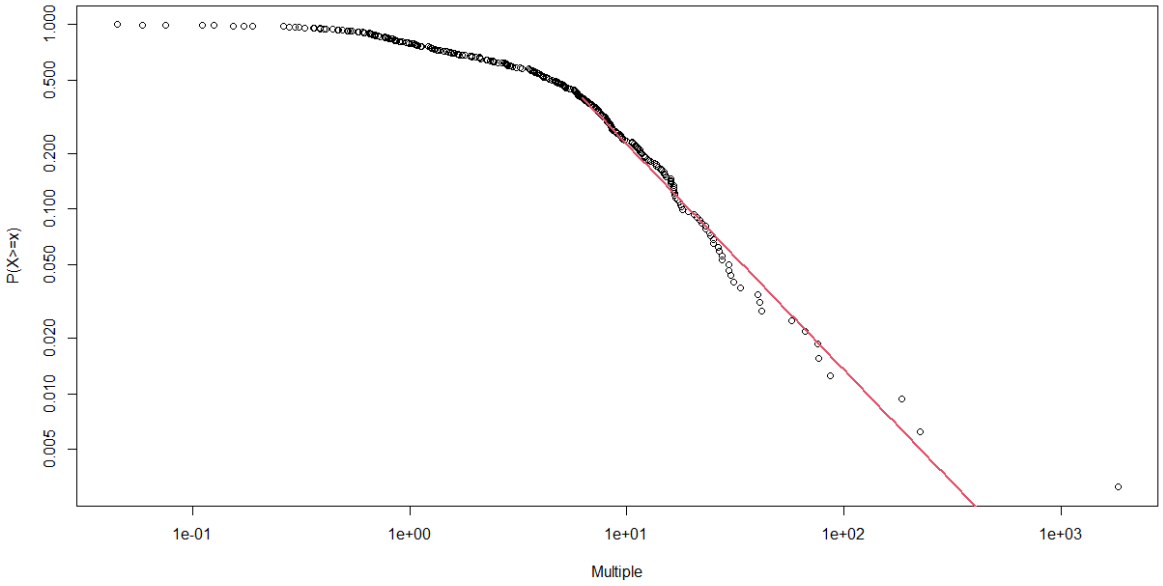

Figure 0.58 - *Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from the US region.*

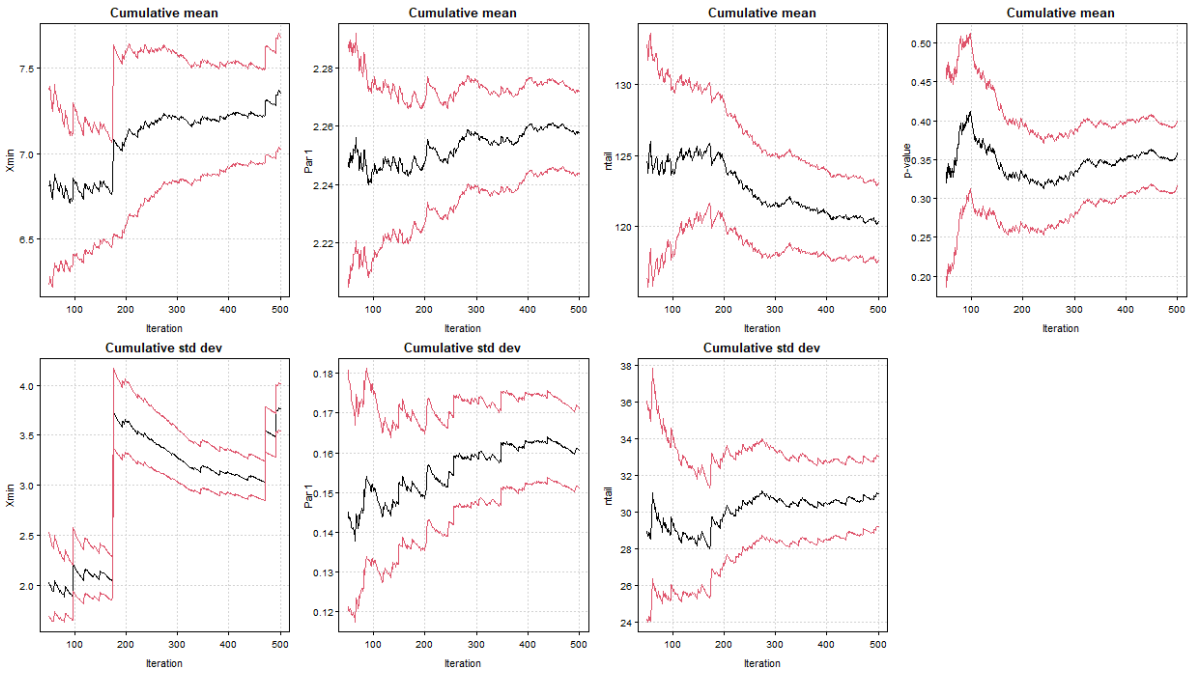

Figure 0.59 *– Bootstrapping results for the estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in all regions and investors from the US region.*

D.3.4 EU region – All investors

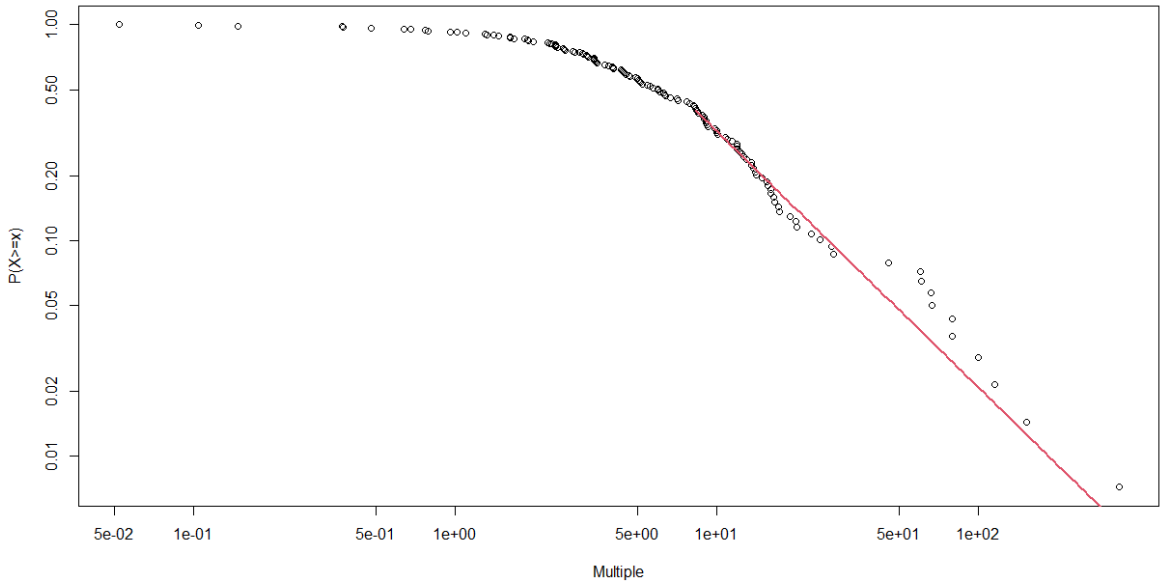

Figure 0.60 - *Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the EU region and investors from all regions.*

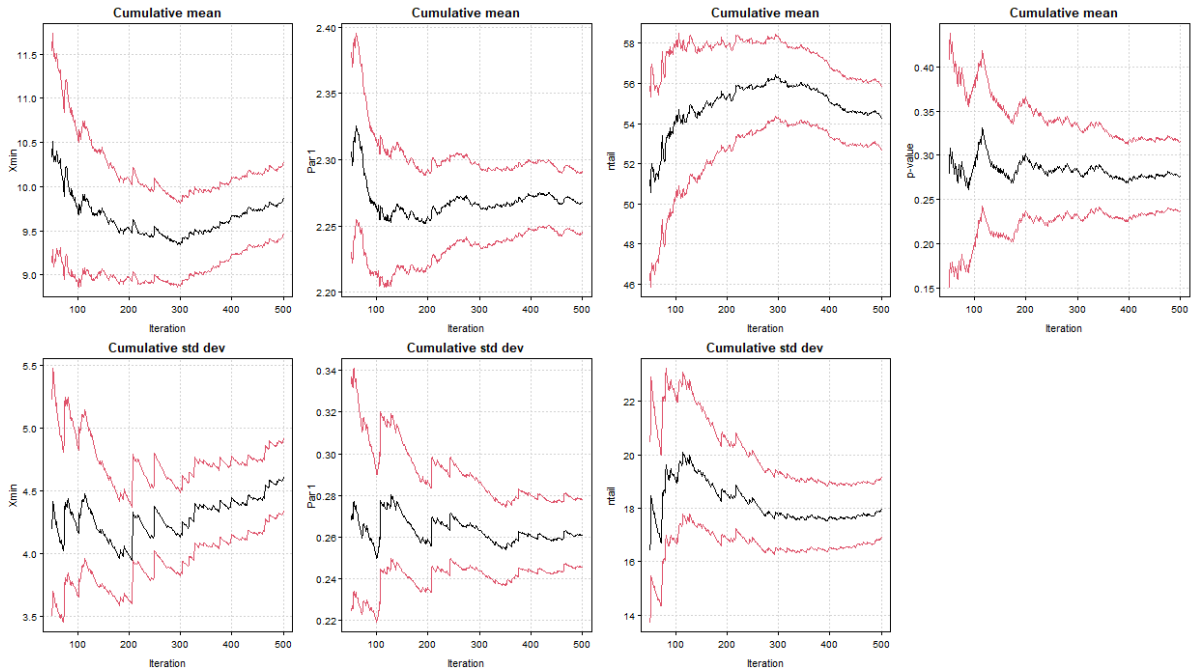

Figure 0.61 - *Bootstrapping results for the estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the EU region and investors from all regions.*

D.3.5 EU region – EU Investors

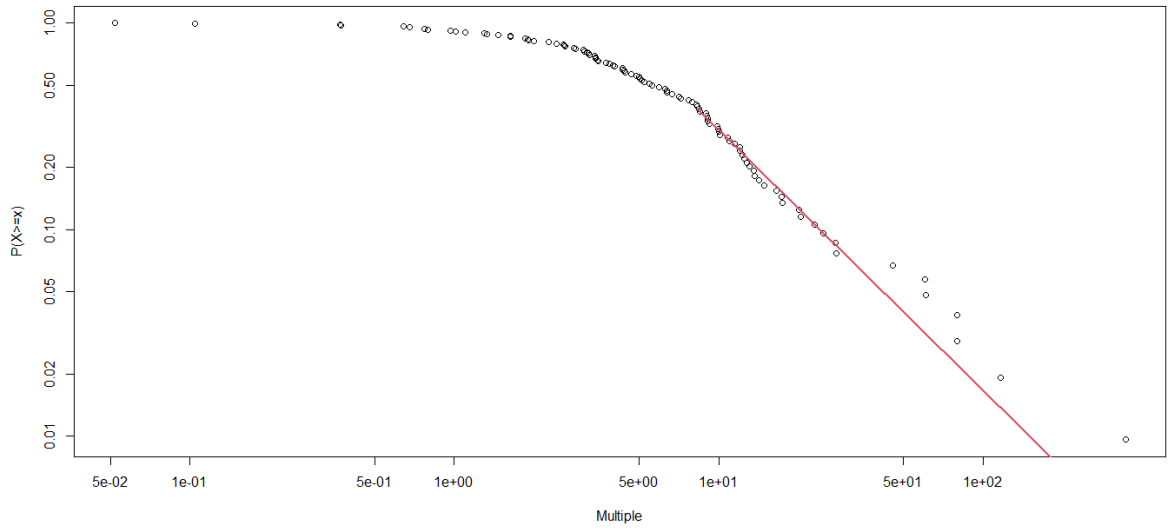

Figure 0.62 - *Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

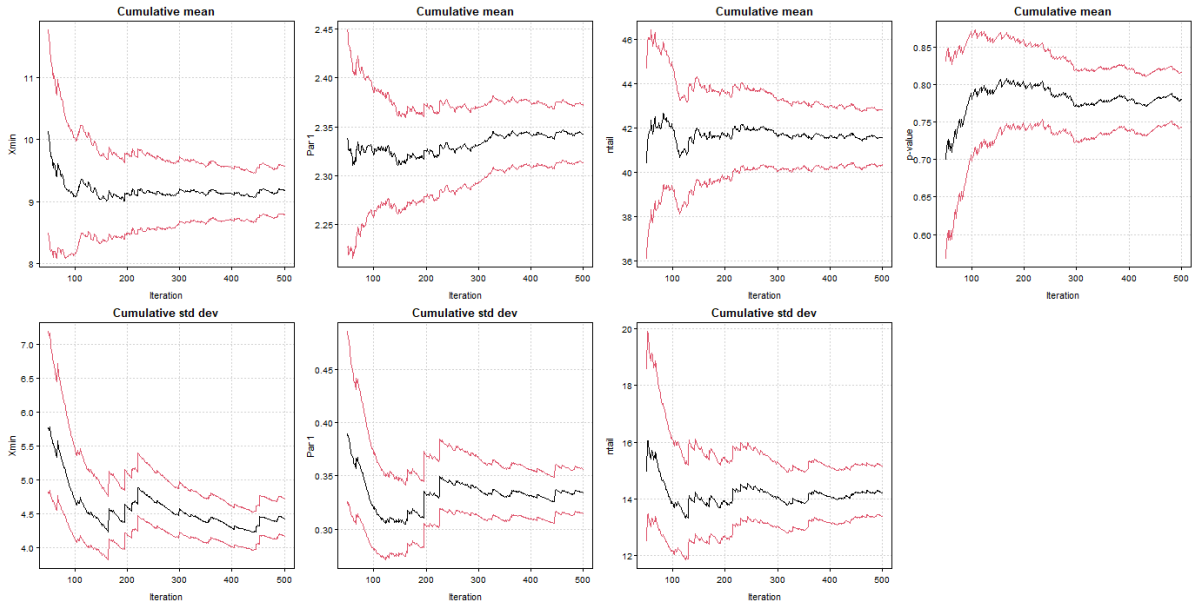

Figure 0.63 *- Bootstrapping results for the Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

D.3.6 US region – All investors

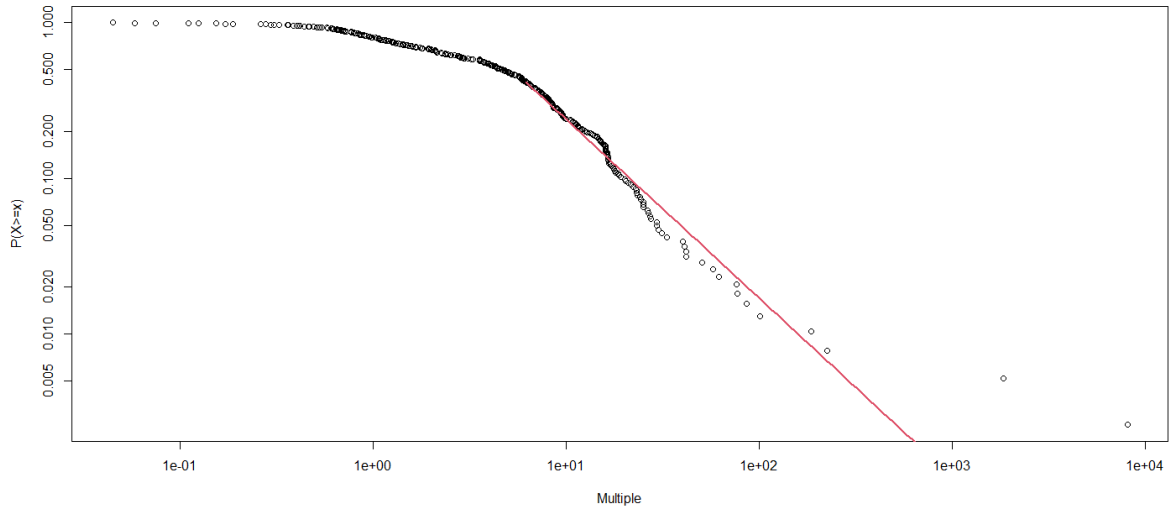

Figure 0.64 *- Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the EU region and investors from all regions.*

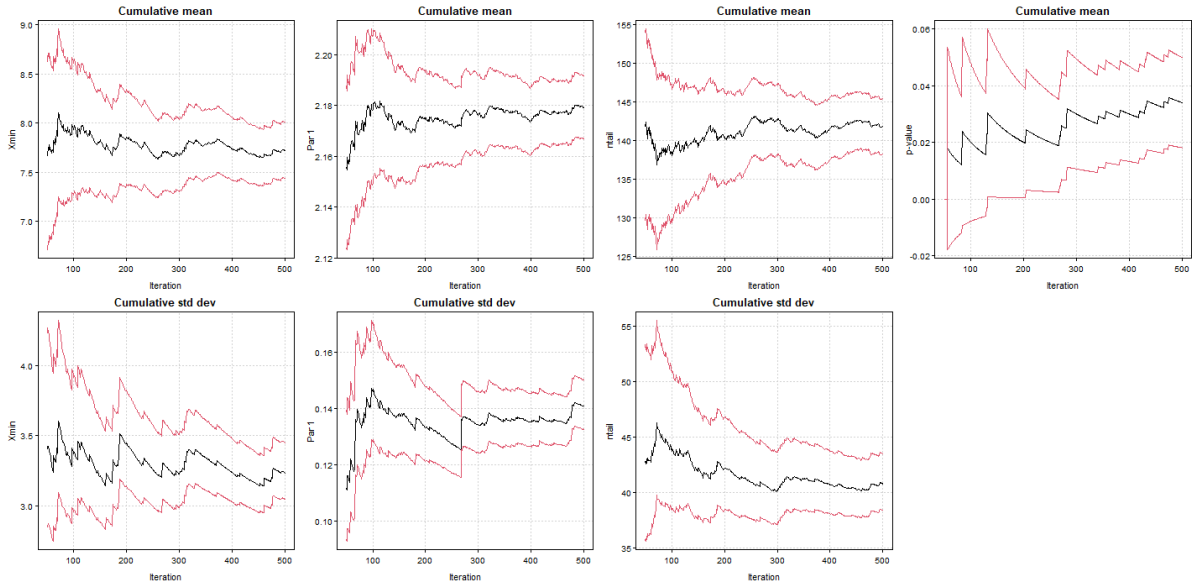

Figure 0.65 *- Bootstrapping results for the Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the EU region and investors from the EU region.*

D.3.7 US region – US investor

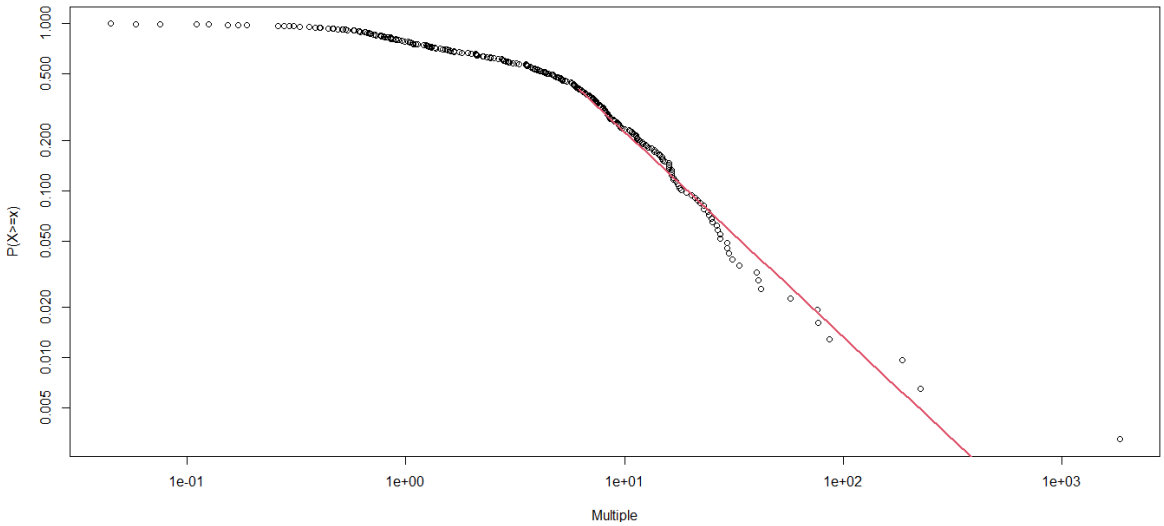

Figure 0.66 *- Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the US region and investors from the US region.*

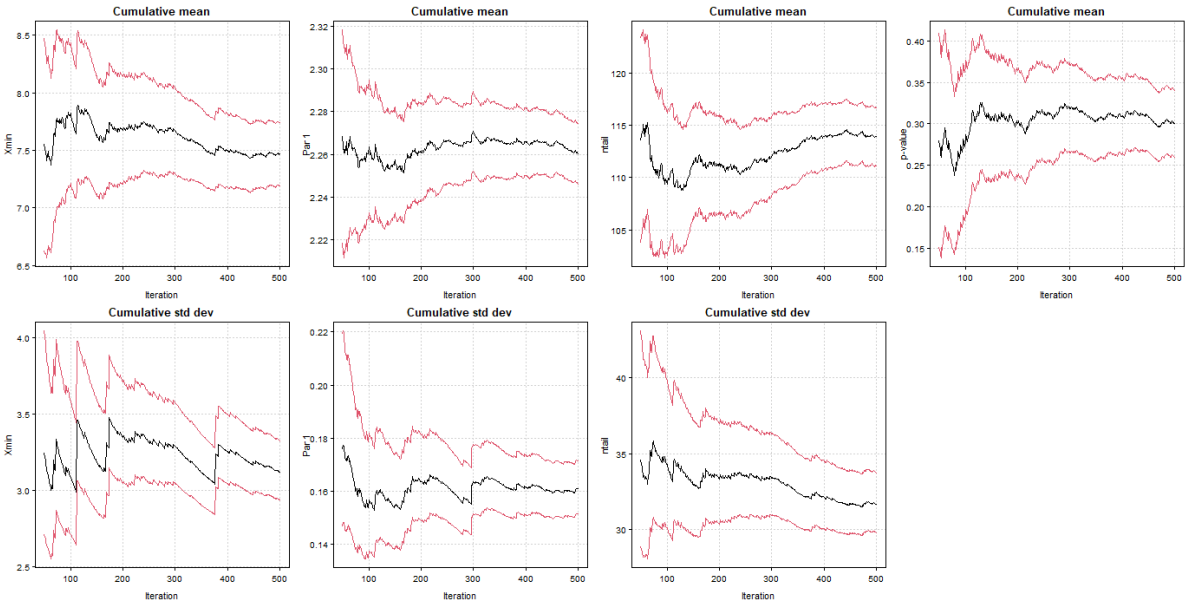

Figure 0.67 *– Bootstrapping results for the estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the US region and investors from the US region.*

D.3.8 US region – EU investors

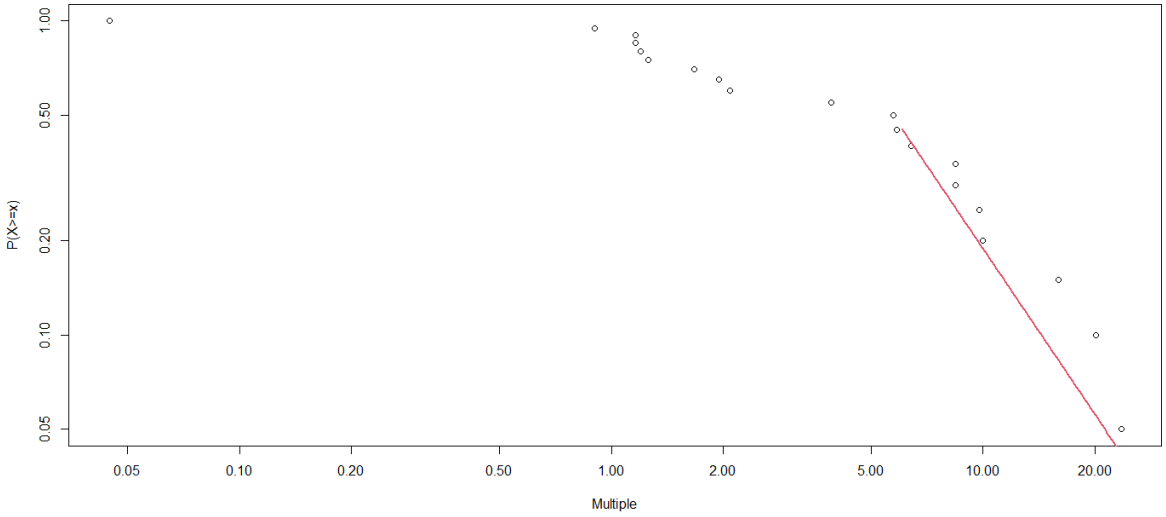

Figure 0.68 *- Estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the US region and investors from the EU region.*

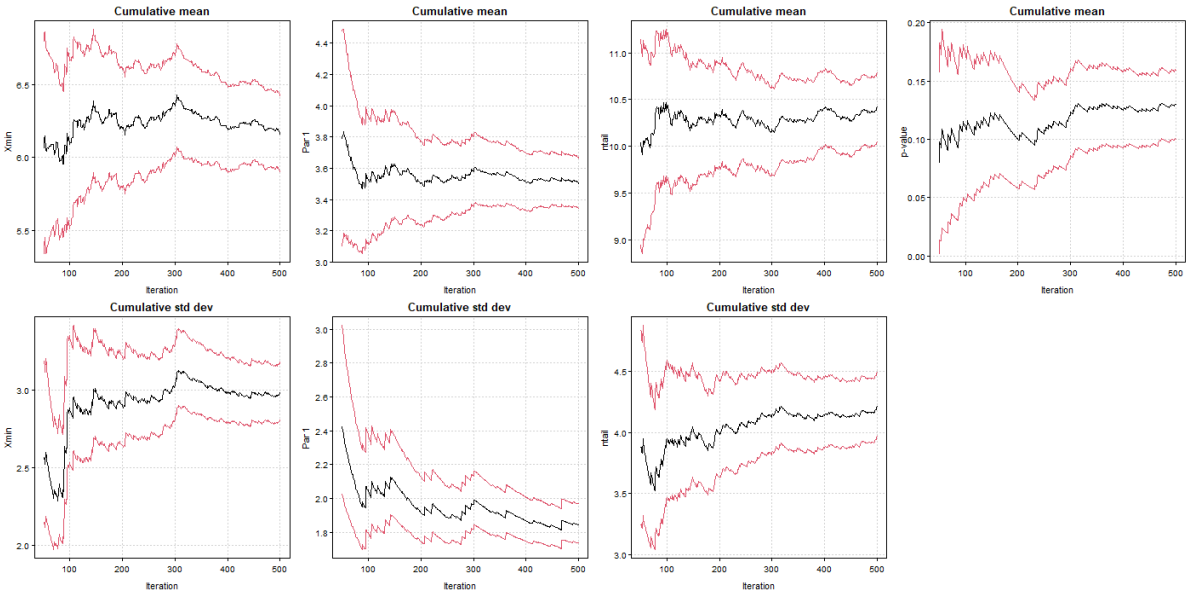

Figure 0.69 *- Bootstrapping results for the estimated power-law distribution of startups' MOICs. Sample composition: startups with headquarters in the US region and investors from the EU region.*

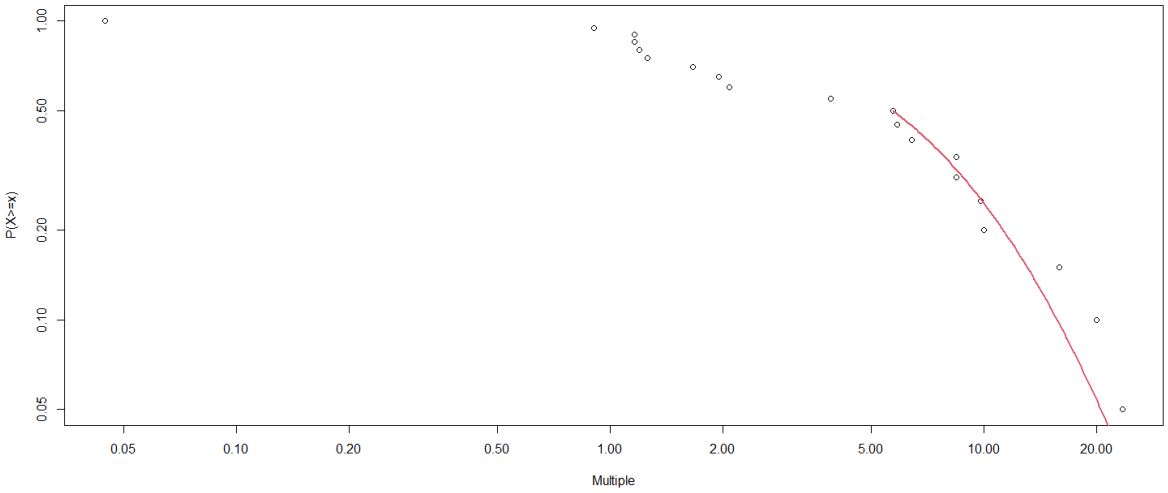

Figure 0.70 *- Estimated log-normal distribution of startups' MOICs. Sample composition: startups with headquarters in the US region and investors from the EU region.*

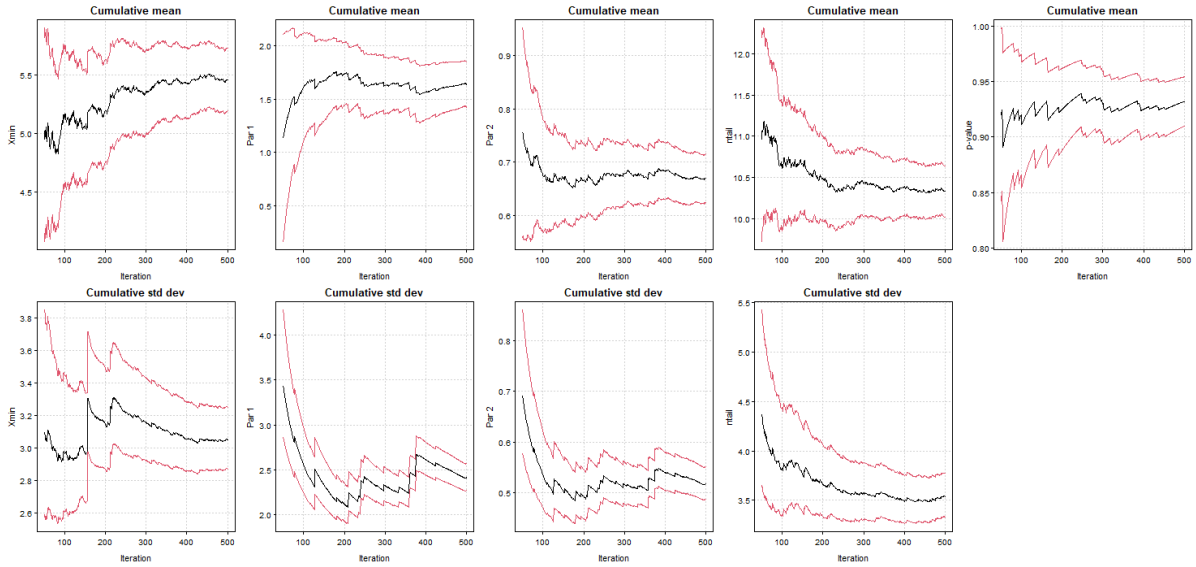

Figure 0.71 *- Bootstrapping results for the estimated log-normal distribution of startups' MOICs. Sample composition: startups with headquarters in the US region and investors from the EU region.*