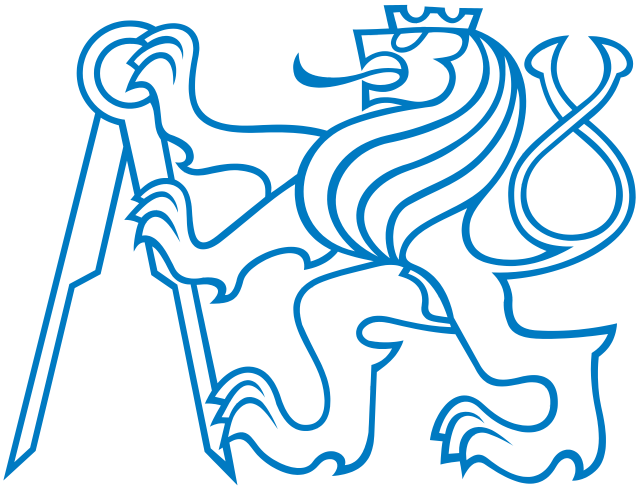
CZECH TECHNICAL UNIVERSITY IN PRAGUE

FACULTY OF ELECTRICAL ENGINEERING

DEPARTMENT OF CYBERNETICS



BACHELOR THESIS

Zipf's Law and Zeta Distribution

Vojtěch Adalbert Delong

Supervisor: Ing. Tomáš Kroupa, Ph.D.

Prague 2011

Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Cybernetics

BACHELOR PROJECT ASSIGNMENT

Student: Vojtěch D e l o n g

Study programme: Electrical Engineering and Information Technology

Specialisation: Cybernetics and Measurement

Title of Bachelor Project: Zipf´s Law and Zeta Distribution

Guidelines:

1. Acquaint yourself with Zipf's law and zeta distribution of a random variable.
2. On real data, using statistical methods judge the applicability of the zeta distribution for modeling phenomena often used in literature as representation of Zipf's law (e.g. words frequencies in a text, urban areas distributions).

Bibliography/Sources: Will be provided by the supervisor.

Bachelor Project Supervisor: Ing. Tomáš Kroupa, Ph.D.

Valid until: the end of the winter semester of academic year 2010/2011

L.S.

prof. Ing. Vladimír Mařík, DrSc.
Head of Department

doc. Ing. Boris Šimák, CSc.
Head

Prague, November 30, 2009

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: Vojtěch D e l o n g

Studijní program: Elektrotechnika a informatika (bakalářský), strukturovaný

Obor: Kybernetika a měření

Název tématu: Zipfův zákon a dzéta rozdělení

Pokyny pro vypracování:

1. Seznamte se s Zipfovým zákonem a rozdělením dzéta diskrétní náhodné veličiny.
2. Na reálných datech kriticky posuďte pomocí statistických technik vhodnost dzéta rozdělení pro modelování jevů, které jsou často uváděny v literatuře jako vyjádření Zipfova zákona (např. frekvence slov v textu, velikost měst).

Seznam odborné literatury: Dodá vedoucí práce.

Vedoucí bakalářské práce: Ing. Tomáš Kroupa, Ph.D.

Platnost zadání: do konce zimního semestru 2010/2011

prof. Ing. Vladimír Mařík, DrSc.
vedoucí katedry

doc. Ing. Boris Šimák, CSc.
děkan

V Praze dne 30. 11. 2009

Statement of originality

To the best of my knowledge and belief all of the material presented in this thesis is of my own original work with the use of the sources as referenced.

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW) uvedené v přiloženém seznamu.

In Prague 5.1.2011 Signature

Acknowledgements

I would like to thank my supervisor Tomáš Kroupa for his great support, extensive patience despite my numerous hospitalizations, the opportunity to work on such an interesting theoretical problem and the access to numerous valuable learning materials. I also would like to thank my lecturer Dmitriy Morozov for the lessons in discrete and Radon measures and also for his remarks and suggestions.

Abstract

Performing statistical hypothesis tests is an important method for analyzing real data sets occuring in various areas of the society. This thesis studies the properties of Zipf's law, its relationship to zeta distribution and the connection with heavy-tail phenomena occuring extensively in economics or demographics. Section 1 contains a short look at the problem in question and also a brief historical explanation of the origins of Zipf's law. In section 2, properties of several related terms are described and a method of the performed analysis is introduced. A practical part of this work is in section 3 where the analysis is performed and results presented. Section 4 summarizes the results for both theoretical and empirical part of the document.

Student's own contribution includes a consistent description of the theoretical part and also developing scripts for the analysis.

Abstrakt

Statistické testování nulových hypotéz je důležitou metodou v analýzách reálných dat, která se objevují v různých odvětvích společnosti. Tato práce studuje vlastnosti Zipfova zákona, jeho vztah k dzéta distribuci a také jeho spojení s fenomény tzv. těžkých konců, které se ve velké míře objevují v ekonomických a demografických studiích. Sekce 1 obsahuje krátký úvod do zmiňovaného problému a také lehký náhled na historickou příčinu počátků Zipfova zákona. V sekci 2 jsou popsány vlastnosti několika příbuzných termínů a je také uvedena metoda prováděné analýzy. Praktická část práce je obsažena v sekci 3, kde je provedena vlastní analýza a jsou předvedeny dílčí výsledky. Sekce 4 sumarizuje výsledky jak teoretické, tak empirické části tohoto dokumentu.

Vlastním příspěvkem studenta je konzistentní popis teoretické části problému a také vývoj programových skriptů pro vlastní analýzu.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Statistics in large complex systems

Often in real world we encounter situations where terms such as 'many', 'some of them' or 'average' are used. It does not matter whether we talk about cars in a parking place, people in a crowd or clouds. It is only logical to expect larger groups of subjects to show unpredicted behavior, not native to individual isolated subjects. For instance, charged particle in an eletrical field behaves in a completely different way than a quasineutral set of charged particles in the same field. Our experience gained by observing individuals no more applies and we have to look at the problem more 'macroscopically'.

Probability and statistics have proven valuable in concieving theories of 'many'. Concept of mean values gives us idea what is an average price of some product, variance suggests what varieties of age a certain society has. Advanced theories were presented by **Pierre Simon de Laplace** and later **Andrei N. Kolmogorov** [8]. One of the most known results of theories concerning probability is *probability distribution*. The idea allows us to decide how single experiments help construct properties of systems at larger scales and also has some interesting consequences (see section 2).

Thinking more specific examples, let us consider a distribution of urban areas in a country. The system (now meaning villages, towns and cities in the country, including e.g. how it develops) is affected by numerous factors, such as agricultural politics, urbanisation intesity etc. As individual, a town develops on the basis of fundings received from government budget, industrial income... simply put, a single town in some bordered area expands and increases the number of inhabitants according to the fundings up to the limitations of the borders. Several towns sharing a single region develop up to the point of interfering, then the model no longer works as the system gains completely new features (not mentioning the possibility of correlated amount of fundings received). It is the interference factor that often decides which town will be the largest, which will stagnate and to what point will the development converge.

Processing information about these systems using statistical methods can be used to analyze different systems with similar features to perform reasonable predictions about the behavior of the other of the same kind. For instance, computing a probability

distribution for a network data flux, considering size of files transfered, could help predicting the file size ratios transfered on a new untested network with similar topology. Designing of the transfer protocols should consider these expectations so as to avoid serious misconceptions, such as insufficient management of transfering large amounts of small and middle-sized files despite their statistical significance (e.g. cluster computing). Another possible point of view is determining whether the distributions are sensitive to certain factors. Experimenting on changing parameters of the system to see if it had some major effect on the distribution might lead to robust predictive system designs. Further in the text, we will focus on a specific phenomenons, originating in *Zipf's law* and developing into study of so-called *heavy-tailed* distributions. (see section 2).

1.2 Zipf's law origin

At halftime of the 20th century, american linguist **George Kingsley Zipf** with several of his students studied a number of texts for his quantitative linguistic analysis [15]. He used several texts written in Chinese, Yiddish and English (Shakespeare's language in particular, as he analyzed Hamlet). Zipf observed a strange phenomenon that should the core words in text be sorted by their frequency f into a decreasing sequence (with an index r), it occurs that from some r_0 on, the sequence roughly satisfies following equation [13]:

$$rf = \mathrm{c}, \quad r \geq r_0, \mathrm{c} \geq 0 \tag{1.1}$$

where c denotes some constant value, characteristic for the sequence.

Since the publishing of this observation, several studies of the same principle were performed, although using different kind of data - e.g. town distribution by the number of inhabitants or connection frequencies on the internet[1][5].

The particular form of the equation 1.1 restricts Zipf's law to simple harmonic series. Using logarithmic scale (see section 2) offers a natural generalization[13] in the exponents of the fraction

$$r^B f = \mathrm{c} \tag{1.2}$$

The result is a form of *power law*[14]. Further we will study its connection to *Pareto distribution* by performing an analysis on several presented data sets (see section 3).

For more specific conclusions about real systems, a specific method needs to be applied on real data. In this thesis, a method of choice is *QQ plotting* to decide whether the real data are distributed in certain manner (see subsection 2.6), namely Pareto[10].

1.3 Delimination

The presented problem is somewhat extensive, so we will focus on a certain part of it. Namely, In section 2 we first describe the problem theoretically and create a basis for an actual analysis of some data in section 3. Mathematical statistics is a difficult area of mathematics and often the answers may be questionable, unless built on a solid pedestal. What conclusions our performing might offer, we present in section 4.

We focus on Czech demographics both for the reason that the respective data analysis might be unique and to present the work in a familiar circumstances. So as the data be reliable, we must choose validable data sets possibly from the databases of the accredited institutes.

2 Theoretical framework

2.1 Needed fundamentals and definitions

Mathematical expressions used in this document have to fall under certain set of rules. As the notation is somewhat unclear and varies in different publications, a precedence notation, furthermore held, is included in appendix A.

Theory construct uses Kolmogorov probability model and respective definitions described thoroughly in [8]. Essential definitions are mentioned in this subsection, as the notation and terms are required to consistently formulate the goals and conclusions of this thesis.

2.1.1 Set essentials

Power set Given an arbitrary set M, a set of all subsets of M is called a *power set* over M and denoted 2^M. Any subset of 2^M is called a *family of subsets* over M.

σ-algebra A set $\mathcal{A}$ that is a subset of 2^M of some arbitrary set M is called a *σ-algebra over M*, if it has following properties:

1. $\emptyset \in \mathcal{A}$
2. $x \in \mathcal{A} \implies \bar{x} \in \mathcal{A}$, within M
3. $(\forall n \in \mathbb{N} : A_n \in \mathcal{A}) \implies \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$

Consider some $\mathcal{S} \subseteq 2^M$. The intersection of all σ-algebras over M that contain $\mathcal{S}$ is also a σ-algebra, is denoted $\mathfrak{U}(\mathcal{S})$ and is said to *be generated* by $\mathcal{S}$.

Bounded intervals on $\mathbb{R}$ A set of all bounded intervals on $\mathbb{R}$ with their finite unions together with an empty set generate a σ-algebra over $\mathbb{R}$[4]. This σ-algebra is denoted $\mathcal{B}(\mathbb{R})$ and called a *Borel σ-algebra* over $\mathbb{R}$.

2.1.2 Basic probability

Sample set Ω denotes set of all possible outcomes considered(e.g. towns in a country). It is called a *sample set* and it is often a set of mathematical representations of real objects, such as in dice throwing, it is the set of all six possible throw results.

Event set Any σ-algebra over Ω can be an *event set*, depending on the choice and context. Semantics are interpretable by designer, such as in the sample set of dice throw results, event set could be the σ-algebra generated by odd/even number events.

Probability measure Let $\mathcal{A}$ be an arbitrary event set over Ω. *Probability measure* is a real positive set function $P(a \in \mathcal{A})$ defined on $\mathcal{A}$ with a range $[0, 1]$ that satisfies following conditions:

$$P(\Omega) = 1 \tag{2.1}$$

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n), \text{ if } \{A_n\} \text{ is pairwise disjoint} \tag{2.2}$$

The second property is called σ-additivity

Probability space The trio $(\Omega, \mathcal{A}, P)$, where $\mathcal{A}$ is some event set over Ω and P a probability measure defined on $\mathcal{A}$, is called a *probability space*.

2.1.3 Random experiments

Random variable A *random variable* on a σ-algebra $\mathcal{A}$ over Ω is a map $X : \Omega \to \mathbb{R}$, such that for every interval $I \subseteq \mathbb{R} : C = \{\omega \in \Omega; X(\omega) \in I\} \in \mathcal{A}$.

Random variable realization After performing a random experiment with an outcome $\omega \in \Omega$, the value $X(\omega)$ is called *random variable realization*. This notes the difference from random variable, as random variable is actually a function[8].

2.1.4 Distributions and inverse

Following definitions are essential to introducing the QQ plotting method, namely cumulative distribution function, because its inverse function defines quantile function used in the plots, and the very definition of the inverse function, as it needs to be defined also for functions that are not injective.

Probability distribution *Probability distribution* P_X of the random variable X is a function $P_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$ defined as

$$P_X(I) = P_X[X \in I] := P(\{\omega \in \Omega; X(\omega) \in I\}) \tag{2.3}$$

Cumulative distribution function Definitions of a distribution function are not unambiguous. We choose a definition of a *cumulative distribution function* as $F_X : \mathbb{R} \to [0, 1]$ from [8]:

$$F_X(u) = P[X \leq u] := P_X((-\infty, u]) \tag{2.4}$$

The definition implies several properties of F_X:

1. $F(x)$ is a nondecreasing function
2. $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$
3. F_X is right-continuous

Respective proofs are in [8] and are based on σ-additivity of probability measure and properties of measure-zero sets.

Inverse function The generally accepted definition of inverse functions is insufficient for the purposes of this text, as the functions of interest presented here do not meet the necessary condition for such inverse functions to exist, i.e. they might not be injective[1]. An alternative definition is as follows[10]: Let $H : \mathbb{R} \to (a, b)$ be a nondecreasing function on $\mathbb{R}$ with a range $-\infty \leq a < b \leq \infty$. Then we define inverse $H^{\leftarrow} : (a, b) \to \mathbb{R}$ of H as

$$H^{\leftarrow}(y) = \inf\{s \in \mathrm{dom}(H) : H(s) \geq y\} \tag{2.5}$$

Probability density function Given a random variable X, if there is a function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$\forall u \in \mathrm{dom}(F_X) : F_X(u) = \int_{-\infty}^{u} f_X(v)\mathrm{d}v \tag{2.6}$$

the cumulative distribution function F_X is *absolutely continuous*[8]. We call such f_x *probability density function.* If at least one exists, there is an infinite number of functions that meet the same condition(as the integral in 2.6 is a Lebesque integral)[2].

[1] Many basic mathematical analysis textbooks and calculus courses define inverse functions only for injective functions, for an example [6].

[2] It is the functions that differ on a set of measure zero[4]. Such property gives a certain level of freedom - of right/left continuity - in constructing probability density function of a Pareto distribution.

Discrete distribution Given a random variable X, if there is a countable subset A of $\mathbb{R}$ such that $P_X(X \in A) = 1$, we say that the probability distribution is *discrete*[8]. A still belongs to $\mathcal{B}(R)$ as every point in $\mathbb{R}$ is a bounded closed interval of zero length and $\mathcal{B}(R)$ is closed under countable unions.

Riemann zeta function Generally, Riemann zeta function is defined as $\zeta : \mathbb{C} \to \mathbb{C}$, such that [2]

$$\zeta(z) = \sum_{l=1}^{\infty} \frac{1}{l^z}, \Re(z) > 1 \tag{2.7}$$

Only the real part of the domain of ζ is relevant for purposes of this text, so we can restrict the function to $\zeta_{\mathbb{R}} : (1, \infty) \to \mathbb{R}$ as

$$\zeta_{\mathbb{R}}(x) = \sum_{l=1}^{\infty} \frac{1}{l^x}, x > 1 \tag{2.8}$$

Probability mass function If a random variable X has a discrete probability distribution, a real function $d_X : \mathbb{R} \to [0, 1)$ defined as

$$d_X(u) = P_X(X = u) \tag{2.9}$$

is called a *probability mass function*[8].

Zeta distribution For some $s \in (1, \infty)$, we define [9] *zeta distribution* as a function $f_s : \mathbb{N}^+ \to \mathbb{R}^+$ expressed as

$$f_s(k) = \frac{1}{\zeta_{\mathbb{R}}(s) k^s} \tag{2.10}$$

Zeta mass function Let X be a random variable with a discrete distribution such that its probability mass fuction is

$$d_X(u) = \begin{cases} f_s(u) & u \in \mathbb{N}^+ \\ 0 & otherwise \end{cases} \tag{2.11}$$

Then d_X is called *zeta mass function* and X is *zeta-distributed.*

Power-law Under the term *power-law*, we will understand function $p_d : \mathbb{R}^+ \to \mathbb{R}^+$, such that[9]

$$p_d(x) = \frac{\mathrm{K}}{x^\alpha} \qquad \alpha, K \in \mathbb{R}^+ \tag{2.12}$$

Pareto distribution We say that a random variable X is with a *Pareto distribution*, if its cumulative distribution function is (for certain positive parameters β, w, x_l)[10]:

$$F_X(x) = \begin{cases} 1 - wx^{-\beta} & x \geq x_l \\ 0 & x < x_l \end{cases} \tag{2.13}$$

2.2 Power-law, zeta distribution and zeta mass

Domain of the zeta distribution is $\mathbb{N}^+$. However, definitions of the cumulative distribution functions(CDF), probability density functions(PDF) and probability mass functions(PMF) are functions of a real variable. If the zeta distribution is to be studied in terms of these definitions, we first need to connect zeta to such functions somehow. In this subsection, we will try to find some similarities between the zeta distribution, zeta mass function and the power-law.

2.2.1 Power-law continuity

We now prove that power-law in the form 2.12 is continuous on $\mathbb{R}^+$. Obviously, it is defined on $\mathbb{R}^+$, so we only need to prove that limit of $p_d(x)$ in every $x_0 \in \mathbb{R}^+$ is equal to the function value. First, continuity of x^α:

$$\lim_{x \to x_0} x^\alpha = \lim_{x \to x_0} \mathrm{e}^{\alpha \ln x} \tag{2.14}$$

$\lim_{x \to x_0} \ln x = \ln x_0$[1] on $\mathbb{R}^+$ and $\lim_{y \to y_0} \mathrm{e}^y = \mathrm{e}^{y_0}$ on $\mathbb{R}$, we apply the chain rule for limits, therefore $\lim_{x \to x_0} x^\alpha = x_0^\alpha$. After the application of algebraic limit theorem, that yields $\lim_{x \to x_0} p_d(x) = p_d(x_0)$, so $p_d(x)$ is continuous on $\mathbb{R}^+$.

2.2.2 Normalization

For zeta distribution, the argument s is greater than 1 because of the fact that if and only if $s > 1$, the series in $\zeta_{\mathbb{R}}(s)$ converges and therefore $\zeta_{\mathbb{R}}(s)$ is defined[2]. Reason for its value in the definition of zeta distribution is that it normalizes the sum over k:

$$\sum_{k=1}^{\infty} f_s(k) = \sum_{k=1}^{\infty} \frac{1}{\zeta_{\mathbb{R}}(s) k^s} = \frac{1}{\zeta_{\mathbb{R}}(s)} \sum_{k=1}^{\infty} \frac{1}{k^s} = \frac{1}{\zeta_{\mathbb{R}}(s)} \zeta_{\mathbb{R}}(s) = 1 \tag{2.15}$$

Despite the fact that power-law is continuous and therefore Newton's integral exists on $\mathbb{R}^+$, normalization of the power-law is somewhat tricky. The antiderivative of general

[1] In equation 2.14, the identity is implied by domain equivalence of $\ln x$ and $p_d(x)$

$p_d(x)$ is depending on α and divides the primitive functions into two classes:

$$\int p_d(x)\mathrm{d}x = \int \frac{K}{x^\alpha}\mathrm{d}x = K\int \frac{1}{x^\alpha}\mathrm{d}x = \begin{cases} K\ln x + c & \alpha = 1 \\ K\frac{x^{-\alpha+1}}{(1-\alpha)} + c & \alpha \in \mathbb{R}^+\backslash\{1\} \end{cases} \tag{2.16}$$

In case $\alpha \in (0,1]$, for some $a \in \mathbb{R}^+$ the limit $\lim\limits_{b\to\infty}\int_a^b p_d(x)\mathrm{d}x = \infty$ and in all cases for some $b \in \mathbb{R}^+$ the limit $\lim\limits_{a\to 0+}\int_a^b p_d(x)\mathrm{d}x = \infty$. Both these divergences dissapear, when the condition $\alpha > 1$ is met and if $p_d(x)$ is zeroed on some punctured right neighborhood of $\dot{U}_{x_{\min}}(0+)$. This new map we now extend to $\mathbb{R}$ and shall denote $p(x) : \mathbb{R} \to [0,\infty)$:

$$p(x) = \begin{cases} \frac{K}{x^\alpha} & x \geq x_{\min} \\ 0 & x < x_{\min} \end{cases} \quad \alpha > 1, x_{\min} > 0 \tag{2.17}$$

The reasons why we did not define the power-law as $p(x)$ instead of 2.12 are several. First, power-law often occurs in physics (gravitational law, Coulomb's law) in the form of 2.12, $p(x)$ is only right continuous at $x = x_{\min}$ and its construction is convenient in connection with Pareto distribution(see subsection 2.3). In order to find a normalization constant, we compute the integral of $p(x)$:

$$S = \int_{-\infty}^{\infty}\frac{K}{x^\alpha}\mathrm{d}x = \int_{x_{\min}}^{\infty}\frac{K}{x^\alpha}\mathrm{d}x = K\left[\frac{1}{(1-\alpha)x^{\alpha-1}}\right]_{x_{\min}}^{\infty} = \frac{K}{(\alpha-1)x_{\min}^{\alpha-1}} \tag{2.18}$$

Because the value of the integral is a positive number, we can normalize $p(x)$ to

$$p_n(x) = \frac{p(x)}{S} = \frac{K}{x^\alpha}\frac{(\alpha-1)x_{\min}^{\alpha-1}}{K} = \frac{\alpha-1}{x_{\min}}\left(\frac{x_{\min}}{x}\right)^\alpha \tag{2.19}$$

Normalization of the zeta mass function in the sense of the Lebesque integral is impossible(see below). The alternative is the normalization of the sum of the values over $\mathrm{supp}(d_X)$, however, that coincides with the normalization of the zeta distribution and needs not to be discussed again.

2.2.3 Mass functions and zeta mass

It should be noted that all probability mass functions have at most countable support. All probability mass functions then differ from a zero function on a set of measure zero, so the integral in eq. 2.6 would be therefore zero $\forall u \in \mathbb{R}$. No probability mass function can form a cumulative distribution function in the sense of a common Lebesque integral, another procedure is required.[8].

Comulative distribution function implied by a probability mass function will be constructed as follows: Let $T = \mathrm{supp}(\mathrm{d_X})$, the elements indexed as T_i. We assign the values of the respective CDF using the equation

$$\forall u \in T : F_X(u) = \sum_{T_i \leq u} d_X(T_i), T_i \in T, u \in \mathbb{R} \tag{2.20}$$

Proof of this relation clears from the σ-additivity of the probability measure, as the function values of d_X are the probability distribution values at pairwise disjoint singletons in $\mathbb{R}$ and the sum can be identified with the sum in the eq. 2.2 with $d_X(T_i) = P(A_n) = P(\{\omega \in \Omega; X(\omega) = T_i\})$. F_X is right continuous and also constant at the other points, which follows the CDF properties.

Zeta mass function is a special case of the above and following eq. 2.20, we obtain

$$\forall k \in \mathbb{N}^+ : F_X(k) = \sum_{l=1}^{k} d_X(l) = \sum_{l=1}^{k} f_s(l) = \sum_{l=1}^{k} \frac{1}{\zeta_{\mathbb{R}}(s) l^s} \tag{2.21}$$

And for other u leaving $F_X(u)$ constant with regard to the right-continuity of F_X.

2.2.4 Power-law and zeta distributed data relationship

Comparing definitions 2.10 and 2.12, we see that zeta distribution f_s is a restriction map of power law $p_d(x)$ with parameters $\alpha = s$ and $K = \zeta_{\mathbb{R}}^{-1}(s)$. From mathematical point of view, only this much is to say, although with the implications of the fact.

In order to be able to discuss correlation between zeta distribution and power-law, the semantics of the problem are essential. Respective cumulative distribution functions of both zeta and power-law distributions are key to presented data analysis, so we will focus on comparing them. Cumulative distribution function(CDF) F_X(u) of the zeta mass function is a stairs-like function, rising by $f_s(k)$ at each $u = k, k \in \mathbb{N}^+$. For computation, the equation 2.21 only needs to be followed. On any interval $[k, k+1)$ it is constant. We can approximately illustrate the comparison on cumulative distribution functions of series $\frac{1}{k^2}, k \in \mathbb{N}^+$ and its extension $\frac{1}{x^2}, x \geq 1$. They are shown in fig. 2.1 (regard the right-continuity).

Clearly the discrete case is pointwise greater than CDF for the continuous extension. We now illustrate that for such extension of the discrete distribution, this is always so.

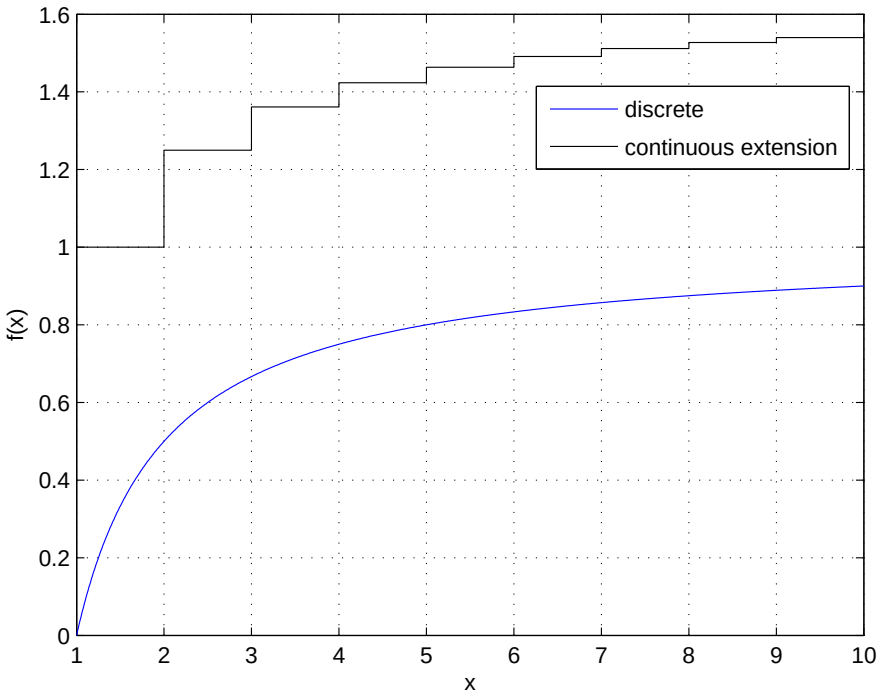


Figure 2.1: Unnormalized CDF of zeta/power-law for $\alpha = 2$

Figure 2.2 shows how CDF increase in both cases. Continuous CDF increases by a value equal to area under the curve $\frac{1}{x^2}$ on every interval $[k, k+1], k \in \mathbb{N}^+$. That is equal to integral (for general $\alpha > 1$)

$$\Delta_p(k) = \int_k^{k+1} \frac{1}{x^\alpha} \mathrm{d}x = \left[\frac{1}{(1-\alpha)x^{\alpha-1}} \right]_k^{k+1} \tag{2.22}$$

Discrete CDF increases at k by an area pictured by the red rectangle in fig. 2.2 – on some interval $[k, k+1)$ – and remains constant to the point $k+1$(proceeding iteratively). Speaking in calculus terms, discrete construction of a CDF is equivalent to some upper Darboux sum of the continuous extension with steps equal to the 'distance' of 1.

Thus we have concluded that CDF for a zeta distribution and power-law as its extension are not generally equal. Additionally, they are not equal at any point within the intersected support of both $f_s(k)$ and $p_d(x)$. Testing data on power-law distributions thus does not imply that data are zeta-distributed. However, looking at the approximate differentials of both graphes in fig 2.1, for relatively large x one asks what is the difference between them. Both can be approximated with the corresponding

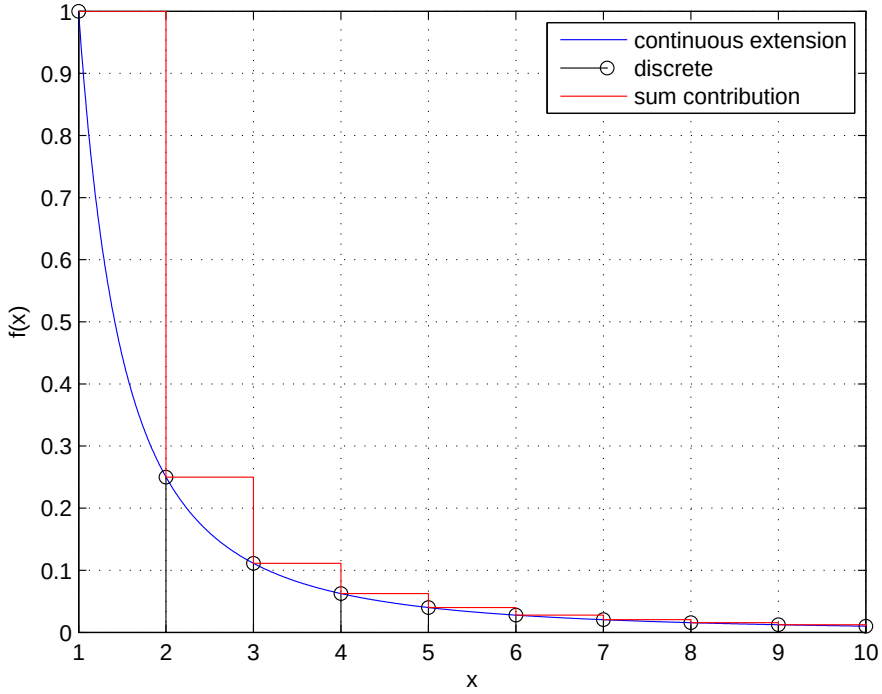


Figure 2.2: CDF increase of zeta/power-law for $\alpha = 2$

increases of discrete and continuous CDFs on intervals of the length 1. We compute the difference of increase for discrete and continuous case on general interval $[k, k+1)$. The increase for the discrete case is

$$\Delta_\zeta(k) = \frac{1}{k^\alpha} \tag{2.23}$$

That is the area of a rectangle in the upper Darboux sum. Difference between the increases is then

$$I(k) = \Delta_\zeta(k) - \Delta_p(k) = \frac{(k+1)^{\alpha-1}(1-k^{-1}(\alpha-1)) - k^{\alpha-1}}{(1-\alpha)k^{\alpha-1}(k+1)^{\alpha-1}} \tag{2.24}$$

For $k \gg 1$, the expression yields

$$I(k \gg 1) \approx \frac{1}{k^\alpha} \tag{2.25}$$

This directly implies that both CDFs do not equal on the intersected support of the respective functions. One could argue that neither of the functions are normalized, however presented derivations only illustrate the fact that generally these cumulative distribution functions are not equivalent. However, they converge to the same 'tail'.

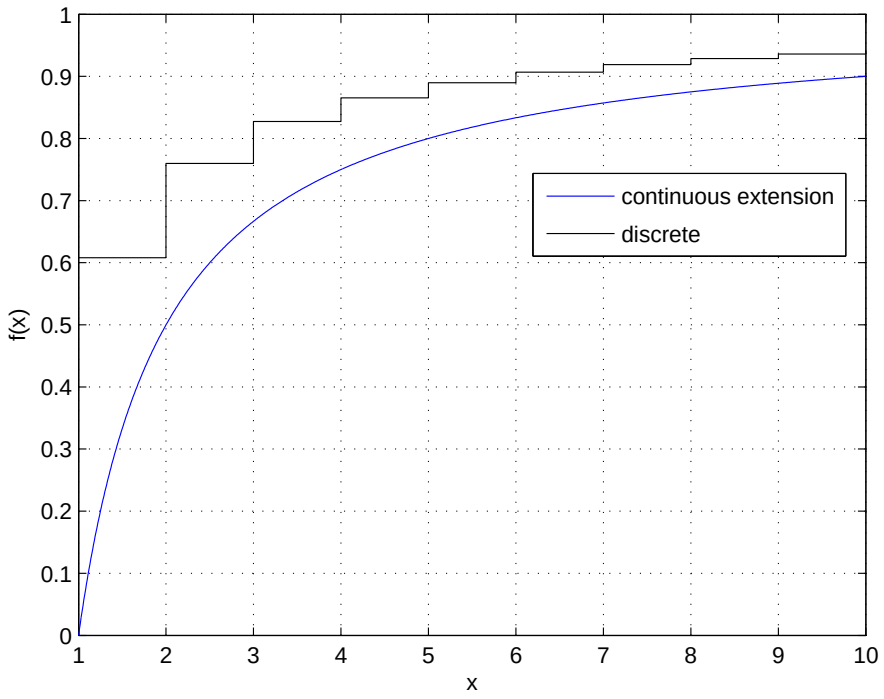


Figure 2.3: CDF normalized for $\alpha = 2$

It will be seen that testing the data positively on power-law tail(i.e. the test results in positive for $X > X_0$ for some $X_0 > 0$) implies that the data are also zeta-distributed there. This is a clue how we can reach a conclusive statement about the character of the distribution of some real data. We say that real data are zeta distributed on the tail, if the respective CDF 'follows' power-law CDF *from some point*[10][13]. More precise explanation of the term 'real data' and its connection to the theory is in subsection 2.5.

2.2.5 Power-law median

Concerning the integral of $p_n(x)$, we may simply estimate the median. It will be such $M_X > x_{min}$ that is a solution of an equation

$$\int_{-\infty}^{M_X} p_n(x)\mathrm{d}x = \int_{M_X}^{\infty} p_n(x)\mathrm{d}x \tag{2.26}$$

Knowing the antiderivative, we compute

$$[x^{-\alpha+1}]_{x_{\min}}^{M_X} = [x^{-\alpha+1}]_{M_X}^{\infty}$$
$$M_X = x_{\min} \mathrm{e}^{\frac{\ln 2}{\alpha-1}} \tag{2.27}$$

It is obvious that for $\alpha \to 1+$ the expression diverges to infinity. Even for values of α near 1 the median is quite large. This suggests that the data being distributed as the power-law for such value of α bear a significant measure of probability on its tail(see subsubsection 2.3.2).

2.3 Pareto distribution

Definition 2.13 mentions only CDF of some probability distribution. We now show that terms power-law and Pareto are replaceable. This is useful, because we may study zeta distribution and power-law as a **heavy-tail phenomenon**[10].

2.3.1 Pareto as an integral

Normalized power-law function $p_n(x)$ defined in equation 2.19 is integrable on any interval on $\mathbb{R}$. We can compute the integral

$$F(u) = \int_{-\infty}^{u} p_n(x)\mathrm{d}x = (\alpha-1)x_{\min}^{\alpha-1}\int_{x_{\min}}^{u} x^{-\alpha}\mathrm{d}x = 1 - \left(\frac{x_{\min}}{u}\right)^{\alpha-1}, u \geq x_{\min} \tag{2.28}$$

That obviously satisfies the requirements of Pareto distribution (see definition 2.13). Note that the exponent satisfies conditions both for parameter α in power-law and positive β in Pareto CDF. The coefficient $w = x_{\min}^{\alpha-1} = x_{\min}^{\beta}$ 'connects' the zero part of Pareto CDF to the rest continuously. In addition, the identity implies that Pareto CDF is absolutely continuous. From now on, we may consider power-law and Pareto distributions as the same for all terms and purposes[9].

2.3.2 Pareto as a heavy-tail

According to the definition of a heavy-tailed distribution in [10], the probability distribution is *heavy-tailed* if it has Pareto tail, i.e. from some point it 'follows' the Pareto distribution. The term 'follows' is ambiguous and de facto means that after perfoming a certain fit/hypothesis test[3], this must give a positive result[8]. This is what we will

further consider the term 'follow' to mean.

The very term 'heavy tail' reflects the properties of the function

$$\bar{F}_X(u) = P[X > u] := 1 - F_X(u) \tag{2.29}$$

where $F_X(u)$ is a Pareto CDF. $\bar{F}_X(u) = \left(\frac{x_{\min}}{u}\right)^{\alpha-1}$ obviously converges to zero as $u \to \infty$, however, more 'slowly' than other distributions, e.g. normal. In precise terms, $\bar{F}_X$ does not belong to any *Schwartz space*[4]. This means that for such distribution, there is a polynomial $a(u)$ so that a function $\eta(u) := a(u)\bar{F}_X(u)$ does not converge to zero for $u \to \infty$. No such polynomial exists e.g. for normal distribution, hence Pareto $\bar{F}_X$ falls more 'heavily' within its 'tail' (i.e. on some subset of its domain $(k, \infty), k \geq x_l$).

Pareto implies that heavy-tailed data sets with relatively large values of random variable X still carry a relative statistical significance[10]. Because of that, even such extremes in distributions must be taken into account, unlike e.g. in the estimation of errors in the normally distributed data[8].

2.4 From Zipf's law to Pareto

In this subsection, we present several statements that specify the problem to a narrow area of interest. We will build on statements and derivations mentioned above.

2.4.1 Logarithmic scale

Let us consider the simpliest form of Zipf's law, equation 1.1: $rf = c$. Applying natural logarithm[1] yields

$$\ln r + \ln f = \ln c \tag{2.30}$$

In logarithmic scales, the dependence of frequency on rank is linear, as $\ln c$ is a constant. The generalization mentioned in subsection 1.2 means simply that the line can be linearily rescaled by a factor of B[13], changing the law to

$$B \ln r + \ln f = \ln c \tag{2.31}$$

[1] Considering f being non-zero, only intervals where Zipf's law holds and thus the logarithm is defined are taken into account. The observation holds only up to the limits of the core language lexicon, namely its finiteness.

To assure convergence of the function to zero(high rank means lower frequency), we choose B to be greater than zero. If we now exponentiate the equation, we end up with the equation $r^B f = c$, i.e. eq. 1.2.

Notice that if $B > 1$, the function $f(r)$ is equivalent to a zeta distribution with a parameter $s = B$ and normalization $c = \zeta_{\mathbb{R}}^{-1}(B)$. Thus in this case, considering this generalized Zipf's law, we may lay an equivalency between the zeta distribution and such Zipf's law.

2.4.2 Connection between the terms

So far we only vaguely discussed properties of the equation characterizing Zipf's law. Equation 1.2 is thus only some relation without links to the terms such as probability mass function or Pareto. Semantics of it are that for the rank r, there is some frequency with which the given subject(e.g.word, firm with a certain turnover) occurs. The frequency is basically the amount of times certain random variable value X_i occured in the whole data set, but normalized (possibly). This is somewhat similar to constructing Laplace probability[8]. Intuitively, because these data sets are at most countable(theoretically) and summed probabilities yield 1, it is only natural to consider these to be *probability mass functions*[9] with a special case of Zipf's law in the form of a zeta distribution[1].

2.4.3 Heavy tail in Zipf's law

If we solve eq. 1.2 towards f, we will obtain a familiar relation:

$$f = \frac{c}{r^B} \qquad c, B > 0, r \in \mathbb{N}^+ \tag{2.32}$$

As mentioned in the subsubsection 2.2.2, equation 2.32 is normalizable only if $B > 1$. We only generalized Zipf's law using logarithmic scale, so the value of B in the log-scale was previously chosen only for f to be a decreasing function, converging to 0.

There are several arguments that help to solve this problem. Simply, we could restrict B to the interval $(1, \infty)$. Indeed, in real systems this is mostly the case[9]. However, it is generally possible that B falls to the interval $(0, 1]$. The respective CDF

[1] Of course the domain needs to be extended, see definition of zeta distribution and zeta mass function in subsection 2.1.

still needs to have the limit $\lim_{u\to\infty} F(u) = 1$, so if such case occurs, from some point it must inevitably break the relation and converge to zero more rapidly.

Because of the finiteness of the data sets, we will perform the hypothesis tests up to the points where data values still occur, as further extension would be meaningless. Because of that, In case of a positive match, we will accept the null hypothesis on the whole tail, as there is no evidence against any null hypothesis where no data exist. For the case where $B > 1$, we formally consider eq. 2.32 to be following zeta distribution. Because of the nature of used statistical hypothesis test, different case cannot be found(see subsection 2.6).Otherwise, we omit any such conclusion and if such case is found using a statistical hypothesis test, only the affilitation to the family of the Schwartz functions will be discussed.

2.5 Data handling

As the empirical(real) data are often sets of random variable realizations, Ω is bounded to the respective samples belonging to these values. Ω then can contain individual cities, pupils at school etc. Numerical samples then define a random variable X on Ω.

An intuitive way to assign measure on $\mathcal{B}(\mathbb{R})$ is to assign points in $\mathbb{R}$ the number of members in Ω that are mapped by X to the point value and normalize by $\frac{1}{n}$, where n is the number of members in Ω. This suggests a way of construction of the empirical cumulative distribution functions, while following definition 2.4(see below). Such measure is by definition discrete.

2.5.1 Empirical CDF

CDF constructed from a data set will be denoted $\hat{F}_n$ and estimated as following: Let Y be a data set with values $X_1, X_2, ..., X_n$, where $n, X_i \in \mathbb{N}^{+1}$. We sort Y increasingly and denote the new indexes as $X_{1:n}, X_{2:n}, ..., X_{n:n}$. Then we take such subset $\Lambda \subseteq Y$ that each integer value occuring in Y is also in Λ, but only once. For all $\lambda \in \Lambda$ then

[1]Positive integers, as decimal values in the data for purposes of this document are irrelevant and at least some data are recommended to exist.

compute $\hat{F}_n(\lambda) := \frac{1}{n} \sum_{X_{i:n} \leq \lambda} 1$. For Λ sorted in the same way as Y, we define

$$\hat{F}_n((\lambda_{k:m}, \lambda_{k+1:m})) := \hat{F}_n(\lambda_{k:m}), \quad \hat{F}_n((-\infty, \lambda_{1:m})) := 0, \quad \hat{F}_n((\lambda_{m:m}, \infty)) := 1 \quad (2.33)$$

This is an equivalent procedure to the ones described in [3] and [10], only modified so as to gain an actual algorithm to run on stored data files.

2.5.2 Test performed

Furthermore, we will focus on a certain hypothesis test, called *QQ plotting*(see below, subsection 2.6). We will attempt to test whether a data set is heavy-tailed and if so, additionally estimate the coefficient β of the Pareto distribution. The null hypothesis will be a typical heavy-tailed distribution. Because of the nature of the hypothesis test, the null hypothesis will not possess the same parameters of the distribution, yet it will serve as a scale reference. We test on Pareto because of the theorem that if data are Pareto distributed on the tail, they uphold Zipf's law there[11], in this case generalized Zipf's law[1]. Accepting Pareto null hypothesis means also accepting that the data set also upholds Zipf's law and in this sense also is zeta distributed.

2.6 QQ plotting

Many statistical hypothesis tests of certain distributions that are often a first choice include χ^2 or Kolmogorov test[3][8]. For a power-law distribution, there is another test, often used in economic studies - *Hill's estimator*[10]. *QQ plotting* is an alternative method, providing a good illustration at the cost of precise direct mathematical conclusions. Its result is a *QQ plot*, which afterwards needs to be further examined. Q stands for *quantile*, which relates with the *quantile functions* of given cumulative distribution functions.

2.6.1 Quantiles and quantile function

At first glance, the very definition of a quantile function is quite simple. Basic term *quantile* refers to a single value of X_0 that is mapped by $F_X(X_0)$ to a given

[1] The theorem actually states that the increase rates of the cumulative distribution functions are equivalent for sufficiently large number of samples.

value in $(0,1)$, thus forming a map $q_{F_X} : (0,1) \to \mathbb{R}$[1]. For an injective CDF, there is always but one such value. Otherwise, if the quantiles are to form a function, we must also choose but one value. The definitions in case of discrete distributions vary in different texts, however our motivation will be that the quantile $b = q_{F_X}(a)$ should be the smallest point in the support of the probability measure that contributed to $F_X(u) \geq a, u \geq b$. Loosely speaking, it is the smallest point that accumulated the value of CDF to reach or exceed a certain value a. This is consistent with the definition 2.5 and we can say that the quantile function is defined as

$$F_X^{\leftarrow}(a) := \inf\{u \in \mathrm{dom}(F_X) : F_X(u) \geq a\} \tag{2.34}$$

The definition is good for both continuous injective cumulative distribution functions and CDF of discrete distributions (for the continuous injective case it is an ordinary inverse on $(0,1)$).

2.6.2 Constructing a QQ plot

Clasiccal QQ plot is a graphic representation of two quantile functions on a two-dimensional plane, i.e. given two quantile functions $F_{X1}^{\leftarrow}(a), F_{X2}^{\leftarrow}(a)$, the QQ plot is a plot of ordered pairs $(F_{X1}^{\leftarrow}(a_0), F_{X2}^{\leftarrow}(a_0))$ for all a_0 in some subset $A \subseteq \mathrm{dom}(F_{X1}^{\leftarrow}) = \mathrm{dom}(F_{X2}^{\leftarrow})$[10]. Basically, we plot the values of both quantile functions with the same argument a. Because the method is often used in analyzing the empirical data, it is convenient to choose A to reflect the fact the data sets are finite and the definitions in subsubsection 2.2.3 create the characteristic 'stairs' in the cumulative distribution function. As the CDF is normalized by $\frac{1}{n}$, where n is the number of samples, the smallest height of such 'stair' is exactly $\frac{1}{n}$. No more than n values of a is needed, as it is the largest information we could hope to obtain from the empirical data set.

The height of the stair is only a motivation for choosing A. According to [10], A should be a sequence $\{\frac{i}{n+1}\}_{i=1}^{n}$, therefore the QQ plot would be ($F_X^{\leftarrow}$ being a null hypothesis quantile function, $\hat{F}_X^{\leftarrow}$ the empirical quantiles):

$$\left\{\left(F_X^{\leftarrow}\left(\frac{i}{n+1}\right), \hat{F}_X^{\leftarrow}\left(\frac{i}{n+1}\right)\right) : i = 1, 2, ..., n\right\} \tag{2.35}$$

[1] The domain $(0,1)$ needs to be an open interval for a good reason. If F_X is injective, the values at $\{0,1\}$ would inevitably diverge to infinities, as F_X is defined on $\mathbb{R}$ and if $F_X(u) = 1$ for any finite u, F_X would no longer be injective - reflectively for $F_X(u) = 0$. It is better to avoid such problems and if needed, to extend quantile function conveniently

The motivation for the values of $\frac{i}{n+1}$ is said to be partly historical[10], yet the following argument should suffice. At every point with a non-zero measure, the CDF increases by the measure of that singleton. Measure constructed following procedure described in the subsection 2.5 implies that such measure is at least $\frac{1}{n}$. Because of that, $\hat{F}_X^{\leftarrow}$ is constant on intervals $\left(\frac{i-1}{n}, \frac{i}{n}\right]$[1]. It is logical for the values of a to iterate between the intervals so a falls within them[2]. The expression 2.35 is therefore acceptable (detailed discussion also in [10]).

Previous statements lead to a conclusion that the value of the quantile $\hat{F}_X^{\leftarrow}\left(\frac{i}{n+1}\right)$ equals $X_{i:n}$ [10] (or the quantile of any value within the above interval for that matter). The set 2.35 can be substituted, resulting in a general QQ plot:

$$^{3}\left\{\left(F_X^{\leftarrow}\left(\frac{i}{n+1}\right), X_{i:n}\right) : i = 1, 2, ..., n\right\} \tag{2.36}$$

If the values in the pairs are equal, all points will lie on some line in the two-dimensional plane of the plot. That results in a logic implication and key statement about QQ plotting: There is no evidence real data do not follow the null hypothesis, if the respective QQ plot *looks roughly linear* [10]. It expresses the fact that such test does not actually *prove* the null hypothesis. By definition, probability distribution of any real data hardly equal to any probability distributions of interest, neither discrete nor others, if only due to the differences between cardinalities of the respective sample sets. The conclusion is somewhat loosened in the requirement that the QQ plot should look(by naked eye) at least approximately linear. Of course further analysis of the linearity might be in order, but the apparent linear shape concludes the QQ hypothesis plot.

2.6.3 Location-scale families

Plots of two non-equal quantile functions generally do not have to produce a linear QQ plot. However, there are certain classes of quantile functions that do have this

[1] If $i = n$, the expression $\frac{i}{n} = 1$ does not belong to the domain of $\hat{F}_X^{\leftarrow}$. However, the empirical CDF is equal to 1 in the point of the respective quantile and is further constant, so without fear we may consider even such a value. Secondly, $\forall i > 0 : \frac{i}{n+1} < \frac{i}{n}$.

[2] We might want to avoid the upper endpoint just for the reasons of the edge of the formal domain of $\hat{F}_X^{\leftarrow}$.

[3] The theorem actually states that the increase rate of the cumulative distribution functions are equivalent for sufficiently large values of X

property. This can be expressed as following identity for a class of the cumulative distribution functions:

$$\mathcal{F}_{\mu,\sigma} := \left\{ F_{\mu,\sigma}(x) = F_{0,1}\left(\frac{x-\mu}{\sigma}\right) \right\}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}\backslash\{0\} \tag{2.37}$$

$F_{0,1}(x)$ is a certain CDF, every other member of $\mathcal{F}_{\mu,\sigma}$ is parametrized by an offset μ and a scale σ. $\mathcal{F}_{\mu,\sigma}$ is called a *location-scale family* and the quantile function of any its member is[1]

$$F^{\leftarrow}_{\mu,\sigma}(q) = \mu + \sigma F^{\leftarrow}_{0,1}(q) \tag{2.38}$$

If the data are distributed with a certain CDF $F_{\mu,\sigma}$ that is a member of some location-scale family containing also a known function $F_{0,1}$, QQ plot may be adjusted by $F_{0,1}$ to both determine the affiliation to the location-scale family and estimate the values of μ and σ. By the substitution in the QQ plot definition 2.36 using identity 2.38 we obtain

$$\left\{ \left(\mu + \sigma F^{\leftarrow}_{0,1}\left(\frac{i}{n+1}\right), X_{i:n}\right) : i = 1, 2, ..., n \right\} \tag{2.39}$$

This is an implication from $F_{0,1}$ being in the same location-scale family as $F_{\mu,\sigma}$. The conclusion is that should $F_{0,1}$ be known, above QQ plot is linear for every respective member of the family, provided the data set is distributed with $F_{\mu,\sigma}$. Parameters μ and σ are simply the linear coefficient and the offset of both quantile functions in the plot.

2.6.4 Adaptation for the Pareto case

Testing a heavy-tail on any real data is somewhat problematic. The normalization of the cumulative distribution functions and therefore of the quantiles as well results only in scaling of either of the axes of the plot. However, Pareto cumulative distribution function has two parameters that we do not know, considering that we first need to know whether the data are heavy-tailed or not. Especially the parameter β of the Pareto CDF is somewhat troubling, as it directly influences the shape of the cumulative distribution function. Parameter x_l creates an offset where CDF is still zero and also serves as a normalization parameter. Therefore, even heavy-tailed data set would not generally produce a linear QQ plot against every Pareto distribution with arbitrary parameters β and x_l. Conclusion of this is either we need

[1] The relation originates in the fact that the quantiles are scaled with an offset[10]

to know the parameters before we attempt to produce any QQ plot or we need to adapt the method to an equivalent so that the QQ plots would look linear for heavy-tailed data.

Consider a cumulative distribution function $F_X(u) = P[X \leq u]$ that is of a Pareto distribution. The function $\bar{F}_X = 1 - F_X$ is then

$$\bar{F}_X(u) = \begin{cases} \left(\frac{x_l}{u}\right)^{\beta} = \left(\frac{u}{x_l}\right)^{-\beta} & u \geq x_l \\ 1 & u < x_l \end{cases} \tag{2.40}$$

The definition of cumulative distribution function in eq. 2.4 allows certain operations within the inequality in the expression $P[a \leq b]$ ($P[a > b]$ respectively) - e.g. substitution. Consider a transformation of $X \geq x_l : X \longrightarrow \beta \ln \frac{X}{x_l}$ [10]. Probability for such CDF is then expressed as (for $y > 0$)

$$P\left[\beta \ln \frac{X}{x_l} > y\right] := P\left(\left\{\omega \in \Omega : \beta \ln \frac{X(\omega)}{x_l} > y\right\}\right) \tag{2.41}$$

It is noteworthy that $X \geq x_l$ to properly express the logarithm and the probability distribution of X is still of Pareto distribution. The inequality could be transfered to a convenient form:

$$P\left[\beta \ln \frac{X}{x_l} > y\right] = P\left[\frac{X}{x_l} > \mathrm{e}^{\frac{y}{\beta}}\right] = P\left[X > x_l \mathrm{e}^{\frac{y}{\beta}}\right] \tag{2.42}$$

Right side of the last inequality is then a substitution for u should $u \geq x_l$ in the definition of $\bar{F}_X$. Knowing this, the result is following[10]:

$$P\left[X > x_l \mathrm{e}^{\frac{y}{\beta}}\right] = \left(\frac{x_l \mathrm{e}^{\frac{y}{\beta}}}{x_l}\right)^{-\beta} = \mathrm{e}^{-y} \tag{2.43}$$

This lemma for the Pareto distribution is a preparation for investigating of the log-scale in the QQ plot. Using this, we may now estimate the expression

$$\begin{aligned} P[\ln X > z] &= P\left[\tfrac{\ln X - \ln x_l}{\beta^{-1}} > \tfrac{z - \ln x_l}{\beta^{-1}}\right] = \\ &= P\left[\beta \ln \tfrac{X}{x_l} > \tfrac{z - \ln x_l}{\beta^{-1}}\right] = \mathrm{e}^{-\frac{z - \ln x_l}{\beta^{-1}}} \end{aligned} \tag{2.44}$$

Motivation for adding parameters β and x_l is to identify heavy-tailed distributions in a location-scale family. Let W_1 be a function

$$W_1(t) := 1 - \mathrm{e}^{-t} \tag{2.45}$$

In a log-scale, we may derive the respective cumulative distribution function:

$$\begin{gathered} P[\ln X > z] = \mathrm{e}^{-\frac{z-\ln x_l}{\beta^{-1}}} = 1 - (1 - \mathrm{e}^{-\frac{z-\ln x_l}{\beta^{-1}}}) = 1 - W_1(\tfrac{z-\ln x_l}{\beta^{-1}}) \\ P[\ln X \leq z] = W_1\left(\tfrac{z-\ln x_l}{\beta^{-1}}\right) \end{gathered} \tag{2.46}$$

Both parameters are separated and identified with a scaling and an offset of one axis in the plot. $W_1\left(\frac{z-\ln x_l}{\beta^{-1}}\right)$ presents a member of a location-scale family with $\mu = \ln x_l$ and $\sigma = \beta^{-1}$. This is an implication for $P[X \leq x]$ being a Pareto CDF - if the data are heavy-tailed, QQ plot of $\ln X_{i:n}$ against $W_1^{\leftarrow}\left(\frac{i}{n+1}\right)$ should look linear regardless of the parameters x_l and β [10].

Estimated quantile for W_1 is

$$\begin{gathered} q = 1 - \mathrm{e}^{-x} \\ x = W_1^{\leftarrow}(q) = -\ln(1-q) \end{gathered} \tag{2.47}$$

This leads to a compact statement regarding tests for Pareto distributions by the QQ plots[10]: QQ plot method does not disprove that the presented data set is Pareto distributed, if the respective plot of $\ln X_{i:n}$:

$$\left\{\left(-\ln\left(1 - \frac{i}{n+1}\right), \ln X_{i:n}\right) : i = 1, 2, ..., n\right\} \tag{2.48}$$

looks linear from some point i_0. With a certain degree of belief, we might conclude that the data actually are distributed in this manner (accepting the null hypothesis).

The term 'roughly linear' is somewhat vague in the sense of a proper mathematical analysis. Decision about whether data follow or not a certain distribution falls entirely into the competence of the analyst. A helpful tool might be generating a sufficient(order of 100) number of quantile sets using a method of 'shooting' at the respective cumulative distribution function[7]. The algorithm generates random numbers in the open interval $(0, 1)$ and determines the quantiles of these numbers for the respective CDF. Plots of these data sets covering the original QQ plot show where the original plot 'sticks out' from the cover. An example of such 'cover' using uniform distribution is in fig. 2.4.

The plot makes a band around an imaginative line that denotes the range of 'acceptable' deviation from the reference quantile function. Its width is varying to reflect the possible shape deviations in different parts of the plot, i.e. deviation in the edges of the plot is more 'serious' than around the center.

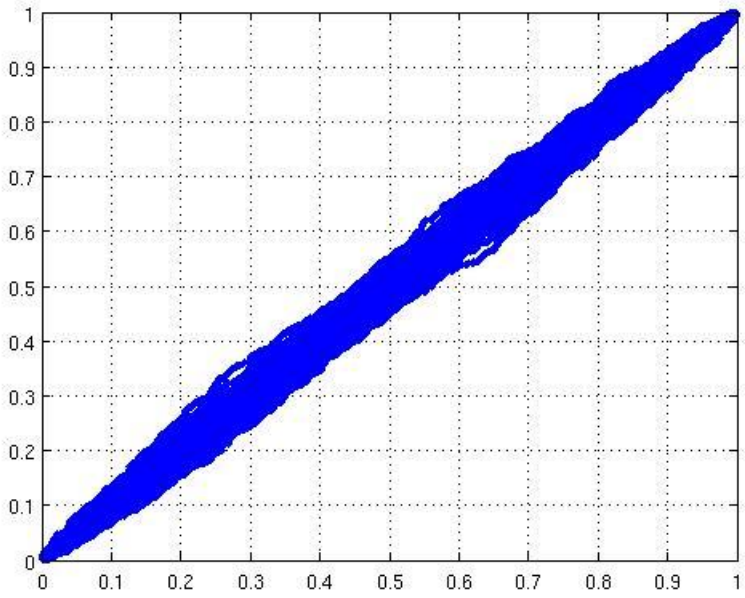


Figure 2.4: Uniform distribution QQ test cover

2.7 Modus operandi

The main advantage of the QQ plot is its illustrativeness. However, as was mentioned above, it lacks a clear statement about the null hypothesis in question. Main dilema is whether to state that the null hypothesis was not disproved or the null hypothesis is correct. Of course the first statement is more precise, as we cannot be 100% sure that the data indeed are distributed according to the hypothesis (Pareto in this case). Nevertheless, we might choose to accept the null hypothesis on the basis of the QQ plot, provided we set a certain rule on which basis we decide. We now choose such rule: if the QQ plot looks scattered, we decide only according to the simulated data sets test. Should the plot resemble a continuous trajectory(a drawn line between the discrete points if necessary), we first judge the appearance. If the plot is clearly linear for the naked eye, we only use the simulated data cover for additional illustration and accept the null hypothesis. If the plot is clearly not linear (e.g. an arc), we deny the null hypothesis and also use the simulated data cover only for an illustration. Otherwise, we rely on the simulated data cover test.

It should be noted that we are testing the tails of the real data distributions, not the entire data sets. It is because of the properties of the zeta distribution

and the power-law described in the subsection 2.2.4. We need to examine the shape of the QQ plots per intervals. It will be shown that the real data may look linear only on certain parts of the plot.

2.7.1 On other tests

It was mentioned that the QQ plotting method is an alternative statistical hypothesis test. For economic and statistical studies, there are several different estimators, namely Hill's estimator, *POT method* or *Pickand's estimator*. Due to the uncertainity of the results of statistical estimators, a result of a single estimator is not considered sufficient and should not be preffered over the others [10]. However, the theoretical background of other estimators, particulary the Hill's estimator, is quite extensive and beyond the capacity of this text, and so only the QQ plotting method will be performed, leaving a suggestion for further analyses.

Most of the data presented in this document were analyzed using log-log probability plots or Hill's estimator, namely towns population or first or family names, although in different context(country, age) [5][9] In these cases, a certain result is expected. There is, however, a data set that is unique and characteristic for the Czech Republic (see subsection 3.5) and that should be analyzed more extensively. The data set reflects the forming of Czech tertiary education - academic titles distributed in the population.

2.7.2 On performing the analysis

The analysis part of this thesis performs a single task of making the QQ plots out of the chosen data files with an afterwards discussion. In no sense the output will be a statement that the null hypothesis is correct, at best we only will accept it as a sufficient model. We are particularily interested in the null hypothesis for the tails, i.e. for the values that exceed a certain threshold.

Data stored and formatted on the attached CD are processed using MATLAB. The focus is on the QQ plots, however, the data files contain data variables suitable for other analyses, e.g. the mentioned Hill's estimator.

3 Data analysis

The applied part of this work includes a statistical analysis of the presented data. Because the data sets needed to be reliable, these were searched and taken from the database on the websites of the **Czech Statistical Office**(CZSO) and the **Ministry of the Interior of the Czech Republic**(MI). Respective sources are summarized in tab. 3.1 and the url links of the files in tab. 3.2.

#	Data description	Reference date	Institution	Homepage
1	Towns population	1.1.2001	CZSO	`http://www.czso.cz/`
2	Towns population	1.1.2010	CZSO	`http://www.czso.cz/`
3	Men first names	1.5.2009	MI	`http://www.mvcr.cz/`
4	Academic titles	19.5.2006	MI	`http://www.mvcr.cz/`

Table 3.1: Data sources(links actual to 21 December 2010)

#	Url address
1	`http://www.czso.cz/csu/2010edicniplan.nsf/t/06003C3DD7/$File/13011003.xls`
2	`http://www.czso.cz/csu/2001edicniplan.nsf/t/130032A03F/$File/obce.xls`
3	`http://www.mvcr.cz/soubor/cet-jm-mall090501-xls.aspx`
4	`http://aplikace.mvcr.cz/archiv2008/sprava/informat/cetnost/2006d/cet_tit_sum.xls`

Table 3.2: Data file names(links actual to 21 December 2010)

All datafiles were converted into files with sorted data sequences. All analyzes were done using MATLAB 2009a (CTU students' license) for Unix operating systems.

Input data files are simply sorted sequences of values related to each data file. Loosely speaking, each value in the data set is a random variable realization acquired upon collecting the statistics. Respective sample set depends on the datafile, however for the analysis purposes only the sequences are required, as they form a discrete measure described in the subsection 2.5 and in such way declare respective probability mass functions and cumulative distribution functions. An example might be a sample set

of all towns in the Czech Republic with a random variable represented by the number of inhabitants.

3.1 Scripting

All user-defined programming scripts used are in the set of m-files for MATLAB that were developed for the purpose of this thesis. They include the scripts for computing the probability mass functions, cumulative distribution functions, scripts for estimating the QQ plots, full-chain analysis of the data sets and a script for testing of the 'rough linearity'. The data are handled in the number of variables common for all data files. Essential variables are listed:

- **data_cdf**: Matrix variable of 2 columns, first the domain of the cumulative distribution function, second the mapped values.
- **data_pmf**: Matrix variable of 2 columns, first the domain of the probability mass function, second the mapped values.
- **data_qqplot**: Matrix variable of 2 columns, first the values of the data set values natural logarithms, second the quantiles of the reference quantile family.
- **data_raw**: 2 columns variable, first the data set sorted upwards, second downwards.
- **data_raw_down**: A column of the data set values sorted upwards.
- **data_raw_up**: A column of the data set values sorted downwards.
- **data_size**: Number of samples.
- **maximum**: Maximum of the data values.
- **minimum**: Minimum of the data values.
- **mu**: Parameter μ of the location-scale family.
- **name**: Identificator of the data set(data file core name).
- **sigma**: Parameter σ of the location-scale family.

These values are stored in the *.mat files `dataname_raw.mat` (only `data_raw` and `name`) and `dataname_final.mat` (all). Descriptions of the operation algorithm of the scripts follow:

pmf The function iteratively adds 1 to $A \in \{1, 2, ..., n\}$ where n is the size of the data set.

cdf Using pmf as input, cdf first estimates the support of pmf and then adds up the values to make a cumulative distribution function.

makeqq Natural logarithm is applied on the data set values and then the values of reference quantiles by the index are assigned.

theorquantile Given a range of a matrix, the output is a matrix of the simulated values of the reference quantile function(the 'shooting' method).

qqtesting Creates the 'cover' for the QQ plots in form of a matrix.

fullchain(data#anal) Loads a data file, estimates pmf, cdf, QQ plot of the data set and shows the plots.

3.2 #1 - Population of the towns in the Czech Republic(2001)

The first data set was the population in the Czech towns to date of 1 January 2001. Sample set Ω is represented by the towns in Czech Republic and the random variable is defined by the number of inhabitants. Specific markers were the numbers of inhabitants in the most populated and least populated towns - Prague with 1,180,131 inhabitants and the village 'Březina' with 8 inhabitants. A full-chain analysis ran with 6258 samples. All probability mass functions and cumulative distribution functions use log-scale in the random variable axis, partly because also the QQ plot uses logarithmic values of X and partly due to better clarity of the figures.

It is obvious that on certain interval the probability mass function in fig. 3.1 certainly is not decreasing. Also, the QQ plot suggests that until certain point the data are not Pareto distributed. However, for the values of $\ln X_{i:n} > 1 (i \geq 4000)$ the QQ

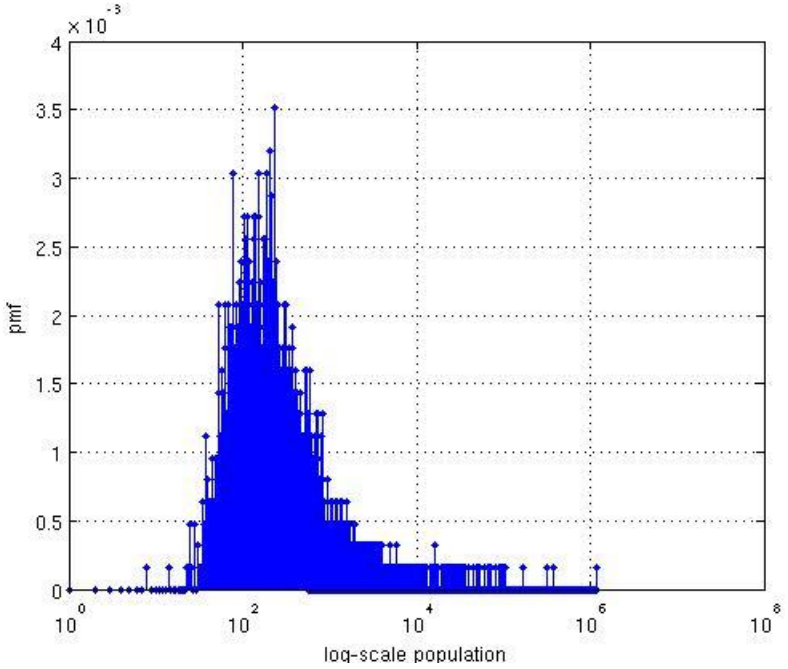


Figure 3.1: Probability mass function - data set #1

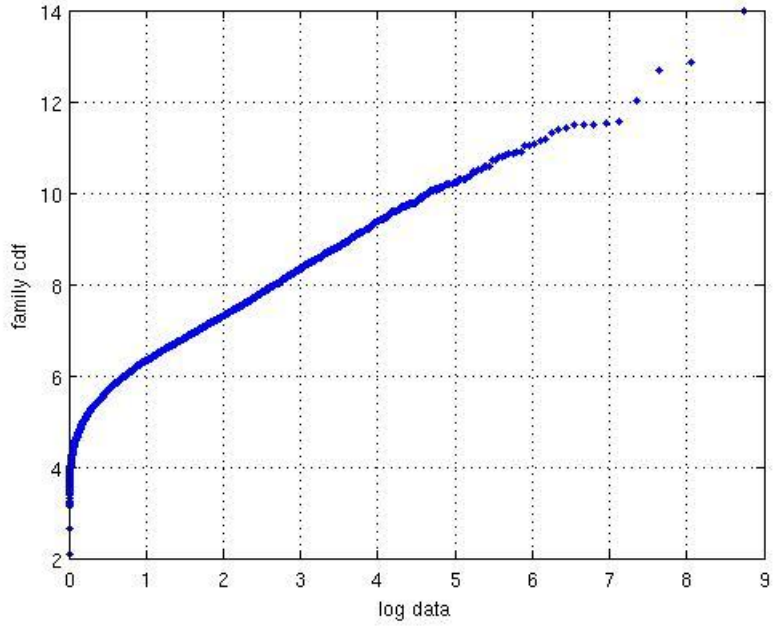


Figure 3.2: QQ plot - data set #1

plot shows a clear linearity. The deviances for larger values are due to the character of the distribution, as the values of $-\ln(1-q)$ are sensitive to perturbations in q as it approaches 1. This is shown in the QQ plot test by simulated data in fig. 3.3.

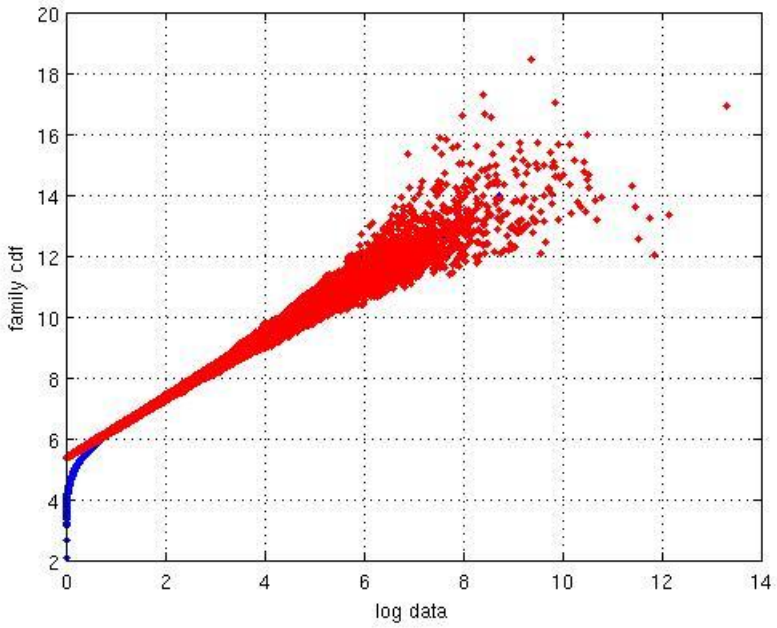


Figure 3.3: QQ plot simulated test - data set #1

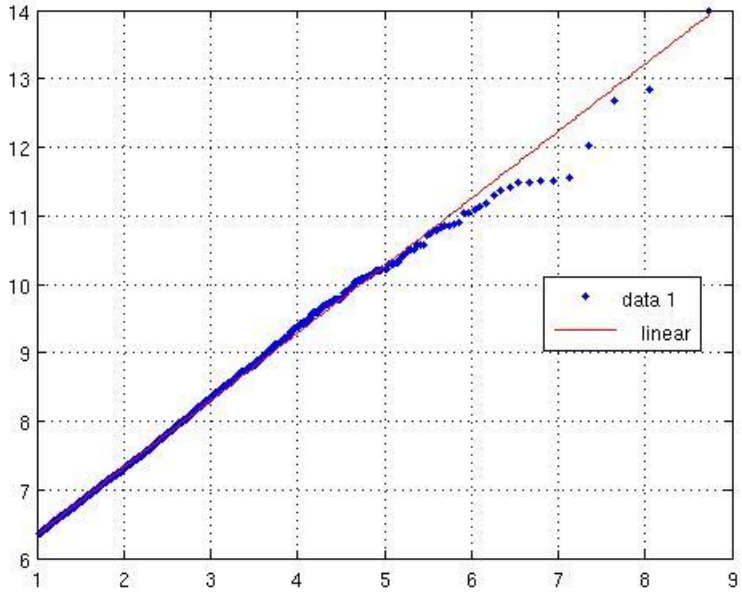


Figure 3.4: QQ plot linearization - data set #1

The covering band clearly spreads for higher values, therefore deviances at the far

No. of samples	min	max	μ	σ	x_l	β	Decision
6,258	8	1,180,131	5.3579	0.9818	212.2838	1.0185	+

Table 3.3: Summarized parameters - data set #1

edge are more acceptable. Linearization to determine the parameters for the cover was done using basic fitting tool in MATLAB. Parameters obtained through analysis are summarized in tab. 3.3 (values rounded to four decimal places).

μ and σ are the parameters of the location-scale family the null hypothesis belongs to, x_l and β the respective parameters of the Pareto distribution. Value for 'decision' is a statement whether the QQ plot is in favor of the null hypothesis or not (+ for positive).

3.3 #2 - Population of the towns in the Czech Republic(2010)

As a second data set the towns population in the Czech Republic was chosen again with the data actual to 1 January 2010. The motivation was to determine whether the coefficients characterizing the null hypothesis will change significantly due to the urbanization. During the decade, 8 towns(villages) lost their legal status. Extremes changed to a minimum of 3 inhabitants in Březina and 1,249,026 inhabitants in Prague. Results are again plotted and summarized.

No. of samples	min	max	μ	σ	x_l	β	Decision
6,250	3	1,249,026	5.4720	0.9563	237.9422	1.0457	+

Table 3.4: Summarized parameters - data set #2

Focusing on the problem at hand, only slight changes in parameters μ and σ were observed, suggesting that provided the null hypotheses are accepted, population flow focused into the rural areas over the decade in question.

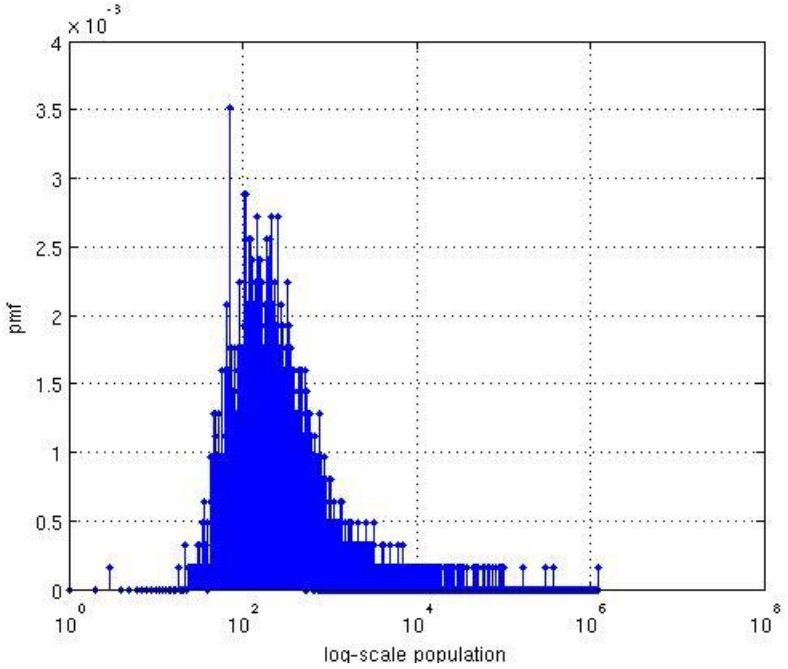


Figure 3.5: Probability mass function - data set #2

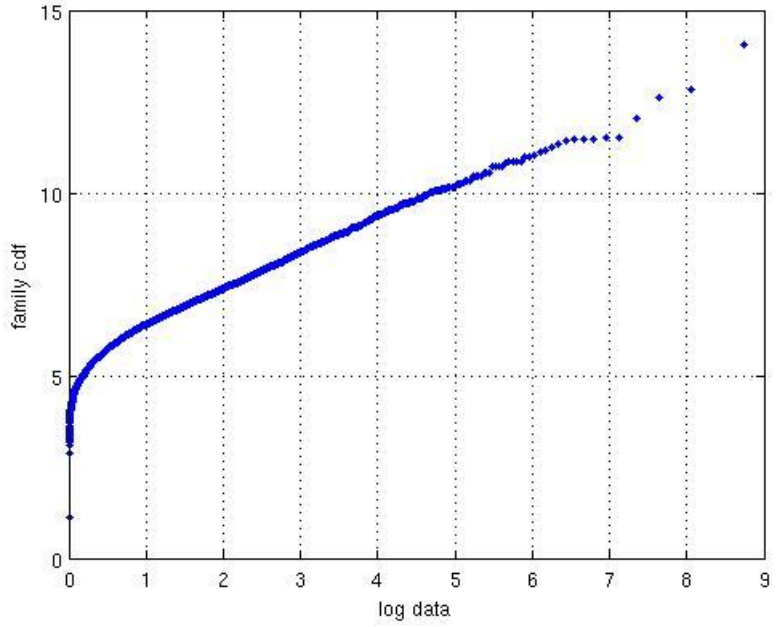


Figure 3.6: QQ plot - data set #2

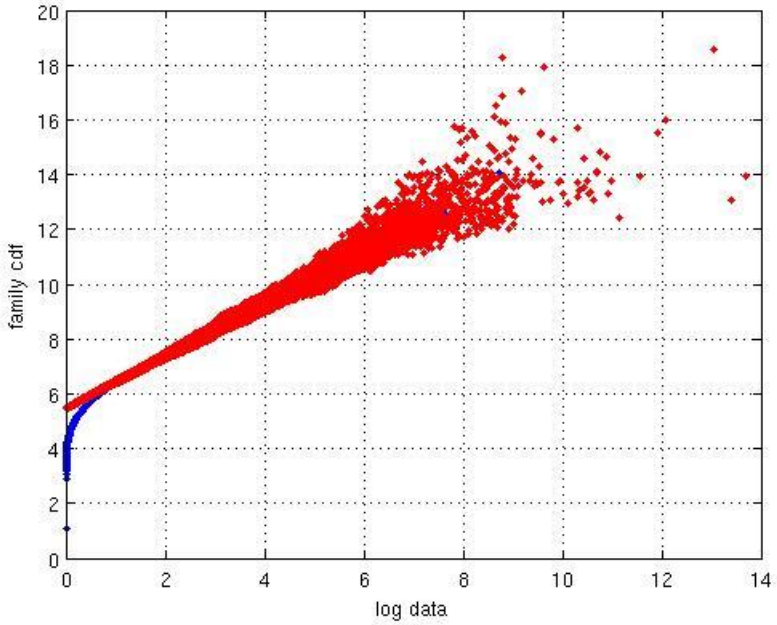


Figure 3.7: QQ plot simulated test - data set #2

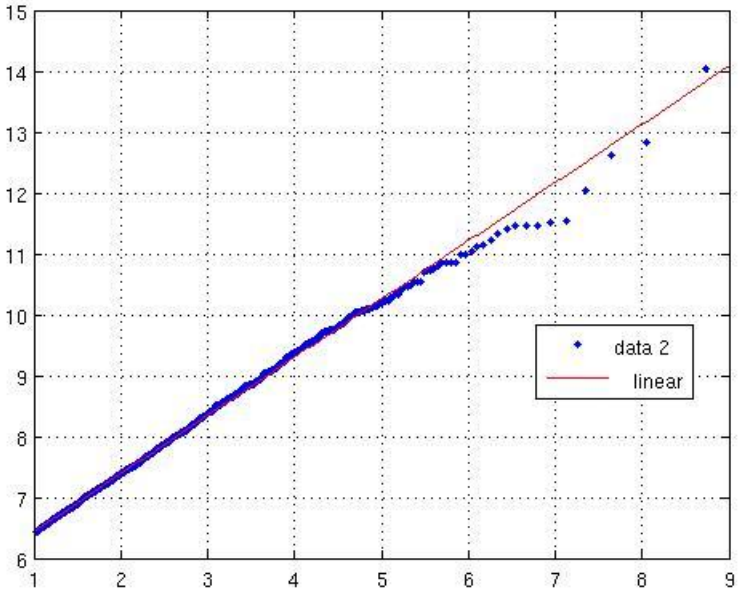


Figure 3.8: QQ plot linearization - data set #2

3.4 #3 - Male first names in the Czech Republic(2009)

Male first names of the inhabitants that have a valid legal residence in the Czech Republic is the sample set for the third data set. Ω is a set of all different registered first names and the random variable is defined as the number of individuals that have such first name. This set is characteristic by a high number of samples - 61,904. Minimum sample is 1 and the same value is shared by a large number of first names. It is noteworthy that even foreign names not domestic to the Czech Republic are included, possibly influencing the result. Full chain analysis follows.

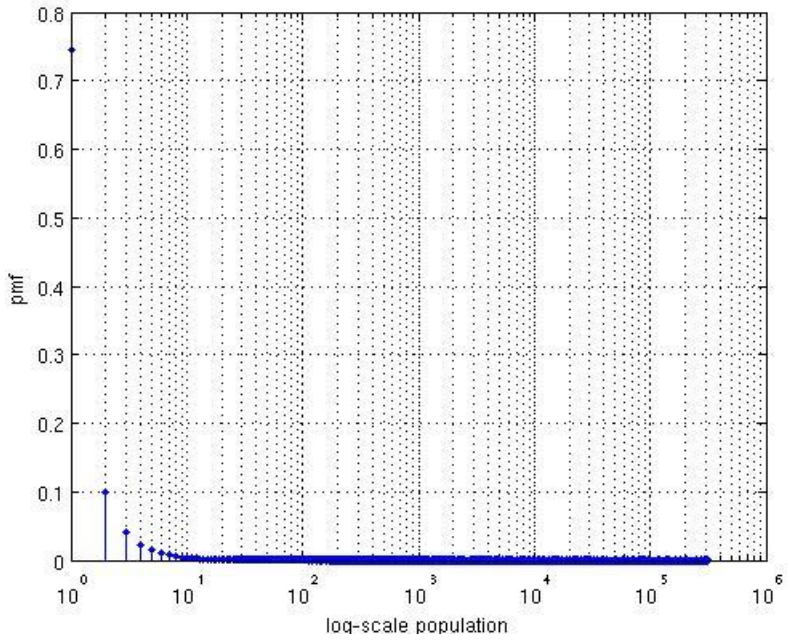


Figure 3.9: Probability mass function - data set #3

No. of samples	min	max	μ	σ	x_l	β	Decision
61,904	1	315,369	-3.6188	1.8056	0.0268	0.5538	-

Table 3.5: Summarized parameters - data set #3

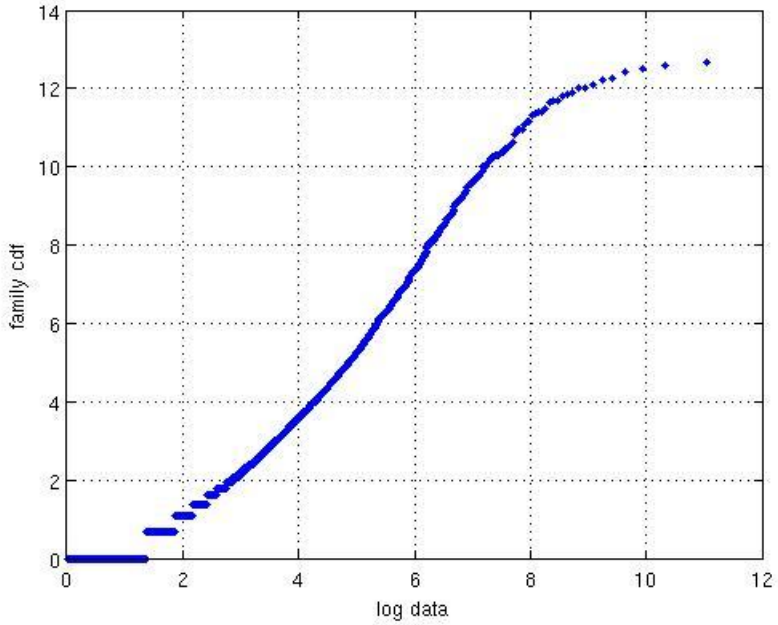


Figure 3.10: QQ plot - data set #3

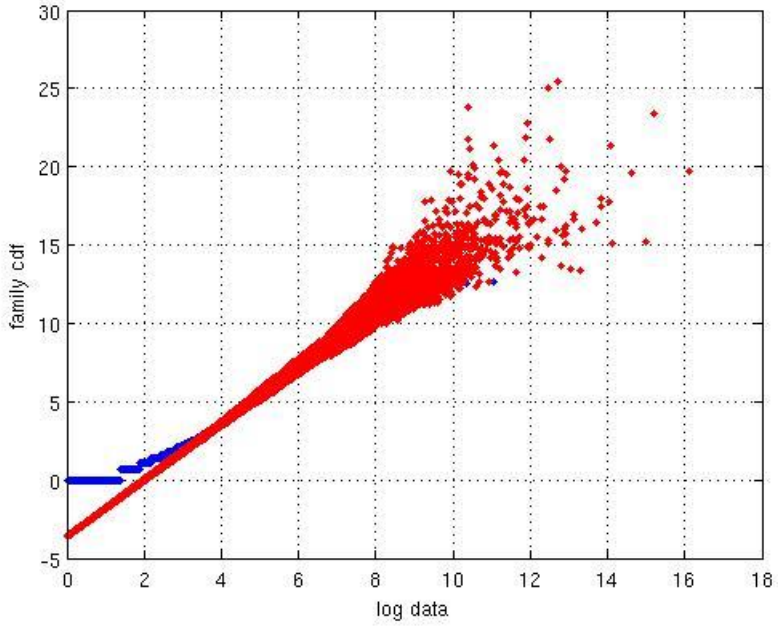


Figure 3.11: QQ plot simulated test - data set #3

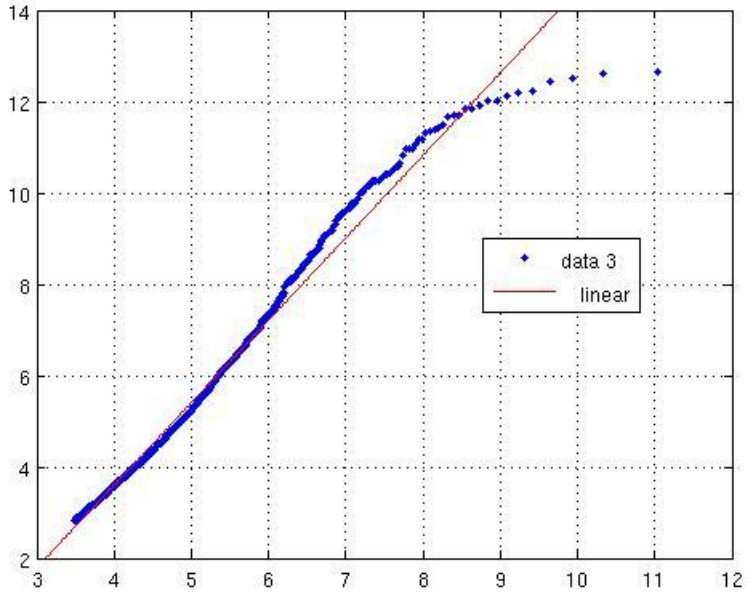


Figure 3.12: QQ plot linearization - data set #3

The QQ plot in fig. 3.10 obviously departs from the linearity for higher values of X. This leads to a conclusion that the data set is not Pareto distributed even if the testing plot in fig. 3.11 envelops the QQ plot (however, one can see the QQ points at the very edge of the band).

3.5 #4 - Academic titles in the Czech Republic(2006)

The fourth and last data set is chosen from the statistics of the university education in the Czech Republic. System of terciary education has not been unified in every country, therefore study of its results may prove interesting. All registered titles are the sample set, number of people who have valid residence in the Czech Republic and do possess the specific title define the random variable. The data set is the smallest presented here with only 603 samples, however even such statistics seem to show a clear result.

No. of samples	min	max	μ	σ	x_l	β	Decision
603	1	368,596	-0.7707	2.2109	0.4627	0.4523	+

Table 3.6: Summarized parameters - data set #4

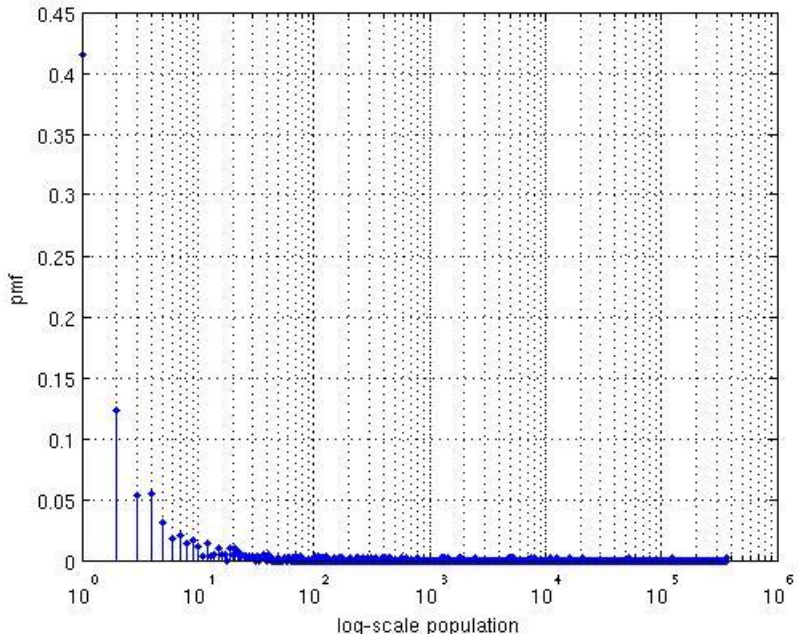


Figure 3.13: Probability mass function - data set #4

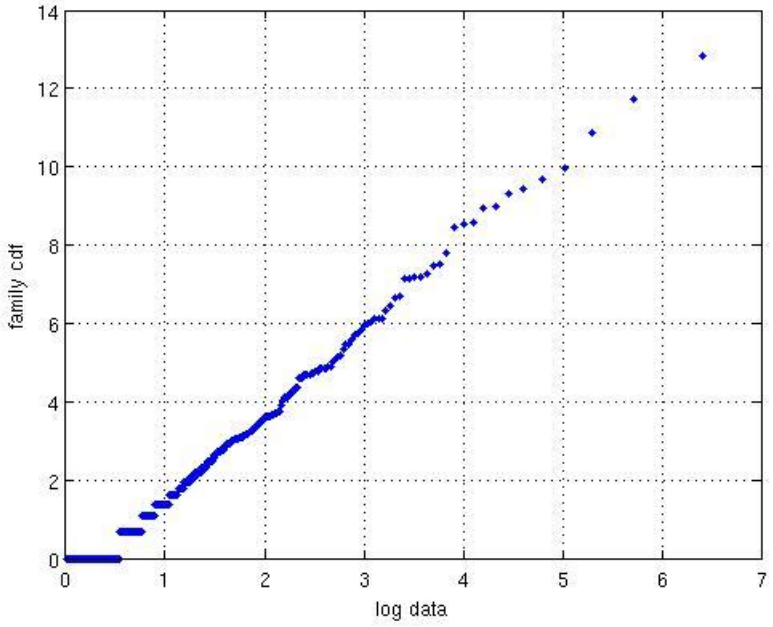


Figure 3.14: QQ plot - data set #4

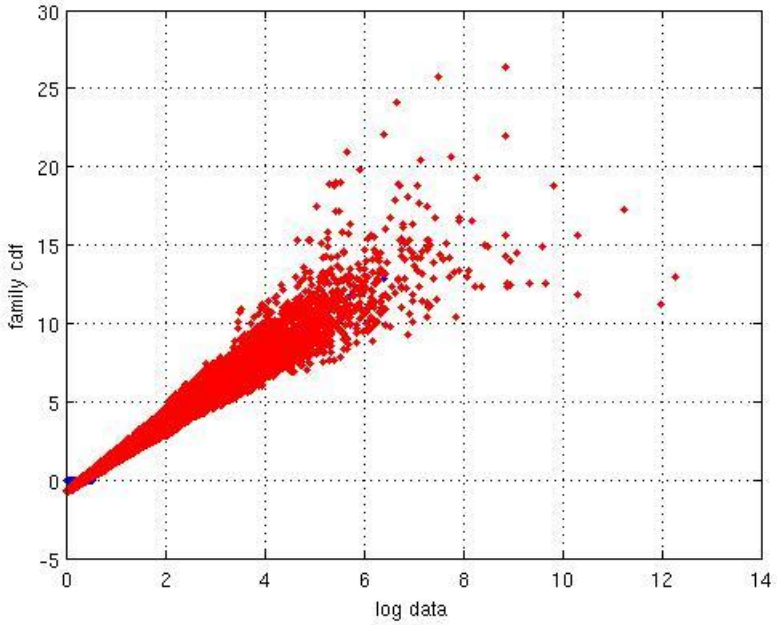


Figure 3.15: QQ plot simulated test - data set #4

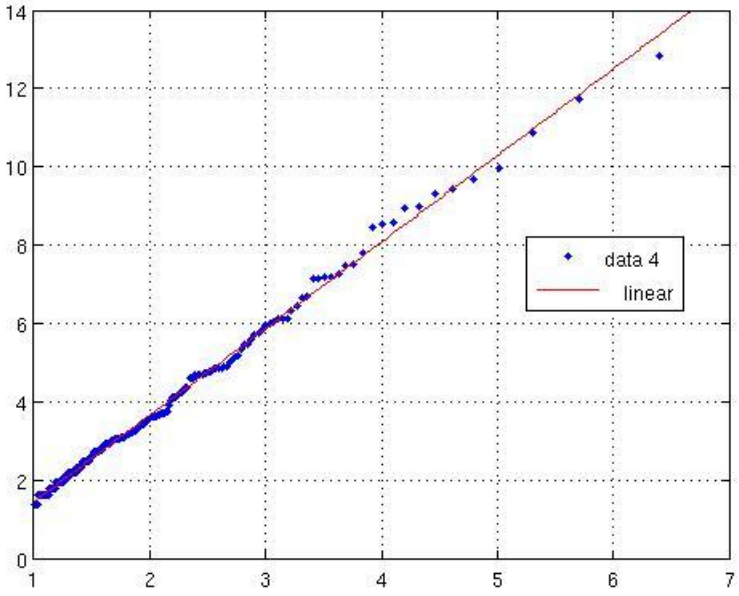


Figure 3.16: QQ plot linearization - data set #4

The QQ plot is clearly in favor of a Pareto distribution, maintaining the linearity even for the upper values of the data set.

4 Conclusions

In the section 2, we discussed and described properties of the zeta distribution and Pareto distribution. We also studied the power-law to identify with Pareto and the connection between zeta distributed and Pareto distributed data. Described method of QQ plotting was used to analyzing of the downloaded data in the section 3. In this part, we discuss the results and conclusions of those two sections.

4.1 QQ plot evaluation

The way of using the QQ plotting method in this thesis does not allow to study cases where the parameter α of the respective power-law distribution would equal 1. It only could near this case if the QQ plot reached very large value of σ. It is because the null hypothesis quantile function is monotonne and increasing and the sorted data are non-decreasing. However, there might occur a case where $\sigma = 0$. In this case, the data also are not Pareto distributed, as the parameter $\beta \in \mathbb{R}^+$ and such value of σ would cause β to diverge to infinity. Such case would occur in the analysis of the data set #4, if we took only the few first members of the data set.

In the case of Pareto distribution, the cover QQ plots show that the quantile functions $F^{\leftarrow}(a)$ are particularily sensitive in changing the value of a if it is somewhere in the near neigbourhood of the point 1. In the QQ plot it shows an increased scattering. This might be useful in analyzing the Monte Carlo performance, where quantiles of a certain distribution are generated. Simulating random numbers, especially where the rate of CDF is low, These covers may show where caution might be in order, because values too scattered indicate a lack of samples(the need of the mentioned 'band').

4.2 Zipf's law, zeta, power-law and Pareto tails in the theory

Pareto distribution, as well as the power-law in form of the normalized function $p_n(x)$, are characterized by two parameters: Pareto in β, x_l and power-law in $\alpha = \beta + 1$, $x_{\min} = x_l$. The original problem concerned Zipf's law in form of 1.1. However, this particular form complies neither with the definition of power-law, nor Pareto distribution. It is because the value of $\alpha(\beta$ respectively) would be out of the possible

interval, $\alpha = 1$(such Pareto CDF from the definition 2.13 would be undefinable). A possible way around would be accepting Zipf's law in form described by eq. 1.2 and allowing α to be slightly greater than 1. Note that the highly improbable yet not ruled out case of $\alpha < 1$[9] in no way can be described using the estabilished method (but still can be analyzed using e.g. log-log probability mass function plot, reflecting the logarithm in eq. 2.30).

The form of Zipf's law with generalized exponent indeed is considered also Zipf's law (although differently defined)[12]. We then are forced to depart from the original idea of eq. 1.1 and allow a generalized exponent. Only then we can conclude that the respectively zeta distributed data also uphold Zipf's law (regarding to the value $s = B$). Thus we identified Zipf's law with zeta distribution and earlier the power-law and Pareto distribution. The connection between these two distributions(zeta, Pareto) lies in the somewhat troubling theorem that states these two equivalent on their tails(with the respective values of s and β). If so, the chosen method is correct also in testing the zeta distribution.

4.3 Test results

As a final summary, we create a table with all relevant results of the data analysis:

Data set	# samples	α	β	x_l	Median	NH Med	Dec.
#1	6,258	2.0185	1.0185	212.2838	384	419.2562	+
#2	6,250	2.0457	1.0457	237.9422	417	461.6793	+
#3	61,904	1.5538	0.5538	0.0268	1	0.0937	-
#4	603	1.4523	0.4523	0.4523	2	2.1420	+

Table 4.1: Summary - all data sets

Values of `Median` and `NH Med` refer to the values of the actual median of the data and the null hypothesis median computed for the respective null hypothesis power-law, described in subsubsection 2.2.5. The median values are mentioned as a 'test' how well the null hypothesis describes the actual data set.

4.3.1 #1 and #2 results

There is no surprise that the QQ plot method is in favor of the Pareto null hypothesis. According to what already has been derived and described in the theoretical framework, if the data were zeta(Zipf) distributed, they would also have a Pareto tail. Despite for a different country, the study of towns population in US also showed a positive match with Zipf's law[5]. Also, the order value of median of the data corresponds with the null hypothesis median quite well. For the second data set, there is an increase in the median. This supports the earlier statement suggesting that the population somewhat flowed into the rural areas.

4.3.2 #3 Results

Parameters of the null hypothesis for the third data set are only formal, as the null hypothesis has been denied following the QQ plot. The order of the median also is off, however, it is difficult to determine properly, as the integer values of the data set mean only 0.5 multiples in the value, much larger than the theoretical median of the respective Pareto. Possible further analysis could include focusing on the names in a certain local area or Czech-only names.

4.3.3 #4 Results

This result is one of the most surprising, as the Czech system of possible academic titles is somewhat unique both in Europe and the world. A clear linearity holds until the upper values of the set, despite the clear uncertainty in that part of the plot. Both theoretical and real medians are equal to the minimum error, suggesting further analyses of this particular data set.

4.4 Epilogue

We described several terms and properties concerning Zipf's law. There is a number of papers and other publications discussing this interesting phenomenon, both in empirical and theoretical point of view. However, there are also inconsistent opposing definitions and descriptions that need to be first fully understood, if we are to present any solid result. The analysis results supported several known facts and also intro-

duced a whole apparatus for studying the problem. In the cases where the described definitions and formulations departed from the ones mentioned in any of the referenced sources, this was done only because there was a choice to be made(e.g. the Zipf's law definition). For different premises, the results could also be different, however the answers based on the structure of this thesis and the line described in [10] are quite solid.

If there is anything this text is hoped to achieve, it is raising questions about the presented data sets and encouraging further analyses. The topic is so vast that one paper, thesis or book cannot cover it completely. QQ plot method is only one amongst a large number of possible estimators. Additionaly, there is still a missing link between the varying character of the data and the same Zipf's law that holds for so many of them. The biggest question is then what is this link, how people with different dreams and passions end up being a part of this simple mathematical distribution or how Shakespeare could uphold this law with his unique plays that set a course of speech for next tens of generations of the whole nation.

A Notation

(a, b)	Open interval
$[a, b]$	Closed interval
2^M	A power set of an arbitrary set M
$\mathcal{A}$	Arbitrary σ-algebra
B	Zipf's law exponent parameter
$\mathcal{B}(\mathbb{R})$	A Borel σ-algebra on a set of real numbers
d_X	Probability mass function
$\mathrm{dom}(A)$	Domain of A
f_s	Zeta distribution
$F_X(u), P[X \leq u], P_X((-\infty, u])$	Cumulative distribution function
$\bar{F}_X(u), P[X > u]$	Supplement of the cum. distr. function in $\mathbb{R}$
$F_X^{\leftarrow}$	Quantile function
$\hat{F}_n$	Empirical cumulative distribution function
$F_{\mu,\sigma}$	Location-scale family member
$\mathcal{F}_{\mu,\sigma}$	Location-scale family
$H^{\leftarrow}$	Inverse of H should it be non-injective
ln	Logaritmus naturalis
$\mathbb{N}$	Set of natural numbers with 0
$\mathbb{N}^+$	Set of positive natural numbers, without 0
p_d	Power-law
p	Extension of the power-law
p_n	Normalized power-law
P	Probability measure
P_X	Probability distribution
r	Zipf's law rank
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^+$	Set of positive real numbers
$\mathrm{supp}(d_X)$	Support of d_X
$\mathfrak{U}(\mathcal{S})$	A σ-algebra generated by $\mathcal{S}$
X	Random variable
$X_{i:n}$	ith smallest value of X

α	Exponential parameter of a power-law
β	Pareto exponential parameter
$\zeta(z)$	Riemann zeta function
$\zeta_{\mathbb{R}}(x)$	Riemann zeta function with a real parameter
Λ	Support of the empirical probability mass function
μ	Offset parameter of the location-scale family
σ	Scale parameter of the location-scale family
Ω	Sample set

B CD content

Besides the electronic copy of the thesis in pdf format in the root directory (file `delonvoj-bcthesis-zipf.pdf`), the CD also contains a directory `analysis`. In the directory, there is a number of commented m-scripts containing all mentioned functions used for the analysis. This directory also contains four subdirectories marked `data`*number*`_`*name*`_`*year* denoting

- *number* - number of the data set
- *name* - name of the data set
- *year* - reference year of the data set

This string will now be referenced as the *dirname*, the original data set file names *origname*. Each subdirectory contains these files:

- *dirname*`_final.mat` - MATLAB output data file with all result variables
- *dirname*`(`*origname*`).xls` - original downloaded data file
- *dirname*`_raw.mat` - MATLAB input data file
- *dirname*`_sorted.csv` - CSV file with the data
- *dirname*`_sorted.ods` - ODS OpenOffice file with the data

The `csv` and `ods` files where created in OpenOffice Calc v3.1 for Linux Ubuntu.

References

[1] ADAMIC, Lada A.; HUBERMAN, Bernardo A. Zipf's law and the Internet. *Glottometrics*. 2002, 3, p. 143-150.

[2] AHLFORS, Lars V. *Complex Analysis: an Introduction to the Theory of Analytic Functions of One Complex Variable*. Massachusetts: McGraw-Hill Science/Engineering/Math, 1979. 336 p.

[3] ANDĚL, Jiří. *Statistické metody*. Prague : Matfyzpress, 2007. 299 p.

[4] BLANK, Jiří; EXNER, Pavel; HAVLÍČEK, Miloslav. *Lineární operátory v kvantové fyzice*. Prague : Karolinum, 1993. 678 p.

[5] JIANG, Bin; JIA, Tao. *Zipf's Law for All the Natural Cities in the United States: A Geospatial Perspective*. Gävle University. 2010, 10 p. Available at WWW: ⟨http://arxiv.org/pdf/1006.0814v2⟩.

[6] KOPÁČEK, Jiří. *Matematická analýza nejen pro fyziky (I)*. Prague : Matfyzpress, 2004. 182 p.

[7] KULHÁNEK, Petr. *Statistická fyzika*. Prague : CTU, 2002. 86 p.

[8] NAVARA, Mirko. *Pravděpodobnost a matematická statistika*. Prague : CTU, 2007. 240 p.

[9] NEWMAN, Mark. Power laws, Pareto distributions and Zipf law. *Contemporary Physics*. 2005, 46, p. 323-351. Available at WWW: ⟨http://arxiv.org/abs/cond-mat/0412004v3⟩.

[10] RESNICK, Sidney I. *Heavy-tail phenomena: Probabilistic and Statistical Modeling*. Ithaca, NY : Springer, 2007. 404 p.

[11] SCHUETTE, Paul; SPRUILL, Marcus C. Tail fit and the Zipf-Pareto law. *Extremes*. 2007, Volume 9, 3-4, p. 243-261.

[12] TRIPP, Omer; FEITELSON, Dror. *Zipf's law revisited*. Jerusalem, Israel : Hebrew University, 2007. 15 p.

[13] WYLLYS, Ronald E. Empirical and Theoretical Bases of the Zipf's Law. *Library Trends*. 1981, summer, p. 53-64.

[14] Zeta distribution. In *Wikipedia : the free encyclopedia* [online]. St. Petersburg (Florida) : Wikipedia Foundation, 20 October 2002, last modified on 7 June 2010 [cit. 2011-01-07].
Available at WWW: ⟨http://en.wikipedia.org/wiki/Zeta_distribution⟩.

[15] ZIPF, George K. *Human Behavior and the Principle of Least Effort*. Massachusetts : Addison-Wesley Press, 1949. 573 p.